WILEY

PHYSICS OF SEMICONDUCTOR DEVICES



THIRD EDITION

S. M. SZE
KWOK K. NG

Sandisk Technologies, Inc. - Ex. 1042, Page 000001
IPR2025-01283 (Sandisk Technologies, Inc. v. Longitude Flash Memory Solutions Ltd.)

WILEY-INTERSCIENCE

E & S
TK
7871.85
.S988
2007

Physics of Semiconductor Devices

SZE
NG

Sandisk Technologies, Inc. - Ex. 1042, Page 000002



The *Third Edition* of the standard textbook and reference in the field of semiconductor devices

This classic book has set the standard for advanced study and reference in the semiconductor device field. Now completely updated and reorganized to reflect the tremendous advances in device concepts and performance, this *Third Edition* remains the most detailed and exhaustive single source of information on the most important semiconductor devices. It gives readers immediate access to detailed descriptions of the underlying physics and performance characteristics of all major bipolar, field-effect, microwave, photonic, and sensor devices.

Designed for graduate textbook adoptions and reference needs, this new edition includes:

- A complete update of the latest developments
- New devices such as three-dimensional MOSFETs, MODFETs, resonant-tunneling diodes, semiconductor sensors, quantum-cascade lasers, single-electron transistors, real-space transfer devices, and more
- Materials completely reorganized
- Problem sets at the end of each chapter
- All figures reproduced at the highest quality

Physics of Semiconductor Devices, Third Edition offers engineers, research scientists, faculty, and students a practical basis for understanding the most important devices in use today and for evaluating future device performance and limitations.

S. M. SZE received his PhD in electrical engineering from Stanford University. He was with Bell Telephone Laboratories from 1963–1989, joining the faculty of the Department of Electronics Engineering, National Chiao Tung University (NCTU) in 1990. Dr. Sze is currently Distinguished Chair Professor of NCTU and has served as a visiting professor to many academic institutions. He has made fundamental and pioneering contributions to semiconductor devices; of particular importance is his coinvention of nonvolatile semiconductor memory such as flash memory and EEPROM. Dr. Sze has authored, coauthored, or edited over 200 technical papers and twelve books. His book *Physics of Semiconductor Devices* (Wiley) is one of the most cited works in contemporary engineering and applied science publications (over 15,000 citations from ISI Press). Dr. Sze is the recipient of numerous awards and holds such titles as Life Fellow of the IEEE, Academician of the Academia Sinica, and member of the US National Academy of Engineering.

KWOK K. NG received his PhD from Columbia University in 1979 and BS from Rutgers University in 1975, both in electrical engineering. He joined Bell Laboratories of AT&T in Murray Hill, New Jersey, in 1980, which spun off as part of Lucent Technologies in 1996. He became affiliated with Agere Systems in Allentown, Pennsylvania, as the microelectronics unit became independent in 2001. He has been with MVC in San Jose, California, since 2005. Dr. Ng has also held positions as editor of *IEEE Electron Device Letters* and liaison to IEEE Press. He is the author of the *Complete Guide to Semiconductor Devices, Second Edition* (Wiley).

Cover Design: Michael Rutkowski

Subscribe to our free Electrical Engineering eNewsletter at www.wiley.com/enewsletters

Visit www.wiley.com/electrical



wiley.com





Sandisk Technologies, Inc. - Ex. 1042, Page 000003

University Libraries
Carnegie Mellon University
Pittsburgh, PA 15213-3890

Sandisk Technologies, Inc. - Ex. 1042, Page 000004

Physics of Semiconductor Devices

Third Edition

S. M. Sze
Department of Electronics Engineering
National Chiao Tung University
Hsinchu, Taiwan

and

Kwok K. Ng
Central Laboratory
MVC (a subsidiary of ProMOS Technologies, Taiwan)
San Jose, California

WILEY-INTERSCIENCE

A JOHN WILEY & SONS, INC., PUBLICATION

Sandisk Technologies, Inc. - Ex. 1042, Page 000005

TK
7871.85
.S988
2007

Description of cover photograph:

A scanning electron micrograph of an array of the floating-gate nonvolatile semiconductor memory (NVSM) magnified 100,000 times. NVSM was invented at Bell Telephone Laboratories in 1967. There are more NVSM cells produced annually in the world than any other semiconductor device and, for that matter, any other human-made item. For a discussion of this device, see Chapter 6. Photo courtesy of Macronix International Company, Hsinchu, Taiwan, ROC.

pv

Copyright © 2007 by John Wiley & Sons, Inc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at http://www.wiley.com/go/permission.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic format. For information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data is available.

ISBN-13: 978-0-471-14323-9
ISBN-10: 0-471-14323-5

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

Sandisk Technologies, Inc. - Ex. 1042, Page 000006

Contents

Introduction 1

Part I Semiconductor Physics

Chapter 1 Physics and Properties of Semiconductors—A Review 7

1.1 Introduction, 7
1.2 Crystal Structure, 8
1.3 Energy Bands and Energy Gap, 12
1.4 Carrier Concentration at Thermal Equilibrium, 16
1.5 Carrier-Transport Phenomena, 28
1.6 Phonon, Optical, and Thermal Properties, 50
1.7 Heterojunctions and Nanostructures, 56
1.8 Basic Equations and Examples, 62

Part II Device Building Blocks

Chapter 2 *p-n* Junctions 79

2.1 Introduction, 79
2.2 Depletion Region, 80
2.3 Current-Voltage Characteristics, 90
2.4 Junction Breakdown, 102
2.5 Transient Behavior and Noise, 114
2.6 Terminal Functions, 118
2.7 Heterojunctions, 124

Chapter 3 Metal-Semiconductor Contacts 134

3.1 Introduction, 134
3.2 Formation of Barrier, 135
3.3 Current Transport Processes, 153
3.4 Measurement of Barrier Height, 170
3.5 Device Structures, 181
3.6 Ohmic Contact, 187

Sandisk Technologies, Inc. - Ex. 1042, Page 000007

Chapter 4 Metal-Insulator-Semiconductor Capacitors 197

4.1 Introduction, 197
4.2 Ideal MIS Capacitor, 198
4.3 Silicon MOS Capacitor, 213

Part III Transistors

Chapter 5 Bipolar Transistors 243

5.1 Introduction, 243
5.2 Static Characteristics, 244
5.3 Microwave Characteristics, 262
5.4 Related Device Structures, 275
5.5 Heterojunction Bipolar Transistor, 282

Chapter 6 MOSFETs 293

6.1 Introduction, 293
6.2 Basic Device Characteristics, 297
6.3 Nonuniform Doping and Buried-Channel Device, 320
6.4 Device Scaling and Short-Channel Effects, 328
6.5 MOSFET Structures, 339
6.6 Circuit Applications, 347
6.7 Nonvolatile Memory Devices, 350
6.8 Single-Electron Transistor, 360

Chapter 7 JFETs, MESFETs, and MODFETs 374

7.1 Introduction, 374
7.2 JFET and MESFET, 375
7.3 MODFET, 401

Part IV Negative-Resistance and Power Devices

Chapter 8 Tunnel Devices 417

8.1 Introduction, 417
8.2 Tunnel Diode, 418
8.3 Related Tunnel Devices, 435
8.4 Resonant-Tunneling Diode, 454

Chapter 9 IMPATT Diodes 466

9.1 Introduction, 466

Sandisk Technologies, Inc. - Ex. 1042, Page 000008

9.2 Static Characteristics, 467
9.3 Dynamic Characteristics, 474
9.4 Power and Efficiency, 482
9.5 Noise Behavior, 489
9.6 Device Design and Performance, 493
9.7 BARITT Diode, 497
9.8 TUNNETT Diode, 504

Chapter 10 Transferred-Electron and Real-Space-Transfer Devices 510

10.1 Introduction, 510
10.2 Transferred-Electron Device, 511
10.3 Real-Space-Transfer Devices, 536

Chapter 11 Thyristors and Power Devices 548

11.1 Introduction, 548
11.2 Thyristor Characteristics, 549
11.3 Thyristor Variations, 574
11.4 Other Power Devices, 582

Part V Photonic Devices and Sensors

Chapter 12 LEDs and Lasers 601

12.1 Introduction, 601
12.2 Radiative Transitions, 603
12.3 Light-Emitting Diode (LED), 608
12.4 Laser Physics, 621
12.5 Laser Operating Characteristics, 630
12.6 Specialty Lasers, 651

Chapter 13 Photodetectors and Solar Cells 663

13.1 Introduction, 663
13.2 Photoconductor, 667
13.3 Photodiodes, 671
13.4 Avalanche Photodiode, 683
13.5 Phototransistor, 694
13.6 Charge-Coupled Device (CCD), 697
13.7 Metal-Semiconductor-Metal Photodetector, 712
13.8 Quantum-Well Infrared Photodetector, 716
13.9 Solar Cell, 719

Sandisk Technologies, Inc. - Ex. 1042, Page 000009

Chapter 14 Sensors 743

14.1 Introduction, 743
14.2 Thermal Sensors, 744
14.3 Mechanical Sensors, 750
14.4 Magnetic Sensors, 758
14.5 Chemical Sensors, 765

Appendixes 773

A. List of Symbols, 775
B. International System of Units, 785
C. Unit Prefixes, 786
D. Greek Alphabet, 787
E. Physical Constants, 788
F. Properties of Important Semiconductors, 789
G. Properties of Si and GaAs, 790
H. Properties of SiO_2 and Si_3N_4, 791

Index 793

Sandisk Technologies, Inc. - Ex. 1042, Page 000010

6

MOSFETs

6.1 INTRODUCTION
6.2 BASIC DEVICE CHARACTERISTICS
6.3 NONUNIFORM DOPING AND BURIED-CHANNEL DEVICE
6.4 DEVICE SCALING AND SHORT-CHANNEL EFFECTS
6.5 MOSFET STRUCTURES
6.6 CIRCUIT APPLICATIONS
6.7 NONVOLATILE MEMORY DEVICES
6.8 SINGLE-ELECTRON TRANSISTOR

6.1 INTRODUCTION

The metal-oxide-semiconductor field-effect transistor (MOSFET) is the most-important device for forefront high-density integrated circuits such as microprocessors and semiconductor memories. It is also becoming an important power device. The principle of the surface field-effect transistor was first proposed in the early 1930s by Lilienfeld[1–3] and Heil.[4] It was subsequently studied by Shockley and Pearson[5] in the late 1940s. In 1960, Ligenza and Spitzer produced the first device-quality Si-SiO_2 MOS system using thermal oxidation.[6] The basic MOSFET structure using this Si-SiO_2 system was proposed by Atalla.[7] Subsequently the first MOSFET was reported by Kahng and Atalla in 1960.[8] The detailed early historical development of the MOSFET can be found in Refs. 9–10. The basic device characteristics have been initially studied by Ihantola and Moll,[11] Sah,[12] and Hofstein and Heiman.[13] The technology, application, and device physics have been reviewed by many books.[14–17]

Figure 1 shows the reduction of the gate-length dimension in production ICs since 1970. This dimension has been decreasing at a steady pace and will continue to shrink in the foreseeable future. The reduction of device dimensions is driven by the require-

Sandisk Technologies, Inc. - Ex. 1042, Page 000011

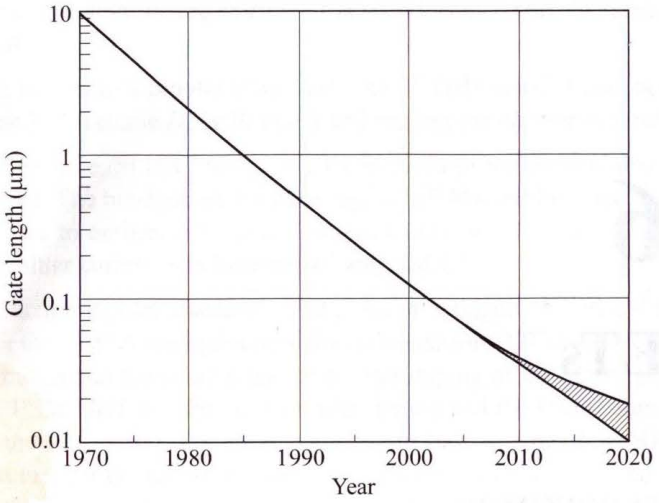


Fig. 1 Minimum gate dimension in commercial integrated circuit as a function of the year of production.

ment for both performance and density. The number of components per integrated circuit chip has grown exponentially. The rate of growth is expected to slow down because of increasing technological challenge and fabrication cost. However, a complexity of 1 billion or more devices per chip had been available around 2000 using 0.1-μm technology.

In this chapter we first consider the basic device characteristics of the so-called long-channel MOSFET; that is, the longitudinal field along the channel is not large enough to cause velocity saturation. In this regime, the carrier velocity is mobility-limited, or under constant mobility. As the channel length becomes shorter, one has to consider short-channel effects due to two-dimensional potential and high-field transport such as velocity saturation and ballistic transport. Many device structures have been proposed to improve MOSFET performance. Some representative advanced structures as well as the nonvolatile semiconductor memory, basically a MOSFET with a multilayer gate structure, will be discussed.

6.1.1 Field-Effect Transistors: Family Tree

The MOSFET is the main member of the family of field-effect transistors. A distinction between the field-effect transistor (FET) and the potential-effect transistor (PET) is warranted here. A transistor in general is a three-terminal device where the channel resistance between two of the contacts is controlled by the third (MOSFETs have the fourth terminal as contact to the substrate). The difference between the FET and the PET is the way the control is coupled to the channel. As shown in Fig. 2, in an FET, the channel is controlled *capacitively* by an electric field (hence the name *field-effect*), and in a PET, the channel's potential is accessed directly (hence the name *potential-effect*). Conventionally in FETs, the channel carriers flow from the source

Sandisk Technologies, Inc. - Ex. 1042, Page 000012

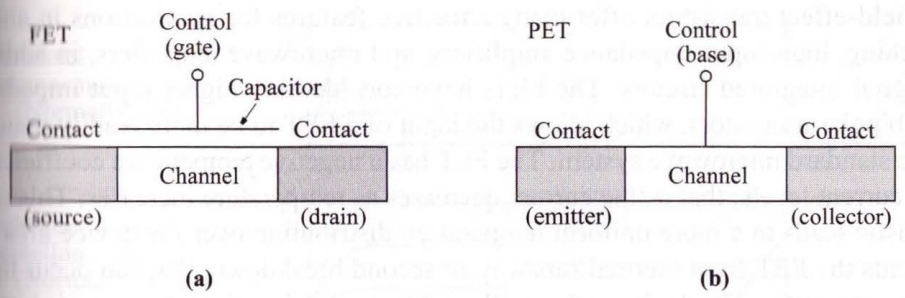


Fig. 2 Distinction between (a) field-effect transistor (FET) and (b) potential-effect transistor (PET).

to the drain, and the control terminal is called the gate, whereas in PETs, these corresponding terminals are called the emitter, collector, and base, respectively. The bipolar transistor is a good representative of the PETs.

A family tree of field-effect transistors is shown in Fig. 3 The three first-level main members are IGFET (insulated-gate FET), JFET (junction FET), and MESFET (metal-semiconductor FET). They are distinguished by the way the gate capacitor is formed. In an IGFET, the gate capacitor is an insulator. In a JFET or a MESFET, the capacitor is formed by the depletion layer of a *p-n* junction or a Schottky barrier, respectively. In the branch of IGFET, we further divide it into MOSFET/MISFET (metal-insulator-semiconductor FET) and HFET (heterojunction FET). In the MOSFET, specifically the insulator is a grown oxide layer, whereas in the MISFET the insulator is a deposited dielectric. In the HFET branch, the gate material is a high-bandgap semiconductor layer grown as a heterojunction which acts as an insulator. Although MOSFETs have been made with various semiconductors such as Ge,[18] Si, and GaAs,[19] and use various oxides and insulators such as SiO_2, Si_3N_4, and Al_2O_3, the most-important system is the SiO_2-Si combination. Hence most of the results in this chapter are obtained from the SiO_2-Si system. The other members, JFETs, MESFETs, and HFETs, will be considered in the following chapter.

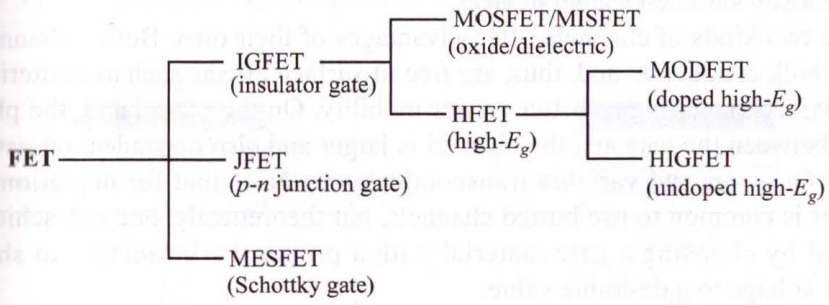


Fig. 3 Family tree of field-effect transistors (FETs).

Sandisk Technologies, Inc. - Ex. 1042, Page 000013

Field-effect transistors offer many attractive features for applications in analog switching, high-input-impedance amplifiers, and microwave amplifiers, in addition to digital integrated circuits. The FETs have considerably higher input impedance than bipolar transistors, which allows the input of a FET to be more readily matched to the standard microwave system. The FET has a negative temperature coefficient at high current levels; that is, the current decreases as temperature increases. This characteristic leads to a more uniform temperature distribution over the device area and prevents the FET from thermal runaway or second breakdown, that can occur in the bipolar transistor. The device is thermally stable, even when the active area is large or when many devices are connected in parallel. Because there is no forward-biased p-n junctions, FETs do not suffer from minority-carrier storage and, consequently, have higher large-signal switching speeds. In addition, the devices are basically square-law or linear devices; intermodulation and cross-modulation products are smaller than those of bipolar transistors.

6.1.2 Versions of Field-Effect Transistors

There are many ways to categorize the versions of FETs. First, according to the type of channel carriers, we have n-channel and p-channel devices. n-channels are formed by electrons and are more conductive with more positive gate bias, while p-channels are formed by holes and are more conductive with more negative gate bias. Furthermore, it is important to describe the state of the transistor with zero gate bias. FETs are called enhancement-mode, or normally-off, if at zero gate bias the channel conductance is very low and we must apply a gate voltage to form a conductive channel. The counterpart is called depletion-mode, or normally-on, when the channel is conductive with zero gate bias and we must apply a gate voltage to turn the transistor off. These four combinations with their I-V characteristics are summarized in Fig. 4.

It is important also to point out the nature of the channel in more details. According to Fig. 5, a channel can be formed by a surface inversion layer or a bulk buried layer. The surface inversion channel is a two-dimensional charge sheet of thickness in the order of 5 nm. The buried channel is much thicker, comparable to the depletion width since when the transistor is turned off, the channel is totally consumed by the surface depletion layer. In the FET family, MESFETs and JFETs are always buried-channel devices, while MODFETs are surface-channel devices. MOSFETs and MISFETs can have both kinds of channels in parallel, but in practice, they are mostly surface-channel devices.

These two kinds of channels offer advantages of their own. Buried channels are based on bulk conduction and, thus, are free of surface effects such as scattering and surface defects, resulting in better carrier mobility. On the other hand, the physical distance between the gate and the channel is larger and also dependent on gate bias, leading to a lower and variable transconductance. Note that for depletion-mode devices, it is common to use buried channels, but theoretically, one can achieve the same goal by choosing a gate material with a proper work function to shift the threshold voltage to a desirable value.

Sandisk Technologies, Inc. - Ex. 1042, Page 000014

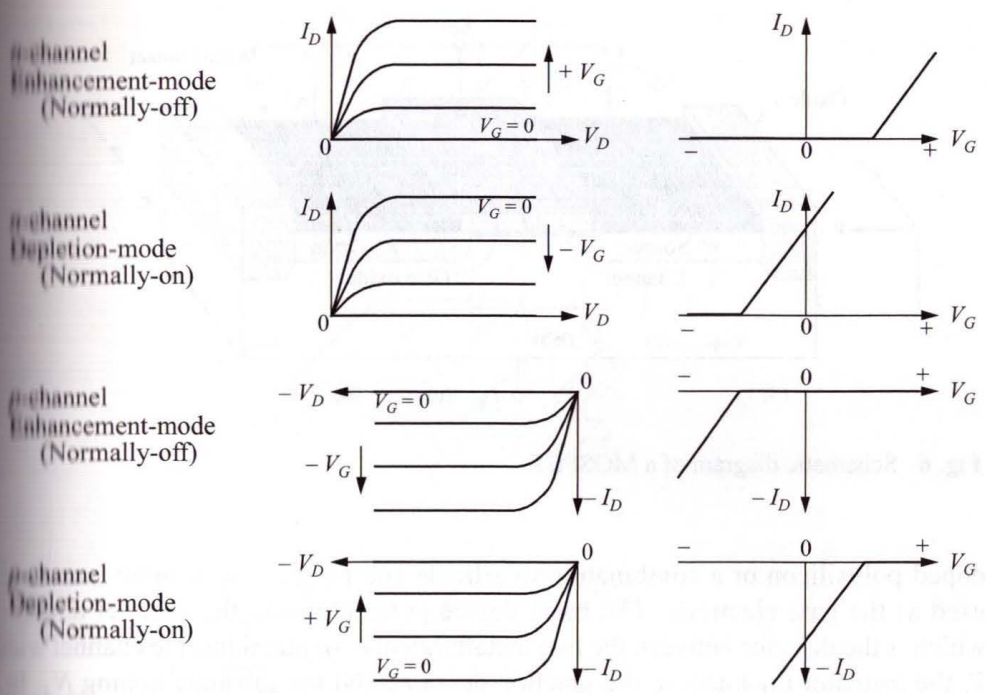


Fig. 4 Versions of MOSFETs; their output and transfer characteristics.

6.2 BASIC DEVICE CHARACTERISTICS

The basic structure of a MOSFET is illustrated in Fig. 6. Throughout this chapter we assume the channel carriers are electrons—an *n*-channel device. All discussion and equations will be applicable to the counterpart *p*-channel devices with appropriate substitution of parameters and the reversal of polarity of the applied voltages. A common MOSFET is a four-terminal device that consists of a *p*-type semiconductor substrate into which two n^+-regions, the source and drain, are formed, usually by ion implantation. The SiO_2 gate dielectric is formed by thermal oxidation of Si for a high-quality SiO_2-Si interface. The metal contact on the insulator is called the gate; heavily

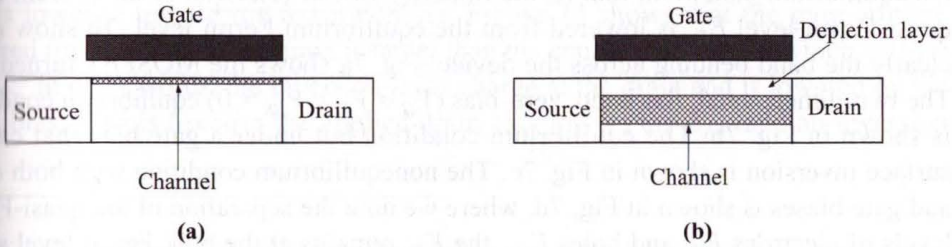


Fig. 5 FET channels: (a) surface inversion channel and (b) buried channel.

Sandisk Technologies, Inc. - Ex. 1042, Page 000015

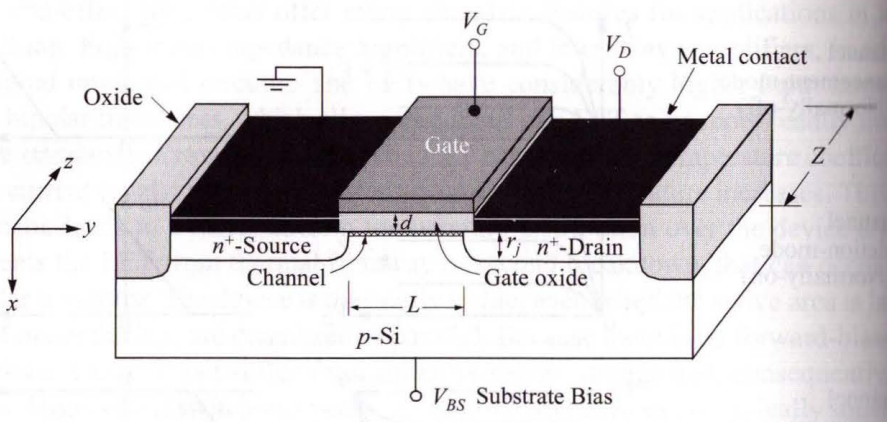


Fig. 6 Schematic diagram of a MOSFET.

doped polysilicon or a combination of silicide and polysilicon is more commonly 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) or a trench filled with insulator to electrically isolate it from adjacent devices.

The source contact will be used as the voltage reference throughout this chapter. When ground or a low voltage is applied to the gate, the main channel is shut off, and the source-to-drain electrodes correspond to two p-n junctions connected back to back. 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 be at the reference voltage or reverse biased; this substrate voltage will also affect the channel conductance.

6.2.1 Inversion Charge in Channel

When a voltage is applied across the source-drain contacts, the MOS structure is in a nonequilibrium condition; that is, the minority-carrier (electron in the present case) quasi-Fermi level E_{Fn} is lowered from the equilibrium Fermi level. To show more clearly the band bending across the device, Fig. 7a shows the MOSFET turned 90°. The two-dimensional, flat-band, zero-bias ($V_G = V_D = V_{BS} = 0$) equilibrium condition is shown in Fig. 7b. The equilibrium condition but under a gate bias that causes surface inversion is shown in Fig. 7c. The nonequilibrium condition with both drain and gate biases is shown in Fig. 7d, where we note the separation of the quasi-Fermi levels of electrons E_{Fn} and holes E_{Fp}; the E_{Fp} remains at the bulk Fermi level while

Sandisk Technologies, Inc. - Ex. 1042, Page 000016

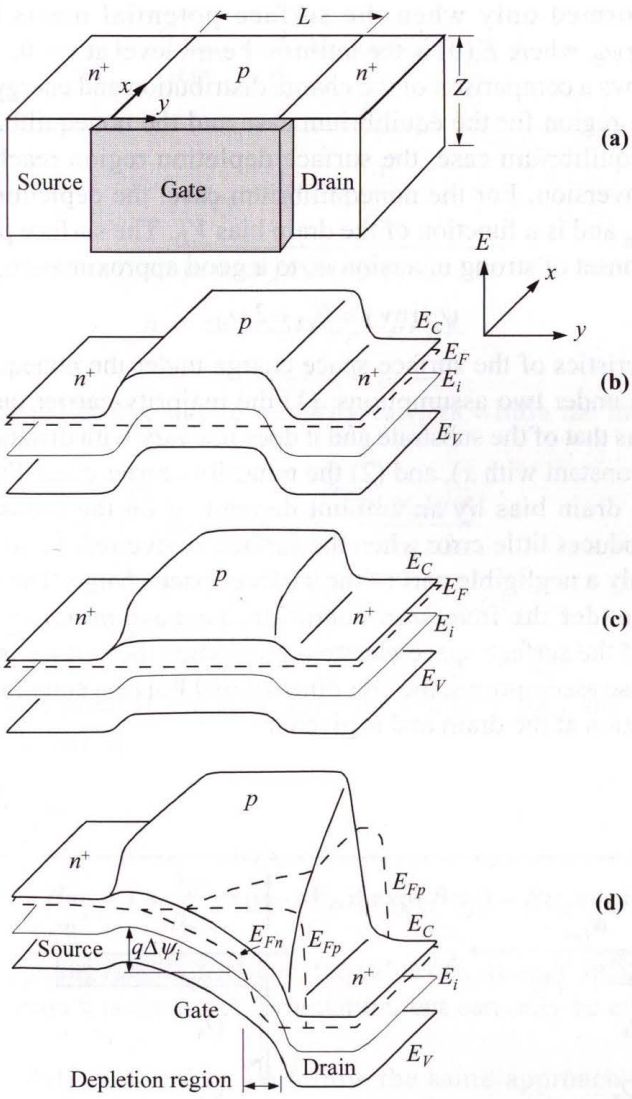


Fig. 7 Two-dimensional band diagram of an *n*-channel MOSFET. (a) Device configuration. (b) Flat-band zero-bias equilibrium condition. (c) Equilibrium condition ($V_D = 0$) under a positive gate bias. (d) Nonequilibrium condition under both gate and drain biases. (After Ref. 20.)

E_{Fn} is lowered toward the drain contact. Figure 7d shows that the gate voltage required for inversion at the drain is larger than the equilibrium case in which $\psi_s(\text{inv}) = 2\psi_B$;[#] in other words, the inversion-layer charge at the drain end is lowered by the drain bias. This is because the applied drain bias lowers the E_{Fn}, and an inversion

[#] The common assumption of $\psi_s = 2\psi_B$ is for the onset of weak inversion. For strong inversion, ψ_s can be larger by a few kT.[15] This can be understood from Fig. 5 of Chapter 4.

Sandisk Technologies, Inc. - Ex. 1042, Page 000017

layer can be formed only when the surface potential meets the criteria of $[E_{Fn} - E_i(0)] > q\psi_B$, where $E_i(0)$ is the intrinsic Fermi level at $x = 0$.

Figure 8 shows a comparison 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, the surface depletion region reaches a maximum width W_{Dm} at inversion. For the nonequilibrium case, the depletion-layer width is deeper than W_{Dm} and is a function of the drain bias V_D. The surface potential $\psi_s(y)$ at the drain at the onset of strong inversion is, to a good approximation, given by

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

The characteristics of the surface space charge under the nonequilibrium condition are derived under two assumptions; (1) the majority-carrier quasi-Fermi level E_{Fp} is the same as that of the substrate and it does not vary with distance from the bulk to the surface (constant with x), and (2) the minority-carrier quasi-Fermi level E_{Fn} is lowered by the drain bias by an amount dependent on the y-position. 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 the surface space-charge region at the drain end is given by

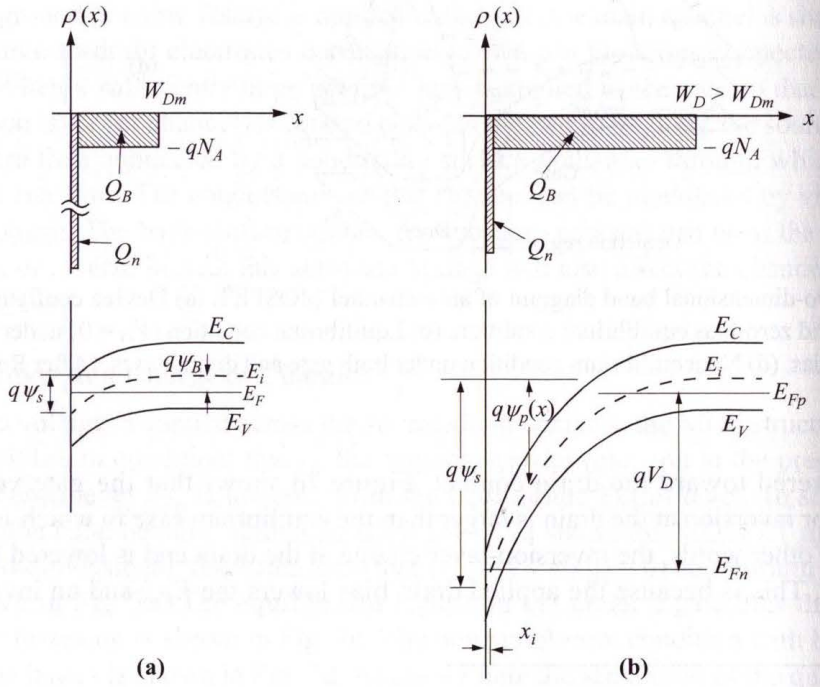


Fig. 8 Comparison of charge distribution and energy-band variation of an inverted p-region in (a) equilibrium and (b) nonequilibrium at the drain end. (After Ref. 21.)

Sandisk Technologies, Inc. - Ex. 1042, Page 000018

$$\frac{d^2\psi_p}{dx^2} = \frac{q}{\varepsilon_s}(N_A - p + n) \tag{2}$$

where

$$p_{po} = N_A = \frac{n_i^2}{n_{po}} \tag{3}$$

$$p = N_A \exp(-\beta\psi_p) \tag{4}$$

$$n = n_{po}\exp(\beta\psi_p - \beta V_D), \tag{5}$$

and $\beta = q/kT$.

Conceptually the charge due to minority carriers within the inversion layer, is given by

charge due to minority carriers

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

$$= q\int_{\psi_s}^{\psi_B} \frac{n_{po}\exp(\beta\psi_p - \beta V_D)d\psi_p}{(\sqrt{2}kT/qL_D)F(\beta\psi_p, V_D, n_{po}/p_{po})} \tag{6}$$

where x_i denotes the point at which $q\psi_p(x) = E_{Fn} - E_i(x) = q\psi_B$, and the function F is defined as (see Chapter 4)

$$F\left(\beta\psi_p, V_D, \frac{n_{po}}{p_{po}}\right) \equiv$$

$$\sqrt{\exp(-\beta\psi_p) + \beta\psi_p - 1 + \frac{n_{po}}{p_{po}}\exp(-\beta V_D)[\exp(\beta\psi_p) - \beta\psi_p\exp(\beta V_D) - 1]}. \tag{7}$$

For the practical doping ranges in silicon, the value of x_i is quite small, of the order of 3 to 30 nm. Equation 6 is the exact formulation, but can only be evaluated numerically.

To get an analytical solution we follow the same approach as in Chapter 4 (Eq. 14). The surface electric field in the x-direction at the drain end is given by

$$\mathcal{E}_s = -\left.\frac{d\psi_p}{dx}\right|_{x=0} = \pm\frac{\sqrt{2}kT}{qL_D}F\left(\beta\psi_s, V_D, \frac{n_{po}}{p_{po}}\right), \tag{8}$$

and the total semiconductor surface charge is then obtained from Gauss' law

total surface charge

$$Q_s = -\varepsilon_s\mathcal{E}_s = \mp\frac{\sqrt{2}\varepsilon_s kT}{qL_D}F\left(\beta\psi_s, V_D, \frac{n_{po}}{p_{po}}\right), \tag{9}$$

where the Debye length is

$$L_D \equiv \sqrt{\frac{kT\varepsilon_s}{N_A q^2}}. \tag{10}$$

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

Sandisk Technologies, Inc. - Ex. 1042, Page 000019

inversion charge after strong inversion

$$Q_n = Q_s - Q_B \tag{11}$$

where the depletion bulk charge is

$$Q_B = -qN_A W_D = -\sqrt{2qN_A\varepsilon_s(V_D + 2\psi_B)}. \tag{12}$$

From Eqs. 9, 11, and 12, the inversion charge Q_n at the drain end can be simplified to

$$|Q_n| \approx \sqrt{2}qN_A L_D\left[\sqrt{\beta\psi_s + \left(\frac{n_{po}}{p_{po}}\right)\exp(\beta\psi_s - \beta V_D)} - \sqrt{\beta\psi_s}\right]. \tag{13}$$

This solution is still difficult to use because at strong inversion, Q_n is very sensitive to the surface potential ψ_s (see Fig. 5 in Chapter 4). Another shortcoming is that the relationship to the terminal bias, that is V_G, is still missing. The charge-sheet model discussed in the following section is simpler and much more useful for deriving the *I-V* characteristics of MOSFETs.

Charge-Sheet Model. In the charge-sheet model,[22] under strong-inversion conditions, the inversion layer is treated as a charge sheet with zero thickness ($x_i = 0$). Consequently, this assumption implies that the potential drop across this charge sheet is also zero. These assumptions do introduce error but within an acceptable level. From Gauss' law, the boundary conditions on both sides of the charge sheet are:

$$\mathscr{E}_{ox}\varepsilon_{ox} = \mathscr{E}_s\varepsilon_s - Q_n. \tag{14}$$

In order to express $Q_n(y)$ throughout the channel, the surface potential is generalized from Eq. 1 to

$$\psi_s(y) \approx \Delta\psi_i(y) + 2\psi_B, \tag{15}$$

where $\Delta\psi_i$ is the channel potential with respect to the source end;

$$\Delta\psi_i(y) \equiv \frac{E_i(x=0,y=0) - E_i(x=0,y)}{q}, \tag{16}$$

(see label in Fig. 7d) and is equal to V_D at the drain end. Note that the electric fields can be expressed as:

$$\mathscr{E}_{ox} = \frac{V_G - \psi_s}{d} = \frac{V_G - (\Delta\psi_i + 2\psi_B)}{d}, \tag{17}$$

$$\mathscr{E}_s = \sqrt{\frac{2qN_A(\Delta\psi_i + 2\psi_B)}{\varepsilon_s}}. \tag{18}$$

In Eq. 17, an ideal MOS system with zero work-function difference is assumed. Equation 18 is simply the maximum field at the edge of the depletion region. Combining Eqs. 14–18 and using $C_{ox} = \varepsilon_{ox}/d$ we obtain

channel charge

$$|Q_n(y)| = [V_G - \Delta\psi_i(y) - 2\psi_B]C_{ox} - \sqrt{2\varepsilon_s qN_A[\Delta\psi_i(y) + 2\psi_B]}. \tag{19}$$

This final form will be used as the channel charge responsible for the current conduction.

Sandisk Technologies, Inc. - Ex. 1042, Page 000020

6.2.1 Current-Voltage Characteristics

We shall now derive the basic MOSFET characteristics under the following idealized conditions: (1) The gate structure corresponds to an ideal MOS capacitor as defined in Chapter 4; that is, there are no interface traps nor mobile oxide charge; (2) only drift current will be considered; (3) doping in the channel is uniform; (4) reverse leakage current is negligible; and (5) 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). This last condition corresponds to the so-called gradual-channel approximation. Note that in condition-(1), the requirements of zero fixed oxide charge and work-function difference are removed, and their effects are included in a flat-band voltage V_{FB} required by the gate to produce the flat-band condition. Consequently V_G is replaced by $V_G - V_{FB}$ for the inversion charge, giving

$$|Q_n(y)| = [V_G - V_{FB} - \Delta\psi_i(y) - 2\psi_B]C_{ox} - \sqrt{2\varepsilon_s q N_A[\Delta\psi_i(y) + 2\psi_B]}. \quad (20)$$

Under such idealized conditions, the channel current at any y-position is given by

$$I_D(y) = Z|Q_n(y)|v(y) \quad (21)$$

where $v(y)$ is the average carrier velocity. Since the current has to be continuous and constant throughout the channel, integration of Eq. 21 from 0 to L gives

$$I_D = \frac{Z}{L}\int_0^L |Q_n(y)|v(y)dy. \quad (22)$$

The carrier velocity $v(y)$ is a function of the y-position since the longitudinal field $\mathscr{E}_y(y)$ is a variable. Because of this, the relationship between $v(y)$ and $\mathscr{E}_y(y)$ is important to evaluate Eq. 22. We first consider the case where $\mathscr{E}_y(y)$ is low such that the mobility is constant. For shorter channel lengths, higher field causes velocity saturation and ultimately ballistic transport. These interesting effects will be discussed later.

Constant Mobility. Under this assumption, substitutions of $v = \mathscr{E}\mu$ and Eq. 20 into Eq. 22 gives

$$I_D = \frac{Z\mu_n}{L}\int_0^L |Q_n(y)|\mathscr{E}(y)dy = \frac{Z\mu_n}{L}\int_0^L |Q_n(y)|\frac{d\Delta\psi_i(y)}{dy}dy = \frac{Z\mu_n}{L}\int_0^{V_D} |Q_n(\Delta\psi_i)|d\Delta\psi_i$$

$$= \frac{Z}{L}\mu_n C_{ox}\left\{\left(V_G - V_{FB} - 2\psi_B - \frac{V_D}{2}\right)V_D - \frac{2}{3}\frac{\sqrt{2\varepsilon_s q N_A}}{C_{ox}}[(V_D + 2\psi_B)^{3/2} - (2\psi_B)^{3/2}]\right\}. \quad (23)$$

Equation 23 predicts that for a given V_G the drain current first increases linearly with drain voltage (the linear region), then gradually levels off (the nonlinear region), and finally approaching a saturated value (the saturation region). The basic output characteristics of an idealized MOSFET are shown in Fig. 9. The dashed line on the right indicates the locus of the drain voltage (V_{Dsat}) at which the current reaches a maximum value I_{Dsat}. For small V_D, the I_D is linear with V_D. Inbetween the two dashed lines, we designate this as the nonlinear region.

Sandisk Technologies, Inc. - Ex. 1042, Page 000021

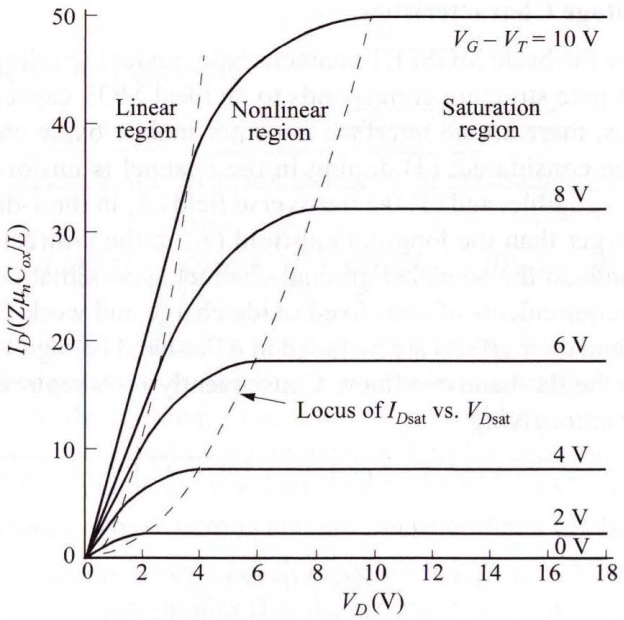


Fig. 9 Idealized drain characteristics (I_D vs. V_D) of a MOSFET. The dashed lines separate the linear, nonlinear, and saturation regions.

A qualitative discussion of the device operation can be helpful, with the aid of Fig. 10. Let us consider that a positive voltage is applied to the gate, large enough to cause an inversion at the semiconductor surface. If a small drain voltage is applied, a current will flow from the source to the drain through the conducting channel. The channel acts as a resistor, and the drain current I_D is proportional to the drain voltage V_D. This is the linear region. As the drain voltage increases, the current deviates from the linear relationship since the charge near the drain end is reduced by the channel potential $\Delta\psi_i$ (Eq. 20). It eventually reaches a point at which the inversion charge at the drain end $Q_n(L)$ is reduced to nearly zero. This location of $Q_n \approx 0$ is called the pinch-off point, Fig. 10b. [In reality $Q_n(L)$ is not zero for current continuity, but small because of its high field and high carrier velocity.] Beyond this drain bias, the drain current remains essentially the same, because for $V_D > V_{Dsat}$, the pinch-off point starts to move toward the source, but the voltage at this pinch-off point remains the same (V_{Dsat}). Thus, the number of carriers arriving at the pinch-off point from the source, and hence the current, remains essentially the same, apart from a decrease in L to the value L' (Fig. 10c). This change of effective channel length will increase the drain current only when the shortened amount is a substantial fraction of the channel length. This will be considered in the section of short-channel effects.

We shall now consider the current equations for the three cases of linear, nonlinear, and saturation regions. In the linear region, with a small V_D, using power series around V_D and taking only the initial terms, Eq. 23 reduces to

Sandisk Technologies, Inc. - Ex. 1042, Page 000022

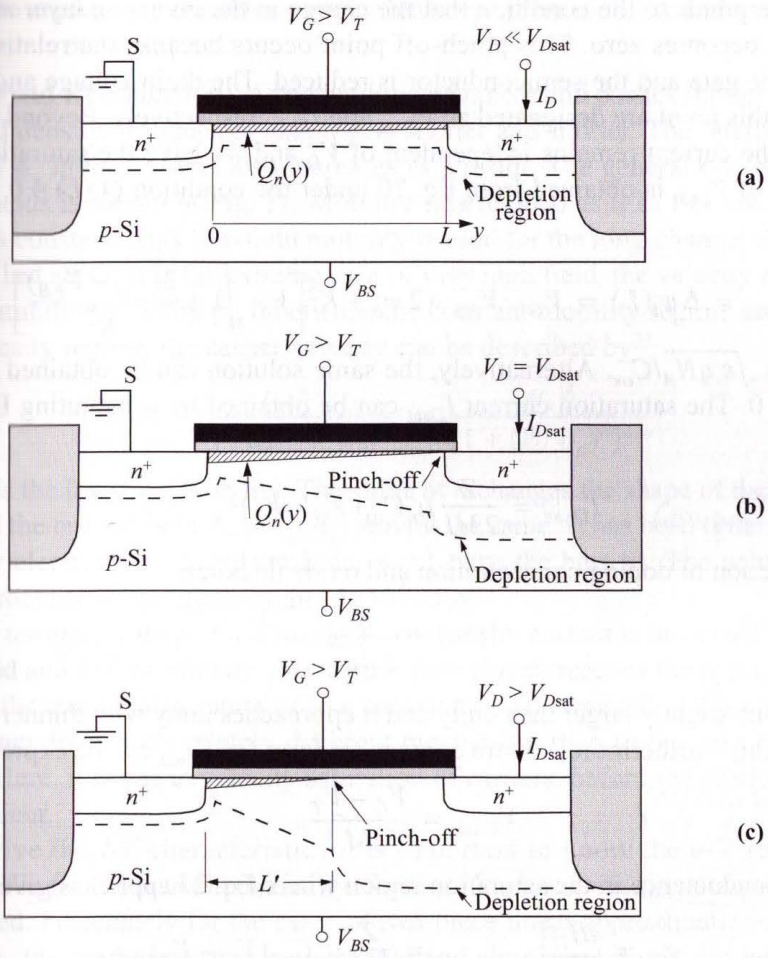


Fig. 10 MOSFET operated (a) in the linear region (low V_D), (b) at onset of saturation, and (c) beyond saturation (effective channel length is reduced).

$$\begin{cases} I_D = \dfrac{Z}{L}\mu_n C_{ox}\left\{\left(V_G - V_{FB} - 2\psi_B - \dfrac{V_D}{2}\right)V_D - \dfrac{2}{3}\dfrac{\sqrt{2\varepsilon_s q N_A}}{C_{ox}}\left(3\sqrt{\dfrac{\psi_B}{2}} V_D\right)\right\} \\ \quad = \dfrac{Z}{L}\mu_n C_{ox}\left(V_G - V_T - \dfrac{V_D}{2}\right)V_D \qquad \text{for} \quad V_D \ll (V_G - V_T) \end{cases} , \tag{24}$$

where V_T is the threshold voltage, one of the most-important parameters, given by

$$V_T = V_{FB} + 2\psi_B + \frac{\sqrt{2\varepsilon_s q N_A (2\psi_B)}}{C_{ox}} . \tag{25}$$

The threshold voltage will be discussed in more details in the next section.

Careful examination of Eq. 23 indicates that the current initially increases but then goes through a peak and then drops with V_D. This drop of current is not physical,

Sandisk Technologies, Inc. - Ex. 1042, Page 000023

but it corresponds to the condition that the charge in the inversion layer at the drain end $Q_n(L)$ becomes zero. This pinch-off point occurs because the relative voltage between the gate and the semiconductor is reduced. The drain voltage and the drain current at this point are designated as V_{Dsat} and I_{Dsat}, respectively. Beyond the pinch-off point the current remains independent of V_D and we have the saturation region. The value of V_{Dsat} is obtained from Eq. 20 under the condition $Q_n(L) = 0$. The solution yields

$$V_{Dsat} = \Delta\psi_i(L) = V_G - V_{FB} - 2\psi_B + K^2\left[1 - \sqrt{1 + \frac{2(V_G - V_{FB})}{K^2}}\right] \tag{36}$$

where $K \equiv \sqrt{\varepsilon_s q N_A}/C_{ox}$. Alternatively, the same solution can be obtained by setting $dI_D/dV_D = 0$. The saturation current I_{Dsat} can be obtained by substituting Eq. 26 into Eq. 23:

$$I_{Dsat} = \frac{Z}{2ML}\mu_n C_{ox}(V_G - V_T)^2 . \tag{37}$$

M is a function of doping concentration and oxide thickness

$$M \equiv 1 + \frac{K}{2\sqrt{\psi_B}} . \tag{38}$$

It has a value slightly larger than unity and it approaches unity with thinner oxide and lower doping. Furthermore, a more convenient form for V_{Dsat} can be expressed as

$$V_{Dsat} = \frac{V_G - V_T}{M} . \tag{39}$$

The transconductance in the saturation region where Eq. 27 applies is given by

$$g_m = \left.\frac{dI_D}{dV_G}\right|_{V_D > V_{Dsat}} = \frac{Z}{ML}\mu_n C_{ox}(V_G - V_T) . \tag{40}$$

It can be seen here that in this saturation region, for constant mobility, the current is a square-law function according to Eq. 27, indicated by the increasing current steps between gate bias shown in Fig. 9.

Finally, the nonlinear region inbetween these two extreme cases can be described well by

$$I_D = \frac{Z}{L}\mu_n C_{ox}\left(V_G - V_T - \frac{MV_D}{2}\right)V_D . \tag{41}$$

Equation 20 for the inversion charge is an exact expression. An approximation of the following form, taking advantage of the definition of threshold voltage, can be made:

$$|Q_n(y)| = C_{ox}[V_G - V_T - M\Delta\psi_i(y)] . \tag{42}$$

Substitution of this into Eq. 22 yields a general expression which is the same as Eq. 31 for the whole three regions of operation. As seen, the only slight departure from the previous results lies in the linear region. This simplified charge expression

Sandisk Technologies, Inc. - Ex. 1042, Page 000024

is helpful to analyze the conditions under field-dependent mobility and velocity saturation which is discussed next.

Velocity-Field Relationship. As technology advances and pushes for device performance and density, the channel length gets shorter and shorter. The internal longitudinal field $\mathscr{E}_y$ in the channel also increases as a result. The general v-$\mathscr{E}$ relationship for high fields is shown in Fig. 11. Mobility μ is defined as $v/\mathscr{E}$. For low fields, the mobility is constant. This low-field mobility is used for the long-channel characteristics in the last section. In the extreme case of very high field, the velocity approaches a value, saturation velocity v_s. Inbetween the constant-mobility regime and the saturation-velocity regime, the carrier velocity can be described by[23]

$$v(\mathscr{E}) = \frac{\mu_n \mathscr{E}}{[1 + (\mu_n \mathscr{E}/v_s)^n]^{1/n}} = \frac{\mu_n \mathscr{E}}{[1 + (\mathscr{E}/\mathscr{E}_c)^n]^{1/n}} \tag{33}$$

where μ_n is the low-field mobility. The value of n changes the shape of the curve, but v_s and the critical field $\mathscr{E}_c$ ($\equiv v_s/\mu_n$) remain the same. It has been observed that in silicon for electrons $n = 2$ and for holes $n = 1$ have the best fit. The value of v_s for silicon at room temperature is around 1×10^7 cm/s.

As the terminal voltage V_D is increased from zero, current is increased because of higher field and higher velocity. Eventually the velocity reaches the maximum value of v_s, and the current also saturates to a constant value. Notice that this current saturation comes from a completely different mechanism than in the case of constant mobility. Here, it is due to velocity saturation of carriers, before the pinch-off condition can occur.

To derive the I-V characteristics it is important to know the v-$\mathscr{E}$ relationship (Fig. 11). We find that mathematically, for Eq. 33 with $n = 2$, the analysis is rather complicated. Fortunately for the cases of two-piece linear approximation and Eq. 33 with $n = 1$, the mathematics is manageable and simple solutions can be obtained. Since these two extremes mostly cover the realistic bounds for different kinds of carriers, we will consider both assumptions.

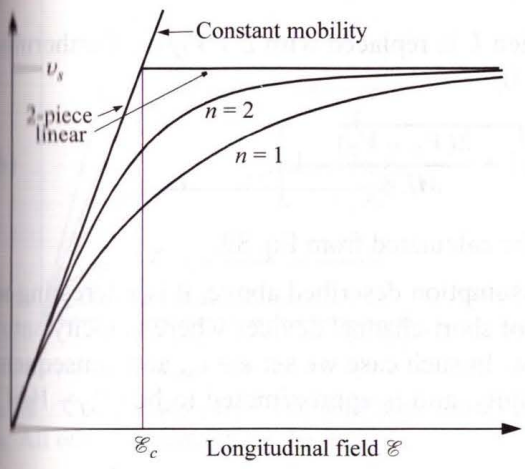


Fig. 11 v-$\mathscr{E}$ relationship (Eq. 33) for $n = 1$ and 2, and two-piece linear approximation. The critical field $\mathscr{E}_c \equiv v_s/\mu$, where μ is low-field mobility, is also indicated.

Sandisk Technologies, Inc. - Ex. 1042, Page 000025

Field-Dependent Mobility: Two-Piece Linear Approximation. In the two-piece linear approximation, the constant-mobility model is valid up to the point when the maximum field near the drain exceeds $\mathscr{E}_c$. Conversely, Eq. 23 is valid up to a new V_{Dsat} value which occurs earlier than the constant-mobility model, so the only task is to find that point. Substituting Eq. 32 into Eq. 21 gives

$$I_D(y) = ZC_{ox}\mu_n\mathscr{E}(V_G - V_T - M\Delta\psi_i). \tag{34}$$

Since we know the maximum field occurs at the drain, the current will saturate when the drain bias is increased to a value where $\mathscr{E}(L) = \mathscr{E}_c$. Equation 34 leads to the following condition:

$$I_{Dsat} = ZC_{ox}\mu_n\mathscr{E}_c(V_G - V_T - MV_{Dsat}). \tag{35}$$

Here we need one more equation to solve for two unknowns. Using Eqs. 32 and 21 we obtain an expression similar to Eq. 31 given as

$$I_{Dsat} = \frac{ZC_{ox}\mu_n}{L}\left(V_G - V_T - \frac{MV_{Dsat}}{2}\right)V_{Dsat}. \tag{36}$$

Equating Eqs. 36 and 35, we can solve for V_{Dsat} as

$$V_{Dsat} = L\mathscr{E}_c + \frac{V_G - V_T}{M} - \sqrt{(L\mathscr{E}_c)^2 + \left(\frac{V_G - V_T}{M}\right)^2}. \tag{37}$$

Since V_{Dsat} here is always smaller than $(V_G - V_T)/M$ (Eq. 29), the field-dependent mobility always gives a lower I_{Dsat}.

Field-Dependent Mobility: Empirical Formula. Next we consider the v-$\mathscr{E}$ relationship of Eq. 33 with $n = 1$. Substituting this into Eq. 21 gives

$$I_D\left(\mathscr{E}_c + \frac{d\Delta\psi_i}{dy}\right) = ZC_{ox}\mu_n\mathscr{E}_c(V_G - V_T - M\Delta\psi_i)\frac{d\Delta\psi_i}{dy}. \tag{38}$$

Notice that the right-hand side of the equation is similar to the constant-mobility model. Integrating the above equation from source to drain gives

$$I_D = \frac{ZC_{ox}\mu_n\mathscr{E}_c}{L\mathscr{E}_c + V_D}\left(V_G - V_T - \frac{MV_D}{2}\right)V_D. \tag{39}$$

This equation is similar to Eq. 31 when L is replaced with $L + V_D/\mathscr{E}_c$. Furthermore, V_{Dsat} is obtained by setting $dI_D/dV_D = 0$,

$$V_{Dsat} = L\mathscr{E}_c\left[\sqrt{1 + \frac{2(V_G - V_T)}{ML\mathscr{E}_c}} - 1\right]. \tag{40}$$

Again, once V_{Dsat} is known, I_{Dsat} can be calculated from Eq. 39.

Velocity Saturation. Using either assumption described above, it is interesting and insightful to look at the extreme case of short-channel devices where velocity saturation completely limits the current flow. In such case we set $v = v_s$, and consequently Q_n has to be fixed for current continuity, and is approximated to be $(V_G - V_T)C_{ox}$. Equation 22 then becomes

Sandisk Technologies, Inc. - Ex. 1042, Page 000026

$$I_{Dsat} = \frac{Z}{L}\int_0^L |Q_n(y)|v(y)dy = \frac{Z}{L}|Q_n|v_s L$$
$$= Z(V_G - V_T)C_{ox}v_s \ . \tag{41}$$

The transconductance becomes

$$g_m \equiv \frac{dI_{Dsat}}{dV_G} = ZC_{ox}v_s \tag{42}$$

and it is independent of gate bias.

To compare models based on constant mobility and velocity saturation, we show I-V curves of identical devices in Fig. 12. Several observations can be made. First, I_{Dsat} and V_{Dsat} are both lowered by velocity saturation, while the linear regions remain similar. The g_m (which is the current difference between V_G steps) also becomes a constant, independent of V_G. Finally Eq. 41 shows an interesting phenomena that the saturation current no longer depends on the channel length.

Experimental data confirm that such simple theory is quite satisfactory. In reality, as Fig. 11 indicates, the carrier velocity never reaches exactly v_s. Also the lateral field is not uniform throughout the channel. It is more difficult for the lower field near the source to reach $\mathscr{E}_c$ and that presents a bottle-neck for the maximum current. A better agreement can often be reached by adding a pre-factor with a value of $\approx 0.5 - 1.0$ for Eqs. 41 and 42.

Ballistic Transport. In the above section, velocity saturation is a steady-state, equilibrium phenomena at high field, when many scattering events are allowed to happen. However, in ultra-short channel lengths whose dimensions are on the order of or shorter than the mean free path, channel carriers do not suffer from scattering. They can gain energy from the field without losing it to the lattice through scattering, and can acquire a velocity much higher than the saturation velocity. This effect is called

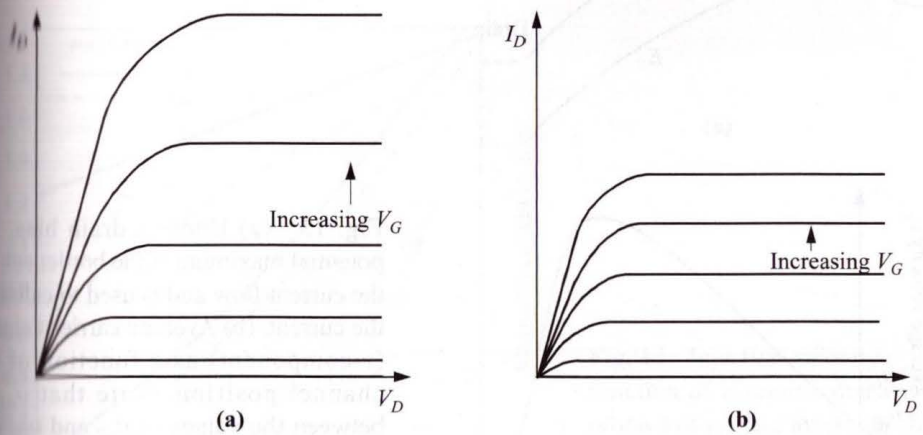


Fig. 12 Comparison of *I-V* characteristics for (a) constant mobility and (b) velocity saturation. All other parameters are the same.

Sandisk Technologies, Inc. - Ex. 1042, Page 000027

ballistic transport (or velocity overshoot by some), introduced in Section 1.5.3. The ballistic transport is important since it points out that the current and transconductance can be higher than that of saturation velocity, giving an additional incentive for shrinking the channel length. The basic theory and insight in this topic are presented in Refs. 24–28.

Computer simulations show that in these devices, the field and velocity are very nonuniform. These quantities are qualitatively shown in Fig. 13. Notice that the longitudinal field along the channel (dE_C/dy) varies monotonically, being the highest at the drain end. Ballisticity always starts at the drain end, where the velocity can exceed the value of saturation velocity v_s. (For silicon at room temperature, $v_s \approx v_{th}$, the thermal velocity.) At positions closer to the source, the velocity decreases. In order to have current continuity, the channel potential and inversion charge must adjust themselves such that the product of velocity and charge would remain constant throughout the channel. By such argument, the bottleneck for the current flow, at the extreme of ultra-short channel length, would be at the position of maximum charge and minimum field, which means the potential maximum near the source end, indicated in Fig. 13a.

In analyzing the saturation current in the ballistic regime, we go back to the general equation of Eq. 22, and apply it to this maximum-potential point. We start with the generalized form

$$I_{Dsat} = Z|Q_n|v_{eff} \tag{41}$$

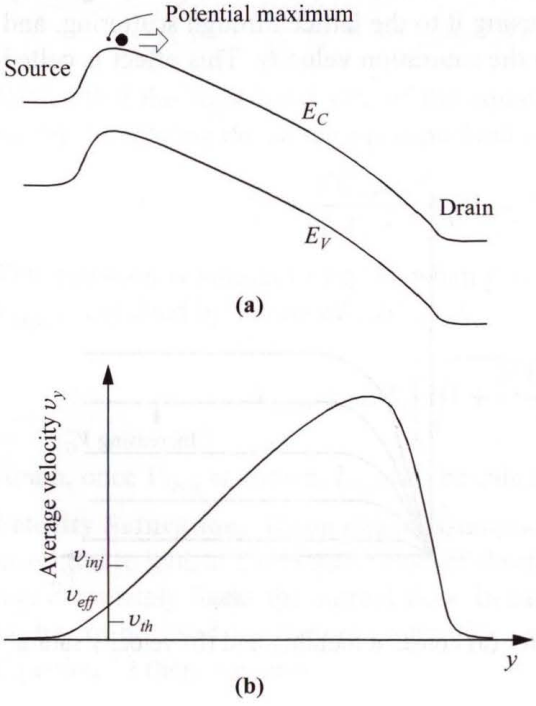


Fig. 13 (a) Under a drain bias, the potential maximum is the bottleneck for the current flow and is used to calculate the current. (b) Average carrier velocity (y-component) as a function of the channel position. Note that v_{eff} is between the values of v_{inj} and v_{th}, and that the maximum velocity near the drain can be higher than v_{inj}.

Sandisk Technologies, Inc. - Ex. 1042, Page 000028

where $|Q_n|$ has the maximum value at the source as $C_{ox}(V_G - V_T)$, and v_{eff} is an effective average carrier velocity that should match the final experimental saturation current. So in this simple expression, the only critical parameter is v_{eff}.

The maximum value of v_{eff}, according to classical thermal equilibrium, is simply the thermal velocity v_{th} [$= (2kT/\pi m^*)^{1/2}$]. Close examination of the system reveals that for higher inversion charge density, the random velocity can exceed this thermal limit. This is a quantum-mechanical effect, called carrier degeneracy,[25] where the mean carrier energy is pushed to a higher state than the thermal energy. This higher value is called the injection velocity v_{inj}, related to the Fermi energy with respect to the quantized energy E_n inside the potential well where carriers reside, given by[24]

$$v_{inj} = \sqrt{\frac{2kT}{\pi m^*}} \frac{F_{1/2}[(E_F - E_n)/kT]}{\ln\{1 + \exp[(E_F - E_n)/kT]\}}, \tag{44}$$

where $F_{1/2}$ is the Fermi-Dirac integral (see Section 1.4.1). With a small inversion charge or $E_F - E_n$, Eq. 44 reduces to $\sqrt{2kT/\pi m^*}$, and $v_{inj} = v_{th}$. If the inversion charge is high, Eq. 44 can be simplified to

$$v_{inj} = \frac{8\hbar}{3m^*}\sqrt{\frac{|Q_n|}{2\pi q}} = \frac{8\hbar}{3m^*}\sqrt{\frac{C_{ox}(V_G - V_T)}{2\pi q}}, \tag{45}$$

and is a function of the inversion charge or gate overdrive. Theoretical v_{inj} as a function of the inversion charge is shown in Fig. 14. The maximum current, which is a product of $Q_n v_{inj}$, gives the ultimate current drive of the ballistic MOSFET and is also plotted in the same figure.

The saturation current of Eq. 43 can be rewritten as

$$\begin{aligned} I_{Dsat} &= r_n Z |Q_n| v_{inj} \\ &= \frac{8 r_n Z \hbar}{3m^*} \frac{[C_{ox}(V_G - V_T)]^{3/2}}{\sqrt{2\pi q}} \end{aligned} \tag{46}$$

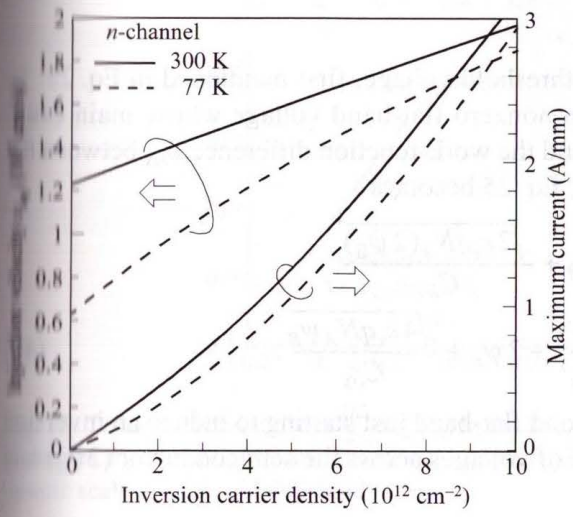


Fig. 14 Injection velocity v_{inj} as a function of inversion density. The product of v_{inj} and inversion charge gives the maximum current. (After Ref. 24.)

Sandisk Technologies, Inc. - Ex. 1042, Page 000029

where r_n is the *index of ballisticity* ($= v_{eff}/v_{inj}$). In the extreme of ballistic transport $r_n = 1$, and it sets the ultimate current drive for $L \to 0$. The transconductance is given by

$$g_m = \frac{4 r_n Z\hbar}{m^*}\sqrt{\frac{C_{ox}(V_G - V_T)}{2\pi q}}. \tag{47}$$

It is seen here that both I_{Dsat} and g_m are independent of channel length L.

The index of ballisticity is also interpreted by back scattering R of channel carriers at the drain back to the source. Furthermore, since mobility is also a consequence of scattering, there should be some relationship between r_n and the low-field mobility μ_n. It has been shown that:[26]

$$\begin{aligned} r_n &= \frac{v_{eff}}{v_{inj}} = \frac{1-R}{1+R} \\ &= \left[\frac{1}{v_{inj}} + \frac{1}{\mu_n \mathscr{E}(0^+)}\right]^{-1}, \end{aligned} \tag{48}$$

where $\mathscr{E}(0^+)$ is the field at a potential kT down from the maximum toward the drain. With this interpretation, some experimental data and simulation trends are nicely explained. It has been seen that at lower temperature, I_{Dsat} is increased, and that at the same temperature, higher low-field mobility always gives higher I_{Dsat} even in the ballistic regime. Both of these can be explained through improvement of μ_n in Eq. 48.

It should be emphasized that in this model, at or near the maximum-potential point, the field is too low to cause ballistic transport, so v_{inj} sets the maximum current even though a location near the drain can have ballistic transport. The high ballistic velocities near the drain cannot produce a higher current than v_{inj} can support, but it helps to achieve this maximum value set by v_{inj} by rebalancing the whole system.

It is interesting to compare the V_G dependence of I_{Dsat} for different channel lengths. In the long-channel, constant-mobility regime, $I_{Dsat} \propto (V_G - V_T)^2$. In the short-channel, saturation-velocity regime, $I_{Dsat} \propto (V_G - V_T)$. And in the limit of ballistic regime, $I_{Dsat} \propto (V_G - V_T)^{3/2}$.

6.2.3 Threshold Voltage

We now return to the discussion of threshold voltage, first mentioned in Eq. 25. To account for the threshold shift from nonzero flat-band voltage whose main cause comes from fixed oxide charges Q_f and the work-function difference ϕ_{ms} between the gate material and the semiconductor, Eq. 25 becomes

$$\begin{aligned} V_T &= V_{FB} + 2\psi_B + \frac{\sqrt{2\varepsilon_s q N_A (2\psi_B)}}{C_{ox}} \\ &= \left(\phi_{ms} - \frac{Q_f}{C_{ox}}\right) + 2\psi_B + \frac{\sqrt{4\varepsilon_s q N_A \psi_B}}{C_{ox}}. \end{aligned} \tag{49}$$

Qualitatively, V_T is the gate bias beyond flat-band just starting to induce an inversion charge sheet and is given by the sum of voltages across the semiconductor ($2\psi_B$) and

Sandisk Technologies, Inc. - Ex. 1042, Page 000030

the oxide layer (last term of Eq. 25). The square-root term is the total depletion-layer charge.

When a substrate bias is applied (negative for n-channel or p-substrate), the threshold voltage becomes

$$V_T = V_{FB} + 2\psi_B + \frac{\sqrt{2\varepsilon_s q N_A (2\psi_B - V_{BS})}}{C_{ox}}, \tag{50}$$

and it is shifted by the amount of

$$\Delta V_T = V_T(V_{BS}) - V_T(V_{BS}=0) = \frac{\sqrt{2\varepsilon_s q N_A}}{C_{ox}}(\sqrt{2\psi_B - V_{BS}} - \sqrt{2\psi_B}). \tag{51}$$

In practice it is often necessary to minimize this threshold-voltage shift due to substrate bias. In these cases, low substrate doping and thin oxide thickness are preferred.

To measure the threshold voltage, we use the linear region by applying a small drain bias ($V_D \ll V_G$), and plot I_D versus V_G as shown in Fig. 15a. According to Eq. 24, the extrapolated value at the V_G axis is equal to $V_T + \frac{1}{2}V_D$. Below the threshold voltage, I_D is considered zero in the linear scale, but details can be displayed in the logarithmic scale (Fig. 15b).

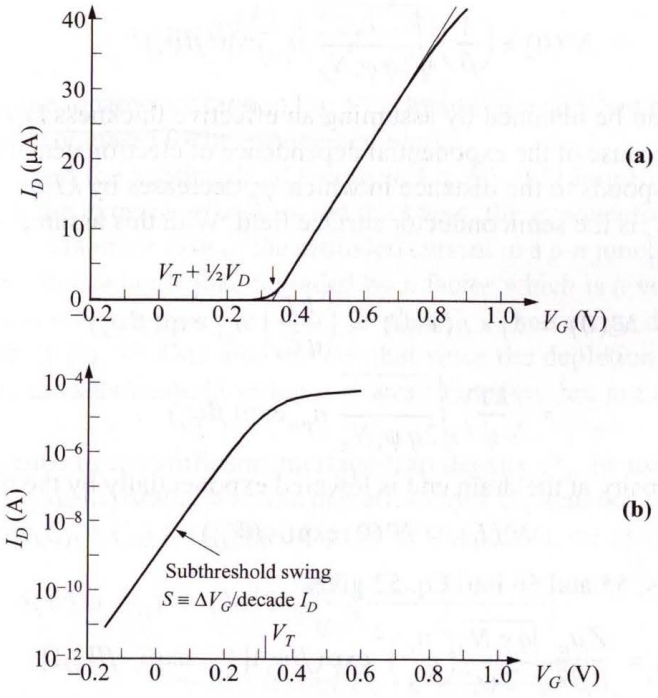


Fig. 15 Transfer characteristics (I_D vs. V_G) in the linear region ($V_D \ll V_G$). (a) I_D in linear scale to deduce V_T. Deviation from linearity at higher V_G is due to lower mobility. (b) I_D in logarithmic scale to show subthreshold swing.

Sandisk Technologies, Inc. - Ex. 1042, Page 000031

6.2.4 Subthreshold Region

When the gate bias is below the threshold and the semiconductor surface is in weak inversion or depletion, the corresponding drain current is called the subthreshold current.[29,30] The subthreshold region tells how sharply the current drops with gate bias and 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.

In weak inversion and depletion, the electron charge is small and, thus, the drift current is low. 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 electron-density gradient in the channel, the diffusion current is given by

$$I_D = -ZqD_n\frac{dN'(y)}{dy} \approx ZqD_n\frac{N'(0)-N'(L)}{L}, \tag{52}$$

where N' is the electron density per unit area, integrated over the depletion width. The electron density at the source end is given by

$$N'(0) = \int_0^{W_D} n(x)\,dx = n_{po}\int_{\psi_s}^{0} \exp(\beta\psi_p)\,d\psi_p. \tag{53}$$

Since the potential distribution inside the depletion region is known, this electron density can be calculated to be[31]

$$N'(0) \approx \left(\frac{1}{\beta}\right)\sqrt{\frac{\varepsilon_s}{2q\psi_s N_A}}\, n_{po}\exp(\beta\psi_s). \tag{54}$$

Similar result can be obtained by assuming an effective thickness (x_i) of the surface charge layer. Because of the exponential dependence of electron density on the potential ψ_p, x_i corresponds to the distance in which ψ_p decreases by kT/q. Therefore, x_i is $kT/q\mathscr{E}_s$ where $\mathscr{E}_s$ is the semiconductor surface field. With this assumption, we get the same expression

$$\begin{aligned} N'(0) &= x_i \times n(x{=}0) = \left(\frac{kT}{q\mathscr{E}_s}\right) n_{po}\exp(\beta\psi_s) \\ &= \left(\frac{kT}{q}\right)\sqrt{\frac{\varepsilon_s}{2q\psi_s N_A}}\, n_{po}\exp(\beta\psi_s) \quad . \end{aligned} \tag{55}$$

The electron density at the drain end is lowered exponentially by the drain bias,

$$N'(L) = N'(0)\exp(-\beta V_D). \tag{56}$$

Substituting Eqs. 55 and 56 into Eq. 52 gives

$$\begin{aligned} I_D &= \frac{Z\mu_n}{L\beta^2}\sqrt{\frac{q\varepsilon_s N_A}{2\psi_s}}\left(\frac{n_i}{N_A}\right)^2 \exp(\beta\psi_s)[1-\exp(-\beta V_D)] \\ &\approx \frac{Z\mu_n}{L\beta^2}\sqrt{\frac{q\varepsilon_s N_A}{2\psi_s}}\left(\frac{n_i}{N_A}\right)^2 \exp(\beta\psi_s) \end{aligned} \tag{57}$$

Sandisk Technologies, Inc. - Ex. 1042, Page 000032

when $V_D \gg kT/q$. Equation 57 indicates that in the subthreshold region the drain current varies exponentially with ψ_s, and for drain voltage V_D larger than $\approx 3kT/q$, the current becomes independent of V_D. Next, in order to relate current to the gate bias, the relationship between V_G and ψ_s is needed.

From Chapter 4 on the MOS capacitor, we have the following relationship (Eq. 33):

$$V_G - V_{FB} = \psi_s + \frac{\sqrt{2\varepsilon_s \psi_s q N_A}}{C_{ox}}. \tag{58}$$

This quadratic equation will not give a simple expression of ψ_s as a function of V_G but once ψ_s is known from Eq. 58, the subthreshold current can be calculated.

The parameter to quantify how sharply the transistor is turned off by the gate voltage is called the subthreshold swing S (inverse of subthreshold slope), defined as the gate-voltage change needed to induce a drain-current change of one order of magnitude. First, from Eq. 58, the relative change of V_G and ψ_s is calculated to be

$$\frac{dV_G}{d\psi_s} = 1 + \frac{1}{C_{ox}}\sqrt{\frac{\varepsilon_s q N_A}{2\psi_s}} = \frac{C_{ox} + C_D}{C_{ox}}. \tag{59}$$

By definition, the subthreshold swing can now be calculated:

$$S \equiv (\ln 10)\frac{dV_G}{d(\ln I_D)} = (\ln 10)\frac{dV_G}{d(\beta\psi_s)}$$

$$= (\ln 10)\left(\frac{kT}{q}\right)\left(\frac{C_{ox} + C_D}{C_{ox}}\right). \tag{60}$$

Note that ψ_s in the square-root term in Eq. 57 is treated as a constant since it is a much weaker function compared to the exponential term.

Having derived the subthreshold swing in Eq. 60, it is intuitive to explain its simple form. In the extreme of zero oxide thickness, the exponential characteristics are identical to the familiar case of the diffusion current in a p-n junction. For nonzero oxide thickness, the swing is just degraded by a factor which is a voltage divider of two capacitors in series, whose ratio is $(C_{ox} + C_D)/C_{ox}$. The voltage divider is exactly the implication of Eq. 59. One also notices that since the depletion width (and C_D) varies with ψ_s, the subthreshold swing is a weak function, but not exactly constant with V_G.

In the presence of a significant interface-trap density D_{it}, its associated capacitance C_{it} $(= q^2 D_{it})$ is in parallel with the depletion-layer capacitance C_D. Using Eq. 60 and substituting $(C_D + C_{it})$ for C_D (see Fig. 14 of Chapter 4), we obtain

$$S(\text{with } D_{it}) = (\ln 10)\left(\frac{kT}{q}\right)\left(\frac{C_{ox} + C_D + C_{it}}{C_{ox}}\right)$$

$$= S(\text{without } D_{it}) \times \frac{C_{ox} + C_D + C_{it}}{C_{ox} + C_D}. \tag{61}$$

If other device parameters such as doping and oxide thickness are known, by measuring the subthreshold swing, the interface-trap density can be obtained. This pro-

Sandisk Technologies, Inc. - Ex. 1042, Page 000033

vides an attractive option in measuring D_{it} besides using the MOS capacitor in which ac measurements have to be made. In general, dc I-V measurements are much easier to make than ac capacitance and conductance, provided a three-terminal transistor structure is available (substrate contact is not critical here).

For a sharp subthreshold slope (small S), it is preferable to have low channel doping, thin oxide thickness, low interface-trap density, and low-temperature operation. When a substrate bias is applied, in addition to shifting the threshold voltage, it increases the value of ψ_s by V_{BS}. Consequently, the depletion-layer capacitance C_d is reduced and therefore S is reduced.

In Fig. 15a, at and near the threshold voltage, the drain current does not turn off as sharply as Eq. 24 predicts. This is due to diffusion current which is the dominant current near and below threshold, and it has been ignored so far, as one of the assumptions made at the beginning of Section 6.2.2. To consider the effect of the diffusion component, we refer to Fig. 7 for the nonequilibrium condition. The total drain current density including both drift and diffusion components is given by

$$J_D(x,y) = q\mu_n n\mathscr{E}_y + qD_n\frac{dn}{dy} = D_n n(x,y)\frac{dE_{Fn}}{dy}. \tag{62}$$

The drain current based on the gradual-channel approximation is

$$I_D = Z\int_0^{x_i} J_D(x,y)dx = \frac{ZD_n}{L}\int_0^L \frac{dE_{Fn}}{dy}\int_0^{x_i} n(x,y)dxdy$$

$$= \frac{Z}{L}\frac{\varepsilon_s\mu_n}{L_D}\int_0^{V_D}\int_{\psi_B}^{\psi_s}\frac{\exp(\beta\psi_p - \beta\Delta\psi_i)}{F(\beta\psi_p, \Delta\psi_i, n_{po}/p_{po})}d\psi_p d\Delta\psi_i \quad . \tag{63}$$

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

$$V_G - V_{FB} = -\frac{Q_s}{C_{ox}} + \psi_s$$

$$= \frac{2\varepsilon_s kT}{C_{ox}qL_D}F\left(\beta\psi_s, \Delta\psi_i, \frac{n_{po}}{p_{po}}\right) + \psi_s . \tag{64}$$

Equation 63 reduces to Eq. 23 for gate voltages well above threshold. The latter, however, becomes inaccurate for gate voltages near and below threshold and near the pinch-off point. For a particular device with known physical dimensions and other device parameters, Eq. 63 can be calculated numerically to give accurate results for the entire range of drain voltage, from the linear region to the saturation region.

6.2.5 Mobility Behavior

Because channel carriers are confined to a thin inversion layer, their drift velocity v and mobility μ are expected to be influenced by the thickness of this inversion layer. When a small longitudinal field $\mathscr{E}_y$ is applied (parallel to the semiconductor surface), the drift velocity varies linearly with $\mathscr{E}_y$ and the proportionality constant is the low-field mobility. Experimental measurements on Si inversion layers show that this low-

Sandisk Technologies, Inc. - Ex. 1042, Page 000034

field mobility, while independent of $\mathscr{E}_y$, is a unique function of the transverse field $\mathscr{E}_x$ that is perpendicular to the current flow.[32] This dependence is not directly on the oxide thickness or doping density, but through their impact of $\mathscr{E}_x$ in the inversion layer. The measured results are shown in Fig. 16. When many devices with different oxide thicknesses and doping levels are measured, the mobility is found to correlate well with a single parameter that is related to $\mathscr{E}_x$. At a given temperature, mobility decreases with an increasing *effective* transverse field, defined as the field averaged over the electron distribution in the inversion layer, given by

$$(\mathscr{E}_x)_{\text{eff}} = \frac{1}{\varepsilon_s}\left(Q_B + \frac{1}{2}Q_n\right). \tag{65}$$

Physically it means an average inversion carrier experiences the full effect of the depletion-layer charge Q_B, but only half of the inversion-layer charge Q_n. Note that this effective mobility is valid for the current expressions in Eqs. 24 and 27, but will be slightly different from that in the g_m expression (such as Eq. 30) in which constant mobility has been assumed.

When the longitudinal field increases, the v-$\mathscr{E}$ relationship starts to deviate from linearity. This field-dependent mobility has been discussed and generally described by Eq. 33 and shown in Fig, 11. Measured electron drift velocity as a function of $\mathscr{E}_y$ for various $\mathscr{E}_x$ is shown in Fig. 17. Since mobility at any field is defined as the ratio of $v/\mathscr{E}_y$, it decreases monotonically with $\mathscr{E}_y$. Eventually velocity saturation occurs and results in a value similar to that of bulk silicon. The influence of $\mathscr{E}_x$ on low-field

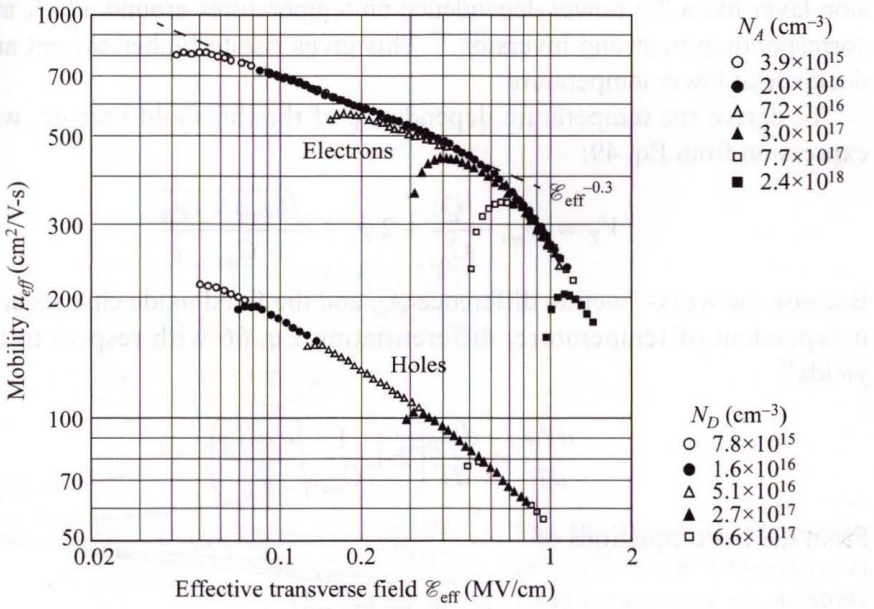


Fig. 16 Electron and hole inversion-layer mobilities vs. effective transverse field, at room temperature on Si (100) surface. (After Ref. 33.)

Sandisk Technologies, Inc. - Ex. 1042, Page 000035

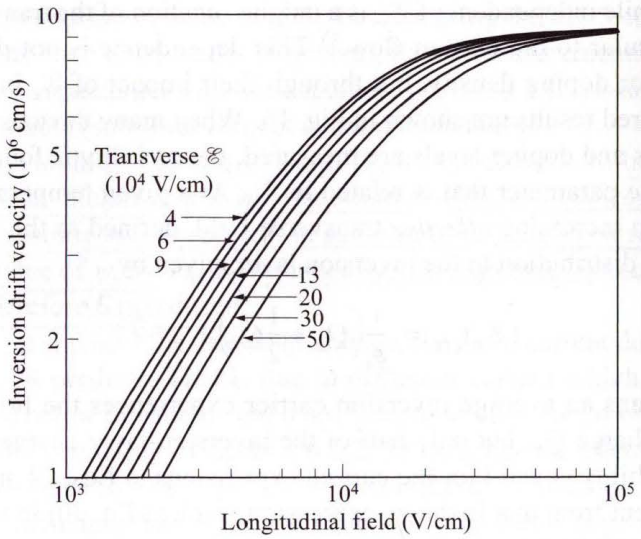


Fig. 17 Electron surface drift velocity vs. longitudinal field for various transverse fields. The slope at low longitudinal field is mobility. (After Ref. 34.)

mobility from Fig. 16 is also reflected in this figure. It can be seen here also that the saturation velocity v_s is independent of the low-field mobility or $\mathscr{E}_x$.

6.2.6 Temperature Dependence

Temperature affects device parameters and performance, in particular mobility, threshold voltage, and subthreshold characteristics. The effective mobility in inversion layer has a T^{-2} power dependence on temperatures around 300 K at gate biases corresponding to strong inversion.[32] This gives rise to higher current and transconductance at lower temperature.

To derive the temperature dependence of the threshold voltage, we repeat the expression from Eq. 49:

$$V_T = \phi_{ms} - \frac{Q_f}{C_{ox}} + 2\psi_B + \frac{\sqrt{4\varepsilon_s q N_A \psi_B}}{C_{ox}}. \tag{66}$$

Because the work-function difference ϕ_{ms} and the fixed oxide charges are essentially independent of temperature, differentiating Eq. 66 with respect to temperature yields[35]

$$\frac{dV_T}{dT} = \frac{d\psi_B}{dT}\left(2 + \frac{1}{C_{ox}}\sqrt{\frac{\varepsilon_s q N_A}{\psi_B}}\right). \tag{67}$$

From the basic equations of

$$\psi_B = \frac{kT}{q}\ln\left(\frac{N_A}{n_i}\right), \tag{68}$$

Sandisk Technologies, Inc. - Ex. 1042, Page 000036

$$n_i^2 \propto T^3 \exp\left(\frac{-E_{g0}}{kT}\right), \tag{69}$$

where E_{g0} is the energy gap at $T = 0$, we obtain

$$\frac{d\psi_B}{dT} \approx \frac{1}{T}\left(\psi_B - \frac{E_{g0}}{2q}\right). \tag{70}$$

Figure 18 shows the results of such calculations at room temperature as a function of substrate doping for various values of oxide thickness. Note that depending on the oxide thickness, the quantity $|dV_T/dT|$ can increase or increase with the substrate doping.

As temperature decreases, the MOSFET characteristics improve, especially in the subthreshold region. Figure 19 shows the transfer characteristics of a long-channel MOSFET ($L = 9$ μm) with temperature as a parameter. Note that as temperature decreases from 296 K to 77 K, the 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. 18. 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 four. This improvement comes mainly from the kT/q term in Eq. 60. Other improvements at 77 K include higher mobility, thus, higher current and transconductance, lower power consumption, lower junction leakage current, and lower metal-line resistance. The major disadvantages are that the MOSFET must be immersed in a suitable inert coolant (e.g., liquid nitrogen), and that a low-temperature setup requires additional equipment and special care.

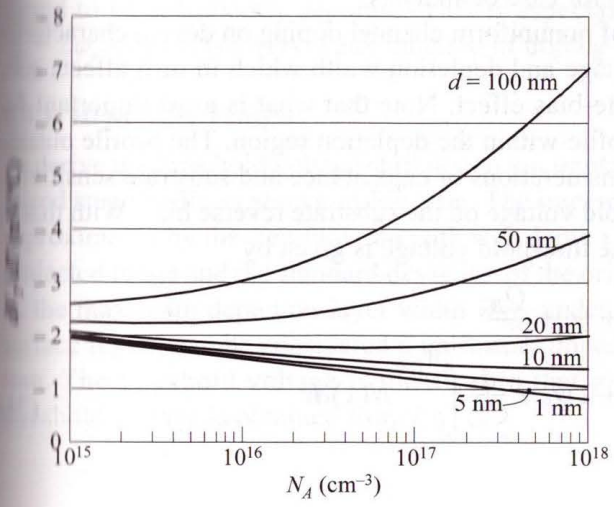


Fig. 18 Threshold-voltage shift (dV_T/dT) of a Si-SiO_2 system at room temperature vs. substrate doping, with oxide thickness d as a parameter.

Sandisk Technologies, Inc. - Ex. 1042, Page 000037

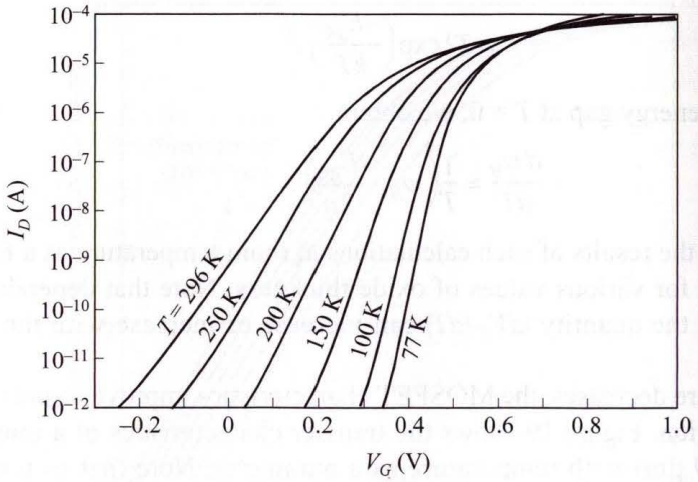


Fig. 19 Subthreshold characteristics for a long-channel MOSFET ($L = 9$ μm) with temperature as a parameter. (After Ref. 36.)

6.3 NONUNIFORM DOPING AND BURIED-CHANNEL DEVICE

In Section 6.2 doping concentration in the channel is assumed to be constant. In practical devices, however, the doping is generally nonuniform because in modern MOSFET technology, ion implantation is used extensively to taylor the doping profile and improve the device performance for specific applications. For example, a lighter doping at deeper region reduces drain-substrate capacitance and also the substrate-bias effect. On the other hand, a lighter doping level near the Si-SiO_2 interface lowers the threshold voltage, reduces the field and improves mobility, and higher level at deeper region reduces punch-through between source and drain. These two general cases, namely *high-low* and *low-high* profiles, are depicted in Fig. 20, with their step-profile approximations for ease of analysis.

We consider next the effect of nonuniform channel doping on device characteristics, especially on threshold voltage and depletion width which in turn affects subthreshold swing and the substrate-bias effect. Note that what is most important for determining V_T is the doping profile within the depletion region. The profile outside the depletion is important for considerations of capacitance and substrate sensitivity, that is, dependence of the threshold voltage on the substrate reverse bias. With that in mind, the general equation for the threshold voltage is given by

$$\begin{aligned} V_T &= V_{FB} + \psi_s + \frac{Q_B}{C_{ox}} \\ &= V_{FB} + 2\psi_B + \frac{q}{C_{ox}} \int_0^{W_{Dm}} N(x)\,dx \end{aligned}, \tag{71}$$

Sandisk Technologies, Inc. - Ex. 1042, Page 000038

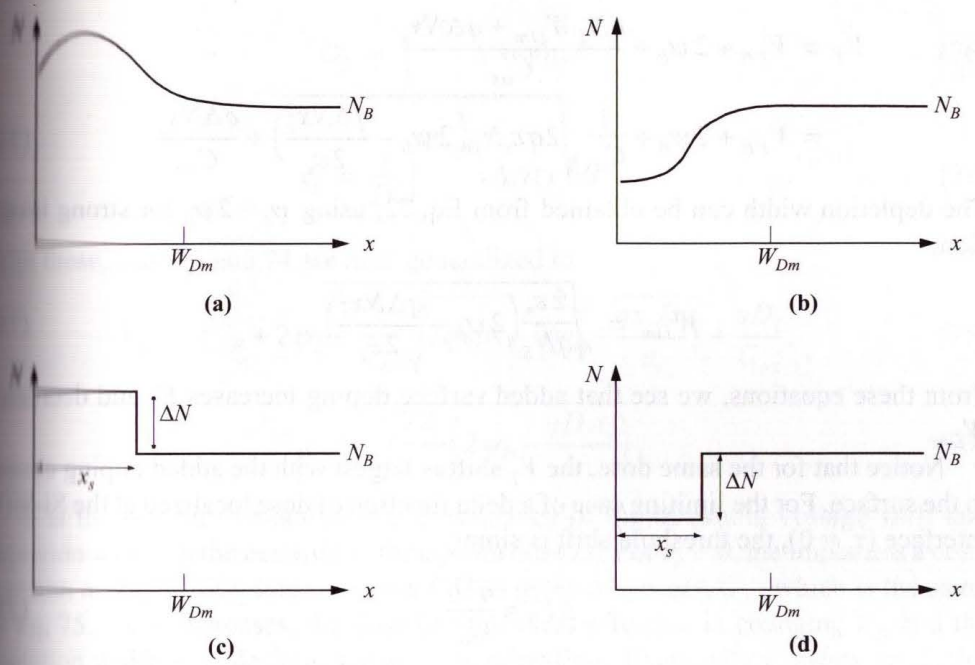


Fig. 20 Nonuniform channel doping profiles. (a) High-low profile. (b) Low-high (retrograde) profile. (c) – (d) Their approximations using step profiles.

where Q_B is the depletion-layer charge. The limit for integration, that is the maximum depletion width W_{Dm}, is needed and determined by the Poisson equation with the onset of strong inversion being the boundary condition,

$$\psi_s = 2\psi_B = \frac{q}{\varepsilon_s}\int_0^{W_{Dm}} xN(x)dx. \tag{72}$$

Note that for a nonuniform profile, the definitions of ψ_B and V_{FB} become nontrivial and complicated. Fortunately, using the background doping of N_B for these values is found to be sufficiently accurate. This is especially true for the surface potential $\psi_s = 2\psi_B$ since it is a very weak function of doping level.

6.3.1 High-Low Profile

To derive the threshold voltage shift due to ion implantation, we shall consider an idealized step profile as shown in Fig. 20c. The implant profile, after thermal anneal, is approximated by the step function with step depth x_s, roughly equal to the sum of the projected range and the standard deviation of the original implant. For a wider x_s, that is, the maximum depletion-layer width W_{Dm} under heavy inversion is within x_s, the surface region can be considered a uniformly doped region with a higher concentration. The threshold voltage is identical to that given by Eq. 50. If $W_{Dm} > x_s$, the threshold voltage is obtained from Eq. 71,

Sandisk Technologies, Inc. - Ex. 1042, Page 000039

$$V_T = V_{FB} + 2\psi_B + \frac{qN_B W_{Dm} + q\Delta N x_s}{C_{ox}}$$
$$= V_{FB} + 2\psi_B + \frac{1}{C_{ox}}\sqrt{2q\varepsilon_s N_B\left(2\psi_B - \frac{q\Delta N x_s^2}{2\varepsilon_s}\right)} + \frac{q\Delta N x_s}{C_{ox}}. \quad (73)$$

The depletion width can be obtained from Eq. 72, using $\psi_s = 2\psi_B$ for strong inversion:

$$W_{Dm} = \sqrt{\frac{2\varepsilon_s}{qN_B}\left(2\psi_B - \frac{q\Delta N x_s^2}{2\varepsilon_s}\right)}. \quad (74)$$

From these equations, we see that added surface doping increases V_T and decreases W_{Dm}.

Notice that for the same dose, the V_T shift is largest with the added doping closest to the surface. For the limiting case of a delta function of dose localized at the Si-SiO_2 interface ($x_s = 0$), the threshold shift is simply

$$\Delta V_T \approx \frac{qD_I}{C_{ox}}, \quad (75)$$

where D_I is the total dose $\Delta N x_s$. Such approach is called threshold adjust, which has the same effect as changing the work-function difference ϕ_{ms} or changing the total fixed oxide charge.

The step-profile approach described above can give first-order results for the threshold voltage. To obtain a more accurate V_T, we have to consider the actual doping profile, because the step width x_s is not well defined for nonuniform doping. A schematic diagram for the nonuniform implanted doping $N(x)$ is shown in Fig. 21. For a typical case, the threshold voltage depends on the implanted dose D_I and the centroid of the dose x_c. Therefore, the actual implant can be replaced by a delta-function located at $x = x_c$ as shown;

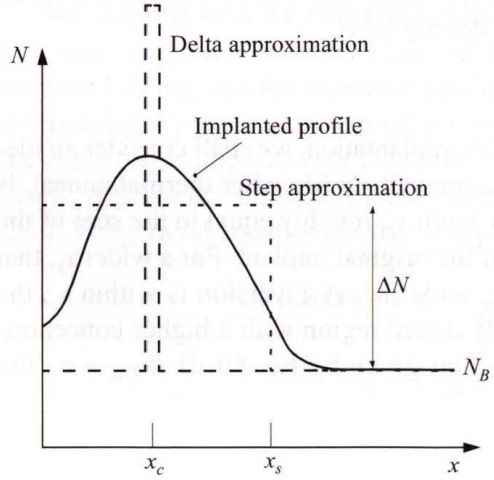


Fig. 21 Approximation of an actual implanted profile by step and delta profiles

Sandisk Technologies, Inc. - Ex. 1042, Page 000040

$$D_I = \int_0^{W_{Dm}} \Delta N(x)dx, \tag{76}$$

$$x_c = \frac{1}{D_I}\int_0^{W_{Dm}} x\Delta N(x)dx. \tag{77}$$

With these, Eqs. 73 and 74 are now generalized to:

$$V_T = V_{FB} + 2\psi_B + \frac{1}{C_{ox}}\sqrt{2q\varepsilon_s N_B\left(2\psi_B - \frac{qx_c D_I}{\varepsilon_s}\right)} + \frac{qD_I}{C_{ox}}, \tag{78}$$

$$W_{Dm} = \sqrt{\frac{2\varepsilon_s}{qN_B}\left(2\psi_B - \frac{qD_I x_c}{\varepsilon_s}\right)}. \tag{79}$$

It is interesting to examine the dependence of the threshold voltage shift and depletion width on the centroid x_c for a given dose D_I. For $x_c = 0$, the implant is a delta function at the Si-SiO_2 interface, and Eq. 78 gives $\Delta V_T = qD_I/C_{ox}$, which is the same as Eq. 75. As x_c increases, the dose becomes less effective in changing V_T, and the depletion width W_{Dm} decreases also at the same time. Eventually x_c meets the depletion edge, and then W_{Dm} becomes clamped to and increases with the implant centroid x_c. The condition for which x_c starts to be equal to W_{Dm} can be obtained from Eq. 79:

$$D_I(x_c = W_{Dm}) = \frac{N_B(W_{Dm0}^2 - x_c^2)}{2x_c}, \tag{80}$$

where W_{Dm0} is the original W_{Dm} with background doping N_B. Eventually, as x_c moves beyond the W_{Dm0}, it no longer has any effect on threshold voltage and depletion width.

To consider the subthreshold swing and the substrate sensitivity, in Section 6.2.4 we have interpreted the subthreshold swing by comparing the gate-oxide capacitance C_{ox} to depletion capacitance C_D. So, once the depletion width is known, the subthreshold swing can be calculated. For the high-low profile, the added doping decreases W_{Dm}, increases C_D, and results in a larger (less steep) subthreshold swing. The substrate sensitivity can be calculated also by substituting $2\psi_B$ with $2\psi_B + V_{BS}$ in calculating V_T.

6.3.2 Low-High Profile

Analysis of the low-high profile (Fig. 20b), also called the retrograde profile, is similar to the high-low case with a ΔN being subtracted from the background doping. The appropriate equations for the threshold voltage and depletion width become, with just a change of signs:

Sandisk Technologies, Inc. - Ex. 1042, Page 000041

$$V_T = V_{FB} + 2\psi_B + \frac{qN_B W_{Dm} - q\Delta N x_s}{C_{ox}}$$
$$= V_{FB} + 2\psi_B + \frac{1}{C_{ox}}\sqrt{2q\varepsilon_s N_B\left(2\psi_B + \frac{q\Delta N x_s^2}{2\varepsilon_s}\right)} - \frac{q\Delta N x_s}{C_{ox}}, \quad (81)$$

and

$$W_{Dm} = \sqrt{\frac{2\varepsilon_s}{qN_B}\left(2\psi_B + \frac{q\Delta N x_s^2}{2\varepsilon_s}\right)}. \quad (82)$$

The threshold voltage is, thus, decreased and the depletion width is increased by a dip at the surface doping.

6.3.3 Buried-Channel Device

In the extreme case of the low-high profile, the surface doping can be of the opposite type of the substrate. When this happens, and if part of the surface doped layer is not fully depleted, that is, there exists some neutral region, current can conduct through this buried layer. We call this type of device a buried-channel device.[37–40] Figure 22a shows a cross section of such a buried-*n*-channel MOSFET. The gate voltage can change the surface depletion layer, thus controlling the net opening of the channel thickness and controlling the current flow. With a large positive gate bias, the channel is fully open, and an addition surface inversion layer can be induced at the surface similar to a regular surface channel, resulting in two channels in parallel.

The surface inversion channel has been the subject of discussion and needs no further elaboration. We now focus on the buried channel whose doping and dimensions are shown in Fig. 22b and whose energy-band diagrams are shown in Fig. 21. The net channel thickness is reduced from x_s by the amounts of surface depletion W_{Ds} and the bottom *p*-*n* junction depletion W_{Dn}. The surface depletion as a function of V_G is the same as Eq. 27 of Chapter 4, and is repeated and modified here as

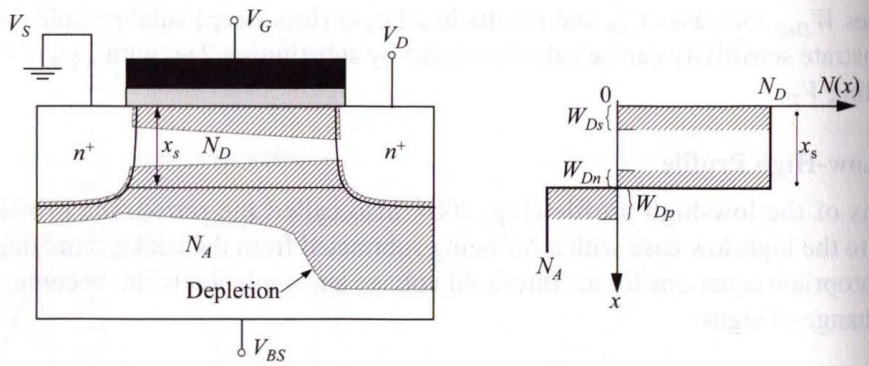


Fig. 22 (a) Schematic of a buried-channel MOSFET under bias. (b) Its doping profile and depletion regions.

Sandisk Technologies, Inc. - Ex. 1042, Page 000042

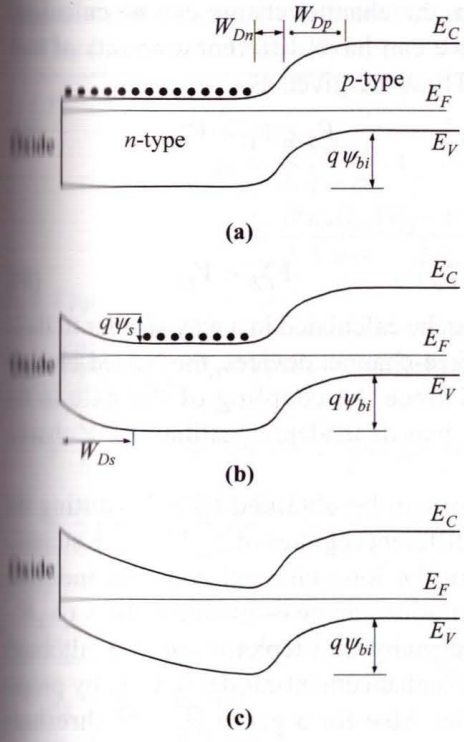


Fig. 23 Energy-band diagrams of a buried-channel MOSFET, for the bias conditions of (a) flat band ($V_G = V_{FB}^*$), (b) surface depletion, and (c) threshold ($V_G = V_T$).

$$W_{Ds} = \sqrt{\frac{2\varepsilon_s}{qN_D}(V_{FB}^* - V_G) + \frac{\varepsilon_s^2}{C_{ox}^2}} - \frac{\varepsilon_s}{C_{ox}}. \tag{83}$$

Note that the flat-band voltage V_{FB}^* takes on slightly different meaning here (Fig. 23a). We refer now to the condition that the surface n-layer has flat band, as opposed to the p-substrate. The new flat-band voltage is redefined as

$$V_{FB}^* = V_{FB} + \psi_{bi} \tag{84}$$

where V_{FB} keeps the p-substrate as reference. The bottom depletion width is from the p-n junction theory, given by

$$W_{Dn} = \sqrt{\frac{2\varepsilon_s \psi_{bi}}{qN_D}\left(\frac{N_A}{N_D + N_A}\right)}. \tag{85}$$

Of special interest is the threshold voltage V_T at which gate bias the channel width is totally consumed by both depletion regions. Setting the condition of

$$x_s = W_{Ds} + W_{Dn}, \tag{86}$$

the threshold voltage is obtained;[40]

$$V_T = V_{FB}^* - qN_D x_s\left(\frac{x_s}{2\varepsilon_s} + \frac{1}{C_{ox}}\right) + \left(\frac{x_s}{\varepsilon_s} + \frac{1}{C_{ox}}\right)\sqrt{\frac{2q\varepsilon_s N_D N_A \psi_{bi}}{N_D + N_A}} - \frac{N_A \psi_{bi}}{N_D + N_A}. \tag{87}$$

Sandisk Technologies, Inc. - Ex. 1042, Page 000043

Once the channel dimensions are known, the channel charge can be calculated easily. Depending on the gate-bias range, we can have different amounts of bulk charge Q_B and surface inversion charge Q_I. These are given as;

$$Q = Q_B = (x_s - W_{Ds} - W_{Dn})N_D \ , \qquad V_T < V_G < V_{FB}^* \ , \tag{85}$$

and

$$\begin{aligned} Q &= Q_B + Q_I \\ &= (x_s - W_{Dn})N_D + C_{ox}(V_G - V_{FB}^*) \ , \qquad V_{FB}^* < V_G \ . \end{aligned} \tag{86}$$

Given the channel charge, the drain current can be calculated in a way similar to those previously derived. But compared to the surface-channel devices, the buried-channel MOSFET equations are more complicated since the coupling of the gate to the channel (or net gate capacitance) is now gate-bias dependent. Qualitative I-V characteristics are shown in Fig. 24.

More-exact solution of the drain current can be obtained by substituting the charge into Eq. 22. The results, divided into different regimes of V_G bias, are summarized in Table 1. These results are based on the long-channel constant-mobility model. Current saturation due to velocity saturation can be estimated to be $\approx Qv_sW$.

A buried-channel MOSFET is usually normally-on (depletion-mode), although theoretically it can be made as a normally-off (enhancement-mode) device, by proper choice of metal work function, for example. Also for a given N_D, the threshold voltage becomes more negative with increased buried-channel depth x_s. However, because there exists a maximum depletion width in an MOS system, if the doping density N_D or/and the buried-channel depth x_s are sufficiently large, W_{Ds} can reach a maximum value without pinching off the channel. A limit on the channel profile thus exists, otherwise the transistor cannot be turned off. This condition is bound by a combination of x_s and N_D;

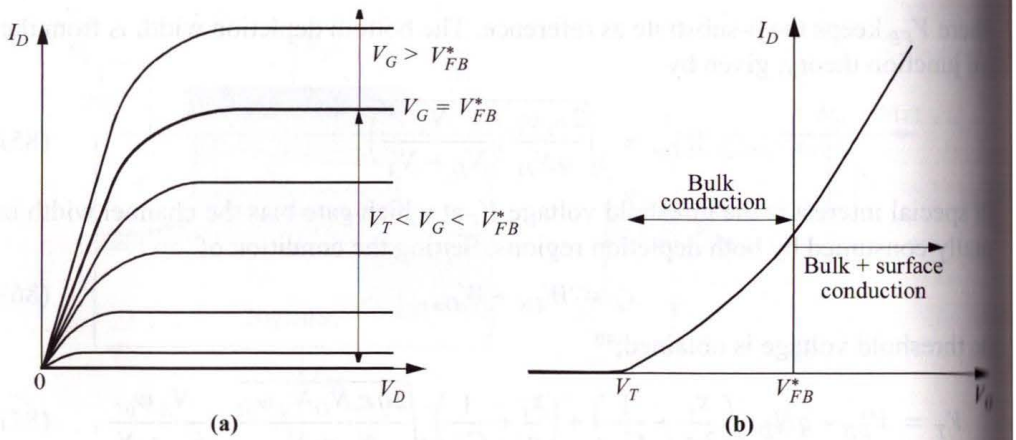


Fig. 24 Buried-channel MOSFET: (a) Output characteristics. (b) Transfer characteristics (I_D vs. V_G) in linear region (small V_D), showing threshold voltage V_T and flat-band voltage V_{FB}^*.

Sandisk Technologies, Inc. - Ex. 1042, Page 000044

Table 1 Current Equations for Buried-Channel MOSFETs, Based on Long-Channel Constant Mobility (After Refs. 15 and 37)

$V_T \le V_G \le V_{FB}^*$

$$I_D = \frac{W}{L}\frac{\mu_B C_{ox}}{1+\sigma}\left[(V_G - V_T)V_D - \frac{1}{2}\alpha V_D^2\right], \qquad V_D \le V_{Dsat}$$
$$= \frac{W}{L}\frac{\mu_B C_{ox}}{1+\sigma}\frac{(V_G - V_T)^2}{2\alpha}. \qquad V_D \ge V_{Dsat}$$

$V_G \le V_{FB}^*$

$$I_B = \frac{W}{L}\frac{\mu_B C_{ox}}{1+\sigma}\left\{(V_G - V_T)V_D - \frac{1}{2}\alpha V_D^2 + (r-1)\left[(V_G - V_{FB}^*)V_D - \frac{1}{2}V_D^2\right]\right\}, \qquad V_D < V_G - V_{FB}^*$$
$$= \frac{W}{L}\frac{\mu_B C_{ox}}{1+\sigma}\left[(V_G - V_T)V_D - \frac{1}{2}\alpha V_D^2 + \frac{1}{2}(r-1)(V_G - V_{FB}^*)^2\right], \qquad V_G - V_{FB}^* \le V_D < V_{Dsat}$$
$$= \frac{W}{L}\frac{\mu_B C_{ox}}{1+\sigma}\left[\frac{(V_G - V_T)^2}{2\alpha} + \frac{1}{2}(r-1)(V_G - V_{FB}^*)^2\right]. \qquad V_D \ge V_{Dsat}$$

where

$$V_{Dsat} = (V_G - V_T)/\alpha \qquad \sigma = \frac{C_{ox}x_s}{\varepsilon_s}\left(\frac{C_{ox}x_s}{2\varepsilon_s} + 1\right) \qquad \alpha = 1 + (1+\sigma)\frac{\gamma}{4\sqrt{\psi_{bi}}}$$

μ_B = bulk mobility
μ_s = surface mobility

$$r = (1+\sigma)\frac{\mu_s}{\mu_B} \qquad \gamma = \frac{\sqrt{2\varepsilon_s q N_A}}{C_{ox}}$$

$$x_s|_{max} = \sqrt{\frac{2\varepsilon_s}{qN_D}}\left(\sqrt{2\psi_B} + \sqrt{\frac{N_A\psi_{bi}}{N_D + N_A}}\right). \tag{90}$$

In the buried-channel devices, the substrate-bias effect is more direct. It can be viewed as a bottom gate. The effects are calculated in the above equations if ψ_{bi} is replaced with $\psi_{bi} - V_{BS}$ (V_{BS} is negative). In particular, V_T (Eq. 87) and W_{Dn} (Eq. 85) can be shifted with a substrate bias, to the extent that the transistor can be turned on and off and between depletion-mode and enhancement-mode.

We now turn to the subthreshold current of buried-channel devices. At a sufficiently large negative gate bias, the channel will be pinched off, that is, when $x_x = W_{Do} + W_{Dn}$ (Fig. 23c). The conduction below the threshold voltage is due to the presence of a region of partially depleted electrons, wherein the current is carried primarily by diffusion of electrons. The resulting subthreshold (sub-pinch-off) current for a buried-channel MOSFET is, thus, directly analogous to the subthreshold current for a surface-channel MOSFET. The subthreshold current will vary exponentially with the gate voltage, and the subthreshold swing S is given by the capacitive divider ratio again of Eq. 60, except now different capacitances have to be used. From Fig. 23c, the maximum electron concentration occurs at the location of $x \approx x_s - W_{Dn}$. So C_D of Eq. 60 should be replaced with the depletion capacitance of the substrate p-n junction

Sandisk Technologies, Inc. - Ex. 1042, Page 000045

$[\varepsilon_s/(W_{Dn} + W_{Dp})]$, and C_{ox} should be replaced with C_{ox} in series with the surface depletion capacitance ε_s/W_{Ds}. These substitutions give an expression of[39]

$$S = (\ln 10)\frac{kT}{q}\left[1 + \frac{\varepsilon_{ox}W_{Ds} + \varepsilon_s d}{\varepsilon_{ox}(W_{Dn} + W_{Dp})}\right], \tag{91}$$

where all depletion layers W_{Ds}, W_{Dn}, and W_{Dp} correspond to the condition at threshold ($V_G = V_T$). The subthreshold swing is usually larger than that of conventional surface-channel devices.

The buried-channel device is expected to have higher carrier mobility than surface-channel devices since carriers are free of surface scattering and other surface effects. They are also less affected by the short-channel effects (to be discussed next) such as hot-carrier-induced reliability problems. On the other hand, since the net distance between the gate and the channel is further away and is gate-bias dependent, the transconductance is smaller and variable. Note that if the gate is replaced by a Schottky junction or a *p-n* junction, the device become a MESFET or a JFET correspondingly, both to be discussed in the next chapter.

6.4 DEVICE SCALING AND SHORT-CHANNEL EFFECTS

Since 1959, the beginning of the integrated-circuit era, the minimum feature length has been reduced by more than two orders of magnitude. We expect the minimum dimension will continue to shrink in the foreseeable future, as illustrated in Fig. 1. As the MOSFET dimensions shrink, they need to be designed properly to preserve the long-channel behavior as much as possible. As the channel length decreases, the depletion widths of the source and drain become comparable to the channel length and punch-through between the drain and source will eventually occur. This requires higher channel doping. A higher channel doping will increase the threshold voltage, and in order to control a reasonable threshold voltage, a thinner oxide is necessary. One sees that the device parameters are interrelated, and certain scaling rules are used to optimize the device performance.

Even with the best scaling rules, as the channel length is reduced, departures from long-channel behavior are inevitable. These departures, the short-channel effects, arise as results of a two-dimensional potential distribution and high electric fields in the channel region. The potential distribution in the channel now depends on both the transverse field $\mathscr{E}_x$ (controlled by the gate voltage and the back-substrate bias) and the longitudinal field $\mathscr{E}_y$ (controlled by the drain bias). In other words, the potential distribution becomes two-dimensional, and the gradual-channel approximation (that is, $\mathscr{E}_x \gg \mathscr{E}_y$) is no longer valid. This two-dimensional potential results in many forms of undesirable electrical behavior.

As the electric field is increased, the channel mobility becomes field-dependent, and eventually velocity saturation occurs. (The mobility behavior was discussed in Section 6.2.5.) When the field is increased further, carrier multiplication near the drain occurs, leading to substrate current and parasitic bipolar-transistor action. High

Sandisk Technologies, Inc. - Ex. 1042, Page 000046

Table 2 MOSFET Scaling

Parameter	Scaling factor: Constant-$\mathscr{E}$	Scaling factor: Actual	Limitation
L	$1/\kappa$	/	/
$\mathscr{E}$	1	>1	/
d	$1/\kappa$	$>1/\kappa$	Tunneling, defects
r_j	$1/\kappa$	$>1/\kappa$	Resistance
V_T	$1/\kappa$	$\gg 1/\kappa$	Off current
V_D	$1/\kappa$	$\gg 1/\kappa$	System, V_T
N_A	κ	$<\kappa$	Junction breakdown

In ideal constant-field scaling parameters are scaled by the same factor. In reality the scaling factors are limited by other reasons and skewed.

fields also cause hot-carrier injection into the oxide leading to oxide charging and subsequent threshold-voltage shift and transconductance degradation.

These aforementioned phenomena will cause short-channel effects which can be summarized as follows: (1) V_T is not constant with L, (2) I_D does not saturate with V_D bias, both above and below threshold; (3) I_D is not proportional to $1/L$; and (4) device characteristics degrade with operation time. Because short-channel effects complicate device operation and degrade device performance, these effects should be eliminated or minimized so that a physical short-channel device can preserve the *electrical* long-channel behavior. In this section, we discuss MOSFET scaling and the short-channel effects that accompany device miniaturization. [Item-(3) is related to high-field mobility or velocity saturation, and has already been treated in Section 6.2.2.]

6.4.1 Device Scaling

The most-ideal scaling rule to avoid short-channel effects is simply to scale down all dimensions and voltages of a long-channel MOSFET so that the internal electric fields are kept the same.[41] This constant-field scaling is shown in Table 2 and Fig. 25.This approach offers a conceptually simple picture for device miniaturization. All dimensions, including channel length and width, oxide thickness, and junction

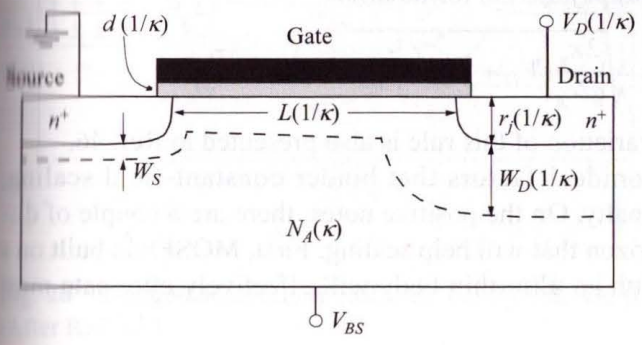


Fig. 25 Physical parameters for MOSFET scaling. Scaling factors for constant-field are indicated.

Sandisk Technologies, Inc. - Ex. 1042, Page 000047

depth, are shrunk by the same *scaling factor* κ. The doping level is increased by κ and all voltages are reduced by κ, leading to a reduction of the junction depletion width by about κ. Note that the subthreshold swing S remains essentially the same since S is proportional to $1 + C_D/C_{ox}$ and both capacitances are scaled up by the same factor κ.

Unfortunately such an ideal scaling rule is hindered by other factors that are fundamentally not scalable. First, the junction built-in voltage and the surface potential for the onset of weak inversion do not scale (only $\approx 10\%$ change for 10 times increase in dopings). The range of gate voltage between depletion and strong inversion is approximately 0.5 V. These limitations stem from the fact that both the energy gap and thermal energy kT remain constant. The gate oxide thickness has the technological difficulty of defects as it approaches the low-nm scale. Tunneling through the oxide is another fundamental limitation. The quantum-mechanical effect discussed in Section 4.3.6 degrades the gate capacitance due to the fact that carriers are located at a finite distance ($\approx$1 nm) from the interface. The source and drain series resistance increases when r_j is decreased. This is especially detrimental when the current of the device increases at the same time. The channel doping cannot be increased indefinitely due to p-n junction breakdown. The threshold voltage cannot be scaled due to the off-current consideration, even with a fixed subthreshold swing. The supply voltage has been historically slow in scaling because of the system consideration, and also the push for higher speed. These limitations in scaling are summarized in Table 2, resulting in the actual nonideal scaling factors which are shown relative to the constant-field scaling. With these limitations, the field is no longer kept the same, and it increases with smaller gate lengths.

With the above practical limitations, other scaling rules have been proposed. These include constant-voltage scaling,[42] quasi-constant-voltage scaling,[43] and generalized scaling.[44] One other scaling rule with a unique feature of having flexible scaling factors has been proposed.[45] This allows the various device parameters to be adjusted independently as long as the overall behavior is preserved. Therefore, all device parameters do not have to be scaled by the same factor κ. The expression for minimum channel length for which long-channel behavior can be observed is found to follow a simple empirical relation:[45]

$$L \geq C_1[r_j d(W_S + W_D)^2]^{1/3} \tag{92}$$

where C_1 is a constant, and $W_S + W_D$ is the sum of the source and drain depletion widths in a one-dimensional abrupt junction formulation:

$$W_D = \sqrt{\frac{2\varepsilon_s}{qN_A}(V_D + \psi_{bi} - V_{BS})}. \tag{93}$$

For $V_D = 0$, W_D equals W_S. A variation of this rule is also presented in Ref. 46.

We have discussed the nonideal factors that hinder constant-field scaling, resulting in some form of a penalty. On the positive notes, there are a couple of disruptive technologies on the horizon that will help scaling. First, MOSFETs built on a three-dimensional structure with an ultra-thin body will effectively eliminate most

Sandisk Technologies, Inc. - Ex. 1042, Page 000048

the conduction path for punch-through, and the requirement on channel doping can be relaxed (see Section 6.5.5). Second, research activities looking for gate dielectrics with high dielectric constants have been intense. Such a high-K gate dielectric can relax the physical thickness, improving the defect density and reducing the field for tunnelling. Both of these technologies can help to circumvent or delay the short-channel effects for a particular generation of channel length.

6.4.2 Charge Sharing from Source/Drain

Analysis of the channel charge so far is 1-dimensional, that is, both the inversion charge and depletion charge is completely balanced by the charge on the gate, so they can be treated as charge density. Detailed 2-dimensional examination at the ends of the channel reveals that some of the depletion charge is balanced by the n^+-source and drain, as shown in Fig. 26a. Departure from long-channel behavior can be shown to happen by applying the charge conservation principle to the region bounded by the gate, the channel, and the source/drain,[47]

$$Q'_M + Q'_n + Q'_B = 0, \tag{94}$$

where Q'_M is the total charge on the gate, Q'_n is the total inversion-layer charge, and Q'_B is the total ionized impurity in the depletion region. This of course assumes that all oxide and interface charges are zero. The threshold voltage, which can be viewed as voltage required to deplete the total bulk charge Q'_B in the maximum depletion width, is given by

$$V_T = V_{FB} + 2\psi_B + \frac{Q'_B}{C_{ox}A}, \tag{95}$$

where A is the gate area $Z \times L$. For long-channel devices, $Q'_B = qAN_AW_{Dm}$, where W_{Dm} is the maximum depletion-layer width,

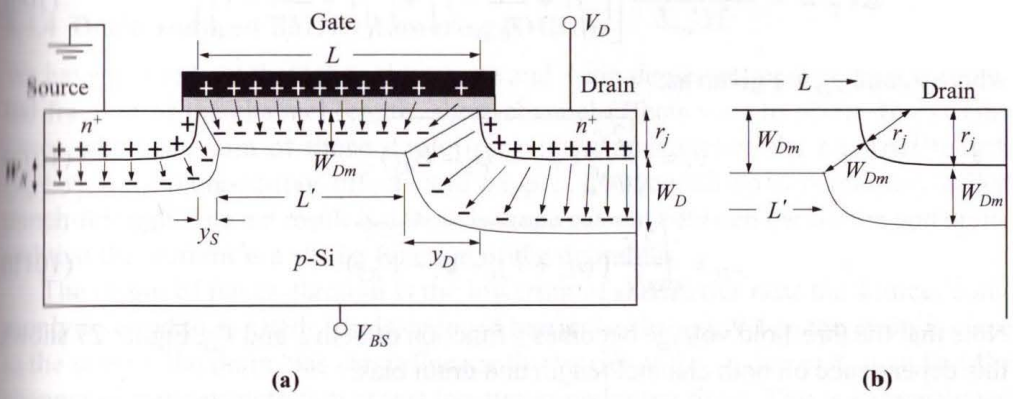


Fig. 26 Charge-conservation model for (a) $V_D > 0$, and (b) $V_D = 0$ where $W_D \approx W_S \approx W_{Dm}$. (After Ref. 47.)

Sandisk Technologies, Inc. - Ex. 1042, Page 000049

$$W_{Dm} = \sqrt{\frac{2\varepsilon_s(2\psi_B - V_{BS})}{qN_A}},$$

and 1-D analysis is sufficient.

For short-channel devices, the full effect of Q'_B on the threshold voltage is reduced, because near the source and drain ends of the channel, some field lines originating from the bulk charges under the channel region terminate at the source and drain instead of the gate (Fig. 26a). Note that the horizontal depletion-layer widths y_S and y_D are smaller than the vertical depletion-layer widths W_S and W_D, respectively because the transverse field strongly influences the potential distribution at the surface.

First-order estimation of the threshold voltage can be made by considering the charge partition. The total bulk depletion charge can be estimated by the trapezoid

$$Q'_B = ZqN_A W_{Dm}\left(\frac{L + L'}{2}\right).$$

For small drain bias, we can assume that $W_D \approx W_S \approx W_{Dm}$, and by straightforward trigonometric analysis (Fig. 26b),

$$L' = L - 2(\sqrt{r_j^2 + 2W_{Dm}r_j} - r_j).$$

The threshold-voltage shift from long-channel behavior is then given by

$$\Delta V_T = \frac{1}{C_{ox}}\left(\frac{Q'_B}{ZL} - qN_A W_{Dm}\right) = -\frac{qN_A W_{Dm}}{C_{ox}}\left(1 - \frac{L + L'}{2L}\right)$$

$$= -\frac{qN_A W_{Dm} r_j}{C_{ox} L}\left(\sqrt{1 + \frac{2W_{Dm}}{r_j}} - 1\right).$$

The negative sign means V_T is lowered and the transistor is easier to turn on. To take into account the effect of the drain voltage and the substrate bias, Eq. 99 can be modified to read [48]

$$\Delta V_T = -\frac{qN_A W_{Dm} r_j}{2C_{ox} L}\left[\left(\sqrt{1 + \frac{2y_S}{r_j}} - 1\right) + \left(\sqrt{1 + \frac{2y_D}{r_j}} - 1\right)\right],$$

where y_S and y_D are given as

$$y_S \approx \sqrt{\frac{2\varepsilon_s}{qN_A}(\psi_{bi} - \psi_s - V_{BS})},$$

$$y_D \approx \sqrt{\frac{2\varepsilon_s}{qN_A}(\psi_{bi} + V_D - \psi_s - V_{BS})}.$$

Note that the threshold voltage becomes a function of both L and V_D. Figure 27 shows this dependence on both channel length and drain bias.

Sandisk Technologies, Inc. - Ex. 1042, Page 000050

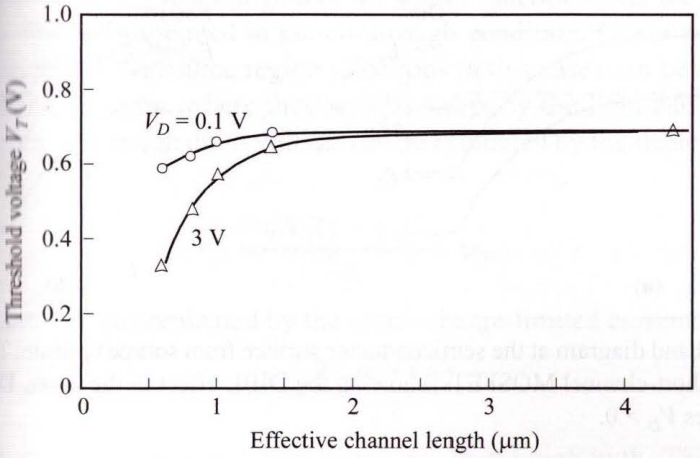


Fig. 17 Dependence of threshold voltage on channel length and drain bias. (After Ref. 49.)

6.4.3 Channel-Length Modulation

Figure 26a also shows that y_D is a high-field region where carriers are swept out efficiently, y_S is a transitional region where the electron concentration is higher than that in the main channel. So for consideration of the channel drift region, the *effective* channel length is more meaningful, given by

$$L_{\text{eff}} = L' = L - y_S - y_D. \tag{102}$$

This factor contributes to a drain-bias dependence of the effective channel length and partially accounts for the nonsaturating current with drain bias. Nevertheless, the change of channel length affects the current only linearly, whereas the barrier lowering caused by the drain bias, considered next, is much more pronounced since the current has an exponential dependence on the barrier.

6.4.4 Drain-Induced Barrier Lowering (DIBL)

We have pointed out that when the source and drain depletion regions are a substantial fraction of the channel length, short-channel effects start to occur. In extreme cases when the sum of these depletion widths approaches the channel length ($y_S + y_D \approx L$), more-serious effects will happen. This condition is commonly called punch-through. The net result is a large leakage current between the source and drain, and that this current is a strong function of the drain bias.

The origin of punch-through is the lowering of the barrier near the source, commonly referred to as DIBL (drain-induced barrier lowering). When the drain is close to the source, the drain bias can influence the barrier at the source end, such that the channel carrier concentration at that location is no longer fixed. This is demonstrated by the energy bands along the semiconductor surface, shown in Fig. 28. For a long-channel device, a drain bias can change the effective channel length, but the barrier at

Sandisk Technologies, Inc. - Ex. 1042, Page 000051

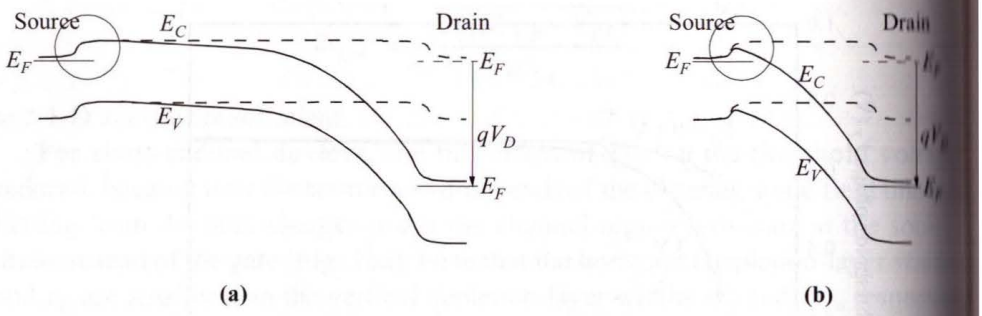


Fig. 28 Energy-band diagram at the semiconductor surface from source to drain, for (a) long-channel and (b) short-channel MOSFETs, showing the DIBL effect in the latter. Dashed lines $V_D = 0$. Solid lines $V_D > 0$.

the source end remains constant. For a short-channel device, this same barrier is no longer fixed. The lowering of the source barrier causes an injection of extra carriers thereby increasing the current substantially. This increase of current shows up in both above-threshold and subthreshold regimes.

Figure 28 shows that punch-through condition occurs at the semiconductor surface. In practical devices, it is common that the substrate concentration is reduced below the depth of the source/drain junction r_j. A reduced substrate doping widens the depletion widths so punch-through can also happen via a path in the bulk.

An example of severe punch-through characteristics above threshold is shown in Fig. 29a. For this device, at $V_D = 0$ the sum of y_S and y_D is 0.26 μm which is larger than the channel length of 0.23 μm. Therefore, the depletion region of the drain junction

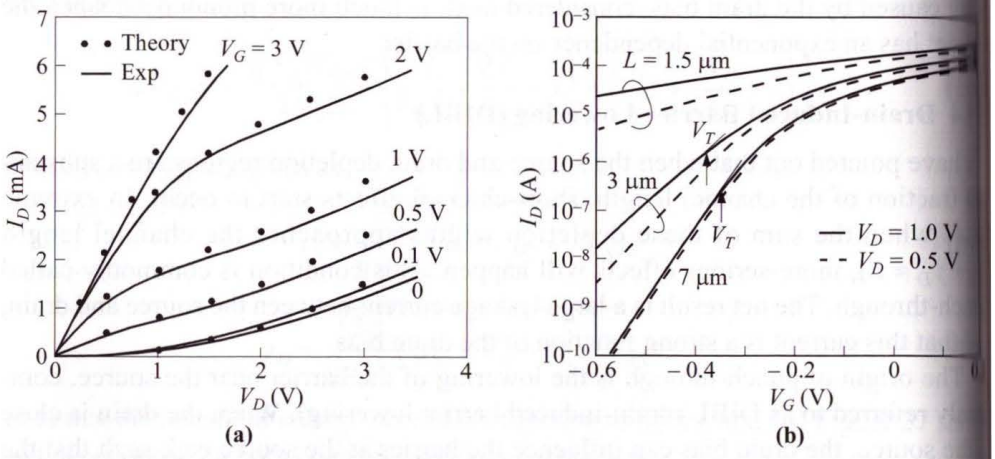


Fig. 29 Drain characteristics of MOSFETs showing DIBL effect. (a) Above threshold. L = 0.23 μm. d = 25.8 nm. $N_A = 7\times10^{16}$ cm^{-3}. (b) Below threshold. d = 13 nm. $N_A = 10^{14}$ cm^{-3}. (After Ref. 50.)

Sandisk Technologies, Inc. - Ex. 1042, Page 000052

has reached the depletion region of the source junction. Over the drain bias range down, the device is operated in punch-through condition. Under such a condition, ajority carriers in the source region (electrons in this case) can be injected into the pleted channel region, where they will be swept by the field and collected at the ain. The punch-through drain voltage can be estimated by the depletion approxima- on to be

$$V_{pt} \approx \frac{qN_A(L-y_S)^2}{2\varepsilon_s} - \psi_{bi}. \tag{103}$$

Drain current will be dominated by the space-charge-limited current:

$$I_D \approx \frac{9\varepsilon_s \mu_n A V_D^2}{8L^3} \tag{104}$$

where A is the cross-sectional area of the punch-through path. The space-charge-imited current increases with V_D^2 and is parallel to the inversion-layer current. The alculated points in the figure are from a 2-dimensional computer calculation incor- porating the punch-through effect and field-dependent mobility effect.

The DIBL effect on subthreshold current is shown in Fig. 29b, for various channel engths. The device with a 7-μm channel length shows long-channel behavior, that is, he subthreshold drain current is independent of drain voltage when $V_D > 3kT/q$ (Eq. 57). For $L = 3$ μm, there is a substantial dependence of current on V_D, with a cor- esponding shift of V_T (which is at the point of current departure of the I-V character- stic from the straight line). The subthreshold swing also increases. For an even horter channel, $L = 1.5$ μm, long-channel behavior is totally lost. The subthreshold wing becomes much worst and the device cannot be turned off any more.

6.4.5 Multiplication and Oxide Reliability

We pointed out earlier that due to nonideal scaling, the internal electric fields in MOSFETs would increase with shorter channel lengths. In this section we discuss the anomalous currents associated with high fields, as well as their impacts. Figure 30 depicts all parasitic currents in addition to the main channel current. Note that the highest field occurs near the drain, and this is the location where most of the anoma- lous currents originate.

First, as the channel carriers (electrons) go through the high-field region, they acquire extra energy from the field without losing it to the lattice. These energetic car- riers are called *hot carriers* whose kinetic energy is measured above the conduction band E_C. This extra energy, if larger than the Si/SiO_2 barrier (3.1 eV), enables them to escape to the oxide layer and to the gate terminal, and gives rise to a gate current.

Another major phenomenon happening in the high-field region is impact ioniza- tion which generates extra electron-hole pairs. These extra electrons go directly to the drain and add to the channel current. The path of the generated holes, however, is more diverse. A small fraction of them are driven to the gate, analogous to the hot electrons mentioned before. The majority of the generated holes flow to the substrate. For short-channel devices, some holes will go to the source. The division of these

Sandisk Technologies, Inc. - Ex. 1042, Page 000053

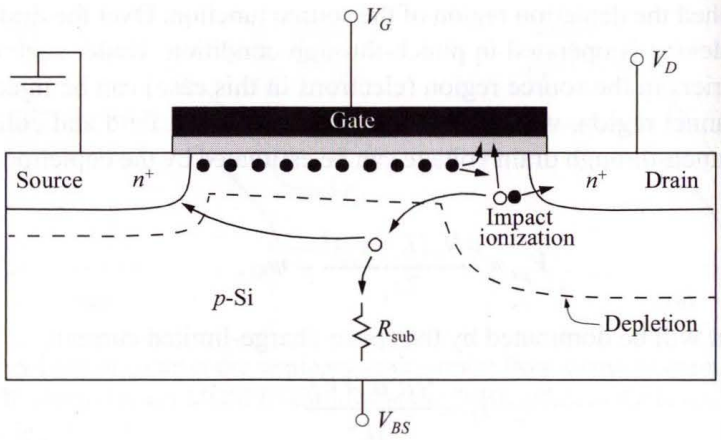


Fig. 30 Current components of a MOSFET under high fields.

paths depends on how good the substrate tie is. A perfect substrate tie ($R_{sub} = 0$) will sink all the hot holes and none of them will go to the source. As explained later, holes going to the gate or source will produce undesirable effects.

An example of the MOSFET terminal currents, including the gate current and substrate current, are shown in Fig. 31. Note that the gate current is due to hot electrons and hot holes *over* the barrier and is different from carriers tunneling *through* the barrier. This hot-carrier gate current peaks approximately at $V_G \approx V_D$. It is gener

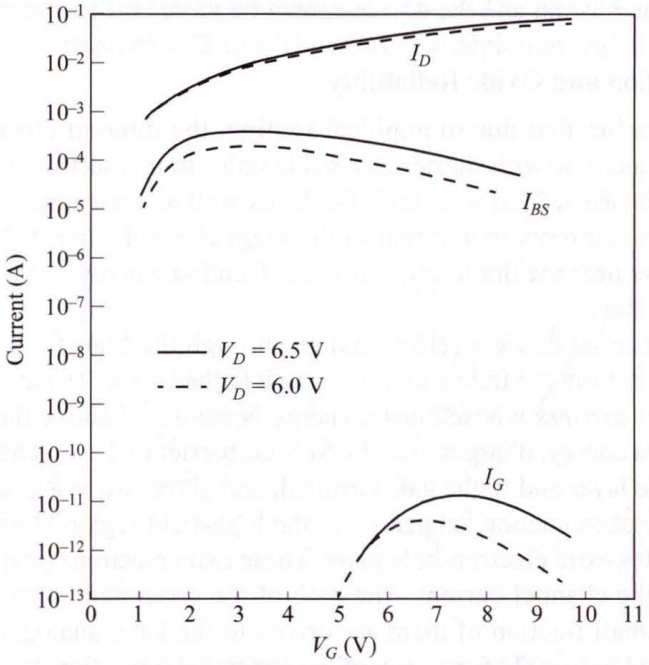


Fig. 31 Drain current, substrate current, and gate current vs. gate voltage of a MOSFET. L/W = 0.8/30 μm. (After Ref. 51.)

Sandisk Technologies, Inc. - Ex. 1042, Page 000054

ally very small, and it by itself does not post a problem since it is negligible compared to the channel current. The impacts are the damages it creates. It is well known that these hot carriers create oxide charges and interface traps.[52] As a result, device characteristics change with operation time. In particular, the threshold voltage shifts, generally to a higher value, and the transconductance g_m degrades because of interface traps and reduced channel mobility. The subthreshold swing becomes larger because of increased interface-trap density. To reduce oxide charging, the density of water-related traps in the oxide should be minimized,[53] because such traps are known to capture hot carriers on their passage. In order to ensure MOSFETs' performance over a reasonable time, it is important to quantify the device lifetime over which the device parameters do not drift outside a given range, for a given bias condition. This is part of the specification of a MOSFET technology.

Figure 31 also shows the general characteristics of the substrate current. The substrate current has a unique bell shape as a function of gate bias.[54] It increases first with V_G, reaches a maximum then decreases. This maximum occurs usually around $V_G \approx V_D/2$. The maximum in I_{BS} can be explained as follows. Assuming that the impact ionization occurs uniformly in the high-field region, the substrate current can be written as

$$I_{BS} \approx I_D \alpha(\mathscr{E}) y_D, \tag{105}$$

where α is the ionization coefficient, the number of electron-hole pairs generated per unit distance, and is a strong function of the electric field; and y_D is the high-field or pinch-off region. For a given V_D, as V_G increases, both I_D and V_{Dsat} increase ($V_{Dsat} \approx V_G - V_T$). When V_{Dsat} increases, the lateral field [$\approx (V_D - V_{Dsat})/y_D$] decreases, causing a reduction of α. Thus we have two conflicting factors. The initial increase of I_{BS} is caused by the increase of drain current with V_G, and at larger V_G, the decrease of I_{BS} is due to the decrease of α. Maximum I_{BS} occurs where the two factors balance each other. Empirically, the substrate current can be expressed as

$$I_{BS} = C_2 I_D (V_D - V_{Dsat}) \exp\left(\frac{-C_3}{V_D - V_{Dsat}}\right), \tag{106}$$

where C_2 and C_3 are constants.

For a short-channel device, that is, a small source-drain separation, avalanched holes have increasing tendency to flow to the source.[55] This hole current constitutes a base current of a parasitic *n-p-n* bipolar-transistor action, where source-substrate-drain is equivalent to the emitter-base-collector configuration. For every hole that goes to the source, there are many electrons injected to the substrate. These electrons will be collected by the drain and show up as additional drain current. This bipolar current gain I_n/I_p is roughly determined by the ratio of N_D/N_A. Another way to look at this is that the substrate current develops a substrate voltage $I_{BS} \times R_{sub}$ which puts the source-substrate *p-n* junction under forward bias, thereby injecting electrons into the substrate. An additional effect is that the higher substrate potential lowers the threshold voltage for the surface channel and increases the surface-channel current. Both effects will increase the total drain current. These effects will become worst with an increase of R_{sub} and shorter L. In the extreme case of a MOSFET without a

Sandisk Technologies, Inc. - Ex. 1042, Page 000055

substrate contact ($R_{sub} = \infty$) such as an SOI or TFT structure (Section 6.5.4), the output curves show sudden rise of I_D with V_D. This is referred to as *kink effect* in the output characteristics.

In more-severe cases, the substrate current can induce source-drain breakdown from this parasitic *n-p-n* bipolar action. Analogous to the analysis of a bipolar breakdown (Section 5.2.3), since the source-drain distance is much shorter than the drain to substrate contact, the base can be treated as open, and the MOSFET source-drain breakdown is given by the parasitic open-base bipolar breakdown (Eq. 47 in Chapter 5)

$$V_{BDS} = V_{BDx}(1 - \alpha_{npn})^{1/n}. \tag{108}$$

Here V_{BDx} is the drain-substrate *p-n* junction breakdown voltage, and n describes the shape of this diode breakdown characteristics. α_{npn} is the common-base current gain given by the product of the base transport factor α_T and emitter efficiency γ. Assuming $\gamma \approx 1$, we have

$$\begin{aligned} \alpha_{npn} &= \alpha_T \gamma \approx \alpha_T \\ &\approx 1 - \frac{L^2}{2L_n^2} \end{aligned} \tag{109}$$

where L (channel length) is the effective base width, and L_n is the electron diffusion length in the substrate. From the above equations the breakdown voltage between source and drain is obtained for a short-channel MOSFET as

$$V_{BDS} \approx \frac{V_{BDx}}{2^{1/n}} \left(\frac{L}{L_n}\right)^{2/n}. \tag{110}$$

Equation 109 has been used to fit the data quite well, provided that the factor n is chosen to be 5.4.[55] The difference in breakdown voltage for different junction curvature can be explained by the dependence of V_{BDx} on r_j, as discussed in Chapter 2. To reduce the parasitic transistor effect, the resistance of the substrate R_{sub} should be minimized so that the product of $I_{BS} \times R_{sub}$ remains smaller than ≈ 0.6 V. Then the breakdown voltage of a short-channel MOSFET will no longer be limited by this parasitic bipolar effect, and higher voltages and more-reliable operation can be expected.

The drain-gate overlap region forms a gated-diode structure. For a thin oxide together with an abrupt junction, avalanche can occur during certain bias conditions and it results in a drain leakage current going to the substrate. Such gated-diode avalanche current is called gate-induced drain leakage (GIDL) and the mechanism has been discussed in more details in Section 2.4.3. For an *n*-channel device, with a fixed drain bias, the normal channel current decreases with decreasing gate bias into the subthreshold regime. At some gate bias, the drain current becomes the GIDL current and it rises again with more negative gate bias. Very often, in short-channel devices this GIDL current already exists at $V_G = 0$, imposing a leakage current component in their off-state.[56]

Sandisk Technologies, Inc. - Ex. 1042, Page 000056

6.5 MOSFET STRUCTURES

Up to now, Si MOSFET has been the workhorse of the electronics industry. As such, the MOSFET channel length and other dimensions have been pushed to shrink for the benefits of performance and density (see Fig. 1). While there is much discussion on what dimensions are the scaling limits,[57] it is certainly true that device scaling is getting increasingly difficult and has diminishing return. There are many possible reasons for hitting the end of scaling. These include: sensitivity of statistical doping fluctuations and surface charges, various forms of short-channel effects, quantum confinement in inversion layer which places a limit on gate capacitance, source/drain series resistance, etc. Most-recent data suggest that channel length below 20 nm is feasible, even with planar technology.[58–59] However, for practical applications, the scaling limit is most likely around 10 nm, even for 3-dimensional structures.

Many device structures have been proposed to control short-channel effects and improve MOSFET performance. We shall now consider the MOSFET structure broken down here into separate parts: channel doping, gate stack, and source/drain design. This is followed by some representative device structures for ultimate performance and special purposes.

6.5.1 Channel Doping Profile

Figure 32 shows the schematic of a typical high-performance MOSFET structure based on planar technology. The channel doping profile has a peak level slightly below the semiconductor surface. Such a retrograde profile is achieved with ion implantation, often of multiple doses and energies. The low concentration at the surface has the advantages of higher mobility due to reduced normal field and low threshold voltage. The high peak concentration below the surface is to control punch-through and other short-channel effects. Lower concentration is typically below the junction depth. It reduces the junction capacitance as well as the substrate-bias effect on threshold voltage.

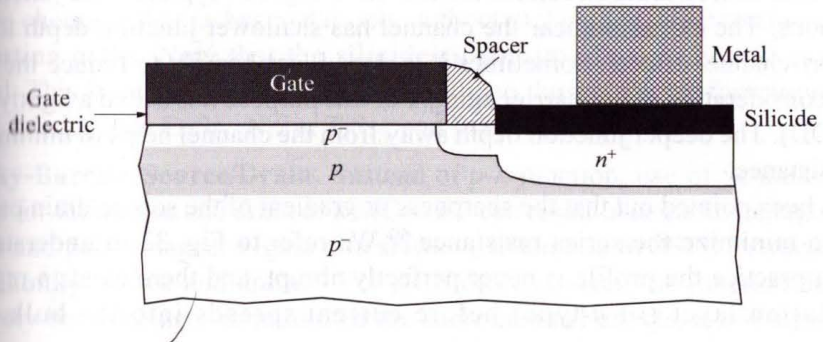


Fig. 32 High-performance MOSFET planar structure with a retrograde channel doping profile, two-step source/drain junction, and self-aligned silicide source/drain contact.

Sandisk Technologies, Inc. - Ex. 1042, Page 000057

6.5.2 Gate Stack

The gate stack consists of the gate dielectric and the gate contact material. The gate dielectric has been exclusively SiO_2 right from the birth of MOSFET. In fact, the ideal Si-SiO_2 interface is the main factor responsible for the success of MOSFET. As the oxide thickness is scaled into the range below ≈ 2 nm, fundamental problem of tunneling and technological difficulty of defects start to demand alternatives. A sensible and popular solution that is actively sought after is a material with high dielectric constant, called high-K dielectric. Such a high-K dielectric can have a thicker physical thickness for the same capacitance, thus reducing its electric field and technological problem related to defects. With the value of dielectric constant considered, the common terminology used is the equivalent oxide thickness [EOT = thickness $\times$ $K(SiO_2)/K$]. Material options being examined are Al_2O_3, HfO_2, ZrO_2, Y_2O_3, La_2O_3, Ta_2O_5, and TiO_2. The dielectric constants for these materials range from 9 to 30, except for TiO_2 which is larger than 80. As seen [$K(SiO_2) = 3.9$], the EOT can easily be extended to below 1 nm if some of the options eventually are proven to be successful. Nevertheless, readers are reminded of the quantum-mechanical effect discussed in Section 4.3.6 which puts a limit on the gate capacitance.

The gate material has been polysilicon for a long time. The advantages of a poly-Si gate is its compatibility with the silicon processing, and its ability to withstand high-temperature anneal that is required after self-aligned source/drain implantation. Another important factor is that the work function can be varied by doping into n-type and p-type. Such flexibility is crucial for a symmetric CMOS technology. One limitation of the poly-Si gate is its relatively high resistance. This does not result in penalty of dc characteristics since the gate terminates on the gate insulator. The penalty shows up in high-frequency parameters such as noise and f_{max} (Section 6.6.1). Another shortcoming of poly-Si gate is the finite depletion width at the oxide interface. This reduces the effective gate capacitance and becomes more severe with thinner oxides. To circumvent the problems of resistance and depletion, gate materials of silicides and metals are obvious choices. Potential candidates are TiN, TaN, W, Mo, and NiSi.

6.5.3 Source/Drain Design

Details of the source/drain structures are shown in Fig. 32. Typically the junction has two sections. The extension near the channel has shallower junction depth to minimize short-channel effects. Sometimes it is doped less heavily to reduce the lateral field for consideration of hot-carrier aging. For this purpose it is called a lightly doped drain (LDD). The deeper junction depth away from the channel helps to minimize the series resistance.

It has been pointed out that the sharpness or gradient of the source/drain profile is critical to minimize the series resistance.[60] We refer to Fig. 33 to understand its origin. In practice the profile is never perfectly abrupt, and there exists a region of accumulation layer (of n-type) before current spreads into the bulk of the

Sandisk Technologies, Inc. - Ex. 1042, Page 000058

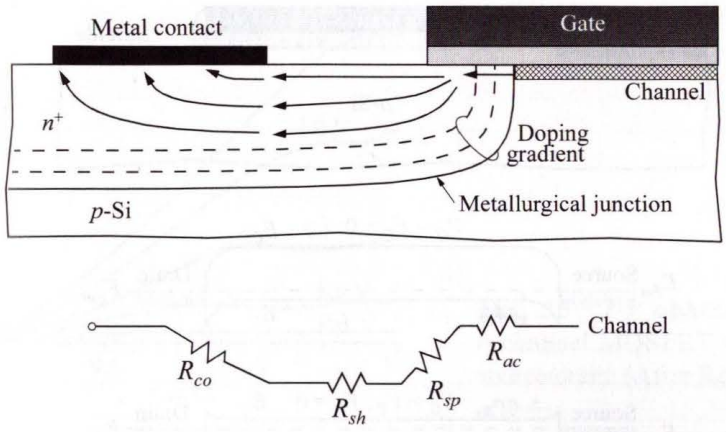


Fig. 33 Detailed analysis of different components of parasitic source/drain series resistance. R_{ac} is accumulation-layer resistance due to doping gradient. R_{sp} = spreading resistance. R_{sh} = sheet resistance. R_{co} = contact resistance. (After Ref. 60.)

source/drain. This accumulation-layer resistance R_{ac} is related to the transition distance before the doping reaches a critical level.

A major milestone for source/drain design is the development of the silicide contact technology which started in the early 1990s. Unlike the metal contact, the silicide can be made self-aligned to the gate, as shown in Fig. 32, thus minimizing the sheet-resistance component (R_{sh}) between the contact and the channel. In this way the silicide has become the metal contact because contact resistance between metal and silicide is very small. This *self-aligned silicide* process has been coined *salicide*. The salicide process is described is follows. After the gate definition, an insulator spacer is formed on the sides of the gate. A metal layer for silicidation is deposited uniformly, which at this stage is shorting the gate and the source/drain. After a thermal reaction at low temperature ($\approx 450°C$), the metal reacts with silicon to form silicide on the source/drain region. Silicide formation on the gate is optional depending on whether the gate is capped with an insulation layer as part of the gate stack. Metal over the spacer region and the field region (between transistors, not shown) remains metal since there is no exposed silicon for reaction. The metal is then removed with a selective chemical that etches metal only without etching silicide, thereby removing the shorting paths. Note that the silicide/silicon interface in Fig. 32 is slightly recessed. This is due to the consumption of silicon during silicide formation. Examples for salicides are $CoSi_2$, $NiSi_2$, $TiSi_2$, and PtSi.

Schottky-Barrier Source/Drain. Instead of *p-n* junction, use of Schottky-barrier contacts for the source and drain of a MOSFET can result in some advantages in fabrication and performance. Figure 34a shows a schematic MOSFET structure with such Schottky source and drain.[61] For a Schottky contact, the junction depth can effectively be made zero to minimize the short-channel effects. *n-p-n* bipolar-transistor action is also absent for undesirable effects such as bipolar breakdown and

Sandisk Technologies, Inc. - Ex. 1042, Page 000059

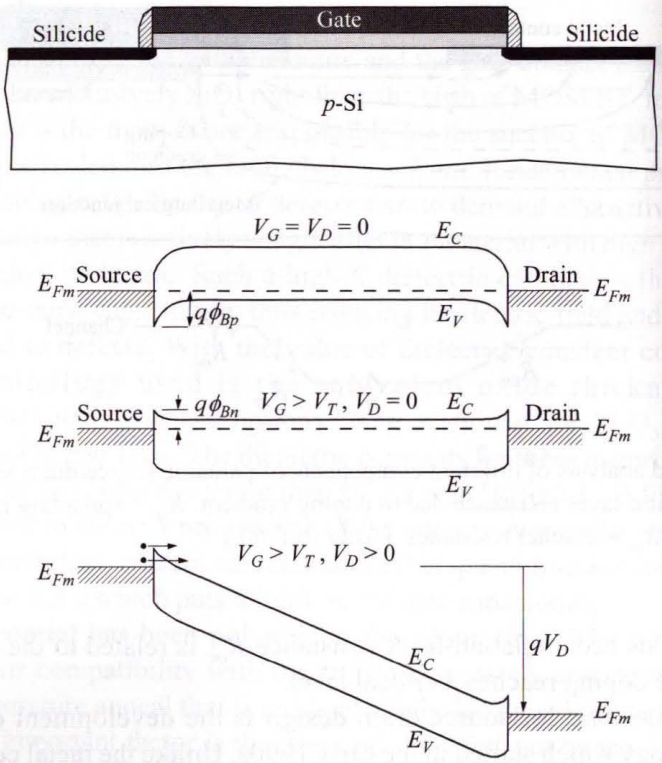


Fig. 34 MOSFET with Schottky-barrier source and drain. (a) Cross-sectional view of device. (b)–(d) Band diagrams along semiconductor surface under various biases.

latch-up phenomenon[62] in CMOS circuits. Eliminating high-temperature implant anneal can promote better quality in the oxides and better control of geometry. In addition, this structure can be made on semiconductors such as CdS where p-n junctions cannot be easily formed.

Figures 34b–d show the working principle of a Schottky source/drain. At thermal equilibrium with $V_G = V_D = 0$, the barrier height of the metal to the p-substrate for holes is $q\phi_{Bp}$ (e.g., 0.84 eV for an ErSi-Si contact).[63] When the gate voltage is above threshold to invert the surface from p-type to n-type, the barrier height between the source and the inversion layer (electrons) is $q\phi_{Bn} = 0.28$ eV. Note that the source contact is reverse biased under operating conditions (Fig. 34d). For a 0.28-eV barrier, the thermionic-type reverse-saturation current density is of the order of 10^3 A/cm^2 at room temperature. To increase current density, metals should be chosen to give the highest majority-carrier barrier such that the minority-carrier barrier height is minimized. Additional current due to tunneling through the barrier should help to improve the supply of channel carriers. At the present, making the structure on a p-type Si substrate for n-channel MOSFET is more difficult compared to p-channel device with n-substrate, because metals and silicides that give large barrier heights on p-type silicon are less common.

Sandisk Technologies, Inc. - Ex. 1042, Page 000060

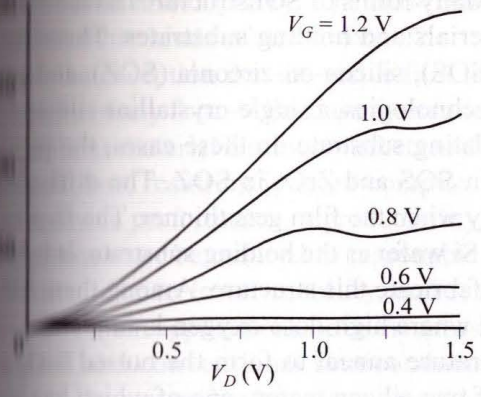


Fig. 35 *I-V* characteristics of *n*-channel MOSFET with Schottky source/drain. (After Ref. 63.)

The disadvantages of the Schottky source/drain are high series resistance due to its finite barrier height, and higher drain leakage current. Typical *I-V* curves show that current is starved at low-drain bias (Fig. 35). Also note that as shown in Fig. 34, the metal or silicide contact has to extend underneath the gate for continuity. This process is much more demanding than a junction source/drain which is done by self-aligned implantation and diffusion.

Raised Source/Drain. An advanced design is the raised source/drain where a heavily doped epitaxial layer is grown over the source/drain regions (Fig. 36a). The purpose is to minimize junction depth to control short-channel effects. Note that an extension underneath the spacer is still needed for continuity. An alternative is the recessed-channel MOSFET where the junction depth r_j is zero or negative (Fig. 36b).[64] The drawback of the recessed-channel structure, especially for submicron devices, is the difficulty in controlling the contour and the oxide thickness at the corners where the threshold voltage is determined. Also, oxide charging may be worsened because more hot-carrier injection will occur.

6.5.4 SOI and Thin-Film Transistor (TFT)

SOI. Unlike thin-film transistor, the top silicon layer of an SOI (silicon-on-insulator) wafer is high-quality single-crystalline material that is suitable for high-performance

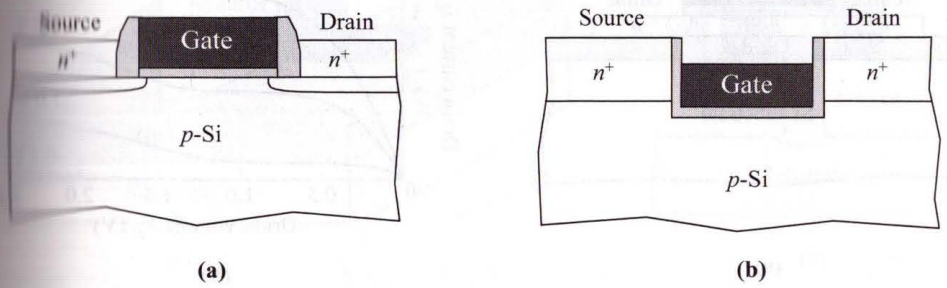


Fig. 36 Means to reduce source/drain junction depth and series resistance. (a) Raised source/drain. (b) Recessed channel.

Sandisk Technologies, Inc. - Ex. 1042, Page 000061

and high-density integrated circuits.[65] Many forms of SOI structures have been d onstrated with different insulator materials and holding substrates. These incl silicon-on-oxide, silicon-on-sapphire (SOS), silicon-on-zirconia (SOZ), and sili on-nothing(air gap). In SOS and SOZ technologies, a single-crystalline silicon fil epitaxially grown on a crystalline insulating substrate. In these cases, the insula are the substrates themselves, Al_2O_3 in SOS and ZrO_2 in SOZ. The difficulti these techniques are the material quality when the film gets thinner. The first opti using oxide as an insulator and another Si wafer as the holding substrate, is by fa most popular. There are many ways to fabricate this structure. Among them SIM (separation by implantation of oxygen) where high-dose oxygen is implanted on silicon wafer followed by high-temperature anneal to form the buried SiO_2 la Another technique involves bonding of two silicon wafers one of which has an o dized layer followed by thinning or removal of the majority of the top wafer until thin silicon layer is left. One technique uses lateral epitaxial growth of silicon over oxide layer, starting from a seed opening to the substrate. Another uses laser recry tallization, transforming amorphous silicon deposited onto the oxide layer in single-crystalline material, or into poly-crystalline form with large grain size.

Figure 37a shows a schematic diagram of an *n*-channel MOSFET made on an SI substrate, with its typical *I-V* characteristics shown in Fig. 37b. The kinks associat with floating body without a substrate tie are noticeable.

The advantages of the SOI substrate include improved MOSFET scaling due to thin body. A thin body can alleviate most problem with punch-through such that th channel can be doped lightly. The subthreshold swing is known to be improved. Th buried oxide layer serves as good isolation to reduce capacitance to the substra giving rise to higher speed. As shown in Fig. 37a, device isolation is much easi simply by removing the surrounding thin film. This can significantly improve th circuit density. This type of isolation, as opposed to junction isolation in planar tec nology, also eliminates latch-up phenomenon in CMOS circuits. The disadvantage of SOI is higher wafer cost, potentially inferior material properties, the kink effe and worse heat conduction because of the oxide layer.

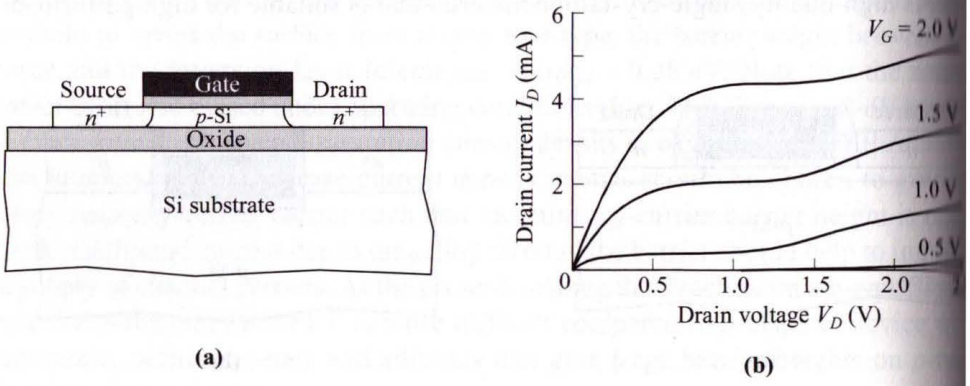


Fig. 37 (a) Typical structure of MOSFET on SOI wafer, and (b) its drain characteristi (After Ref. 66.)

Sandisk Technologies, Inc. - Ex. 1042, Page 000062

Thin-Film Transistor (TFT). The thin-film transistor usually refers to MOSFET as opposed to other kinds of transistors. The structure is similar to MOSFET built on SOI with the exception that the active film is a deposited thin film and that the substrate can be of any form.[67] Because the semiconductor layer is formed by deposition, the amorphous material has more defects and imperfections than in single-crystalline semiconductors, resulting in more complicated transport processes in the TFT. To improve device performance, reproducibility, and reliability, the bulk and interface-trap densities must be reduced to reasonable levels. In a TFT, the current is always very limited due to lower mobility, and leakage current is always higher due to defects. The main applications lie in the areas where a large-area or flexible substrate is required and conventional semiconductor processing is not feasible. A good example is large-area display where an array of transistors is required to control the array of lighting elements. In such applications, device performance such as current or speed is not critical.

6.5.5 Three-Dimensional Structures

In device scaling, the optimum design is with MOSFET built on a body of ultra-thin layer such that the body is fully depleted under the whole bias range. A design to achieve this more efficiently is to have a surround gate structure that encloses the body layer from at least two sides. Two examples of these 3-dimensional structures are shown in Fig. 38. They can be classified according to their current-flow pattern; the horizontal transistor[68–69] and vertical transistor.[70] While both of these are very challenging from a fabrication point of view, the horizontal scheme is more compatible with SOI technology and more data are reported in the literature. A new set of difficulties arise from the fact that the majority or all of the channel surface is on a vertical wall, for both of these structures. This fact presents great challenges in achieving a smooth channel surface from etching and growth or deposition of gate dielectrics on these surfaces. Formation of the source/drain junction is no longer

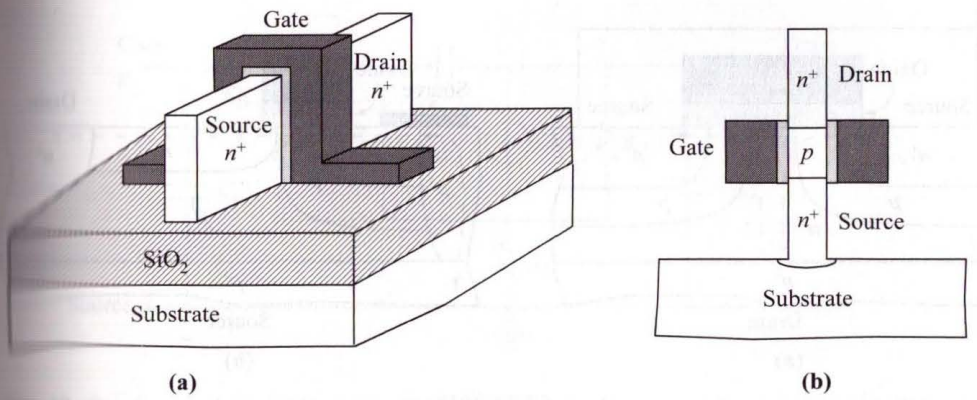


Fig. 38 Schematic 3-dimensional MOSFETs. (a) Horizontal structure. (b) Vertical structure. Note commonality of surround gate and thin body.

Sandisk Technologies, Inc. - Ex. 1042, Page 000063

trivial by means of ion implantation. Salicide formation will also be much more difficult. Whether one of these turns out to be the device choice in the future remains to be seen.

6.5.6 Power MOSFETs

In general, power MOSFETs employ thicker oxides, deeper junctions, and have longer channel lengths. These generally post a penalty on device performance such as transconductance (g_m) and speed (f_T). Nevertheless, power applications from MOSFETs have been on the rise, for example, due to the increasing demand of cellular phones and cellular base stations which require extra-high voltage. We will present two power structures that are designed for RF power applications.

DMOS. As the name implies, in the DMOS (double-diffused MOS) transistor shown in Fig. 39a the channel length is determined by the higher diffusion rate of the p-dopant (e.g., boron) compared to the n^+-dopant (e.g., phosphorus) of the source. This technique can yield very short channels without depending on a lithographic mask. The p-diffusion serves as channel doping and has good punch-through control. The channel is followed by a lightly doped n^--drift region. This drift region is long compared to the channel, and it minimizes the peak electric field in this region by maintaining a uniform field.[71] Usually the drain is located at the substrate contact. The field near the drain is the same as in the drift region, so avalanche breakdown, multiplication, and oxide charging are lessened compared to conventional MOSFETs.

However, the threshold voltage V_T is more difficult to control in a DMOS transistor because the channel doping is no longer constant along its length.[72] Since V_T is determined by the local doping concentration along the semiconductor surface, varying doping level leads to variations in V_T as a function of distance and bias. Also, the localization of punch-through control by a thin p-shield region requires a high-

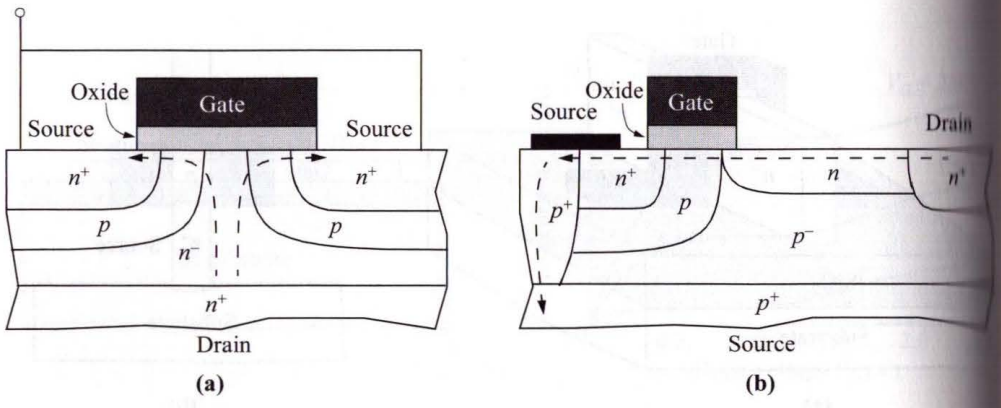


Fig. 39 (a) Vertical DMOS transistor and (b) LDMOS transistor. Current path is indicated by dashed line. In LDMOS transistors, it is common to connect the source to the substrate to reduce inductance of the bonding wire.

Sandisk Technologies, Inc. - Ex. 1042, Page 000064

doping level compared to a conventional structure and it leads to poorer turn-off behavior for DMOS transistors.

LDMOS. The major difference of the LDMOS (laterally diffused MOS) transistor (Fig. 39b) from a DMOS transistor is that it has a lateral current-flow pattern. The drift region here is an implanted horizontal region. Such a horizontal arrangement enables the p^+-substrate to deplete this drift region at high drain bias. Yet at low drain bias its higher doping gives lower series resistance. This drift region, thus, behaves as a nonlinear resistor. At low drain bias, its resistance is determined by $1/nq\mu$. At high drain bias, this region is fully depleted so a large voltage drop can be supported. This concept is called RESURF (reduced surface field) technology.[73] Because of this feature, the drift region can be doped with a higher concentration compared to the DMOS transistor for a lower on-resistance. Another advantage of the LDMOS transistor is that the source can be tied internally to the substrate by a deep p-type diffusion. This avoids using a bond wire that has high inductance to the source. The LDMOS transistor can thus perform at higher speed.

6.6 CIRCUIT APPLICATIONS

6.6.1 Equivalent Circuit and Microwave Performance

The MOSFET is ideally a transconductance amplifier with an infinite input resistance and a current generator at the output. In practice, however, we have other nonideal circuit elements. An equivalent circuit is shown in Fig. 40 for the common-source connection. The gate resistance R_G is associated with the gate contact material over the oxide. The input resistance R_{in} is caused by tunneling current through the thin gate insulator, and it also includes any conductance through defects. This of course is a function of the oxide thickness. For a thermally grown silicon dioxide layer, this leakage current between the gate and the channel is negligibly small; thus, the input resistance is very high, one of the main advantages of a MOSFET. For oxides below

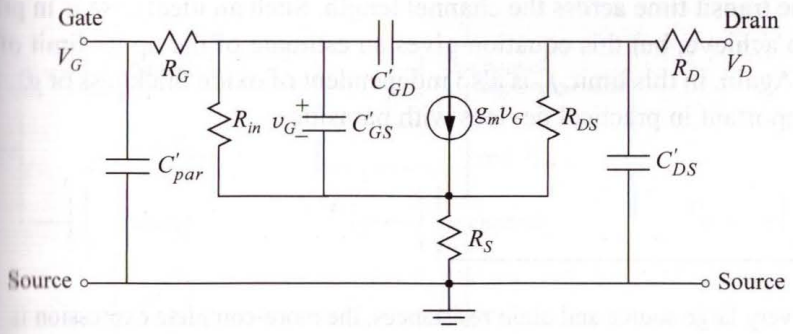


Fig. 40 Small-signal equivalent circuit of MOSFET for the common-source configuration. v_G is the small signal of V_G. Capacitance symbols with prime designate total capacitance in Farad as opposed to capacitance per unit area.

Sandisk Technologies, Inc. - Ex. 1042, Page 000065

a thickness of ≈ 5 nm, the tunneling current starts to become an important factor. The gate capacitance C'_G ($= C'_{GS} + C'_{GD}$) is mostly due to C_{ox} multiplied by the active channel area $Z \times L$. In practical devices, the gate extends somewhat above the source and drain regions, and these overlap capacitances add to the total C'_G. This fringing effect is also an important contribution to the feedback capacitance C'_{GD}. The drain output resistance R_{DS} is due to the fact that the drain current does not truly saturate with the drain bias. This effect is especially pronounced for short-channel devices, as part of the short-channel effects discussed earlier. The output capacitance C'_{DS} consists mostly of the two p-n junction capacitances connected in series through the semiconductor bulk.

In the saturation region, V_D and thus R_D has little effect on the drain saturation current. The R_S affects the effective gate bias, and the extrinsic transconductance is given by

$$g_{mx} = \frac{g_m}{1 + R_S g_m}. \tag{11}$$

In analyzing the microwave performance, we follow the same procedure as in Section 5.3.1 for obtaining the cutoff frequency f_T, defined as the frequency for unity current gain (ratio of drain current to gate current),

$$f_T = \frac{g_m}{2\pi(C'_G + C'_{par})} = \frac{g_m}{2\pi(ZLC_{ox} + C'_{par})}, \tag{11}$$

where C'_{par} is the total input parasitic capacitance. (See footnote* for a more-complete f_T expression with excessive R_S and R_D.) It is interesting to note that if C'_{par} is only due to the gate-drain and gate-source overlap regions, it has the same oxide thickness dependence as g_m, and f_T would be independent of C_{ox} or oxide thickness. In the ideal case that there is zero parasitics, it can be shown that

$$f_T = \frac{g_m}{2\pi ZLC_{ox}} = \frac{v}{2\pi L} = \frac{1}{2\pi\tau_t} \tag{11}$$

where τ_t is the transit time across the channel length. Such an ideal case is in practice impossible to achieve, but this equation gives an estimate of the upper limit of f_T by using $v = v_s$. Again, in this limit, f_T is also independent of oxide thickness or g_m. However, g_m is important in practical devices with parasitics.

---

* In cases of very large source and drain resistances, the more-complete expression is

$$f_T = \frac{g_m}{2\pi\left[C'_G\left(1 + \frac{R_D + R_S}{R_{DS}}\right) + C'_{GD}g_m(R_D + R_S) + C'_{par}\right]}.$$

Sandisk Technologies, Inc. - Ex. 1042, Page 000066

Another figure-of-merit for microwave performance is the maximum frequency of oscillation f_{max}, the frequency at which the unilateral gain becomes unity. It is given by

$$f_{max} = \sqrt{\frac{f_T}{8\pi R_G C'_{GD}}}. \tag{113}$$

So for high-frequency performance, the most-important device parameters are g_m, C_{gs} and all other parasitic capacitances.

6.6.3 Basic Circuit Blocks

In this section we present the basic digital-circuit building blocks in both logic and memory circuits. The most-basic unit for a logic circuit is the inverter. Different configurations for MOSFET inverters are shown in Fig. 41. By far the most common is the CMOS (complementary MOS) inverter where both *n*-channel and *p*-channel transistors are used. This logic consumes very low dc power because when the input is either high or low, one of the transistors in series is off so that there is very little steady-state current (subthreshold current) passing through them. In fact, this is one of the main advantages and applications of MOSFETs where the insulated gate can withstand input voltage of any polarity. Such an arrangement is much more difficult with bipolar transistors or MESFETs without putting a large resistor in front of the input. In an NMOS logic (Fig. 41b), the load of the *p*-channel transistor is replaced with a depletion-mode *n*-channel transistor. The advantage is a simpler technology since a *p*-channel device is not required at the expense of higher dc power. This depletion-mode device with the gate tied to the source is basically a two-terminal nonlinear resistor, which is an improvement compared to a simple resistor load shown in Fig. 41c.

Two basic MOSFET memory cells, for SRAM (static random-access memory) and DRAM (dynamic random-assess memory) circuits, are shown in Fig. 42. The

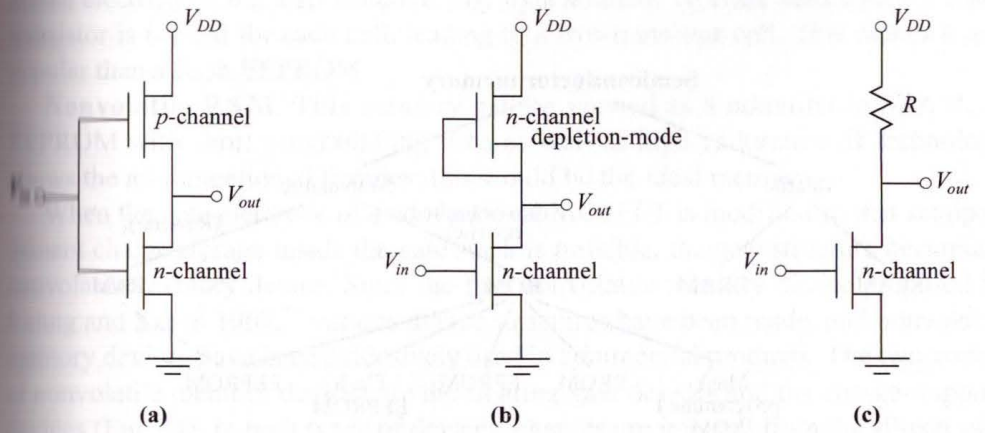


Fig. 41 Versions of inverters: (a) CMOS logic. (b) NMOS logic with depletion-mode–transistor load, and (c) NMOS logic with resistor load.

Sandisk Technologies, Inc. - Ex. 1042, Page 000067

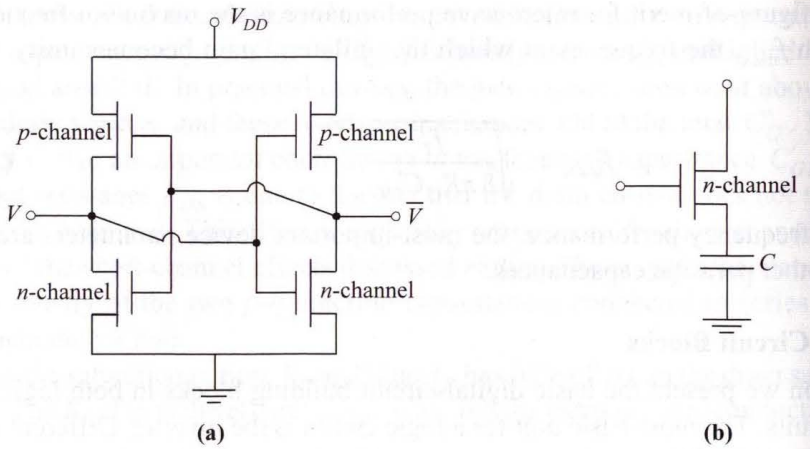


Fig. 42 Basic memory cells in (a) SRAM and (b) DRAM circuits.

SRAM cell has two CMOS inverters connected back to back. It is a latch and a stable cell but it requires four transistors (six including controls for word line and bit line). The DRAM cell only requires one transistor and, thus, has very high memory density. Its memory information is stored as a charge across the capacitor. Since there is finite leakage of charge in the nonideal capacitor, the cell needs to be refreshed periodically, typically at a frequency of ≈ 100 Hz.

6.7 NONVOLATILE MEMORY DEVICES

Semiconductor memory devices are classified in Fig. 43. The first division is based on their ability to maintain their states when the power is disconnected. As the names imply, a volatile memory loses the content, but a nonvolatile memory does not need voltage to maintain the data.[74–76]

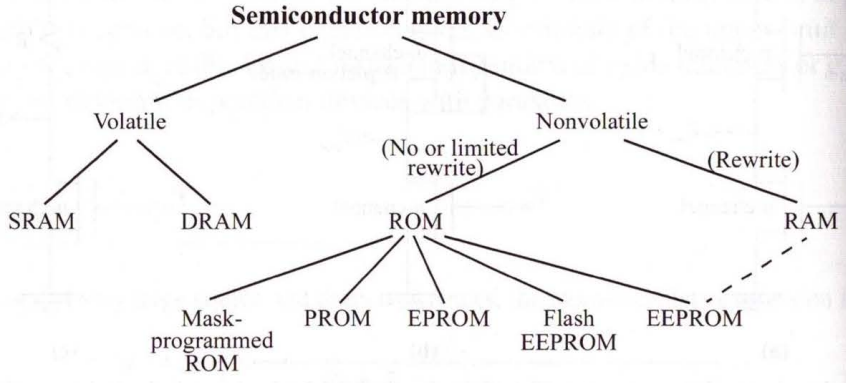


Fig. 43 Classification of semiconductor memories.

Sandisk Technologies, Inc. - Ex. 1042, Page 000068

Before getting into each type of nonvolatile semiconductor memory, we should first clarify the difference between a *RAM* and a *ROM*. A RAM (random-access memory) has an *x-y* address for each cell, which distinguishes it from other serial memories such as magnetic memory. Strictly speaking, a ROM (read-only memory) also has random-access capability since the addressing architecture is similar. In fact the read processes of the RAM and ROM are almost identical. More appropriately, a RAM is sometimes called a read-write memory. However, the nonvolatile ROM has long started to develop some extent of rewriting capability. So the main difference now between a RAM and a ROM is the ease and frequency of erasing and programming. A RAM has almost equal opportunity of rewrite and read. A ROM in general has much more frequent read than rewrite. It itself has a spectrum of rewriting capability, ranging from a pure ROM without any writing capability to a full-feature EEPROM. Because a ROM is smaller in size and more cost-effective than a RAM, it is used whenever frequent rewriting is not required. With this background, different types of nonvolatile memories are explained below:

Mask-programmed ROM: The memory content is fixed by the manufacturer and is not programmable once it is fabricated. Sometimes mask-programmed ROM is simply referred to as ROM.

PROM: Programmable ROM is sometimes called field-programmable ROM or fusible-link ROM. The connectivity of the array is custom programmed using the technique of fusing or antifusing. After programming, the memory works as a ROM.

EPROM: In an electrically programmable ROM, programming is performed by hot-electron injection or tunneling to the floating gate, and it requires biases on both the drain and the control gate. Global erase is by exposure to a UV light or x-ray. Selective erase is not possible.

Flash EEPROM: A flash EEPROM, as opposed to a full-feature EEPROM below, can be erased electrically but only by a large block of cells simultaneously. It loses byte selectivity but maintains a one-transistor cell. It is, thus, a compromise between an EPROM and a full-feature EEPROM.

EEPROM: In an electrically erasable/programmable ROM, not only can it be erased electrically, but also selectively by byte address. To erase selectively, a select transistor is needed for each cell, leading to a two-transistor cell. This makes it less popular than a flash EEPROM.

Nonvolatile RAM: This memory can be viewed as a nonvolatile SRAM, or EEPROM with short programming time as well as high endurance. If technology allows the aforementioned features, this would be the ideal memory.

When the gate electrode of a conventional MOSFET is modified so that semipermanent charge storage inside the gate stack is possible, the new structure becomes a nonvolatile memory device. Since the first nonvolatile memory device proposed by Kahng and Sze in 1967,[77] various device structures have been made, and nonvolatile memory devices have been extensively used in commercial products. The two groups of nonvolatile memory devices are the floating-gate devices and the charge-trapping devices (Fig. 44). In both types of devices, charges are injected from the silicon substrate across the first insulator and stored in the floating gate or at the nitride-oxide

Sandisk Technologies, Inc. - Ex. 1042, Page 000069

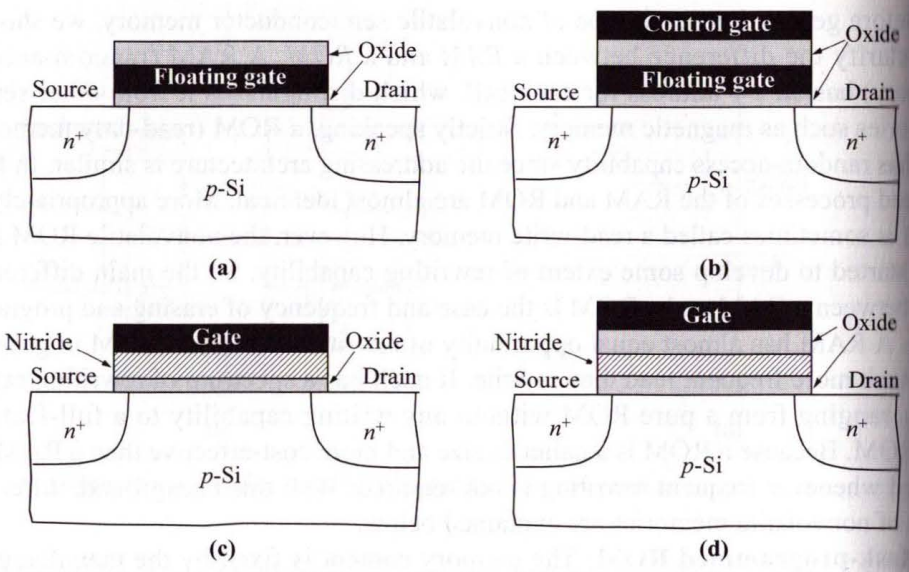


Fig. 44 Variations of nonvolatile memory devices: Floating-gate devices as (a) FAMOS transistor and (b) stacked-gate transistor. Charge-trapping devices as (c) MNOS transistor and (d) SONOS transistor.

interface. The stored charge gives rise to a threshold-voltage shift, and the device is at a *high-threshold state* (programmed). For a well-designed memory device, the charge retention time can be over 100 years. To return to the *low-threshold state* (erased), a gate voltage or other means (such as ultraviolet light) can be applied to erase the stored charge.

6.7.1 Floating-Gate Devices

In a floating-gate memory device, charge is injected to the floating gate to change the threshold voltage. The two modes of programming are hot-carrier injection and Fowler-Nordheim tunneling. Figure 45a shows the mechanisms of hot-carrier injection. Near the drain, the lateral field is at its highest level. The channel carriers (electrons) acquire energy from the field and become hot carriers. When their energy is higher than the barrier of the Si/SiO_2 interface, they can be injected to the floating gate. At the same time, the high field also induces impact ionization. These generated secondary hot electrons can also be injected to the floating gate. The hot-carrier injection currents give rise to the equivalence of gate current in a regular MOSFET, and as shown in Fig. 31. This gate current peaks at $V_{FG} \approx V_D$ where V_{FG} is the potential of the floating gate.

Figure 45b shows the original method of hot-carrier injection using drain-substrate avalanche. In this scheme, the floating-gate potential is more negative such that hot holes are injected instead.* This injection scheme is found to be less efficient and is no longer used in practice.

Sandisk Technologies, Inc. - Ex. 1042, Page 000070

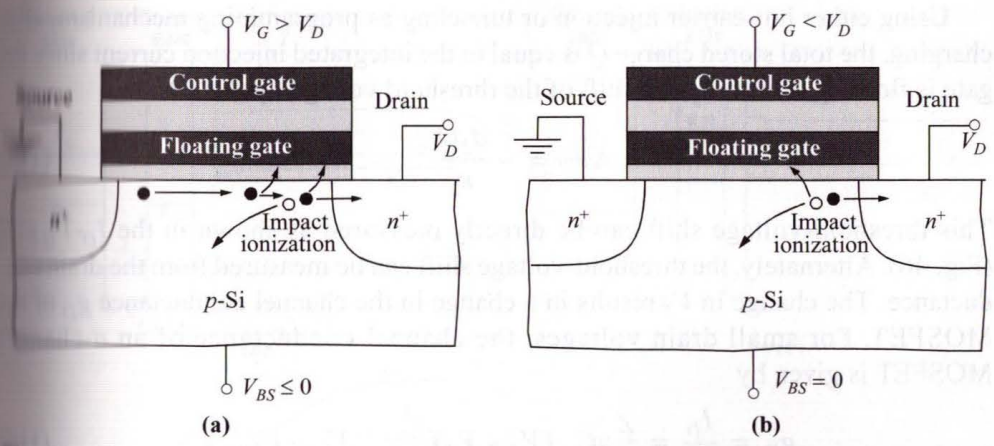


Fig. 45 Charging of the floating gate by hot carriers. (a) Hot electrons from channel and impact ionization. (b) Hot holes from drain avalanche. Note difference in gate bias between the two figures.

Besides hot-carrier injection, electrons can be injected by tunneling. In this programming mode, the electric field across the bottom oxide layer is most critical. On application of a positive voltage V_G to the control gate, an electric field is established in each of the two insulators (Fig. 44b). We have, from Gauss' law, that

$$\varepsilon_1 \mathscr{E}_1 = \varepsilon_2 \mathscr{E}_2 + Q \tag{114}$$

and

$$V_G = V_1 + V_2 = d_1 \mathscr{E}_1 + d_2 \mathscr{E}_2, \tag{115}$$

where the subscripts 1 and 2 correspond to the bottom and top oxide layer respectively, and Q (negative) is the stored charge on the floating gate. From Eqs. 114–115 we obtain

$$\mathscr{E}_1 = \frac{V_G}{d_1 + d_2(\varepsilon_1/\varepsilon_2)} + \frac{Q}{\varepsilon_1 + \varepsilon_2(d_1/d_2)}. \tag{116}$$

The current transport in insulators is generally a strong function of the electric field. When the transport is Fowler-Nordheim tunneling, the current density has the form

$$J = C_4 \mathscr{E}_1^2 \exp\left(\frac{-\mathscr{E}_0}{\mathscr{E}_1}\right) \tag{117}$$

where C_4 and $\mathscr{E}_0$ are constants in terms of effective mass and barrier height. This type of current transport occurs in SiO_2 and Al_2O_3 as discussed in Chapters 4 and 8.

---

1 The original devices were p-channel such that hot electrons are injected under this scheme which are relatively more efficient than hot holes. We use the same n-channel for better comparison.

Sandisk Technologies, Inc. - Ex. 1042, Page 000071

Using either hot-carrier injection or tunneling as programming mechanism, after charging, the total stored charge Q is equal to the integrated injection current since the gate is floating. This causes a shift of the threshold voltage by the amount

$$\Delta V_T = -\frac{d_2 Q}{\varepsilon_2}. \tag{119}$$

This threshold-voltage shift can be directly measured as shown in the I_D-V_G plot (Fig. 46). Alternately, the threshold-voltage shift can be measured from the drain conductance. The change in V_T results in a change in the channel conductance g_D of the MOSFET. For small drain voltages, the channel conductance of an n-channel MOSFET is given by

$$g_D = \frac{I_D}{V_D} = \frac{Z}{L}\mu C_{ox}(V_G - V_T), \qquad V_G > V_T. \tag{119}$$

After altering the charge on the floating gate by Q (negative charge), the g_D-V_G plot shifts to the right by ΔV_T.

To erase the stored charge, a negative bias is put on the control gate or a positive bias on the source/drain. The process is the reverse of the tunneling process described above, and the stored electrons tunnel out of the floating gate to the substrate.

The programming and erasing sequence of a floating-gate memory can be understood with the energy-band diagrams in Fig. 47. In Fig. 47b, electron injection can be due to hot carriers over the barrier or tunneling through the barrier. Figure 47c shows that the accumulated negative charge at the floating gate raises the threshold voltage compared to its initial condition in Fig. 47a. The erase is carried out by electron tunneling from the floating gate back to the substrate (Fig. 47d).

In both programming and erasing operations, it is important to modulate the floating-gate potential efficiently by the control-gate applied voltage. An important parameter in the floating-gate memory is the coupling ratio which determines the portion of the control-gate voltage that gets coupled to the floating gate capacitively. This coupling ratio is determined by the capacitance ratio

$$R_{CG} = \frac{C_2'}{C_1' + C_2'} \tag{120}$$

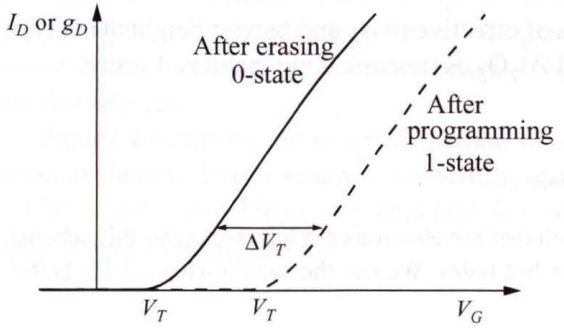


Fig. 46 Drain-current characteristics of a stacked-gate n-channel memory transistor, showing the change of threshold voltage after erasing and programming.

Sandisk Technologies, Inc. - Ex. 1042, Page 000072

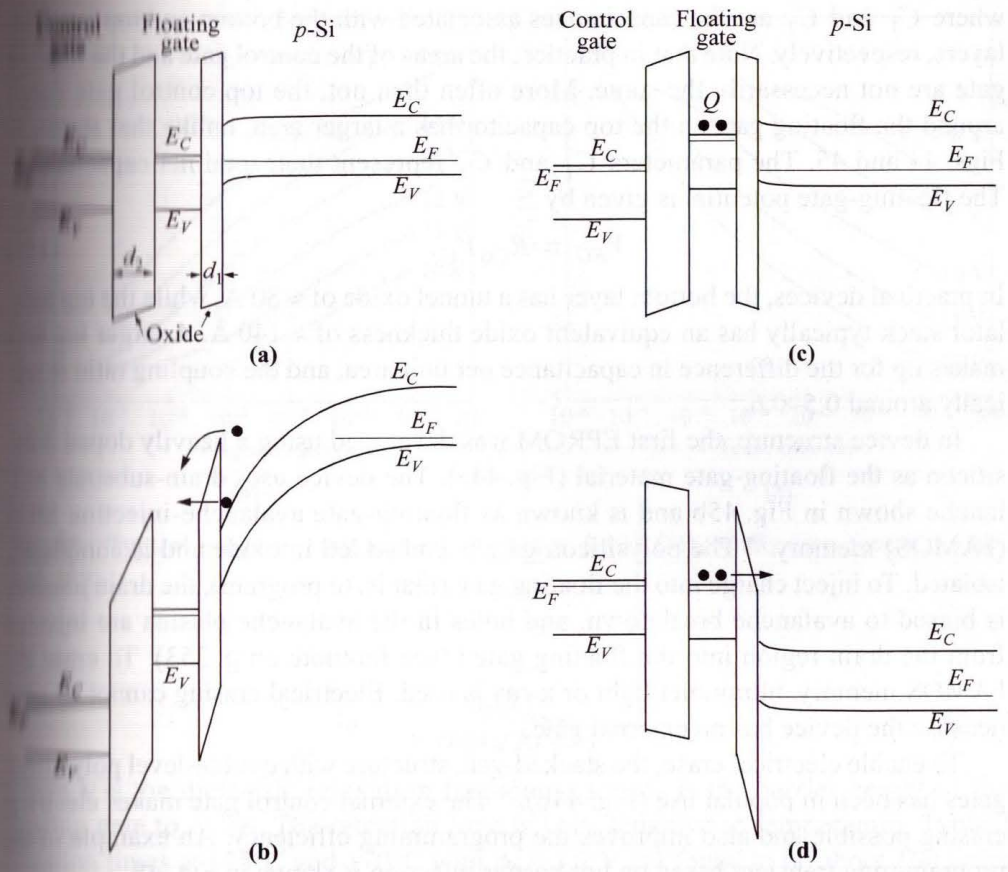


Fig. 47 Energy-band diagrams for a stacked-gate memory transistor at different stages of operation. (a) Initial stage. (b) Charging by hot electrons or electron tunneling. (c) After charging, the floating-gate having charge Q (negative) is at higher potential and V_T is increased. (d) Erasing by electron tunneling.

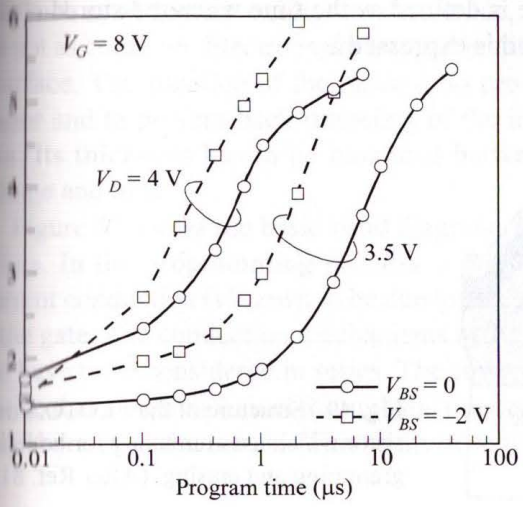


Fig. 48 Programming of floating-gate memory using hot-electron injection. (After Ref. 80.)

Sandisk Technologies, Inc. - Ex. 1042, Page 000073

where C_1' and C_2' are the capacitances associated with the bottom and top insulator layers, respectively. Note that in practice, the areas of the control gate and the floating gate are not necessarily the same. More often than not, the top control gate wraps around the floating gate so the top capacitor has a larger area, unlike that shown in Figs. 44 and 45. The parameters C_1' and C_2' represent their total net capacitances. The floating-gate potential is given by

$$V_{FG} = R_{CG} V_G. \tag{12}$$

In practical devices, the bottom layer has a tunnel oxide of ≈ 80 Å, while the top insulator stack typically has an equivalent oxide thickness of ≈ 140 Å. A larger top area makes up for the difference in capacitance per unit area, and the coupling ratio is typically around 0.5–0.6.

In device structure, the first EPROM was developed using a heavily doped polysilicon as the floating-gate material (Fig. 44a). The device uses drain-substrate avalanche shown in Fig. 45b and is known as floating-gate avalanche-injection MOS (FAMOS) memory.[78] The polysilicon gate is embedded in oxide and is completely isolated. To inject charge into the floating gate (that is, to program), the drain junction is biased to avalanche breakdown, and holes in the avalanche plasma are injected from the drain region into the floating gate (*see footnote on p. 353). To erase the FAMOS memory, ultraviolet light or x-ray is used. Electrical erasing cannot be used because the device has no external gate.

To enable electrical erase, the stacked-gate structure with double-level polysilicon gates has been in popular use (Fig. 44b).[79] The external control gate makes electrical erasing possible and also improves the programming efficiency. An example of the programming transient based on hot-carrier injection is shown in Fig. 48.

In EEPROM circuits, it is more common to use tunneling as an injection mechanism for programming. A successful commercial device, called FLOTOX (floating gate tunnel oxide) transistor, confines the tunneling process to a small area over the drain, as shown in Fig. 49. Typical programming and erasing transients for the FLOTOX transistors are shown n Fig. 50.

After programming, by definition, a long retention time is required for nonvolatile memory operation. The retention time is defined as the time when the stored charge decreases to 50% of its initial value and is expressed by

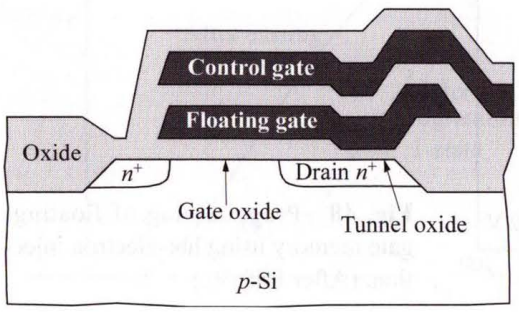


Fig. 49 Structure of the FLOTOX transistor which uses tunneling for both programming and erasing. (After Ref. 81.)

Sandisk Technologies, Inc. - Ex. 1042, Page 000074

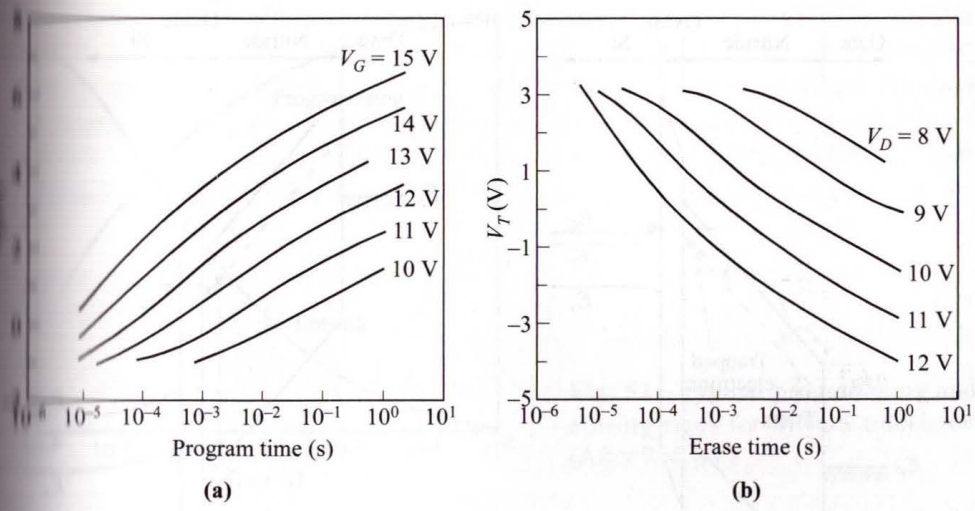


Fig. 50 Typical programming and erasing times for FLOTOX memory device. (After Ref. 76.)

$$t_R = \frac{\ln(2)}{\nu \exp(q\phi_B/kT)} \tag{122}$$

where ν is the dielectric relaxation frequency, and ϕ_B is the barrier height of the floating gate to oxide. The retention time is very sensitive to temperature. Typical retention times at 125°C and 170°C with $\phi_B = 1.7$ V are found to be about 100 years and 1 year respectively.[82]

6.7.2 Charge-Trapping Devices

MNOS Transistor. As a memory device, in the MNOS (metal-nitride-oxide-silicon) transistor, the silicon-nitride layer is used as an efficient material to trap electrons as current passes through the dielectric.[83] Other insulators in place of the silicon-nitride film such as aluminum oxide, tantalum oxide, and titanium oxide have been used but are not as common. Electrons are trapped in the nitride layer close to the oxide-nitride interface. The function of the oxide is to provide a good interface to the semiconductor and to prevent back-tunneling of the injected charge for better charge retention. Its thickness has to be balanced between retention time and programming voltage and time.

Figure 51 shows the basic band diagrams for the programming and erasing operations. In the programming process, a large positive bias is applied to the gate. Current conduction is known to be due to electrons that are emitted from the substrate to the gate. The conduction mechanisms in the two dielectric layers are very different and have to be considered in series. The current through the oxide J_{ox} is by tunneling. Notice that electrons tunnel through the trapezoidal oxide barrier, followed by a triangular barrier in the nitride. This form of tunneling has been identified as modified

Sandisk Technologies, Inc. - Ex. 1042, Page 000075

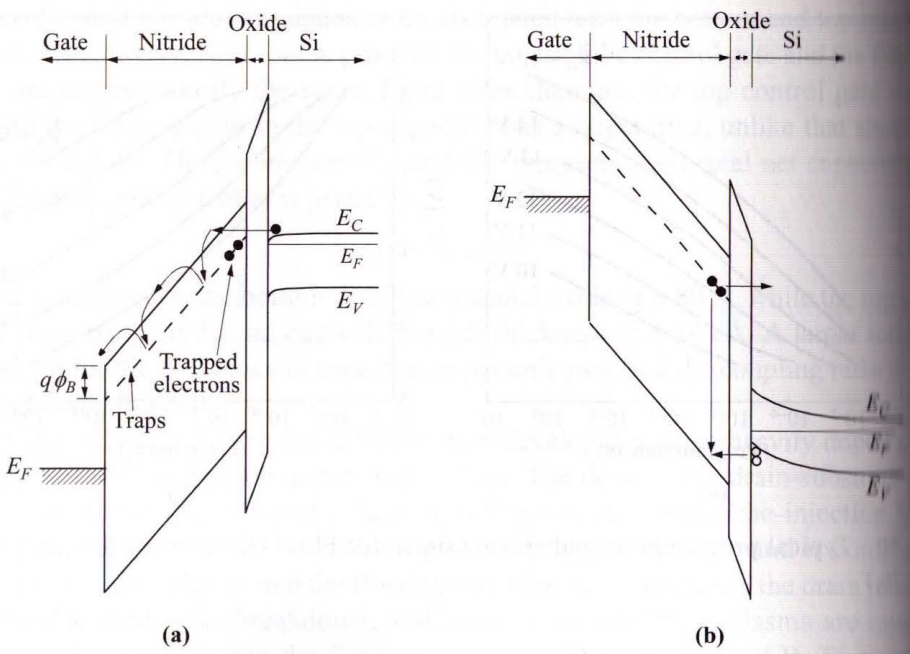


Fig. 51 Rewriting of MNOS memory. (a) Programming: electrons tunnel through oxide and are trapped in the nitride. (b) Erasing: holes tunnel through oxide to neutralize the trapped electrons, and tunneling of trapped electrons.

Fowler-Nordheim tunneling, as opposed to Fowler-Nordheim tunneling through a single triangular barrier. It has the following form of

$$J_{ox} = C_5 \mathscr{E}_{ox}^2 \exp\left(-\frac{C_6}{\mathscr{E}_{ox}}\right), \tag{123}$$

where $\mathscr{E}_{ox}$ is the field in the oxide layer, and C_5 and C_6 are constants. The current through the nitride layer J_n is controlled by Frenkel-Poole transport, which has the form

$$J_n = C_7 \mathscr{E}_n \exp\left[\frac{-q(\phi_B - \sqrt{q\mathscr{E}_n/\pi\varepsilon_n})}{kT}\right] \tag{124}$$

where $\mathscr{E}_n$ and ε_n are the electric field and permittivity in nitride, ϕ_B is the trap level below the conduction band (≈ 1.3 V), and C_7 is a constant [$= 3\times10^{-9}$ $(\Omega\text{-cm})^{-1}$].

It is known that at the beginning of the programming process, modified Fowler-Nordheim tunneling is capable of a higher current, and current conduction is limited by Frenkel-Poole transport through the nitride layer. When the negative charge starts to build up, the oxide field decreases and the modified Fowler-Nordheim tunneling starts to limit the current. The threshold voltage as a function of programming pulse width is shown in Fig. 52. Initially, the threshold voltage changes linearly with time, followed by a logarithmic dependence, and finally it tends to saturate. This programming speed is largely affected by the choice of oxide thickness; a thinner oxide allows

Sandisk Technologies, Inc. - Ex. 1042, Page 000076

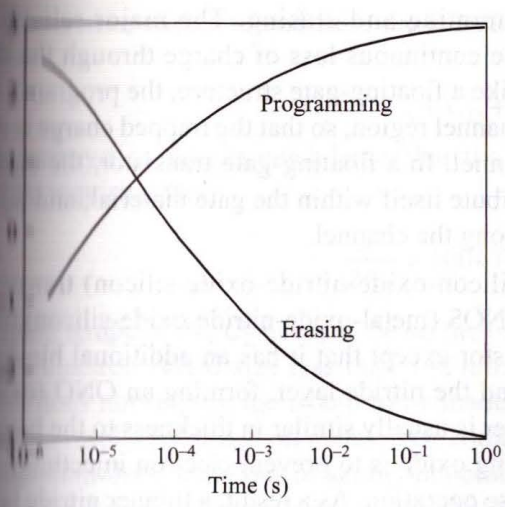


Fig. 52 Typical programming and erasing rates for MNOS transistor. (After Ref. 83.)

shorter programming time. Programming speed has to be balanced with charge retention since too thin an oxide will allow the trapped charge to tunnel back to the silicon substrate.

The total gate capacitance C_G of the dual dielectrics is equal to the serial combination of their capacitances

$$C_G = \frac{1}{(1/C_n) + (1/C_{ox})} = \frac{C_{ox}C_n}{C_{ox} + C_n}, \tag{125}$$

where the capacitances $C_{ox} = \varepsilon_{ox}/d_{ox}$ and $C_n = \varepsilon_n/d_n$ correspond to the oxide and nitride layers respectively. The amount of trapped charge density Q near the nitride-oxide interface depends on the trapping efficiency of the nitride and is proportional to the integrated Frenkel-Poole current having passed through it. The final threshold-voltage shift is given by

$$\Delta V_T = -\frac{Q}{C_n}. \tag{126}$$

In the erasing process, a large negative bias is applied to the gate (Fig. 51b). Traditionally, the discharge process is believed due to the tunneling of trapped electrons back to the silicon substrate. New evidence shows that the major process is due to tunneling of holes from the substrate to neutralize the trapped electrons. The discharge process as a function of pulse width is also shown in Fig. 52.

The advantages of the MNOS transistor include reasonable speed for programming and erasing, so it is a candidate as a nonvolatile RAM device. It also has superior radiation hardness, due to minimal oxide thickness and the absence of a floating gate. The drawbacks of the MNOS transistor are large programming and erasing voltages and nonuniform threshold voltage from device to device. The passage of tunneling current gradually increases the interface-trap density at the semiconductor surface and also causes a loss of trapping efficiency due to leakage or tunneling of trapped electrons back to the substrate. These result in a narrowing threshold voltage

Sandisk Technologies, Inc. - Ex. 1042, Page 000077

window after many cycles of programming and erasing. The major reliability problem of the MNOS transistor is the continuous loss of charge through the thin oxide. It should be pointed out that unlike a floating-gate structure, the programming current has to pass through the entire channel region, so that the trapped charge is distributed uniformly throughout the channel. In a floating-gate transistor, the charge injected to the floating gate can redistribute itself within the gate material, and injection can take place locally anywhere along the channel.

SONOS Transistor. The SONOS (silicon-oxide-nitride-oxide-silicon) transistor (Fig. 44d) is sometimes called the MONOS (metal-oxide-nitride-oxide-silicon) transistor. It is similar to an MNOS transistor except that it has an additional blocking oxide layer placed between the gate and the nitride layer, forming an ONO (oxide-nitride-oxide) stack. This top oxide layer is usually similar in thickness to the bottom oxide layer. The function of the blocking oxide is to prevent electron injection from the metal to the nitride layer during erase operation. As a result, a thinner nitride layer can be used, leading to lower programming voltage as well as better charge retention. The SONOS transistor now replaces the older MNOS configuration, but the operation principle remains the same.

6.8 SINGLE-ELECTRON TRANSISTOR

With the continuing advancement of technology to nano-scale device geometry, there are new experimental observations that have not been possible before. One of them is the charge-quantization effect in a single-electron transistor (SET),[84] first observed in 1987.[85] The structure of an SET is represented by the schematic circuit diagram in Fig. 53a. It has a central single-electron island that has to be extremely small. The island is connected between the source and drain via capacitors through which tunneling occurs to conduct current. The third terminal is the insulated gate and its purpose is to control the current between the source and drain, similar to the case of an FET.

The opportunity to observe quantization of charge comes directly from the small dimension of the single-electron island. The minimum energy needed to transport a

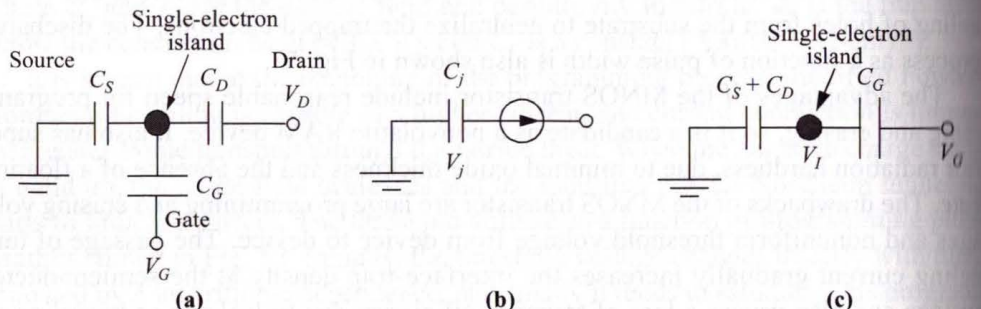


Fig. 53 Circuit representations of (a) single-electron transistor, (b) charging a tunneling capacitor, and (c) single-electron box.

Sandisk Technologies, Inc. - Ex. 1042, Page 000078

single electron charge to and from the island is $q^2/2C_\Sigma$, where C_Σ is its total capacitance,

$$C_\Sigma = C_S + C_D + C_G. \tag{127}$$

This energy also must be much larger than the thermal energy for experimental observation, requiring that

$$\frac{q^2}{2C_\Sigma} > 100kT. \tag{128}$$

At room temperature, C_Σ needs to be on the order of aF (10^{-18} F). This necessitates a single-electron island size less than 1–2 nm. It is also interesting to note that this effect does not require the island to be made of semiconductor materials and most reported results are based on metal dots. Although for the limit of small islands such as quantum dots made out of semiconductors, since the total number of electrons within the dots (< 100) is much less than that in metal ($\approx 10^7$), a second effect of quantization of energy levels can also be observed. This causes addition structures in the I_D-V_D characteristics,[86] but does not contribute to the most-important features of an SET and will not be discussed here. As a matter of fact, the SET does not require any semiconductor material, but only needs metal and insulator.

The capacitors between the island and the source or drain are characterized by the tunneling resistances R_{TS} and R_{TD}. These resistances need to be small (thin insulator layers) to conduct a reasonable amount of current. But they are bound in the lower limit by the Uncertainty principle that electrons have to be treated as particles being clearly on either side of the junction. This requires that

$$R_{TS} \approx R_{TD} > \frac{h}{q^2} \tag{129}$$

($h/q^2 = 25.8$ kΩ) and they should be above ≈ 1 MΩ.

The basic *I-V* characteristics of an SET are shown in Fig. 54. First, Fig. 54a shows that at most values of V_G, there is a knee V_D below which current is very much suppressed. This threshold drain voltage, caused by Coulomb blockade, is explained later. Another important feature is that this Coulomb blockade can be varied by the gate voltage. At some values of V_G, the Coulomb blockade totally vanishes. Shown in Fig. 54b, the cycle can be repeated many times and is called a Coulomb-blockade oscillation. This is very different from the gate control of a regular transistor, where the current can be turned on or off monotonically.

To explain these characteristics, it is best to go back to the simplest structure: a tunneling capacitor as shown in Fig. 53b. Here the capacitor is charged by a small current source, so the junction voltage V_j will increase until an electron can tunnel. The basis for the Coulomb blockade is that it requires a certain minimum V_j before there is enough energy for an electron to tunnel. The minimum energy needed is $q^2/2C_j$, which will be the change of energy of the capacitor when an electron tunnels. This is also the same as the energy gained by the electron when tunneling across the capacitor of voltage V_j, giving

Sandisk Technologies, Inc. - Ex. 1042, Page 000079

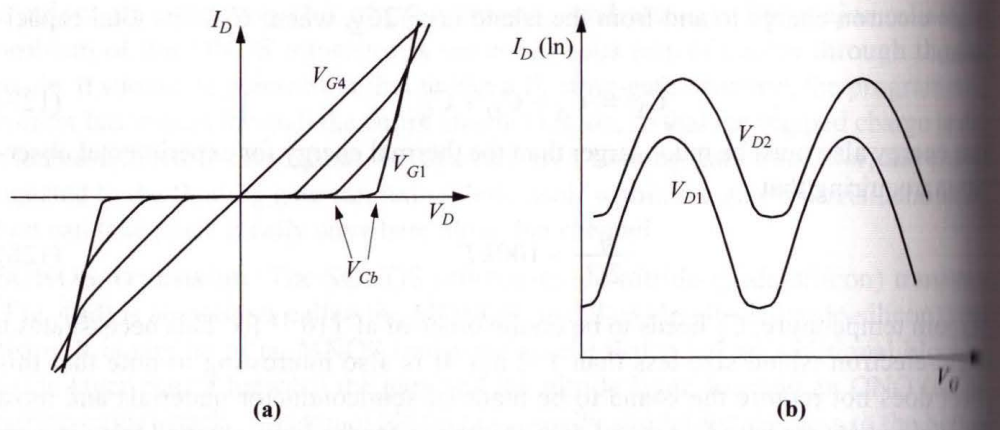


Fig. 54 (a) *I-V* characteristics of SET for various V_G. The Coulomb-blockade voltage can be varied by V_G. (b) Drain current (logarithmic scale) as a function of V_G for various V_D. Note that V_G is shifted by V_D.

$$\frac{q^2}{2C_j} = qV_j. \tag{130}$$

So V_j has to reach $q/2C_j$ before an electron can tunnel. This threshold voltage is the basis for the Coulomb blockade. Alternatively, one can get the same answer by considering the charging energy (E_{ch}) for transferring N_i number of electrons,

$$\begin{aligned} E_{ch} &= \frac{(Q_o - N_i q)^2}{2C_j} - \frac{Q_o^2}{2C_j} \\ &= \frac{N_i^2 q^2}{2C_j} - N_i q V_j \end{aligned} \tag{131}$$

where Q_o is the original charge before tunneling and is equal to V_j/C_j. The criterion for switching to a different state is that E_{ch} has to be negative and at minimum.

Next, we consider a single-electron box where an island is placed between two capacitors, the same as the situation when the source and drain of an SET is tied together (Fig. 53c). As the gate voltage is increased, the island voltage (V_I) is also increased accordingly, although scaled down by a factor of C_G/C_Σ. Similar to the case above, as soon as the tunneling junction gets above a voltage of $q/2C_\Sigma$, one electron starts to tunnel across it. Once an electron has tunneled to the center island, its potential drops by q/C_Σ. Figure 55 shows the charging of the center island and its potential as a function of the gate voltage. It can be seen that the gate voltage at which multiple values of N_i can coexist is at

$$V_G = \frac{q}{C_G}\left(N_i + \frac{1}{2}\right). \tag{132}$$

This condition implies degeneracy: multiple N_i can exist without a change of energy, and one electron can tunnel freely to and from the island. One can imagine that for an

Sandisk Technologies, Inc. - Ex. 1042, Page 000080

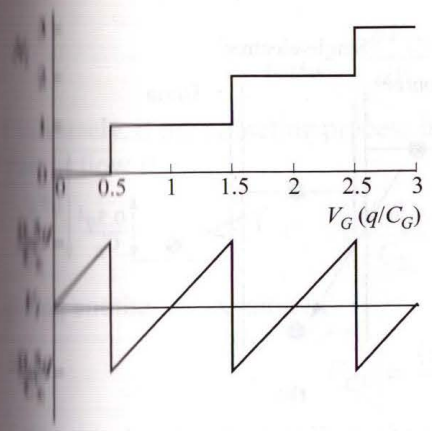


Fig. 55 Charging of single-electron box as a function of V_G (in q/C_G), and the corresponding island voltage V_I.

SET, if a small bias is applied to the drain, electrons can tunnel from the source to the island freely and subsequently from the island to the drain. This corresponds to the condition of V_G where the Coulomb blockade disappears in an SET.

Alternatively, one can derive Eq. 132 by considering the charging energy of a single-electron box,

$$E_{ch} = \frac{N_i^2 q^2}{2C_\Sigma} - \frac{N_i q V_G C_G}{C_\Sigma}. \tag{133}$$

By equating $E_{ch}(N_i + 1) = E_{ch}(N_i)$, the condition of Eq. 132 can be reached. Another approach to understand this is to plot E_{ch} vs. charge ($N_i q$) for different V_G, as shown in Fig. 56. Remembering that N_i takes on only integer values, there are only certain values of V_G where the E_{ch} minimum takes on two values of N_i, a condition of degeneracy. This means that the system can switch between these two states easily without any energy barrier.

We can now return to the SET and explain the two most-important phenomena: the Coulomb blockade and its voltage, and the Coulomb-blockade oscillations. For current to conduct from the source to drain, there are two junctions that electrons have

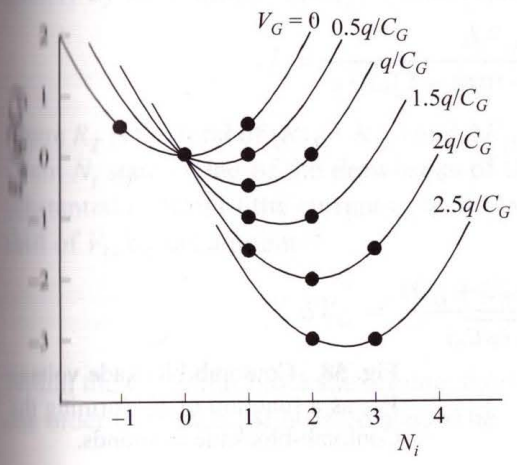


Fig. 56 E_{ch} vs. N_i for different V_G to determine N_i in single-electron box. It can be seen that depending on V_G, the E_{ch} minimum falls on either single or double values of N_i.

Sandisk Technologies, Inc. - Ex. 1042, Page 000081

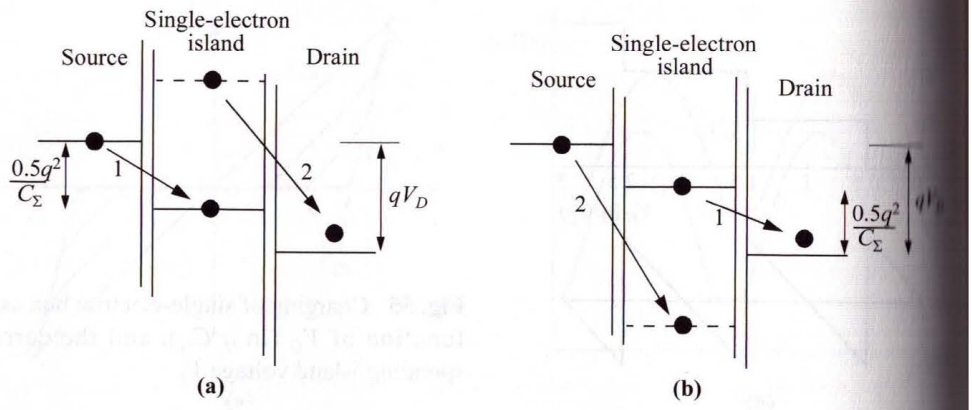


Fig. 57 Energy-band diagrams showing the sequence of tunneling events when the triggering process occurs at the junction (a) between the single-electron island and the source, and (b) between the island and the drain. Event-1 occurs before event-2. Note that the island potential changes by q/C_Σ after each tunneling event.

to tunnel through, but there is only one junction that controls the current flow. Using the energy-band diagrams of Fig. 57, if the bottleneck is at the junction between the source and the single-electron island, an electron will start to tunnel if that junction voltage exceeds $q/2C_\Sigma$, corresponding to the criterion of

$$\frac{V_D C_D}{C_\Sigma} + \frac{V_G C_G}{C_\Sigma} \geq \frac{q}{2C_\Sigma}, \tag{134}$$

giving a minimum value of V_D, or

$$V_{Cb} = \frac{q}{2C_D} - \frac{V_G C_G}{C_D}. \tag{135}$$

This blockade voltage as a function of V_G is shown in Fig. 58 as the line with a negative slope of

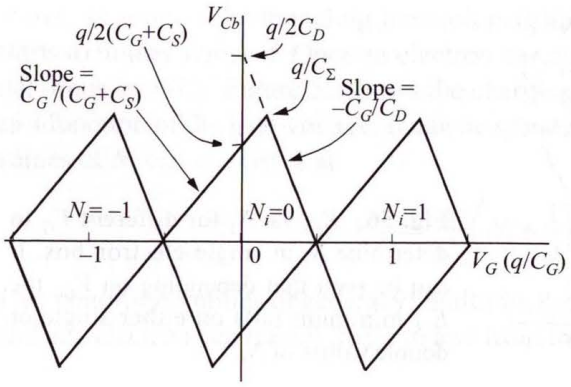


Fig. 58 Coulomb-blockade voltage V_{Cb} as a function of V_G forming the Coulomb-blockade diamonds.

Sandisk Technologies, Inc. - Ex. 1042, Page 000082

$$\frac{dV_{Cb}}{dV_G} = -\frac{C_G}{C_D}. \tag{136}$$

Conversely, if the tunneling process is initiated by the island-drain junction, electrons start to flow if

$$V_D - \left(\frac{V_D C_D}{C_\Sigma} + \frac{V_G C_G}{C_\Sigma}\right) \geq \frac{q}{2C_\Sigma}, \tag{137}$$

giving another expression of

$$V_{Cb} = \frac{(q/2) + V_G C_G}{C_G + C_S}, \tag{138}$$

with a positive slope

$$\frac{dV_{Cb}}{dV_G} = \frac{C_G}{C_G + C_S}. \tag{139}$$

Note that since the current contours follow closely the shape of the Coulomb-blockade diamonds,[87] an SET has both positive and negative transconductance, depending on the V_G range. This is a unique feature that is different from a regular transistor. From the two lines of Eqs. 135 and 138, an intercept of q/C_Σ can be obtained which is the maximum V_{Cb}.

Alternatively, the Coulomb-blockade voltage can be obtained by setting the charging energy negative, either for the island-source junction or the island-drain junction as the current-limiting junction:

$$E_{ch}(N_i{=}1) = \frac{q^2}{2C_\Sigma} - q\left(\frac{V_D C_D}{C_\Sigma} + \frac{V_G C_G}{C_\Sigma}\right) \leq 0, \tag{140}$$

$$E_{ch}(N_i{=}1) = \frac{q^2}{2C_\Sigma} - q\left[V_D - \left(\frac{V_D C_D}{C_\Sigma} + \frac{V_G C_G}{C_\Sigma}\right)\right] \leq 0. \tag{141}$$

These equations lead to the same conclusions as Eqs. 135 and 138, respectively.

Above and below V_{Cb}, the current of an SET has been found to be described adequately by the orthodox theory,[88] which states that the tunneling rate is given by

$$T = \frac{\Delta E_{ch}}{q^2 R_T [1 - \exp(-\Delta E_{ch}/kT)]}, \tag{142}$$

where R_T is the total of $(R_{TS} + R_{TD})$ and ΔE_{ch} is the change of charging energy for different N_i states. One of the drawbacks of the SET is that besides V_G, the drain has substantial control of the current as well. As shown in Fig. 54b, a drain bias causes a shift of V_G by the amount [88]

$$\Delta V_G = \frac{(C_G + C_S - C_D)V_D}{2C_G}. \tag{143}$$

Within the Coulomb-blockade regime, the V_G swing needed to change the current by one order of magnitude is calculated to be

Sandisk Technologies, Inc. - Ex. 1042, Page 000083

$$\Delta V_G \approx (\ln 10)\left(\frac{C_\Sigma}{C_G}\frac{kT}{q}\right). \tag{144}$$

Similarly, the V_D swing needed for the same change is given by

$$\Delta V_D \approx (\ln 10)\left(\frac{2kT}{q}\right). \tag{145}$$

So to have a transistor that has more gate control than drain control requires that the ratio C_G/C_Σ be larger than 0.5.

For application, the SET can perform logic. Since it can possess both positive and negative transconductance, a complementary type of logic can be employed using only one type of device, operated at different regimes. However, the small current and transconductance due to the tunneling nature limits its practicality in real circuits with parasitics. Another problem associated with SETs is the extreme sensitivity to parasitic charge surrounding the single-electron island, which is difficult to control. One potential application of the SET is nonvolatile memory, whose structure is shown in Fig. 44b except that the floating gate is miniaturized to a single-electron island. (More accurately, this memory cell employs a single-electron box or single-electron charging, but does not contain an SET.) A small number of electrons is stored in or emptied in the single-electron island to control the threshold voltage of the MOSFET. The advantage is that since the charge in the floating-gate island is small and discrete, the signal of threshold voltage is quantized and the memory has multiple values.

REFERENCES

1. J. E. Lilienfeld, "Method and Apparatus for Controlling Electric Currents," U.S. Patent 1,745,175. Filed 1926. Granted 1930.
2. J. E. Lilienfeld, "Amplifier for Electric Currents," U.S. Patent 1,877,140. Filed 1928. Granted 1932.
3. J. E. Lilienfeld, "Device for Controlling Electric Current," U.S. Patent 1,900,018. Filed 1928. Granted 1933.
4. 0. Heil, "Improvements in or Relating to Electrical Amplifiers and other Control Arrangements and Devices," British Patent 439,457. Filed and granted 1935.
5. W. Shockley and G. L. Pearson, "Modulation of Conductance of Thin Films of Semiconductors by Surface Charges," *Phys. Rev.*, **74**, 232 (1948).
6. J. R. Ligenza and W. G. Spitzer, "The Mechanisms for Silicon Oxidation in Steam and Oxygen," *J. Phys. Chem. Solids*, **14**, 131 (1960).
7. M. M. Atalla. "Semiconductor Devices Having Dielectric Coatings," U.S. Patent 3,206,670. Filed 1960. Granted 1965.
8. D. Kahng and M. M. Atalla, "Silicon-Silicon Dioxide Field Induced Surface Devices," *IRE-AIEE Solid-State Device Res. Conf.*, (Carnegie Inst. of Tech., Pittsburgh, PA), 1960.
9. D. Kahng, "A Historical Perspective on the Development of MOS Transistors and Related Devices," *IEEE Trans. Electron Dev.*, **ED-23**, 655 (1976).

Sandisk Technologies, Inc. - Ex. 1042, Page 000084

10. C. T. Sah, "Evolution of the MOS Transistor—From Conception to VLSI," *Proc. IEEE*, **76**, 1280 (1988).

11. H. K. J. Ihantola and J. L. Moll, "Design Theory of a Surface Field-Effect Transistor," *Solid-State Electron.*, **7**, 423 (1964).

12. C. T. Sah, "Characteristics of the Metal-Oxide-Semiconductor Transistors," *IEEE Trans. Electron Dev.*, **ED-11**, 324 (1964).

13. S. R. Hofstein and F. P. Heiman, "The Silicon Insulated-Gate Field-Effect Transistor," *Proc. IEEE*, **51**, 1190 (1963).

14. J. R. Brews, "Physics of the MOS Transistor," in D. Kahng, Ed., *Applied Solid State Science*, Suppl. 2A, Academic, New York, 1981.

15. Y. Tsividis, *Operation and Modeling of the MOS Transistor*, 2nd Ed., Oxford University Press, Oxford, 1999.

16. Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*, Cambridge University Press, Cambridge, 1998.

17. R. M. Warner, Jr. and B. L. Grung, *MOSFET Theory and Design*, Oxford University Press, Oxford, 1999.

18. L. L. Chang and H. N. Yu, "The Germanium Insulated-Gate Field-Effect Transistor (FET)," *Proc. IEEE*, **53**, 316 (1965).

19. P. D. Ye, G. D. Wilk, J. Kwo, B. Yang, H. J. L. Gossmann, M. Frei, S. N. G. Chu, J. P. Mannaerts, M. Sergent, M. Hong, K. K. Ng, and J. Bude, "GaAs MOSFET with Oxide Gate Dielectric Grown by Atomic Layer Deposition," *IEEE Electron Dev. Lett.*, **EDL-24**, 209, (2003).

20. H. C. Pao and C. T. Sah, "Effects of Diffusion Current on Characteristics of Metal-Oxide (Insulator)-Semiconductor Transistors (MOST)," *IEEE Trans. Electron Dev.*, **ED-12**, 139 (1965).

21. A. S. Grove and D. J. Fitzgerald, "Surface Effects on *p-n* Junctions: Characteristics of Surface Space-Charge Regions under Nonequilibrium Conditions," *Solid-State Electron.*, **9**, 783 (1966).

22. J. R. Brews, "A Charge-Sheet Model of the MOSFET," *Solid-State Electron.*, **21**, 345 (1978).

23. D. M. Caughey and R. E. Thomas, "Carrier Mobilities in Silicon Empirically Related to Doping and Field," *Proc. IEEE*, **55**, 2192 (1967).

24. K. Natori, "Ballistic Metal-Oxide-Semiconductor Field Effect Transistor," *J. Appl. Phys.*, **76**, 4879 (1994).

25. K. Natori, "Scaling Limit of the MOS Transistor—A Ballistic MOSFET," *IEICE Trans. Electron.*, **E84-C**, 1029 (2001).

26. M. Lundstrom, "Elementary Scattering Theory of the Si MOSFET," *IEEE Electron Dev. Lett.*, **EDL-18**, 361 (1997).

27. F. Assad, Z. Ren, D. Vasileska, S. Datta, and M. Lundstrom, "On the Performance Limits for Si MOSFET's: A Theoretical Study," *IEEE Trans. Electron Dev.*, **ED-47**, 232 (2000).

28. M. Lundstrom, "Essential Physics of Carrier Transport in Nanoscale MOSFETs," *IEEE Trans. Electron Dev.*, **ED-49**, 133 (2002).

29. M. B. Barron, "Low Level Currents in Insulated Gate Field Effect Transistors," *Solid-State Electron.*, **15**, 293 (1972).

Sandisk Technologies, Inc. - Ex. 1042, Page 000085

30. W. M. Gosney, "Subthreshold Drain Leakage Current in MOS Field-Effect Transistors," *IEEE Trans. Electron Dev.*, **ED-19**, 213 (1972).

31. G. W. Taylor, "Subthreshold Conduction in MOSFET's," *IEEE Trans. Electron Dev.*, **ED-25**, 337 (1978).

32. A. G. Sabnis and J. T. Clemens, "Characterization of the Electron Mobility in the Inverted ⟨100⟩ Si Surface," *Tech. Dig. IEEE IEDM*, p.18, 1979.

33. S. Takagi, A. Toriumi, M. Iwase, and H. Tango, "On the Universality of Inversion Layer Mobility in Si MOSFET's: Part I—Effects of Substrate Impurity Concentration," *IEEE Trans. Electron Dev.*, **ED-41**, 2357 (1994).

34. J. A. Cooper, Jr. and D. F. Nelson, "High-Field Drift Velocity of Electrons at the Si-SiO$_2$ Interface as Determined by a Time-of-Flight Technique," *J. Appl. Phys.*, **54**, 1445 (1983).

35. L. Vadasz and A. S. Grove, "Temperature Dependence of MOS Transistor Characteristics Below Saturation," *IEEE Trans. Electron Dev.*, **ED-13**, 863 (1966).

36. F. Gaensslen, V. L. Rideout, E. J. Walker, and J. J. Walker, "Very Small MOSFET's for Low-Temperature Operation," *IEEE Trans. Electron Dev.*, **ED-24**, 218 (1977).

37. G. Merckel, "Ion Implanted MOS Transistors—Depletion Mode Devices," in F. Van de Wiele, W. L. Engle, and P. G. Jespers, Eds., *Process and Device Modeling for IC Design*, Noordhoff, Leyden, 1977.

38. J. S. T. Huang and G. W. Taylor, "Modeling of an Ion-Implanted Silicon-Gate Depletion Mode IGFET," *IEEE Trans. Electron Dev.*, **ED-22**, 995 (1975).

39. T. E. Hendrikson, "A Simplified Model for Subpinchoff Condition in Depletion Mode IGFET's," *IEEE Trans. Electron Dev.*, **ED-25**, 435 (1978).

40. M. J. van der Tol and S. G. Chamberlain, "Potential and Electron Distribution Model for the Buried-Channel MOSFET," *IEEE Trans. Electron Dev.*, **ED-36**, 670 (1989).

41. R. H. Dennard, F. H. Gaensslen, H. Yu, V. L. Rideout, E. Bassons, and A. R. LeBlanc, "Design of Ion-Implanted MOSFET's with Very Small Physical Dimensions," *IEEE J. Solid State Circuits*, **SC-9**, 256 (1974).

42. P. K. Chatterjee, W. R. Hunter, T. C. Holloway, and Y. T. Lin, "The Impact of Scaling Laws on the Choice of *n*-channel or *p*-channel for MOS VLSI," *IEEE Electron Dev. Lett.*, **EDL-1**, 220 (1980).

43. J. Meindl, "Circuit Scaling Limits for Ultra Large Scale Integration," *Digest Int. Solid State Circuits Conf.*, 36, Feb. 1981.

44. G. Baccarani, M. R. Wordeman, and R. H. Dennard, "Generalized Scaling Theory and its Application to a 1/4 Micrometer MOSFET Design," *IEEE Trans. Electron Dev.*, **ED-31**, 452 (1984).

45. J. R. Brews, W. Fichtner, E. H. Nicollian, and S. M. Sze, "Generalized Guide for MOSFET Miniaturization," *IEEE Electron Dev. Lett.*, **EDL-1**, 2 (1980).

46. K. K. Ng, S. A. Eshraghi, and T. D. Stanik, "An Improved Generalized Guide for MOSFET Scaling," *IEEE Trans. Electron Dev.*, **ED-40**, 1895 (1993).

47. L. D. Yau, "A Simple Theory to Predict the Threshold Voltage of Short-Channel IGFET's," *Solid-State Electron.*, **17**, 1059 (1974).

48. W. Fichtner and H. W. Potzl, "MOS Modeling by Analytical Approximations. I. Subthreshold Current and Threshold Voltage," *Int. J. Electron.*, **46**, 33 (1979).

Sandisk Technologies, Inc. - Ex. 1042, Page 000086

49. Y. Taur, G. J. Hu, R. H. Dennard, L. M. Terman, C. Y. Ting, and K. E. Petrillo, "A Self-Aligned 1 μm Channel CMOS Technology with Retrograde *n*-well and Thin Epitaxy," *IEEE Trans. Electron Dev.*, **ED-32**, 203 (1985).

50. W. Fichtner, "Scaling Calculation for MOSFET's," *IEEE Solid State Circuits and Technology Workshop on Scaling and Microlithography*, New York, Apr. 22, 1980.

51. K. K. Ng and G. W. Taylor, "Effects of Hot-Carrier Trapping in *n*- and *p*-Channel MOSFET's," *IEEE Trans. Electron Dev.*, **ED-30**, 871 (1983).

52. T. H. Ning, C. M. Osburn, and H. N. Yu, "Effect of Electron Trapping on IGFET Characteristics," *J. Electron. Mater.*, **6**, 65 (1977).

53. E. H. Nicollian and C. N. Berglund, "Avalanche Injection of Electrons into Insulating SiO_2 Using MOS Structures," *J. Appl. Phys.*, **41**, 3052 (1970).

54. T. Kamata, K. Tanabashi, and K. Kobayashi, "Substrate Current Due to Impact Ionization in MOSFET," *Jpn. J. Appl. Phys.*, **15**, 1127 (1976).

55. E. Sun, J. Moll, J. Berger, and B. Alders, "Breakdown Mechanism in Short-Channel MOS Transistors," *Tech. Dig. IEEE IEDM*, p. 478, 1978.

56. T. Y. Chan, A. T. Wu, P. K. Ko, and C. Hu, "Effects of the Gate-to-Drain/Source Overlap on MOSFET Characteristics," *IEEE Electron Dev. Lett.*, **EDL-8**, 326 (1987).

57. D. J. Frank, R. H. Dennard, E. Nowak, P. M. Solomon, Y. Taur, and H. P. Wong, "Device Scaling Limits of Si MOSFETs and Their Application Dependencies," *Proc. IEEE*, **89**, 259 (2001).

58. B. Yu, H. Wang, A. Joshi, Q. Xiang, E. Ibok, M. Lin, "15nm Gate Length Planar CMOS Transistor," *Tech. Dig. IEEE IEDM*, p.937, 2001.

59. A. Hokazono, K. Ohuchi, M. Takayanagi, Y. Watanabe, S. Magoshi, Y. Kato, T. Shimizu, S. Mori, H. Oguma, T. Sasaki, et al., "14 nm Gate Length CMOSFETs Utilizing Low Thermal Budget Process with Poly-SiGe and Ni Salicide," *Tech. Dig. IEEE IEDM*, p.639, 2002.

60. K. K. Ng and W. T. Lynch, "Analysis of the Gate-Voltage-Dependent Series Resistance of MOSFETs," *IEEE Trans. Electron Dev.*, **ED-33**, 965 (1986).

61. M. P. Lepselter and S. M. Sze, "SB-IGFET: An Insulated-Gate Field-Effect Transistor Using Schottky Barrier Contacts as Source and Drain," *Proc. IEEE*, **56**, 1088 (1968).

62. R. R. Troutman, *Latchup in CMOS Technology: The Problem and its Cure*, Kluwer, Norwell, Massachusetts, 1986.

63. J. Kedzierski, P. Xuan, E. H. Anderson, J. Bokor, T. J. King, and C. Hu, "Complementary Silicide Source/Drain Thin-Body MOSFETs for the 20nm Gate Length Regime," *Tech. Dig. IEEE IEDM*, p.57, 2000.

64. S. Nishimatsu, Y. Kawamoto, H. Masuda, R. Hori, and O. Minato, "Grooved Gate MOSFET," *Jpn. J. Appl. Phys.*, **16**; Suppl. **16-1**, 179 (1977).

65. G. K. Celler and S. Cristoloveanu, "Frontiers of Silicon-on-Insulator," *J. Appl. Phys.*, **93**, 1 (2003).

66. K. A. Jenkins, J. Y. C. Sun, and J. Gautier, "History Dependence of Output Characteristics of Silicon-on-Insulator (SOI) MOSFET's," *IEEE Electron Dev. Lett.*, **EDL-17**, 7 (1996).

67. C. R. Kagan and P. Andry, Eds., *Thin-Film Transistors*, Marcel Dekker, New York, 2003.

68. D. Hisamoto, T. Kaga, and E. Takeda, "Impact of the Vertical 'DELTA' Structure on Planar Device Technology," *IEEE Trans. Electron Dev.*, **ED-38**, 1399 (1991).

Sandisk Technologies, Inc. - Ex. 1042, Page 000087

69. B. S. Doyle, S. Datta, M. Doczy, S. Hareland, B. Jin, J. Kavalieros, T. Linton, A. Murthy, R. Rios, and R. Chau, "High Performance Fully-Depleted Tri-Gate CMOS Transistors," *IEEE Electron Dev. Lett.*, **EDL-24**, 263 (2003).

70. J. M. Hergenrother, G. D. Wilk, T. Nigam, F. P. Klemens, D. Monroe, P. J. Silverman, T. W. Sorsch, B. Busch, M. L. Green, M. R. Baker, et. al., "50 nm Vertical Replacement-Gate (VRG) nMOSFETs with ALD HfO_2 and Al_2O_3 Gate Dielectrics," *Tech. Dig. IEEE IEDM*, p.51, 2001.

71. T. Masuhara and R. S. Muller, "Analytical Technique for the Design of DMOS Transistors," *Jpn. J. Appl. Phys.*, **16**, 173 (1976).

72. M. D. Pocha, A. G. Gonzalez, and R. W. Dutton, "Threshold Voltage Controllability in Double-Diffused MOS Transistors," *IEEE Trans. Electron Dev.*, **ED-21**, 778 (1974).

73. A. W. Ludikhuize, "A Review of RESURF Technology," *Proc. 12th Int. Symp. Power Semiconductor Devices & ICs*, p.11, 2000.

74. P. Cappelletti, C. Golla, P. Olivo, and E. Zanoni, Eds., *Flash Memories*, Kluwer, Norwell, Massachusetts, 1999.

75. C. Hu, Ed., *Nonvolatile Semiconductor Memories: Technologies, Design, and Applications*, IEEE Press, Piscataway, New Jersey, 1991.

76. W. D. Brown and J. E. Brewer, Eds., *Nonvolatile Semiconductor Memory Technology*, IEEE Press, Piscataway, New Jersey, 1998.

77. D. Kahng and S. M. Sze, "A Floating Gate and Its Application to Memory Devices," *Bell Syst. Tech. J.*, **46**, 1283 (1967).

78. D. Frohman-Bentchkowsky, "FAMOS—A New Semiconductor Charge Storage Device," *Solid-State Electron.*, **17**, 517 (1974).

79. H. Iizuka, F. Masuoka, T. Sato, and M. Ishikawa, "Electrically Alterable Avalanche-Injection-Type MOS Read-Only Memory with Stacked-Gate Structures," *IEEE Trans. Electron Dev.*, **ED-23**, 379 (1976).

80. S. Mahapatra, S. Shukuri, and J. Bude, "CHISEL Flash EEPROM—Part I: Performance and Scaling," *IEEE Trans. Electron Dev.*, **ED-49**, 1296 (2002).

81. S. K. Lai and V. K. Dham, "VLSI Electrically Erasable Programmable Read Only Memory," in N. G. Einspruch, Ed., *VLSI handbook*, Academic Press, Orlando, FL, 1985.

82. Y. Nishi and H. Iizuka, "Nonvolatile Memories," in D. Kahng, Ed., *Applied Solid State Science*, Suppl. 2A, Academic, New York, 1981.

83. Y. Kamigaki and S. Minami, "MNOS Nonvolatile Semiconductor Memory Technology Present and Future," *IEICE Trans. Electron.*, **E84-C**, 713 (2001).

84. D. V. Averin and K. K. Likharev, "Coulomb Blockade of Single-Electron Tunneling, and Coherent Oscillations in Small Tunnel Junctions," *J. Low Temp. Phys.*, **62**, 345 (1986).

85. T. A. Fulton and G. J. Dolan, "Observation of Single-Electron Charging Effects in Small Tunnel Junctions," *Phys. Rev. Lett.*, **59**, 109 (1987).

86. M. A. Kastner, "Artificial Atoms," *Physics Today*, 24 (Jan. 1993).

87. Y. A. Pashkin, Y. Nakamura and J. S. Tsai, "Room-Temperature Al Single-Electron Transistor Made by Electron-Beam Lithography," *Appl. Phys. Lett.*, **76**, 2256 (2000).

88. K. Uchida, K. Matsuzawa, J. Koga, R. Ohba, S. Takagi and A. Toriumi, "Analytical Single-Electron Transistor (SET) Model for Design and Analysis of Realistic SET Circuits," *Jpn. J. Appl. Phys.*, **39**, 2321 (2000).

Sandisk Technologies, Inc. - Ex. 1042, Page 000088

PROBLEMS

1. Derive Eq. 23 from Eqs. 20 and 22 (p.303).

2. For a square MOSFET ($Z/L = 1$), I_D is measured to be 18.7 μA at $V_D = 0.4$ V and $V_G = 3$ V. If we require a current of 1.6 mA at $V_D = 0.4$ V and $V_G = 3$ V, what is the minimum width Z of the device? Assume that the polysilicon gate length is 0.6 μm and the n^+ source and drain each diffuses sideways 0.05 μm under the gate.

3. Consider a submicron MOSFET with $L = 0.25$ μm, $Z = 5$ μm, $N_A = 10^{17}$ cm^{-3}, $\mu_n = 500$ cm^2/V-s, $C_{ox} = 3.45\times10^{-7}$ F/cm^2, and $V_T = 0.5$ V, find the channel conductance for $V_G = 1$ V and $V_D = 0.1$ V.

4. For a MOSFET with a channel length of 10 μm under certain biasing conditions, the channel current I_D is 1 mA and the gate current is 1 μA. We want to reduce the gate current to $10^{-6} I_D$ under the same biasing condition and for the same device parameters except the channel length. Find the channel length.

5. For an MOSFET with sufficient drain voltage to be in saturation (under constant-mobility condition), the current is 50 μA at $V_G = 1$ V, and 200 μA at $V_G = 3$ V. Find the threshold voltage.

6. (a) To avoid hot-electron effect in an n-channel MOSFET, we assume an allowed maximum field in the oxide to be 1.45×10^6 V/cm. Find the corresponding surface potential ψ_s in the silicon for a doping concentration of 10^{18} cm^{-3}.
   (b) For an n^+-polysilicon gate, find the threshold voltage of the above MOSFET with $d =$ 8 nm, assuming $Q_{it} = Q_{ox} = Q_f = Q_m = 0$.

7. An n-channel MOSFET is designed to have a threshold voltage of +0.5 V and a gate oxide thickness of 15 nm. Find the channel doping to give the desired V_T if n^+-polysilicon is used as the gate material, and there are no oxide charge, interface-trapped charge, and mobile ions in the device. V_T - eq.25 V_{FB} - eq.49 ϕ_{ms} - chap4. fig22

8. To isolate devices from interacting with each other, each MOSFET is surrounded by a field oxide. If the "field transistor" associated with the field oxide must have a threshold voltage of ≥ 20V, calculate the minimum field-oxide thickness. $N_A = 10^{17}$ cm^{-3}, $Q_f/q = 10^{11}$ cm^{-3}, and an n^+-polysilicon is used for local interconnect as the gate electrode. d = ?

9. An n-channel n^+-polysilicon-SiO_2-Si MOSFET has $N_A = 10^{17}$ cm^{-3}, $Q_f/q = 2\times10^{10}$ cm^{-2}, and $d = 10$ nm. Boron ions are implanted to increase the threshold voltage to +1 V. Find the implant dose, assume that the implanted ions form a sheet of negative charge at the Si-SiO_2 interface. ($\phi_{ms} = -0.98$ V).

10. For an n-MOSFET with $q\psi_B$ of 0.5 eV, the threshold voltage change ΔV_T is 1 V when a substrate bias V_{BS} of –1 V is applied. What is ΔV_T when V_{BS} is –3 V?

11. A MOSFET ($N_A = 10^{17}$ cm^{-3}, $d = 5$ nm) has a threshold voltage of $V_T = 0.5$ V, a subthreshold swing of 100 mV/decade, and a drain current of 0.1 μA at V_T. If we want to reduce the leakage current at $V_G = 0$ to 10^{-13} A, find the reverse substrate-source bias required to achieve the reduction. eqn 5.1

12. The subthreshold current of an ideal MOSFET is given by

$$I_D = A(\beta\psi_s)^{-1/2}\exp(\beta\psi_s),$$

$$\Delta V_T = \frac{\sqrt{2\varepsilon_s q N_A}}{C_{ox}}\left(\sqrt{2\psi_B + V_{BS}} - \sqrt{2\psi_B}\right)$$

Sandisk Technologies, Inc. - Ex. 1042, Page 000089

$$\beta\psi_s = \beta V_G - \frac{a^2}{2\beta}\left[\sqrt{1+\frac{4}{a^2}(\beta V_G - 1)} - 1\right],$$

where ψ_s is the surface potential, $\beta \equiv q/kT$, $a \equiv \sqrt{2}\,(\varepsilon_s/\varepsilon_{ox})(t_{ox}/L_D)$, L_D is Debye length $= \sqrt{\varepsilon_s/qN_A\beta}$, and A = constant. Show that the subthreshold swing S is given by

$$S \equiv (\ln 10)\frac{dV_G}{d(\ln I_D)} = \frac{kT}{q}(\ln 10)\left(1+\frac{C_D}{C_{ox}}\right)$$

where $C_D \equiv \sqrt{q\varepsilon_s N_A/2\psi_s}$, $C_{ox} \equiv \varepsilon_{ox}/t_{ox}$, and $a \gg C_D/C_{ox}$.

13. For a MOSFET with a gate oxide of 10 nm and a substrate doping of 10^{17} cm^{-3}, find the subthreshold swing.

14. For a Si MOSFET with $N_A = 5\times10^{16}$ cm^{-3}, d =10 nm, and an interface-trap density of 10^{11} cm^{-2}, find the subthreshold swing with a grounded substrate terminal.

15. An idealized implanted step doping profile has $N_S = 10^{16}$ cm^{-3}, $N_B = 10^{15}$ cm^{-3}, and x_j = 0.3 μm. Find (1) the implanted dose D_I, (2) the centroid of the dose, and (3) the threshold voltage shift (d = 100 nm) with respect to a uniformly doped N_B case.

16. Derive Eq. 79.

17. Refer to Fig. 21 (p. 322), assume $N_B = 7.5\times10^{15}$ cm^{-3}, d = 35 nm, a reverse back bias of 1 V, and an implanted dose $D_I = 6\times10^{11}$ cm^{-2}, find the depth of the centroid at which the depletion-layer edge is clamped to the implant (in nm).

18. Find the scaled drain current per unit channel width (I_D/Z) for two n-MOSFETs, one with constant-voltage scaling and one with constant-field scaling. Assume the devices are operated under velocity-saturation condition. The initial device parameters are L = 1 μm, d = 10 nm, V_D = 5 V, I_D/Z = 500 μA/μm. The scaling factor is κ = 5.

19. For the constant-voltage scaling approach of MOSFET with a scaling factor κ=10, find the doping concentration of a scaled device if the original device has a doping of 10^{15} cm^{-3} (in cm^{-3}).

20. When the linear dimensions of a MOSFET are scaled down by a factor of 10 based on the constant-field scaling, (a) find the corresponding factor of the scaled switching energy, and (b) find the scaled power-delay product, assuming the product is 1 J for the original large device.

21. A composite structure of a 20 nm Ta_2O_5 ($\varepsilon_i/\varepsilon_o$ = 25) and a 2-nm SiO_2 is sandwiched between a top and bottom electrodes. Find the equivalent SiO_2 thickness (in nm).

22. A DRAM must operate with a minimum refresh time of 4 ms. The storage capacitor in each cell has a capacitance of 50 fF and is fully charged at 5 V. Estimate the worst-case leakage current that the dynamic capacitor node can tolerate (i.e., the charge in the capacitor has dropped to its 50% level).

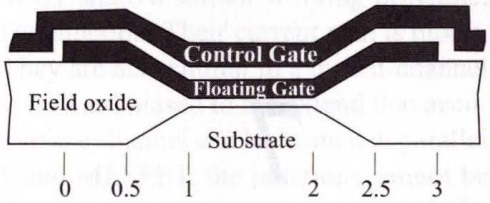
23. For DRAM operation assume that we need a minimum of 10^5 electrons for the MOS storage capacitor. If the capacitor has an area of 0.5 μm×0.5 μm on the wafer surface, an oxide thickness of 5 nm, and is fully charged to 2 V, what is the required minimum depth of a rectangular-trench capacitor?

24. For a floating-gate nonvolatile memory device, the lower insulator has a dielectric constant of 4 and is 10 nm thick. The insulator above the floating gate has a dielectric constant of 10 and is 100 nm thick. If the current density $J_1 = \sigma\mathscr{E}_1$ where $\sigma = 10^{-7}$ S/cm, and the current

Sandisk Technologies, Inc. - Ex. 1042, Page 000090

in the upper insulator is zero, find the threshold voltage shift for a sufficiently long time such that J_1 becomes negligibly small. The applied voltage on the control gate is 10 V.

18. Consider a NVSM cell whose cross section is shown. The channel width is 1 μm. Assume that the bird's beak has the linear wedge shape illustrated. The gate oxide thickness (between substrate and FG) is 35 nm, the interpoly dielectric is an oxide of 50 nm, and the field oxide is 0.6 μm. The physical gate length is 1.2 μm, the metallurgical junction lies 0.15 μm under the gate and the effective channel length is 0.7 μm. The floating gate poly is 0.3 μm thick. Calculate (a) the value of the control gate to floating gate capacitance, (b) the drain to floating gate capacitance, assume half of the channel capacitance to the source and other half to the drain, and (c) if the floating gate to substrate capacitance is 0.14 fF, calculate the control gate to floating gate coupling ration, R_{CG}, and the drain to floating gate coupling ration, R_D.



19. For a silicon nonvolatile memory with a floating gate, the thickness and the dielectric constant for the first insulator (thermally grown SiO_2) are 3 nm and 3.9, and the corresponding values for the second insulator are 30 nm and 30. Estimate the stored charge/cm^2 in the floating gate after a gate voltage of 5.52 V is applied for 1 ms. There is no current conduction through the second insulator, and the current in the first insulator is by Fowler-Nordheim tunneling.

20. A floating-gate nonvolatile semiconductor memory has a total capacitance of 3.71 fF, a control gate to floating gate capacitance of 2.59 fF, a drain to floating gate capacitance of 0.49 fF, a floating gate to substrate capacitance of 0.14 fF. How many electrons are needed to shift the measured threshold by 0.5 V (measured from the control gate)?

21. An EEPROM has $C_{CG} = 2.59$ fF, $C_S = C_D = 0.49$ fF, and $C_B = 0.14$ fF, where they indicate the capacitances between the floating gate and the control gate, source, drain, and substrate, respectively. Assume that when the control gate and the floating gate are shorted together, the device threshold is measured to be 1.5 V. If the control gate is at 12 V, and the drain is at 7 V during programming, to what potential can the floating gate be charged while the programming voltages are present? What threshold would be observed after programming under these biases for a drain bias during reading of 2 V?

Sandisk Technologies, Inc. - Ex. 1042, Page 000091