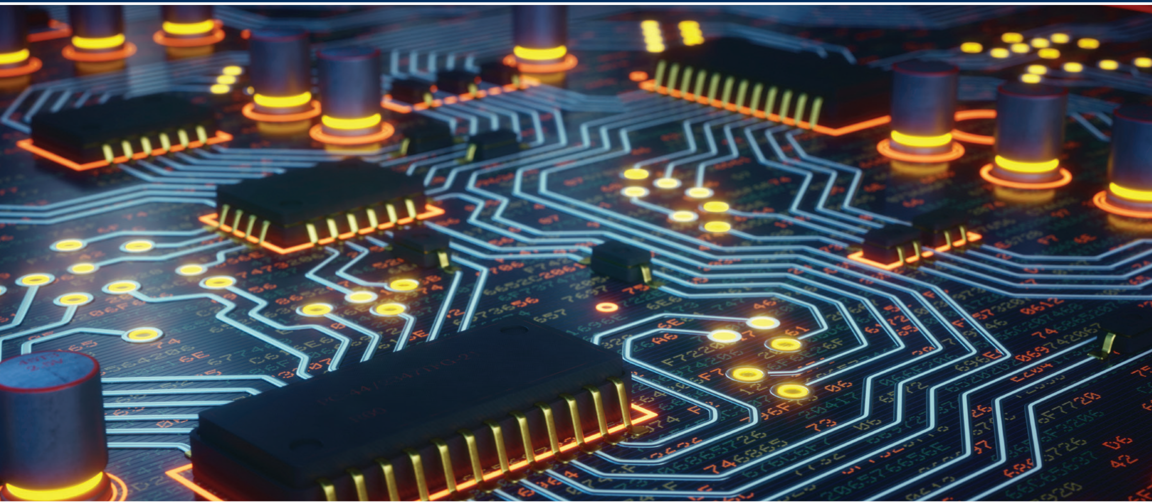
IEEE Press Series on Microelectronic Systems
R. Jacob Baker, Series Editor

Junctionless Field-Effect Transistors

Design, Modeling, and Simulation



Shubham Sahay and Mamidala Jagadesh Kumar

WILEY

JUNCTIONLESS FIELD-EFFECT TRANSISTORS

IEEE Press
445 Hoes Lane
Piscataway, NJ 08854

IEEE Press Editorial Board
Ekram Hossain, *Editor in Chief*

Giancarlo Fortino	Andreas Molisch	Linda Shafer
David Alan Grier	Saeid Nahavandi	Mohammad Shahidehpour
Donald Heirman	Ray Perez	Sarah Spurgeon
Xiaoou Li	Jeffrey Reed	Ahmet Murat Tekalp

JUNCTIONLESS FIELD-EFFECT TRANSISTORS

Design, Modeling, and Simulation

SHUBHAM SAHAY
MAMIDALA JAGADESH KUMAR

IEEE Press Series on Microelectronic Systems

WILEY

Copyright © 2019 by The Institute of Electrical and Electronics Engineers, 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 formats. For more information about Wiley products, visit our web site at www.wiley.com.

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

ISBN 978-1-119-52353-6

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

Dedicated to Saraswati Mata, the Goddess of Learning

CONTENTS

Preface **xi**

1 Introduction to Field-Effect Transistors **1**

1.1 Transistor Action, 2
1.2 Metal-Oxide-Semiconductor Field-Effect Transistors, 4
1.3 MOSFET Circuits: The Need for Complementary MOS, 9
1.4 The Need for CMOS Scaling, 11
1.5 Moore's Law, 13
1.6 Koomey's Law, 13
1.7 Challenges in Scaling the MOSFET, 13
1.8 Conclusion, 23
References, 23

2 Emerging FET Architectures **27**

2.1 Tunnel FETs, 28
2.2 Impact Ionization MOSFET, 34
2.3 Bipolar I-MOS, 39
2.4 Negative Capacitance FETs, 41
2.5 Two-Dimensional FETs, 46
2.6 Nanowire FETs, 49
2.7 Nanotube FETs, 51
2.8 Conclusion, 57
References, 58

3 Fundamentals of Junctionless Field-Effect Transistors 67

3.1 Device Structure, 69
3.2 Operation, 70
3.3 Design Parameters, 80
3.4 Parameters that Affect the Performance, 82
3.5 Beyond Silicon JLFETS: Other Materials, 100
3.6 Challenges, 103
3.7 Conclusion, 110
References, 111

4 Device Architectures to Mitigate Challenges in Junctionless Field-Effect Transistors 125

4.1 Junctionless Accumulation-Mode Field-Effect Transistors, 126
4.2 Realizing Efficient Volume Depletion, 129
4.3 SOI JLFET with a High-κ Box, 131
4.4 Bulk Planar JLFET, 137
4.5 JLFET with a Nonuniform Doping, 140
4.6 JLFET with a Step Doping Profile, 144
4.7 Multigate JLFET, 149
4.8 JLFET with a High-κ Spacer, 153
4.9 JLFET with a Dual Material Gate, 157
4.10 Conclusion, 162
References, 162

5 Gate-Induced Drain Leakage in Junctionless Field-Effect Transistors 173

5.1 Hole Accumulation, 174
5.2 Parasitic BJT Action, 176
5.3 Impact of BTBT-Induced Parasitic BJT Action on Scaling, 177
5.4 Impact of Silicon Film Thickness on GIDL, 179
5.5 Impact of Doping on GIDL, 187
5.6 Impact of Spacer Design on GIDL, 189
5.7 Nature of GIDL in Different NWFET Configurations, 190
5.8 Device Architectures to Mitigate GIDL, 199
5.9 Conclusion, 248
References, 249

6 Impact Ionization in Junctionless Field-Effect Transistors 255

6.1 Impact Ionization, 256
6.2 Floating Body Effects in Silicon-on-Insulator MOSFETs, 256
6.3 Nature of Impact Ionization in JLFETs, 260

6.4 Zero Gate Oxide Thickness Coefficient, 263
6.5 Single Transistor Latch-Up in JLFETs, 266
6.6 Impact of Body Bias on Impact Ionization in JLFETs, 267
6.7 Subband Gap Impact Ionization in DGJLFETS with Asymmetric Operation, 268
6.8 Impact of Gate Misalignment on Impact Ionization in DGJLFETs, 270
6.9 Spacer Design Guideline from Impact Ionization Perspective, 272
6.10 Hysteresis and Snapback in JLFETs, 273
6.11 Impact of Heavy-Ion Irradiation on JLFETs, 275
6.12 Conclusions, 276
References, 276

7 Junctionless Devices Without Any Chemical Doping 281

7.1 Charge Plasma Doping, 282
7.2 Charge Plasma Based p–n Diode, 283
7.3 Junctionless I-MOS FET, 288
7.4 Junctionless Tunnel FETs, 290
7.5 JLTFET on a Highly Doped Silicon Film, 294
7.6 Bipolar Enhanced JLTFET, 294
7.7 Junctionless FETS Without Any Chemical Doping, 297
7.8 Challenges for CPJLFETs, 302
7.9 Electrostatic Doping Based FETs, 312
7.10 Conclusions, 319
References, 319

8 Modeling Junctionless Field-Effect Transistors 327

8.1 Introduction to FET Modeling, 328
8.2 Surface Potential Modeling of JLFETs, 330
8.3 Charge-Based Modeling Approach, 351
8.4 Drain Current Modeling Approach, 355
8.5 Modeling Short-Channel JLFETs, 365
8.6 Modeling Quantum Confinement, 372
8.7 Conclusion, 379
References, 379

9 Simulation of JLFETS Using Sentaurus TCAD 385

9.1 Introduction to TCAD, 386
9.2 Tool Flow, 387
9.3 Sample Input Deck for Long-Channel JLFETS, 391
9.4 Model Calibration, 407
9.5 Model Calibration for Short-Channel JLFETs, 409
9.6 Model Calibration for NWFETS, 422

9.7 Conclusion, 436
References, 436

10 Conclusion and Perspectives **439**

10.1 JLFETS As a Label-Free Biosensor, 441
10.2 JLFETS As Capacitorless DRAM, 443
10.3 Nanowire Junctionless NAND Flash Memory, 444
10.4 Junctionless Polysilicon TFTS with a Hybrid Channel, 447
10.5 JLFETS for 3D Integrated Circuits, 449
10.6 Summary, 450
References, 451

Index **457**

PREFACE

We are living in an era of supercomputing where smartphones, smartwatches, and smart technology have become an inevitable part of our daily life. The research and development in the field of transistors, which forms the basic building block of the computing devices, has driven this "smart" revolution. The dimensions of the transistors have been incessantly scaled down to increase the number of transistors per chip, which has not only reduced the chip area enabling hand-held devices but also increased the functionality and operating frequency and decreased the power dissipation. However, all the modern-day transistors such as metal-oxide-semiconductor field-effect transistors (MOSFETs) (or tunnel field-effect transistors or ferroelectric field-effect transistors, etc.) contain two metallurgical junctions: one at the source–channel interface and other at the channel–drain interface. To further scale down the modern transistors to the sub-10 nm regime and exploit the performance improvements brought by the scaling process, the doping must change abruptly from a high value (typically $\sim 10^{20}$ cm^{-3}) in the source and drain regions to a low value (typically $\sim 10^{14}$–10^{16} cm^{-3}) with complementary dopants in the channel region within a span of a few nanometers ($\sim$1–2 nm). Experimental realization of such an ultrasteep doping profile is extremely difficult even with the industry-standard ion-implantation process. To add to this misery, achieving high dopant activation in the heavily doped source/drain regions requires a high-temperature annealing. The annealing process, in turn, leads to a thermally assisted lateral diffusion of dopant atoms from source/drain regions into the channel region. This further restricts the possibility of realizing ultrasteep doping profiles in MOSFETs. As lateral diffusion is inevitable while annealing, the simultaneous requirement of a high dopant activation and an ultrasteep doping profile puts a complex constraint on the thermal budget. Our lives as device designers

would have been much easier if there have been no metallurgical junctions. Therefore, to alleviate the need for ultrasteep doping profiles, field-effect transistors without any metallurgical junction were proposed to facilitate the scaling down of the conventional MOSFETs. These junctionless FETs (JLFETs) utilize an ultrathin semiconductor film with a gate stack to control its resistance and modulate the current flowing through it. The absence of a metallurgical junction leads to an altogether new conduction mechanism and device properties, which are different from conventional MOSFETs.

Surprisingly, the working principle of the JLFET was conceptualized and patented by Austrian-Hungarian physicist Julius Edgar Lilienfield in 1930 even before the discovery of the point-contact transistor by Shockley, Brattain, and Bardein in 1947. But it was only with the recent advancements in the fabrication technology that nanowire JLFETs were experimentally realized in 2010, inspired by Lilienfield's work. An exhaustive research has been carried out on JLFETs since then. The number of research papers on JLFETs has increased exponentially, and our understanding of JLFETs has also improved significantly over the years. The junctionless architecture, owing to its low cost, low fabrication complexity, and lower thermal budget, has opened up a new domain of exciting possibilities whereby JLFETs could be employed as sensors, memories, such as capacitor-less DRAM, NAND flash memory, display devices, and for biocompatible, optoelectronic, and three-dimensional (3D) sequential integrated circuit applications apart from logic applications. The enormous possibilities offered by the junctionless transistor architecture are exciting opportunities to the researchers to explore and invent novel device structures for a variety of applications ranging from logic circuits to memories, sensors, 3D integration, and display technology. However, due to the lack of a comprehensive textbook, research papers are currently the primary source of knowledge on JLFETs. With a plethora of research papers appearing on JLFETs, gaining a basic understanding of a JLFET and keeping track of the latest research is a challenge.

This book endeavors to be a comprehensive guide for those who are about to begin their study (and research) or have already started working on JLFETs. It provides a one-stop volume for studying JLFETs for someone having a basic knowledge of device physics. The book covers the fundamental physics behind the operation of JLFETs and provides a comparative analysis of different performance metrics of the JLFETs with respect to the MOSFETs. The book unfolds the challenges for JLFETs if they were to replace MOSFETs and incorporates a comprehensive study of the device architectures and designs proposed in the literature to mitigate the challenges and improve the performance of JLFETs. The book also includes a detailed analysis of the junctionless devices realized without the need for conventional chemical doping. In addition, it discusses in detail the different approaches used for analytical or compact modeling of JLFETs for the purpose of circuit design and circuit simulation. Therefore, this book is the first attempt to encompass the research reported on JLFETs on aspects spanning from device architectures and simulations to analytical modeling. Also, every aspect of the JLFET has been compared to the MOSFET so that the material presented in the book allows the entire semiconductor device fraternity to

evaluate the potential of JLFETs and take informed decisions regarding its integration with the prevailing technology in the industry. Another unique feature of this book is that it describes the process of carrying out numerical simulations of JLFETs using the technology computer-aided design (TCAD) tool Sentaurus S-device. TCAD simulations are helpful for studying the behavior of any semiconductor device without getting into the complex process of fabrication and characterization, thus reducing the time to market. The calibrated simulation setup provided in the book would definitely aid the researchers especially the beginners in the field and provide them with an effective tool to analyze, evaluate, and invent new junctionless architectures for different applications, which may serve as a stepping-stone in the early stage of their work. We hope that this book covering the fundamentals of the JLFET along with their analytical modeling and simulation using TCAD would encourage the beginners to pursue research on JLFETs and augment the efforts of the existing researchers to realize a power-efficient JLFET for “green” electronics, which would eventually lead to a better society.

1

INTRODUCTION TO FIELD-EFFECT TRANSISTORS

We are living in an era of information technology where smartphones, smart watches, and smart technology have become an inevitable part of our lives. You might have observed a drastic improvement in the performance of these smart devices. For instance, the shift from single core processors to multicore processors, the increase in CPU's frequency from few MHz to several GHz, the increase in the RAM from few MB to several GB, and so on. All these factors have led to a tremendous increase in the performance of these computing devices. The smart devices found in every household nowadays have a performance metric comparable to the earlier supercomputers. For instance, the Apple watch has twice the processing power of a 1985 Cray-2 supercomputer [1]. In addition, the device size has also shrunk significantly and the focus in the research and development of computing devices has shifted toward mobile devices. Moreover, the functionality per device has also increased considerably. For instance, the present day smartphones not only have processing capabilities of a supercomputer but can also perform the functions of a good quality camera, a Wi-Fi dongle, an X-BOX gaming system, and so on. To summarize, every other person in this modern era has access to low-cost, high-performance gadgets.

Have you ever wondered what drives the "smartness" and the supercomputing capabilities of all the smart technology gadgets? Let us try to understand this from a human body–gadget analogy. Just like the human body is composed of cells as the building block, the electronic gadgets are made up of transistors. In human body, the

Junctionless Field-Effect Transistors: Design, Modeling, and Simulation, First Edition.
Shubham Sahay and Mamidala Jagadesh Kumar.
© 2019 by The Institute of Electrical and Electronics Engineers, Inc. Published 2019 by John Wiley & Sons, Inc.

cells are grouped together to perform a particular function and form an organ. Therefore, the efficiency and the number of different functions that can be performed by the body depends exclusively on these cells. Similarly, the transistors act like a switch and are wired together in a chip (which is similar to the organ from body–gadget analogy) in a specific manner to enable a particular function. The larger the number of transistors in a gadget, the more the number of functions it can perform. The research and development in the field of transistors has driven this "smart" revolution. It is indeed very interesting how such small chunks of silicon chips drive our lives.

1.1 TRANSISTOR ACTION

But what exactly is a transistor? The word transistor was given by its first inventors: Shockley, Brattain, and Bardeen in 1947 [2–5]. At that time, no one would have wondered that this discovery (which actually was an accident) would be driving the lives of common people for generations to come. The transistors are often conceived as a device where the resistance between two terminals may be controlled by the current/voltage at the third terminal. Therefore, transistor refers to any three-terminal device where the current (or voltage) between two terminals may be controlled by the action of voltage (or current) at the third terminal.

In the subsequent sections, we shall see how the most common transistors work from both a qualitative approach and an energy band diagram perspective. The bipolar junction transistors (BJTs) dominated the semiconductor industry until late 1970s. Although BJTs are still used in the high-frequency circuits such as in radio frequency circuits, the throne is captured by the metal-oxide-semiconductor field-effect transistors (MOSFETs) and they continue to drive the semiconductor industry even today. Therefore, we shall discuss the MOSFETs in detail in the next section.

Transistors such as MOSFETs act as switches in the integrated circuits. However, it may be noted that the MOSFETs are not ideal switches (which are expected to consume no power when switched-OFF and deliver a high current instantaneously when switched-ON). The MOSFETs exhibit a small leakage current and, therefore, consume power from the supply even when they are switched-OFF. This power consumption is termed as the static power dissipation (P_s) given as

$$P_S = V_{DD} \cdot I_{OFF} \tag{1.1}$$

where V_{DD} is the supply voltage and I_{OFF} is the leakage current that flows through the transistor when the switch is turned off. Furthermore, the MOSFETs also consume a significant power when switched from the ON-state to OFF-state or vice versa. This power consumption also depends on the frequency of switching of the MOSFETs and is termed as the dynamic power dissipation (P_d) given as

$$P_D = V_{DD}^2 C_L f \alpha \tag{1.2}$$

where V_{DD} is the supply voltage, f is the frequency of operation and α is the switching probability, which simply tells us that the MOSFET is not switched in each cycle, and C_L is the load capacitance. In a wired network of MOSFETs, a MOSFET drives another MOSFET. Therefore, in most cases C_L is the input capacitance of the MOSFET. The interested readers are requested to refer [5] for more details.

Until recent past, the focus of the researchers all over the world was to miniaturize the dimensions of the MOSFETs so as to increase the number of MOSFETs per chip, which would not only reduce the area enabling mobile devices but also increase the number of operations that may be performed by a single chip. Scaling the MOSFET dimensions also reduces the input capacitance and increases its capability (current) to drive another MOSFET in the wired chip network and helps to achieve large frequency of operation due to fast charging of C_L. Although the drive current of MOSFET increases with scaling, the OFF-state current also increases drastically due to the short-channel effects that are triggered by MOSFET gate length scaling. The increase in the OFF-state current results in a significant static power dissipation. While the dynamic power dissipation was a major concern for the researchers until recent past, the scaling trends suggest that the static power dissipation would eventually surpass the dynamic power dissipation if the conventional MOSFETs are scaled aggressively.

A high static power consumption means that the MOSFETs would draw a significantly large power from the supply even when it is switched-OFF. Therefore, the chip would drain the battery or the power source even when the functionality provided by the chip is not being utilized. This is detrimental to the performance of computing devices especially for the hand-held devices like smartphones, which have a limited supply available in the form of a battery. Furthermore, the static power dissipation also heats up the chip and degrades the performance of the gadgets which are designed for room temperature operation. Of course, every consumer wants to have a smart device with an unlimited battery or power supply with no heating effects. To reduce the power dissipation, we can reduce the supply voltage as evident from equations (1.1) and (1.2). However, there lies a fundamental limitation on the MOSFETs which is inherent to the very physics of the device. The current in a MOSFET cannot increase by more than a ten-fold when the input voltage is raised by 60 mV. This limitation is due to the Maxwell–Boltzmann distribution of electrons in matter and is often referred to as the "Boltzmann tyranny." The application of MOSFET as a switch requires that the ON-state to OFF-state current ratio be high so that these states are easily distinguishable ($\sim 10^4$ to 10^6). To achieve an ON-state to OFF-state current ratio of a million, the variation of the input voltage, and therefore the supply voltage, needs to be at least equal to $60 \times \log(10^6) = 360$ mV. This limitation simply implies that if we have an extremely scaled supply voltage, the ratio of the ON-state current to the OFF-state current of the transistor would be very low and the MOSFET would cease to act like a switch. Therefore, the Boltzmann limit hinders the use of the conventional MOSFETs as a switch for ultralow supply voltages.

As a result, the conventional MOSFETs cannot cater the need of yielding an area and power-efficient chip with multiple functionalities. Moreover, scaling the conventional MOSFETs also requires a large investment from the manufacturing point of view. Therefore, present day research focuses on design of a low-cost and highly scalable MOSFET with minimum power dissipation. As you would have noted, the research and development in this context has gradually shifted from an area-driven perspective to a power-driven scenario.

This chapter will help to develop a basic understanding of the conventional MOSFETs. After a subtle discussion of the various modes of operation of these devices, Section 1.3 describes how basic circuits can be formed using MOSFETs. Section 1.3.2 focuses on different types of power dissipations reported earlier in the introduction.

1.2 METAL-OXIDE-SEMICONDUCTOR FIELD-EFFECT TRANSISTORS

To understand a MOSFET, we shall first get an in-depth understanding of a MOS capacitor (Fig. 1.1) which is the heart of a MOSFET, grasp the concept of "field–effect," and then discuss operation of MOSFETs. The MOS capacitor consists

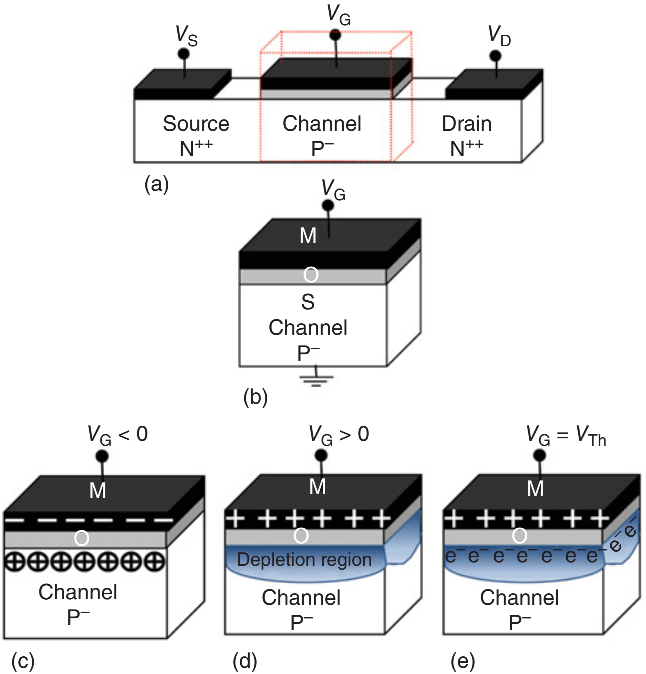


FIGURE 1.1 (a) Three-dimensional view of an n-MOSFET and (b) the MOS capacitor and the operation mode of a MOS capacitor in (c) accumulation regime, (d) depletion regime, and (e) inversion regime.

of three layers as the name suggests: metal-oxide-semiconductor. A thin insulating oxide layer is sandwiched between a metal and a semiconductor. Since the structure has a dielectric inserted between two conducting plates (assuming that the semiconductor is doped or at room temperature), the MOS structure is essentially a capacitor. The MOS capacitor with the p-type doped semiconductor is called a p-type MOS capacitor, whereas the MOS capacitor with an n-type doped semiconductor is called an n-type MOS capacitor.

The capacitance of the MOS structure can be controlled by the gate voltage just like the capacitance of a p–n junction is controlled by the applied bias. However, the range of capacitances exhibited by the MOSFET is large compared to the p–n junction capacitance.

At this point, we would also like to mention that the property of bulk atoms and surface atoms are different. Indeed, in the words of W. Pauli (who gave the Pauli exclusion principle), "God made the bulk; interfaces were invented by the devil" [7]. In MOS devices, all the charge dynamics occur at the surface. Therefore, silicon is the most preferred material for MOS devices as the Si–SiO_2 interface constitutes the best quality semiconductor–insulator interface. Silicon is also abundant on earth in the form of sand (silica).

1.2.1 "Field-Effect" and Operation Modes

To understand the different modes of operation of a MOS capacitor, it is essential to understand the concept of "field-effect" applied to the MOS devices. The field-effect simply means controlling the charge dynamics with the aid of an electric field. Now, let us look at how electric field controls charges in case of a p-type MOS capacitor shown in Fig. 1.1(b).

If we apply a negative voltage on the gate terminal with respect to silicon, an electric field will be generated across the insulator with a direction from the semiconductor to the metal. The applied negative potential on the gate can be conceptualized as depositing negative charges on the gate which attract the majority holes in the p–Si toward the Si–SiO_2 interface. Therefore, the holes would be accumulated at the Si–SiO_2 surface due to the application of a negative bias on the gate. From an electric field perspective, the holes move in the direction of the electric field and accumulate at the Si–SiO_2 interface. As a result, the effective carrier concentration at the interface is increased in the accumulation mode as shown in Fig. 1.1(c).

Now, if a positive voltage is applied to the gate terminal, the electric field direction across the insulator is reversed and points toward the semiconductor from the gate. A positive potential at the gate can also be conceptualized as depositing positive charges on the gate which repel the majority holes close to the Si–SiO_2 interface. The repelled holes move into the bulk leaving behind uncovered negative acceptor ions. Therefore, a depletion region is formed in the semiconductor in the vicinity of the Si–SiO_2 interface. In other words, the electric field pushes holes away from the interface to the bulk, increasing the depletion region width. This region of operation is called the depletion mode.

Now, what happens if the positive voltage is increased even further? One may expect that the depletion region would continue to expand until it spans the entire semiconductor. However, this is not what happens since the minority electrons are also there in the bulk of the p-type semiconductor, which may provide negative charge for the electric field lines from the gate to terminate. Therefore, when we continue to increase the positive voltage, the depletion region increases until a maximum value and then the minority electrons move to the surface from the bulk and start accumulating to facilitate the termination of the electric field lines. From electric field perspective, the field becomes so strong that it pulls the minority electrons to the surface. The value of gate voltage at which the electron concentration at the surface becomes equal to the bulk doping concentration of the p-type semiconductor is called the threshold voltage. At the threshold voltage, the majority carriers change from holes to electrons at the surface. This phenomenon is called inversion, and the electron layer is called the inversion layer.

For the n-type MOS capacitor, accumulation of electrons takes place for positive gate voltages. Upon application of a negative gate voltage, the semiconductor is depleted first and then the depletion region reaches its maximum value. As the magnitude of the negative gate voltage is increased further, the minority holes start moving to the surface and, eventually, an inversion layer of holes is formed. Therefore, the characteristics of a p-type MOS capacitor are just complementary to the n-type MOS capacitor and hence do not require a detailed discussion.

Now, with this background, we shall discuss the structure of a MOSFET.

1.2.2 MOSFET as a Switch

A MOSFET consists of the MOS capacitor appended by the source and drain regions, which makes it a three-terminal device as shown in Fig. 1.1(a). The three terminals of the MOSFET are source (acts as a source of carriers), gate (controls the amount/concentration of carriers), and drain (acts as the sink for carriers). The source and drain are heavily doped, whereas the channel is lightly doped with a polarity opposite to that of source and drain. MOSFETs also utilize a fourth terminal called the body terminal. The body terminal is connected to the channel and is used to manipulate the electron conduction (and the threshold voltage) under special circumstances. Otherwise, it is normally grounded.

As discussed in Section 1.2.1, if a positive bias greater than the threshold voltage is applied, an inversion layer of electrons is formed at the surface of p-type silicon. Now, if a positive voltage is applied at the drain terminal, the electrons of the source would find a low-resistance conduction path via the inversion layer and flow into the drain. Therefore, the electrons would flow from the source to the drain region and MOSFET would act as a closed switch. However, when the applied bias is lower than the threshold voltage, the channel region remains depleted and offers a high-resistance path. Therefore, the electrons in the source find it difficult to reach the drain via the channel and the MOSFET behaves like a open switch. Therefore, the MOSFET acts like a switch which can be switched ON or OFF depending on

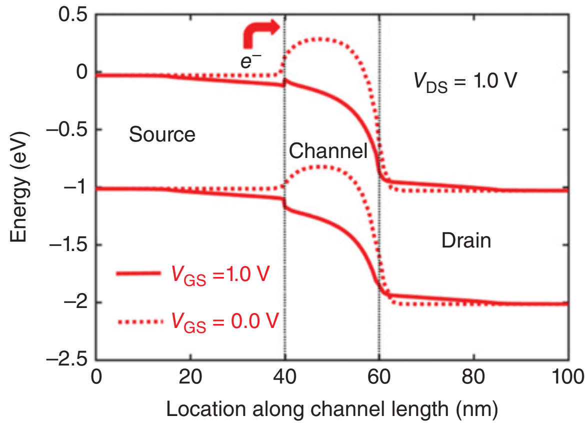


FIGURE 1.2 Energy band profiles of the MOSFET at ON-state (V_{GS} = 1.0 V) and OFF-state (V_{GS} = 0.0 V) showing that the gate voltage modulates the barrier height for source electrons.

the voltage applied to the gate. The drain to source current can, therefore, be controlled by the voltage applied to the gate. Since current through two terminals is being controlled by the voltage at the third terminal, the MOSFET is called a transistor, i.e. a resistor whose resistance may be controlled by the gate. As shown in Fig. 1.2, the gate voltage simply modulates the effective barrier height seen by the source electrons to move into the drain region through the channel.

1.2.3 Transfer Characteristics and Output Characteristics

At this point, we would like to introduce the concept of transfer characteristics and the output characteristics. The output characteristics refer to the relation between the output current (drain current in the case of a MOSFET) with the output voltage (drain voltage). Therefore, the output characteristics in a MOSFET are simply a plot between the drain current versus the drain voltage for a particular gate voltage as shown in Fig. 1.3(a). The drain current first increases linearly with the drain voltage and then gets saturated owing to the pinch-off of the channel region at the drain end. The inversion layer charge increases with increasing gate voltage leading to a larger drain current.

The relationship between the output current (drain current) and the input voltage (gate voltage) for a particular drain voltage is called the transfer characteristics. The drain current is very low below the threshold voltage (subthreshold regime) and is governed by diffusion of carriers from the source to the drain region. As the gate voltage increases above the threshold voltage, an inversion layer forms and the drain current increases significantly (Fig. 1.3(b)). If the transfer characteristics are plotted on a linear scale, the threshold voltage can be extracted by extrapolating the drain current after which the current starts increasing dramatically (Fig. 1.3(b)).

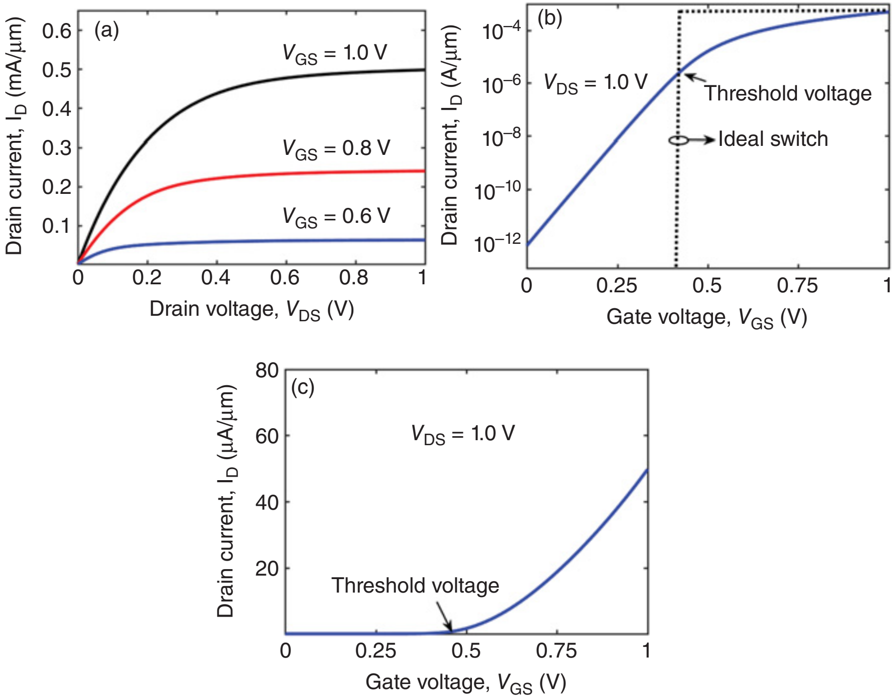


FIGURE 1.3 (a) Output characteristics and transfer characteristics of the MOSFET in (b) log scale and (c) linear scale.

Several methods have been proposed for determination of the threshold voltage. Some of these methods include finding the zeroes of the double derivative of the transfer characteristics (which is equivalent to finding maxima of the derivative), whereas others rely on finding the exact surface potential-gate voltage relationship by solving the Poisson equation in the channel region and equating the surface potential equal to twice the Fermi potential at threshold condition [8]. The Poisson equation simply relates the potential to the charge contained in any region and is defined as

$$\nabla^2\varphi = -\frac{\rho}{\varepsilon_{\mathrm{Si}}} \tag{1.3}$$

Though equation (1.3) appears very simple, it is very difficult to solve analytically and requires numerical solvers or approximations as discussed in Chapter 7. The simplest approach to find the threshold voltage is the constant current method, which defines threshold voltage as the gate voltage for a particular constant drain current, generally $(W/L_g) \times 10^{-7}$ A [8]. This approach is simplest as it does not involve any modulation or extrapolation or numerical solver.

In addition, other important parameters can also be extracted from the transfer characteristics at a drain voltage of V_{DD}. The drain current corresponding to the $V_{GS} = V_{DS} = V_{DD}$ is termed as the ON-state current, and the drain current at

$V_{GS} = 0$, $V_{DS} = V_{DD}$ is termed as the OFF-state current. Furthermore, the subthreshold swing can also be extracted from the transfer characteristics. There are two subthreshold swings for each transfer characteristics: a point subthreshold slope and an average subthreshold slope. The point subthreshold slope is simply the derivative of the transfer characteristics at a given gate voltage, whereas the average subthreshold slope is calculated by taking the mean of the point subthreshold slopes at different gate voltages ranging from the OFF-state voltage ($V_{GS} = 0$) to the threshold voltage ($V_{GS} = V_{Th}$) and is given as

$$SS_{avg} = \frac{V_{Th}}{\log(I_D)_{V_{Th}} - \log(I_{OFF})} \tag{1.4}$$

These are the parameters that are given as the design specification to a device designer. In general, there is a minimum ON-state current to OFF-state current ratio ($I_{ON}/I_{OFF} \sim 10^4$–10^6) and a maximum subthreshold swing, which a MOSFET must satisfy to be used in circuits.

With a background of essential physics of MOSFETs, we can now discuss the implementation of circuits with the help of MOSFET as a switch.

1.3 MOSFET CIRCUITS: THE NEED FOR COMPLEMENTARY MOS

If you remember from our discussion in Section 1.1, it is indeed these MOSFETs in the form of switches which are wired together in the integrated circuits to perform particular functions. Nearly every MOSFET in a circuit has to drive a load capacitance, which may correspond to input capacitance of other MOSFETs or an external load. Therefore, for circuit representation, we connect a load capacitance at the output terminal of the MOSFET. We chose the simplest circuit, i.e. an inverter, to give an insight into the digital circuit implementation using MOSFETs.

An inverter is a NOT gate which essentially inverts the input logic “0” into output logic “1” and vice versa. How can a MOSFET perform this action? Let us consider the case of an n-MOSFET with a load capacitance connected to the drain end (Fig. 1.4(a)), which represents the output load. If the output is initially at logic “1”, i.e. if the load capacitance is initially charged, then the voltage across this capacitor is essentially the drain voltage of the n-MOSFET. Now, if an input logic “0” ($V_G = 0$) is applied, since $V_{GS} < V_{Th}$ (assuming $V_{Th} \sim 0.2\ V_{DD}$), the n-MOSFET remains OFF and, hence, the output logic remains at “1”. However, if an input logic “1” ($V_G = V_{DD}$) is applied to the n-MOSFET, since $V_{GS} > V_{Th}$, the n-MOSFET turns on and current flows from the drain terminal to the source terminal. This current would drain the charges on the output load capacitance to the ground. This can also be viewed as discharging of the output load capacitance through the resistance of the n-MOSFET in the ON-state. Therefore, the load capacitance is discharged to the ground potential which corresponds to the output logic “0”. Therefore, the application of an input logic “0” leads to an output logic “1” and vice versa and the n-MOSFET acts like an inverter.

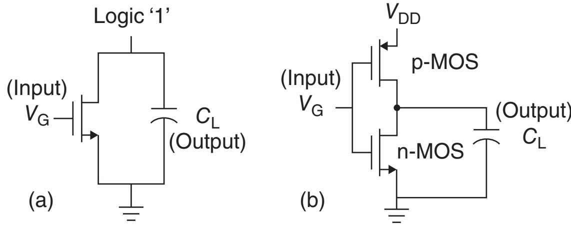


FIGURE 1.4 (a) n-MOSFET connected to a load capacitance, which is initially charged to logic "1", i.e., V_{DD} and (b) schematic view of a complementary metal-oxide-semiconductor (CMOS) inverter.

At this juncture, you may wonder what would happen in case the load capacitance was not charged initially? If the load capacitance is discharged, regardless of whether the input logic is "0" or "1" (n-MOSFET is OFF or ON), the output logic state remains at "0" because there is no means to charge the output load capacitance. Therefore, an n-MOSFET can only perform the inverter operation when the output logic is "1" and fails when the output logic is "0". Therefore, an n-MOSFET satisfies only half logic function and cannot be used alone for making even a simple inverter. Similarly, the p-MOSFET can only charge the load capacitance to logic "1" and works fine if output logic is initially "0" but fails when the output logic is initially "1". Therefore, even the p-MOSFET also satisfies only half logic function.

How can we make a complete inverter logic from MOSFETs if the n-MOSFET and p-MOSFET cannot individually perform the inverter operation? This calls for the need of complementary metal-oxide-semiconductor (CMOS) process. Since the p-MOSFET can charge the output to logic "1" whereas n-MOSFET can discharge the output to logic "0", both charging and discharging paths can be realized if they are used together as shown in Fig. 1.4(b). The p-MOSFET and n-MOSFET complement each other's logical function and can perform complete logic implementation only if used together. The circuit implementations in which both n-MOS and p-MOS are used together are known as CMOS circuits.

Now that we have recognized the importance of the CMOS process, let us analyze the working of a CMOS inverter and how a complete NOT operation is performed.

1.3.1 CMOS Inverter

The schematic of a CMOS inverter is shown in Fig. 1.4(b). When the input voltage is close to 0 V and low, for n-MOSFET, $V_{GS} < V_{Th}$ and it does not conduct. However, for the p-MOSFET, $|V_{GS}| > |V_{Th}|$. Therefore, the p-MOSFET turns ON and conducts. The load capacitance gets charged to V_{DD} via the p-MOSFET. Therefore, the output logic becomes "1". Similarly, when the input voltage becomes high, i.e. close to V_{DD}, for the n-MOSFET, $V_{GS} > V_{Th}$ and it conducts while for the p-MOSFET, $|V_{GS}| < |V_{Th}|$ and it remains switched OFF. Therefore, the load capacitance discharges to output logic "0" via the n-MOSFET. Hence, the inverter action is realized.

1.3.2 Power Dissipation in CMOS Inverter

A CMOS inverter takes in current from the supply only when both n-MOSFET and p-MOSFET are ON simultaneously, resulting in a path from the supply to the ground. Therefore, the current flows in a CMOS inverter only when the input voltage is close to 0.5 V_{DD} [6].

Ideally, we assume that the transistors do not consume any current when they are in the OFF-state. However, from our discussion in Section 1.1, we know that even below threshold voltage, the current is not equal to zero and a finite subthreshold leakage current flows through the MOSFETs. This leads to a power dissipation even when the input and output states of the CMOS inverter remain idle. This is called static power dissipation given by equation (1.1).

Also, every time the CMOS inverter output switches from “0” to “1”, the load capacitor gets charged by V_{DD}. Therefore, the charge deposited on the load capacitance is $V_{DD}C_L$. Now, the energy taken from the supply is simply $C_L V_{DD}^2$. However, when a capacitor is charged with a voltage V_{DD}, the energy stored in the capacitor is only 0.5 $C_L V_{DD}^2$. Therefore, out of the total $C_L V_{DD}^2$ energy taken from the supply, only half is stored in the load capacitance. Since the law of conservation of energy explicitly says that energy can neither be created nor be destroyed, where did the half of the energy go? Actually, the capacitor gets charged via the p-MOSFET, which acts as a resistor, and this half energy is dissipated as heat across this resistor. Now, when the inverter output switches from “1” to “0”, capacitor discharges through the n-MOSFET and the energy stored in the capacitor is dissipated through the n-MOSFET as heat. Hence, in every cycle of switching from “1” to “0” and back from “0” to “1”, a power equal to $C_L V_{DD}^2$ is dissipated in the CMOS inverter. Since power is dissipated only when the inverter switches, this power dissipation is called the dynamic power dissipation. The dynamic power dissipation may be generalized for any CMOS circuit as shown in equation (1.2).

1.4 THE NEED FOR CMOS SCALING

In Section 1.3.2, we analyzed the different power consumption mechanisms in the CMOS inverter. You may wonder what can be done to reduce the power consumption? The static power dissipation depends on the supply voltage V_{DD} and the OFF-state leakage current. Therefore, a reduction in the supply voltage or the leakage current can reduce static power dissipation. Since the digital circuits, for example, the microprocessor runs at a dramatically high frequency (~GHz range), the contribution from the dynamic power dissipation is most significant in the total power dissipation. The dynamic power dissipation depends on α, f, V_{DD}, and C_L. The parameter α depends on the functionality of the digital circuit and cannot be altered. The frequency of operation needs to be increased for faster computation speed. As a result, the power dissipation can only be lowered by reducing V_{DD} and C_L.

However, reducing the supply voltage reduces the ON-state to OFF-state current ratio (I_{ON}/I_{OFF}) due to the fundamental Boltzmann limit on subthreshold swing. For

feasible switching operation, the I_{ON}/I_{OFF} has to be at least 10^4. Therefore, the supply voltage cannot be reduced significantly.

The load capacitance C_L is essentially the input capacitance of similar logic circuits. C_L can be minimized by scaling down the area of MOSFETs. This calls for the need of CMOS scaling. Scaling the length of the MOSFET not only reduces the input capacitance but also increases the speed of the transistor as the ON-state current varies inversely with the gate length. Since the output capacitances are essentially charged and discharged, a higher drive current will increase the rate of charging or discharging. Therefore, scaling the MOSFET gate length leads to a reduced capacitance and facilitates high-frequency operation while increasing the number of MOSFETs and functionalities in a given chip area. CMOS scaling seems to be the best method to achieve a power efficient and high-speed multifunctionality device.

There are two ways in which CMOS scaling can be performed. These scaling techniques are categorized depending upon whether the supply voltage is scaled along with the channel length and width. A constant electric field scaling rule or the full scaling rule implies that the supply voltage is scaled by the same ratio as that of the length and the width. In the fixed-voltage scaling rule, the voltage is not scaled along with the length and width. The constant electric field scaling rule, which is also referred to as the Dennard's scaling rule, is followed in the industry, and the impact of the scaling factor (S) on various parameters is summarized in Table 1.1 [9].

TABLE 1.1 Scaling Factor for Different Parameters Utilizing Dennard's Scaling Rule [9]

Parameter	Scaling factor	Scaled value considering $S = \sqrt{2}$	Relative change
Channel length (L_g) Channel width (W) Gate oxide thickness (t_{OX}) Supply voltage (V_{DD})	$1/S$	0.7	30% ↓
Gate capacitance (C_{gg}) ~ $\left(\frac{WL_g}{t_{ox}}\right)$ Depletion region thickness (x_j) Intrinsic delay (τ) ~ $\left(\frac{C_{gg}V_{DD}}{I_{eff}}\right)$	$1/S$	0.7	30% ↓
Power dissipation (P_D) ~ ($V_{DD} \bullet I_{eff}$) Area (A) ~ ($W \bullet L_g$)	$1/S^2$	0.5	50% ↓
Power-delay product ~ ($P_D \bullet \tau$)	$1/S^3$	0.35	65% ↓
Electric field ~ $\left(\frac{V_{DD}}{t_{ox}}\right)$ Power density ~ $\left(\frac{P_D}{A}\right)$	1	1	0% =
Frequency (f) ~ $\left(\frac{1}{\tau}\right)$	S	1.4	40% ↑

1.5 MOORE'S LAW

At this point, we would also introduce the famous law of CMOS scaling introduced by Gordon Moore, which says that the number of transistors in a chip will double after every one and a half years (18 months). The CMOS industry followed this famous Moore's law for more than 40 years with the help of CMOS scaling [10]. However, the dimensions of the scaled MOSFETs gradually became comparable to the depletion region widths at the source–channel and channel–drain interfaces. Such MOSFETs in which the channel length approaches the source–channel or channel–drain depletion region width are known as short-channel MOSFETs. The MOSFET electrostatics, which we have discussed until now, can be extended to the short-channel MOSFETs with slight modifications, which arise due to the effects that originate only when devices are scaled to the short-channel regime. These short-channel effects are discussed in Section 1.7.1.

1.6 KOOMEY'S LAW

Another parameter for estimating the efficiency of the microprocessors apart from the number of transistors in a chip is the number of computations it performs per unit power consumption. This is a more fundamental property since it relates to the energy efficiency of the microprocessors. This parameter is evaluated as the number of computations performed by the microprocessor every kilo-watt-hour of power consumed by it when it is operating at its peak output frequency.

Interestingly, because of a reduction in the power dissipation and improvement in the operating speed owing to scaling, even the number of computations performed per unit energy consumption follows the same trend as the number of transistors per chip. Even this parameter has nearly doubled every 18 months [11]. This observation was first made by Jonathan G. Koomey and hence came to be known as the Koomey's law. However, unlike the deviation from the Moore's law owing to the short-channel effects, the Koomey's law has remained intact even in the post 2010 scenario and the computing efficiency with respect to power has been doubling every 18 months. Since the physical basis for Koomey's law is more centric to today's energy-efficient computing systems including Internet of things (IoT), servers, and big data systems, it is expected to last longer than the Moore's law.

Now, in the subsequent section, we will discuss about the challenges while scaling the MOSFETs.

1.7 CHALLENGES IN SCALING THE MOSFET

1.7.1 Short-Channel Effects

In Section 1.5, we discussed that if the channel length of a MOSFET is comparable to the length of the source–channel and channel–drain depletion regions, it is termed as the short-channel MOSFET [12]. You may wonder how much exactly is the depletion

region width at the source–channel or channel–drain interface? Typically, the doping concentration of source/drain region is $N_D = 10^{20}$ cm^{-3} and that of the channel is $N_A = 10^{16}$ cm^{-3}. The expression for depletion region width (x_{dep}) for a one-sided p–n junction with doping levels similar to MOSFET is

$$x_{dep} = \frac{2\varepsilon_{Si} V_{bi}}{qN_A} \tag{1.5}$$

$$\text{where} \quad V_{bi} = \frac{kT}{q} \ln\left(\frac{N_A \cdot N_D}{n_i^2}\right), \tag{1.6}$$

Solving this expression, the typical depletion region width at the source–channel or channel–drain interface comes out to be ~350 nm. As the channel lengths are scaled in this regime, several new physical phenomena arise and degrade the performance of MOSFETs. Therefore, in this section, we would give a brief overview of few short-channel effects, which dominate the performance of the MOSFETs in this ultrashort-channel length regime. Interested readers are directed to [13–15] for more detailed analysis of the short-channel effects.

1.7.1.A Threshold Voltage Roll-Off In a MOSFET, the depletion region of the heavily doped source/drain region protrudes into the p-channel and depletes it at the source–channel and channel–drain interface. As the gate length of the MOSFET is scaled to the short-channel regime, the source/drain-induced depletion region widths become a significant proportion of the overall channel length. The charge dynamics of the source/drain induced depletion region in the channel is no longer controlled solely by the "field-effect" of the gate electrode. Therefore, the effective channel charge that may be controlled by the gate electrode reduces significantly. As a result, the amount of gate voltage required to invert the channel region reduces considerably as compared to an undepleted MOS capacitor of similar dimensions. This reduction in the channel charge and the threshold voltage required to invert the channel with gate length scaling is known as threshold voltage roll-off [16, 17].

We already know that the subthreshold current, which contributes to the OFF-state leakage current, is due to diffusion of carriers from source to drain region and varies exponentially with the gate overdrive voltage, i.e., $V_{GS} - V_{Th}$. A lower V_{Th} simply means that the subthreshold leakage current would increase significantly. Therefore, the threshold voltage roll-off due to gate length scaling increases the OFF-state leakage current exponentially. It may be noted that while the increase in the ON-state current due to channel length scaling is linear, the OFF-state current increases exponentially. This leads to a significant reduction in the I_{ON}/I_{OFF} with channel length scaling.

1.7.1.B Drain-Induced Barrier Lowering As shown in Fig. 1.2, the application of a gate voltage simply modulates the source to channel barrier height and alters the