

**ASSET**

International Series on Advances in Solid State Electronics and Technology

Founding Editor: Chih-Tang Sah

# **BSIM4 AND MOSFET MODELING FOR IC SIMULATION**



**Weidong Liu  
Chenming Hu**

**World Scientific**

# **BSIM4 AND MOSFET MODELING FOR IC SIMULATION**

**International Series on Advances in Solid State Electronics and Technology  
(ASSET)**

*Founding Editor:* Chih-Tang Sah

---

*Published:*

Modern Semiconductor Quantum Physics  
*by Li Ming-Fu*

Ionizing Radiation Effects in MOS Oxides  
*by Timothy R. Oldham*

MOSFET Modeling for VLSI Simulation: Theory and Practice  
*by Narain Arora*

MOSFET Modeling for Circuit Analysis and Design  
*by Carlos Galup-Montoro and Márcio Cherem Schneider*

The Physics and Modeling of MOSFETs: Surface-Potential Model HiSIM  
*by Mitiko Miura-Mattausch, Hans Jürgen Mattausch & Tatsuya Ezaki*

Invention of the Integrated Circuits: Untold Important Facts  
*by Arjun N Saxena*

Electromigration in ULSI Interconnections  
*by Cher Ming Tan*

Compact Hierarchical Bipolar Transistor Modeling with HICUM  
*by Michael Schroter and Anjan Chakravorty*

BSIM4 and MOSFET Modeling for IC Simulation  
*by Weidong Liu and Chenming Hu*

**ASSET**

**International Series on Advances in Solid State Electronics and Technology**

**Founding Editor: Chih-Tang Sah**

# **BSIM4 AND MOSFET MODELING FOR IC SIMULATION**

**Weidong Liu**

Synopsys, USA

**Chenming Hu**

University of California, Berkeley, USA



**World Scientific**

NEW JERSEY • LONDON • SINGAPORE • BEIJING • SHANGHAI • HONG KONG • TAIPEI • CHENNAI

*Published by*

World Scientific Publishing Co. Pte. Ltd.

5 Toh Tuck Link, Singapore 596224

*USA office:* 27 Warren Street, Suite 401-402, Hackensack, NJ 07601

*UK office:* 57 Shelton Street, Covent Garden, London WC2H 9HE

**British Library Cataloguing-in-Publication Data**

A catalogue record for this book is available from the British Library.

**International Series on Advances in Solid State Electronics and Technology  
BSIM4 AND MOSFET MODELING FOR IC SIMULATION**

Copyright © 2011 by World Scientific Publishing Co. Pte. Ltd.

*All rights reserved. This book, or parts thereof, may not be reproduced in any form or by any means, electronic or mechanical, including photocopying, recording or any information storage and retrieval system now known or to be invented, without written permission from the Publisher.*

For photocopying of material in this volume, please pay a copying fee through the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, USA. In this case permission to photocopy is not required from the publisher.

ISBN-13 978-981-256-863-2

ISBN-10 981-256-863-8

**Disclaimer:** This book was prepared by the authors. Neither the Publisher nor its Series Editor thereof, nor any of their employees, assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information. The contents, views, and opinions of the authors expressed herein do not necessarily state or reflect those of the Publisher, its Series Editor, and their employees.

Printed in Singapore.

## **Dedication**

The authors dedicate this book to their families for their love and support, and the BSIM users of the past, present and future.

**This page intentionally left blank**

## Foreword

This monograph is the seventh book in this series on modeling of integrated-circuit devices. The purpose of this series is to provide archival references, described by the model originators or authorities, on the mathematically compact and computationally efficient engineering models of transistors and devices that are interconnected on a small, about a square centimeter or smaller area, silicon semiconductor wafer, die, dice, or chip, containing many integrated transistor circuits, which is the ‘brain’ of the equipment used in our daily activities today.

The monograph series idea came about six years ago in 2005 when I was asked by colleagues of Compact Modeling Council and Compact Model Community (CMC) to join them and give a keynote at their 4th annual international Workshop on Compact Modeling on May 10, 2005. I had been absent on transistor modeling research for 40 years since my first PhD student, Henry Pao, finished his PhD thesis in 1966 on the analytical model of the surface-channel silicon field-effect transistor (FET) with the Metal-Oxide gate on Silicon body (MOS). The prerequisite was to talk about a subject of their prime interest. But, I had not heard the term “compact model” and had the faintest idea of its meaning due to 40 years of absence. While searching the literature for references to write the keynote address, it became evident that there were many compact models and originators, but few books had provided descriptions sufficiently detailed for the professionals who must use the compact transistor models to execute a computer-aided-design of the integrated circuits that contain thousands to millions, now more than a billion transistors. A second purpose of this device modeling series is to provide state-of-the-art textbooks for graduate students and reference books for practicing engineers who are the users but may not be the designers of the latest integrated circuits.

After editing six compact modeling monograph volumes in this ASSET (Advances in Solid-State Electronics and Technology) series and reading literature on the faster-than-exponential rise of progresses in other areas of sciences and engineering, it seemed suddenly becoming obvious, if not already a common knowledge, that the word ‘Model’ is the buzz word everywhere. So perhaps it is timely to add a few words in this “Foreword of Opportunity” about “Model”, because in the two 60+year-old transistors which dominate everything of our life today, and even earlier devices, “model” has been the key to their discoveries and inventions. Quantitative modeling was certainly used even earlier in all engineering; even earlier, in evolution, in qualitative and habitual learning, and in the basics of the most basics of foundation of physics, explored by man about his origin and universe.

Our past, present and future are all model based. Some are more quantitative, others more qualitative; some more deterministic, others more statistical; some local, others global; some experience based, others more science (scientific method) based. Weather, archeology, evolution, particles and materials, societal and custom-habit, stock market, biology and medicine, are some of the examples, in addition to the engineering feat, the invention on paper, of the two transistors by Shockley (see

below). Most were compact or incomplete models based on a limited set of inventors' then current and latest knowledge, idea, notion, and imagination. All were already abbreviated or compact, compacted from the 'complete' theory based on intuitive physics, but most were further compacted to meet mathematical tractability by the inventor's own mathematical skill if not more so the inventor's desire to popularize the invention in laymen or more common language rather than abstract-obscure. But the compact models or compact-compact (compact<sup>2</sup>) models are increasingly less compact or increasingly more complicated because of increased computer capability to numerically solve them fast, due to better models that help design faster computers, creating still better models, a regenerative positive feedback loop. Such a loop fuels the increasingly faster than exponential rate of technology growth and advances, which would extrapolate to reach infinity at some finite future time, if not for some unseen, not modeled missing resistive-damping.

This book volume on the MOS transistor compact model BSIM exemplifies the development of a compact model, in this case, the one most used by man in its successful engineering applications on the design of computing, communication, and control circuits, which process electrical signals. BSIM is a compact model on the field effect transistor with insulated gate, the Insulated Gate Field-Effect Transistor (IGFET), more commonly known as the Metal-Oxide-Silicon field effect transistor (MOSFET) or just MOS Transistor (**MOST** which drives home the fact that it is the **most** abundant artificial device produced by man). The models or mathematical models for transistors are based on the current knowledge in the physics and 'model' of matter, the condensed phase which semiconducts electricity by two species of carriers or 'elementary' particles, electrons and holes, with exactly opposite unit of electrical charge, but somewhat different in their inertia masses, not unexpected in this imperfect world and asymmetrical, skewed universe, not unlike the two biological genders of much larger size, containing many more molecular particles.

Field effect transistor (FET, 1952) is one of the two transistors invented by William Shockley about sixty years ago. The other was the bipolar junction transistor (BJT, 1949). In-between, he wrote his classic textbook, *Electrons and Holes*. The two transistors have been the foundation devices of the electronic revolution. Shockley invented them on paper using the simplest physics models, one current channel in each, the diffusion current for the BJT and the drift current for the FET, both in the bulk or bulk-channel of a semiconductor, the Germanium, because crystalline Silicon was difficult to produce in the 1950's due to its high melting point, 1420C, even at the Bell Telephone Laboratories. The complete transistor model, which became obvious to this author not until 60 years later, is the eight channel device, analyzed by this author and his collaborator, Jie Binbin, both at Xiamen University, China, which is the eight combinations of electrical current channels, the drift and diffusion currents of electrons and holes in the surface and the bulk-or-body channels, giving  $2 \times 2 \times 2 = 2^3 = 8$  electrical current channels or electrical charge carrier transport pathways. The two foundation transistors, BJT and FET, could not have been invented if Shockley had not mentally visualized and singled out the two simplest (or most compact) models:

bulk drift (for FET) and bulk diffusion (for BJT) by just one carrier species (electron or hole), that enabled him to mathematically analyze and design the two transistors and to predict their electrical characteristics and performance, all on paper, 60+ years ago.

The comprehensive and authoritative compact modeling of the modern BJT was covered by the latest prior volume of this compact transistor model series, known as the HICUM, authored by Professor Michael Schroter of Germany and Dr. Anjan Chakravorty of India, and published by WSPC of Singapore in the fall of 2010.

The compact modeling of the second transistor type, FET, invented by Shockley in 1952, in the modern form, the silicon MOSFET or silicon MOST, is covered in this volume. The model is known as BSIM, and is in its fourth generation, BSIM4. It is authored by the founder-inventor of the BSIM, Professor Chenming Hu of China and USA, and his able assistant, Dr. Weidong Liu of China and USA. It is published in this ASSET series (this book) by WSPC one year later in the fall of 2011.

The publication sequence of these two volumes on the two transistors, 1949-BJT and 1952-FET in the modern form of Si MOST, followed the invention and the application histories of the two transistors. But this coincidental publication sequence was not planned in this compact modeling monograph series; however, the publication sequence does reflect the volume manufacturability determined by the advances of fabrication technology, especially the equipment, which also determines the application volumes and varieties.

I am especially thankful to the invited authors of the four startup volumes (Narain, Carlos+Márco, Mitiko+Hans+Tatsuya, and Arjun), of the later volumes (Cherming, Michael+Anjan, and Chenming+Weidong of this volume), and of the two to be published volumes (Ching-Hsiang+Yuan-Tsai and Francisco+Adelmo.). They concurred with my objectives and agreed to take up the chore of writing their books during their very busy schedules. Some have delays of one, two or even three years. Nevertheless, their monographs are still the archival records of the state of the art, and the world's authoritative contributions to the device modeling literature, because these authors are the creators, inventors and/or authorities of the models, and because the models are the industry-wide consensus models, used by all circuit designers of recent generations, and expected to be used by the future generations.

The present volume is a detailed comprehensive description of the latest version of the industrial consensus compact model for silicon MOSFET, the BSIM4, used by 90% if not 95% or more of the integrated circuit design and application engineers, and learned by all the electrical and computer engineering graduate students, for generations, two if not more decades. It also contains the historical development from the first BSIM, narrated by its creator, innovator, custodian, and father, Hu Chenming or Calvin Hu, the TSMC Distinguished Chair Professor at the University of California, Berkeley. A most important and salient teaching feature of this book is that the compact model of each of the electron-hole transport phenomena in the MOSFET is developed from the basic device physics of charge carrier transport in semiconductors, governed by the Shockley Equations. The compacted models are not only accurate in

device physics, but also in computational efficiency and accuracy to give the numerical results for representing the characteristics of the transistor which must be provided to the circuit simulator, such as SPICE, to analyze the performance of an integrated circuit, which may contain hundreds to millions of transistors, exceeding one billion recently. The fastest and largest (in memory capacity) computer would not be able to give numerical results in a short enough turn-around time (say a day or even a week) for manufacturability if the one million or more transistors were represented by their original and not-compacted models.

The last chapter gives the current and latest compact model for the upcoming generation of sub-quarter-100nm or 20-nm MOS transistor using the 3-dimensional Fin-like geometry structure, known as the FinFET, invented by Professor Hu.

The notations and terminologies employed in this monograph follow the industrial practice used by the design and manufacturing engineers of the MOSFETs. The monograph descriptions provide the connection of the notations and terminologies to the physics of semiconductors. For examples, silicon bandgap is noted as electron energy gap in silicon; avalanche multiplication and breakdown, as interband impact generation of electron-hole pairs by hot or energetic electron or hole; flicker noise, as 1/f noise from the addition of many trapping noise sources each with a different trapping time constant; white noise or Johnson noise, as random scattering noise with a reciprocal scattering rate in femto second range so the ‘corner’ or noise-power drop-off frequency is in the THz range; and many others.

I would like to thank all the WSPC production staff members at Singapore and their production editor of this monograph, Mr. Tjan Guangwei, and the acquisition editor, VP Ms. Zhai Yubing in New Jersey, for their untiring and dedicated efforts and supports. Special thanks are due NTU Physics Professor Kok-Khoo Phua, the Founder and Chairman of WSPC, for his farsights and foresights, one of which was pursuing me into authoring and editing textbooks and monographs in 1990, when I was rather discouraged if not also disgusted by the tactics of some publishers, which is a proof of my thesis, the Evolutionary Intelligent Design. I also thank Dr. Jie Binbin, Professor of Physics of Xiamen University, for editorial assistance. We also thank our supporters, President Zhu Chongshi, School of Physics Dean Wu Chenxu, and Department of Physics Chair Zhao Hong, all of Xiamen University, China. Finally, the editorial efforts of both of us, Sah Chihtang and Jie Binbin of this ASSET series, have been supported by the CTSAH Associates, Florida, USA, which was founded by the late Linda Su-Nan Chang Sah in mid-1970 at Urbana, Illinois, and which was reactivated on February 27, 2010, at Gainesville, Florida.

### **Sah Chihtang (Chih-Tang Sah)**

**Department of Physics, Xiamen University, China.**

**CTSAH Associates, Florida, USA.**

Initial draft, July 1, 2011, near the North Pole on way to Beijing and Xiamen from Seattle on Delta B767-300ER.

Final version, July 4, 2011 at Yi-Fu Hotel, Xiada campus.

Eighth year on August 13, 2011 and moving forwards.

## Preface

The compact models of semiconductor devices are the bridges between design and manufacturing in the integrated circuit industry. As such, compact models play an important and unique role in the IC technology, which brings gigantic benefits to the world economy and the quality of human life.

The compact models are as old as the computer simulation of integrated circuits with models such as MOS Level 1 and Gummel-Poon bipolar junction transistor models dating back to the 1970's. In the 1990's, the advent of the foundry-fabless partition in the IC industry changed compact modeling in several ways. First, the compact models were suddenly under the limelight because they became the bridge between IC chip foundries and hundreds of design companies that use these foundries. Second, the need for very accurate compact models was heightened because the models would now serve as the "contract" on transistor behaviors between the foundries and their many customers. Third, the compact models became much more complex than MOS Level 1 and Gummel-Poon BJT for deep-submicron and nanometer process technologies and, as a result, many major IC companies sought to reduce the high cost of developing their own compact models by joining force to seek, select, and support *standard* compact models under the banner of Compact Model Council. The first international standard compact model was BSIM (Berkeley Short-channel IGFET Model).

The authors have been honored and privileged to develop and support the BSIM4 model for the IC industry worldwide. Following its predecessor BSIM3v3, the industry's first standard, BSIM4 has served nearly all IC design and manufacturing companies from the 130-nanometer technology node down to the 20-nanometer node as of the

date of this book publishing. Over the years, BSIM4 has contributed to the development of IC products worth untold billions of dollars. During that time, we communicated with innumerable BSIM4 users around the globe. Often the same discussion topics would be brought up in those communications. Thus, we set out to write this book to provide the insights and comprehensive descriptions of BSIM4 and address those topics.

Another aim of this book is to share the knowledge and experiences we have gained from developing the BSIM models so as to further the art of compact modeling for emerging IC technologies.

Behind this book are many unsung heros, counted in hundreds. It is impossible for us to name them all. We would like to thank them collectively. But we particularly would like to give special thanks to the following.

We would like to acknowledge our fellow BSIM4 development team members at the University of California at Berkeley for their invaluable contributions to BSIM4 and to our knowledge. They are Xiaodong Jin, Kanyu M. Cao, Xuemei (Jane) Xi, Wenwei (Morgan) Yang, Chung-Hsun Lin, Mohan Dunga, Darsen Lu, Pin Su, Jin He, Tanvir Morshed, Jeff J. Ou, Mansun Chan, and Ali M. Niknejad.

We acknowledge the discussions, testing, and improvements on BSIM4 from Keith Green, Josef Watts, Judy An, William Liu, Britt Brooks, Min-Chie Jeng, Sally Liu, Yu-Tai Chia, Bing Sheu, Ke-Wei Su, Chung-Kai Lin, T. L. Tsai, C. S. Yeh, J. K. Chen, Jin-Shyong Jan, Annie Kuo, Peiming Lei, Daniel Wan, Waisum Wong, Richard Williams, Lawrence Wagner, Yoo-Mi Lee, Calvin Bittner, Peter Lee, Wenliang Zhang, Takahiro Iizuka, Paul Humphries, Geoffrey Coram, Jean-Paul Morin, Andre Juge, Peter Klein, Ali Icel, Jung-Suk Goo, Changhong Dai, Shiu-Wuu Lee, Anwen Liu, Susan Wu, Eugene Chen, Jeff Watt, Liping Li, Tom Mahatdejkul, Sam Lo, Akira Ito, Ben Gu, Jushan Xie, Ahmed Ramadan, Rick Poore, James Ma, Jonathan Sanders, John O'Donovan, and other BSIM users.

The authors especially would like to acknowledge the colleagues at the Analog Mixed-Signal group, Synopsys, Inc., for their invaluable

support and encouragement given to this book writing and their contributions to making BSIM4 a successful compact model for the IC industry over the past decade.

Special thanks are due to the editing and publishing staff at WSPC. They are Professor Chih-Tang Sah, Ms. Yubing Zhai, Gregory Lee, Jason Lim, Rajesh Babu, Yolande Koh, and Quek Yeow Hwa for their year-long relentless efforts in numerous rounds of rigorous expert review, editing and proofreading of the manuscript. The tremendous attention by Professor Sah to every detail significantly improved the quality of this book.

Weidong Liu  
Silicon Valley, California

Chenming Hu  
Berkeley, California

July 1, 2011

**This page intentionally left blank**

# Contents

<b>Forword</b>	vii
<b>Preface</b>	xi
<b>Chapter 1 BSIM and IC Simulation</b>	1
1.1 Circuit Simulation and Compact Models	1
1.2 BSIM – The Beginning	1
1.3 BSIM3 – A Compact Model Based on New MOSFET Physics	3
1.4 BSIM3v3 – World’s First MOSFET Standard Model	5
1.5 BSIM4 – Aimed for 130nm Down to 20nm Nodes	6
1.6 BSIM SOI	7
1.7 Impact of BSIM	7
1.8 Looking Towards the Future – The Multi-Gate MOSFET Model	8
1.9 The Intent of This Book	8
References	9
<b>Chapter 2 Fundamental MOSFET Physical Effects and Their Models for BSIM4</b>	13
2.1 Introduction and Chapter Objectives	13
2.2 Gate and Channel Geometries and Materials	14
2.2.1 Gate and Channel Lengths and Widths	14
2.2.2 Model Card and Parameter Binning	16
2.2.3 Gate Stack and Substrate Material Model Options	18
2.3 Temperature-Dependence Model Options	21
2.4 Threshold Voltage	22
2.4.1 Long Channel with Uniform Substrate Doping	22
2.4.2 Short-Channel Effect: V <sub>th</sub> Roll-Off and Drain Bias Effects	25
2.4.3 Narrow-Width Effects	31
2.4.4 Non-Uniform Substrate Doping	32
2.4.4.1 Non-Uniform Vertical Doping	33
2.4.4.2 Non-Uniform Lateral Doping: Pocket Implants	36
2.4.5 V <sub>th</sub> Temperature Dependence	39
2.4.6 BSIM4 V <sub>th</sub> Equation	40
2.5 Poly-Silicon Gate Depletion	42
2.6 Bulk-Charge Effects	45

2.7 LDD Resistances	46
2.8 Finite Charge Thickness	51
2.9 Effective Mobility	53
2.10 Layout-Dependent Effects: Mechanical Stress and Proximity Effects	58
2.11 Chapter Summary	66
2.12 Parameter Table	66
References	85
<b>Chapter 3 Channel DC Current and Output Resistance</b>	<b>87</b>
3.1 Introduction and Chapter Objectives	87
3.2 Channel Current Theory	88
3.3 Single Continuous Channel Charge Model	89
3.4 Channel Current in Subthreshold and Linear Operations	93
3.5 Velocity Saturation and Velocity Overshoot	95
3.6 Output Resistance in Saturation Region	98
3.6.1 CLM: Channel Length Modulation	100
3.6.2 DIBL: Drain-Induced Barrier Lowering	103
3.6.3 DITS: Drain-Induced Threshold Voltage Shift Due to Non-Uniform Doping	104
3.6.4 SCBE: Substrate Current Induced Body-Bias Effect	105
3.6.5 Channel Current Model for All Regions of Operation	106
3.7 Source-End Velocity Limit	107
3.8 Chapter Summary	108
3.9 Parameter Table	109
References	113
<b>Chapter 4 Gate Direct-Tunneling and Body Currents</b>	<b>115</b>
4.1 Introduction and Chapter Objectives	115
4.2 Gate Direct-Tunneling Current Theory and Model	116
4.2.1 Tunneling Mechanisms and Current Components	116
4.2.2 Gate Oxide Voltage	120
4.2.3 Gate-Body Tunneling Current $I_{gb}$	121
4.2.4 Gate-Source/Drain Tunneling Through Overlap Regions	123
4.2.5 Gate-Channel Tunneling Current	125
4.2.5.1 $I_{gc0}$ : The $V_{ds} = 0$ Bias Scenario	125
4.2.5.2 $I_{gcs}$ and $I_{gcd}$ Partitioning: The Non-Zero $V_{ds}$ Scenario	127
4.2.6 Characterization and Parameter Extraction	137
4.3 Body Currents	138
4.3.1 Impact Ionization	139
4.3.2 Gate-Induced Source and Drain Leakage	143
4.4 Summary of BSIM4 Branch and Terminal DC Currents	147
4.5 Chapter Summary	148

4.6 Parameter Table	149
References	152
<b>Chapter 5 Charge and Capacitance Models</b>	<b>155</b>
5.1 Introduction and Chapter Objectives	155
5.2 MOSFET Capacitance Theory	157
5.3 Intrinsic Charge and Capacitance Models	167
5.3.1 Charge-Thickness Model (CTM)	168
5.3.2 CAPMOD = 2 Charge Model Formulations	176
5.4 Fringing and Overlap Capacitances	180
5.4.1 Fringing Capacitances	180
5.4.2 Overlap Capacitances	181
5.5 Chapter Summary	183
5.6 Parameter Table	184
References	186
<b>Chapter 6 Non-Quasi-Static and Parasitic Gate and Body Resistances</b>	<b>189</b>
6.1 Introduction and Chapter Objectives	189
6.2 Gate Electrode Resistance	190
6.3 Gate Intrinsic-Input Resistance for Non-Quasi-Static Modeling	193
6.4 Charge-Deficit Transient and AC NQS Models	201
6.4.1 Charge-Deficit Transient NQS Model	201
6.4.2 Charge-Deficit AC NQS Model	211
6.5 Body Resistance Network	214
6.5.1 RBODYMOD = 1: A Local Network	216
6.5.2 RBODYMOD = 2: A Scalable Network	218
6.5.2.1 The 5-R Model	219
6.5.2.2 The 3-R Model	224
6.5.2.3 The 1-R Model	225
6.6 Chapter Summary	226
6.7 Parameter Table	227
References	233
<b>Chapter 7 Noise Models</b>	<b>235</b>
7.1 Introduction and Chapter Objectives	235
7.2 Noise Representations and Parameters	236
7.2.1 Noise and Power Spectral Intensity	236
7.2.2 SPICE Noise Representations	238
7.2.3 Noise Representation and Parameters of a Two-Port Network	239
7.3 BSIM4 Flicker Noise Models	246
7.3.1 The FNOIMOD = 0 Simple Flicker Noise Model	248

7.3.2 The FNOIMOD = 1 Physics-Based, Unified Flicker noise Model	248
7.4 BSIM4 Channel Thermal Noise Models	253
7.4.1 The TNOIMOD = 0 Charge-Based Model	254
7.4.2 The TNOIMOD = 1 Holistic Thermal Noise Model	255
7.5 Other Noise Sources	263
7.6 Chapter Summary	263
7.7 Parameter Table	264
References	266

## **Chapter 8 Source and Drain Parasitics: Layout-Dependence Model**

	<b>269</b>
8.1 Introduction and Chapter Objectives	269
8.2 Connections of a Multi-Transistor Stack	270
8.3 Source and Drain of a Transistor With Multiple Gate Fingers	273
8.4 GEOMOD: The End-Source and End-Drain of a Multi-Finger Transistor	275
8.5 Source and Drain Area and Perimeter Calculation	276
8.6 Saturation Junction Leakage Current and Zero-Bias Capacitance Models	290
8.7 Source and Drain Contact Scenarios and Diffusion Resistances	292
8.8 RGEOMOD: Selecting A Source and Drain Contact Scenario for GEOMOD	296
8.9 Chapter Summary	298
8.10 Parameter Table	298

## **Chapter 9 Junction Diode IV and CV Models**

	<b>303</b>
9.1 Introduction and Chapter Objectives	303
9.2 Physical Mechanisms of Diode DC Currents	303
9.2.1 Shockley-Read-Hall (SRH) Generation and Recombination	305
9.2.2 Trap-Assisted Tunneling (TAT)	307
9.2.3 Band-To-Band Tunneling (BTBT)	308
9.2.4 Diode Breakdown	309
9.3 BSIM4 Diode DC IV Model [4]	311
9.3.1 DIOMOD = 0	311
9.3.2 DIOMOD = 1	315
9.3.3 DIOMOD = 2	316
9.4 BSIM4 Junction Leakage Due to Trap-Assisted Tunneling [4]	321
9.5 BSIM4 Diode Charge and Capacitance [4]	322
9.5.1 BSIM4 Diode CV Model [4]	323
9.6 Diode Temperature-Dependence Model [4]	328
9.6.1 Temperature-Dependence Model for Diode IV	329
9.6.2 Diode CV Temperature-Dependence Model	332

9.7 Chapter Summary	333
9.8 Parameter Table	333
References	343
 <b>Chapter 10 SPICE Implementation Example:</b>	
<b>The Methodology with BSIM4 Transient NQS</b>	
10.1 Introduction and Chapter Objectives	345
10.2 Review of the Charge-Deficit Transient NQS Model	346
10.3 Time Discretization, Equation Linearization and Matrix Stamping	347
10.3.1 Discretization and Linearization of $i_{ch\_qs(t)}$	351
10.3.2 Stamping of $i_{ch\_qs(t)}$	356
10.3.3 Linearization of $i_{C_{\tau_{nqs}}(t)}$	357
10.3.4 Stamping of $i_{C_{\tau_{nqs}}(t)}$	357
10.3.5 Linearization of $i_{R_{\tau_{nqs}}(t)}$	358
10.3.6 Stamping of $i_{R_{\tau_{nqs}}(t)}$	360
10.3.7 Linearization of $i_g(t)$	361
10.3.8 Stamping of $i_g(t)$	363
10.3.9 Linearization of $i_d(t)$	363
10.3.10 Stamping of $i_d(t)$	366
10.3.11 Linearization of $i_s(t)$	368
10.3.12 Stamping of $i_s(t)$	371
10.4 Composite Stamps for Transient NQS Model	373
10.5 Bypass	375
10.6 Convergence Checking	382
10.7 Chapter Summary	385
References	386
 <b>Chapter 11 Multi-Gate Transistor Model</b>	
11.1 Introduction and Chapter Objectives	387
11.2 Advantages of FinFETs Over Planar CMOS	388
11.3 BSIM-CMG	390
11.3.1 The Core Model: Surface Potential Modeling	390
11.3.2 Channel I-V Model	396
11.3.3 Charge and Capacitance Models	399
11.3.4 Modeling of Advanced Physical Effects	403
11.4 Model Validation	406
11.5 Chapter Summary	409
References	409
 <b>Index</b>	411

**This page intentionally left blank**

## Chapter 1

# BSIM and IC Simulation

BSIM (Berkeley Short-channel IGFET Model) became the first international industry standard compact model for the simulation of CMOS (Complementary Metal-Oxide-Semiconductor) integrated circuits in 1997. It is believed that most of the ICs developed worldwide since 1998 were designed using BSIM. BSIM has served a wide range of CMOS technologies and IC applications.

### 1.1 Circuit Simulation and Compact Models

An integrated circuit contains millions to billions of transistors. The functionality and performance of the circuits must be verified by computer simulation before it is committed to expensive fabrication. Circuits are simulated by a method known as SPICE (Simulation Program with Integrated Circuits Emphasis), which was first developed by Professors Ron Rohrer and Don Pederson and their students at the University of California, Berkeley, in the early 1970's. In this method, the differential and algebraic nodal and branch equations (DAE) of an integrated circuit are solved by numerical analysis algorithms. The circuits are usually nonlinear because transistors such as MOSFETs (Metal-Oxide-Semiconductor Field Effect Transistors) are nonlinear devices (in contrast to a bias-independent resistor or capacitor).

For MOSFET devices, the complex behavior of the transistor drain current of the general form of  $I_d(V_g, V_d, V_s, V_b, L, W)$  is accurately represented by a group of analytical equations known as a compact model. If these equations are printed on paper, they may occupy a few pages. However, when they are implemented in SPICE, tens of thousands of computer code lines result. The length and complexity of

the functions used in the equations have significant impact on the circuit simulation time. It is therefore important to optimize both the computational efficiency of the model and its accuracy. In addition to the drain current, the device terminal charges and/or capacitances are also represented by analytical equations.

The model equations inevitably contain many adjustable constants known as model parameters. They are adjusted by modeling engineers to fit the compact model to measured terminal currents, conductances, charges and capacitances of the transistors — a process known as model parameter extraction. This process is performed with the aid of extraction software tools.

Large circuits are designed sometimes by using what is called cell library methodology. A circuit is often assembled from a pre-selected and characterized library of building blocks known as standard cells. A standard cell library typically contains many hundreds or thousands of standard cells such as inverters, NOR and NAND gates, flip flops, as well as other more complex cells. Every cell is characterized with SPICE simulation and compact models. The simulation results are reduced to a macro model — Often an expression of input to output delay, power dissipation, noise, and circuit gain as functions of input voltage ramp rates, frequencies, load capacitances, device parameters, supply voltages, and temperatures.

In other cases, circuits are simulated and designed with SPICE and compact models directly. For instance, an entire memory chip may be simulated with SPICE. Very often, large circuits are divided into smaller blocks each containing hundreds or thousands of transistors for SPICE simulation in order to complete the circuit simulation in hours rather than days or even weeks. In order to speed up the simulation of even larger circuits including SRAM memory circuits, another class of SPICE, “fast SPICE simulators”, are utilized to obtain satisfactory accuracy but with only a fraction of SPICE simulation time. This is accomplished with a combination of techniques including circuit partitioning and building table models of transistor characteristics rather than evaluating the equations of the compact model for every iteration at every time point. However, tables are still built from model equations. Therefore, nearly all ICs are designed with the use of transistor compact models.

## 1.2 BSIM – The Beginning

BSIM stands for Berkeley Short-channel IGFET Model. IGFET (Insulated-Gate Field Effect Transistors) is an older, more generic name for MOSFET transistors. BSIM's genesis can be traced back to 1984 [1 – 2]. This work produced BSIM1. Later Min-Chie Jeng working with Ping Ko and Chenming Hu introduced a successor version BSIM2 [3].

Hu and Ko had a close research collaboration in MOSFET physics and technology. The BSIM research was funded by the Semiconductor Research Cooperation (SRC), a consortium of semiconductor companies that funds university research projects deemed important to the IC industry. Their research into the more fundamental device physics and behaviors of advanced MOSFETs attracted additional industry supports. In this way, they gradually built a collection of models for the  $V_{th}$  (threshold voltage) dependence on biases and gate lengths, mobility degradation, velocity saturation effects, output conductance, unified flicker noise theory, SOI, and gate tunneling leakage. Eventually, these models became the building blocks of later BSIM models.

## 1.3 BSIM3 – A Compact Model Based on New MOSFET Physics

Most IC simulations require the accuracy to be better than a few percent from linear to saturation and from subthreshold to strong inversion covering a current range from pA/ $\mu$ m to mA/ $\mu$ m. This accuracy is achieved by painstaking modeling of many physical phenomena in a modern MOSFET including electrostatic, materials, quantum, and thermal effects. When another student Jian-Hui Huang started to work on a new version of BSIM around 1991, it was decided that the new version would incorporate some of the new original physical models mentioned in the previous sub-section. This approach was a marked departure from all previous compact models. Those models opted for simple equations favoring the reliance on “fitting” the transistor data to simplistic model equations over the use of physical, predictive, and complex models. The decision was a gamble that physics-based models can justify the model computational cost with their better accuracy and robustness. This new

approach began modestly with more accurate models of  $V_{th}$  and the drain saturation voltage,  $V_{dsat}$  [4] and an accurate model of the output conductance of MOSFETs [5]. The output conductance is very important to the accurate simulation of analog circuits because it determines the voltage gains of amplifying circuits. The result was BSIM3 [6]. BSIM3 was soon recognized as being far superior to the other compact models of that time.

For example, the output conductance is no longer explained by just channel length modulation but two additional mechanisms — drain-induced barrier lowering (DIBL) and hot carrier induced body bias effects [5]. Each of these three mechanisms in turn is modeled with insights derived from research on the quasi-two-dimensional analysis of the velocity saturation region near the drain [4], the effect of channel length and drain voltage on the threshold voltage [7], and the hot-electron current [8]. For the first time, a compact model can model the output conductance in a way that is not only accurate but also predictive of the effects of changing the gate oxide thickness, junction depth, and the threshold voltage. On the other hand, Reference [7] is the basis of BSIM3's predictive  $V_{th}$  roll-off model, and Reference [8] is the basis of the substrate current model.

Another example is the gate induced-drain leakage or GIDL [9], the band-to-band tunneling current induced by the gate-to-drain voltage  $V_{gd}$ . Once the mechanism was clearly understood, a simple analytical model became obvious and proved to be accurate.

Yet another example is the flicker noise or 1/f noise. The unified flicker noise model incorporates both the fluctuation in the number of inversion layer charge carriers (the number fluctuation) and the fluctuation in the Coulombic scattering mobility (the mobility fluctuation). They are correlated because both are caused by the fluctuation in the number of trapped charges in the  $\text{SiO}_2$  near the silicon/ $\text{SiO}_2$  interface [10]. This model was characterized in detail using the random telegraphic noise measurements that can only be observed in transistors with very short lengths and narrow widths such that a transistor may contain only one or two oxide traps [11]. These physics studies led to an accurate compact model of the MOSFET flicker noise [12].

## 1.4 BSIM3v3 – World’s First MOSFET Standard Model

BSIM3 was released in 1993. In that year, Ko moved to Hong Kong but stayed involved with BSIM for several more years. Hu continued to improve the physical accuracy by modeling poly-silicon depletion, universal mobility based on  $V_{th}$  and the gate voltage  $V_{gs}$  [13]. More important, the subthreshold and inversion regions of operation are now modeled with a single continuous function,  $V_{gsteff}$ . Verified with careful split-capacitance measurements, this model was a major improvement [14]. Similarly the linear and saturation regions are modeled with a single function, rather than piecewise formulations in practice then, through a continuous effective drain-to-source voltage  $V_{dseff}$ . These changes eliminated glitches in high-order derivatives of MOSFET drain current.

Encouraged by the early favorable reactions from the industry, BSIM3 began to emphasize robustness and usability for various CMOS technologies. BSIM was to be maintained and supported as a tool that billion-dollar companies may depend on for operation. In other words, impact to the industry joined creation of knowledge as an explicit goal of the BSIM project. Efforts were made to scrub thousands of lines of C-code for possible occurrences of numerical problems such as divide-by-zero, square-root-of-negative, round-off, discontinuities and over/underflows that would cause circuit simulation to abort. It also meant timely releases of bug fixes and active communications with users via numerous emails and the BSIM release website, <http://www-device.eecs.berkeley.edu/~bsim3/>. BSIM3v3 won the R&D 100 Award from the R&D Magazine as one of the most significant R&D products of 1995. These efforts culminated in a robust productized release of BSIM3v3 [15, 16, 17].

The semiconductor foundry industry was beginning to grow. A foundry gives its customers, the circuit design companies, basically these inputs to undertake the design process: Compact models such as BSIM, a set of geometrical and electrical design rules, and reference design flows. The foundry is then obligated to deliver working products of the designs that the customers may create with these inputs. To both parties, a very close agreement between the models and all the behavioral details of transistors is of paramount importance. One practical way to eke out the

## 6 BSIM4 AND MOSFET MODELING FOR IC SIMULATION

By Weidong Liu and Chenming Hu

last few percents of accuracy is a pedestrian technique called binning, making model parameters themselves functions of transistor channel lengths and widths, such that a single set of model parameters can apply to all transistors manufactured with the same process recipe.

Inspired by the success of BSIM3v3, several companies started to organize a movement to standardize a compact model. The purpose was to promote the adoption of a standard model by the IC industry. At that time, most large semiconductor companies developed their own transistor models. Often a single company would have many models in active use at the same time. Smaller companies would license models from CAD tool companies. There were many dozens of MOSFET models in use and the resources expended by each company to maintain these models and to fit each new generation of transistors to multiple models were enormous. This problem was aggravated by the dual trends of increasing inter-company and international collaborations and the growing foundry-fabless business model. Both trends would be well served if all the companies use the same model.

In 1996, Sematech, another semiconductor industry consortium, organized a series of workshops to discuss whether and how to select a standard model. This led to the formation of the Compact Model Council (CMC) that defined and executed the year-long process of selecting an industry standard MOSFET model. Several models competed in the selection process. In 1997, BSIM3v3 was selected by CMC as the world's first standard transistor model for IC simulation. The industry rallied around BSIM. Since then, BSIM3v3 has been employed for the 0.5 $\mu$ m, 0.35 $\mu$ m, 0.25 $\mu$ m, 0.18 $\mu$ m, and 0.15 $\mu$ m technology nodes.

### 1.5 BSIM4 – Aimed for 130nm Down to 20nm Nodes

BSIM continues. With the release of BSIM4 [18] in 2000 by Weidong Liu, Xiaodong Jin, Kanyu M. Cao, and Chenming Hu, BSIM was able to support the sub-130nm CMOS technologies and the growth of high-speed analog, mixed-signal, as well as radio-frequency (RF) CMOS integrated circuits that powered new lifestyle including wireless applications of the 21<sup>st</sup> century. A major new model is the physical

holistic noise model for the channel thermal noise and the induced gate noise [19, 20]. The induced gate noise and its correlation with the channel thermal noise are modeled. Accuracy at high frequencies up to the cut-off frequency of transistors is achieved with a simple intrinsic input resistance ( $R_{ii}$ ) model. A substrate resistance network was added. A novel quantum effect model called the charge layer thickness model was introduced [21]. The first gate direct-tunneling leakage current model was also introduced to anticipate the rise of gate leakage currents [22]. The pocket implant effect in advanced MOSFETs was introduced [23]. The layout-dependent effects from mechanical stress and well proximity were modeled. The modeling of high- $k$  metal-gate stacks and non-silicon materials became possible. BSIM4 has been used for the 0.13 $\mu$ m, 90nm, 65nm, 45/40nm, 32/28nm, and 22/20nm technology nodes.

## **1.6 BSIM SOI**

In parallel, the BSIM team developed a compact model for SOI-MOSFETs, BSIMSOI. Naturally, BSIM SOI shares many features and model modules from BSIM3 and BSIM4. It took major efforts to develop a floating-body model [24] as well as a self-heating model [25, 26], which required the use of a thermal sub-circuit to model the history dependencies of the underlying physical phenomena. BSIM SOI has served several major semiconductor companies for SOI-CMOS IC products.

## **1.7 Impact of BSIM**

Upon the selection of BSIM as the industry standard model in 1997, all the foundry companies quickly adopted BSIM. This led to hundreds of fabless companies to design all their products using BSIM since then. Gradually, integrated design-manufacturing (IDM) companies gave up their own proprietary compact models and migrated to BSIM. A few companies continued to use their own proprietary models but also used BSIM so as to facilitate the collaborations with other companies.

Most of the ICs using 0.5 $\mu$ m and newer technologies since 1997 were designed with BSIM. That amounts to a trillion US dollars worth of ICs.

## 1.8 Looking Towards the Future – The Multi-Gate MOSFET Model

As the CMOS technology continues to scale down, the device drain terminal is pulled closer to the middle of the channel. This increases the capacitive coupling between the drain and the channel, producing unwanted/unduly short-channel effects such as standby channel leakage currents. This is the biggest problem facing the nearly five-decade long planar CMOS transistor structure.

FinFETs [27] allow the gate to control the channel from three sides of the channel, hence the term CMG (common multiple gates), which increases the gate control and allowing the gate length to be further scaled down. Since the introduction of FinFETs, this structure has set the world's records of the smallest gate length several times at various research laboratories. The current record is held at three nanometers.

BSIM-CMG [28] is a compact model for the class of common multi-gate FET devices. This model has been extensively validated with advanced multi-gate CMOS technologies, both SOI and bulk. It is reviewed as an example of a compact model to serve multi-gate CMOS technologists and circuit designers to facilitate the transition from the planar CMOS to the multi-gate vertical CMOS era.

## 1.9 The Intent of This Book

Modeling ideal prototypical CMOS devices well is one thing, but modeling myriad real devices of numerous technology generations well is the hallmark of BSIM. BSIM4 is no exception and it is by far the most sophisticated and widely used compact MOSFET model. It has served innumerable device technologists, design automation engineers as well as IC designers around the world for half a dozen CMOS technology nodes.

This book is intended to present and analyze in depth the BSIM4 theory, hands-on techniques, and methodology that can take a compact model from its prototype into a production-worthy version. It covers MOSFET device operation and physics, manufacturing process effects, model formulations, parameter extraction, SPICE implementation, and their implications to integrated circuit design.

This book is written for BSIM users, undergraduate and graduate students in EE, and those that make their professions in those areas.

## References

- [1] Bing. J. Sheu, D. L. Scharfetter, Chenming Hu, and D. O. Pederson, “A compact IGFET charge model,” *IEEE Trans. Circuits and Systems*, vol. CAS-31, no. 8, pp. 745-748, August 1984.
- [2] Bing. J. Sheu, D. L. Scharfetter, P.-K. Ko, and M.-C. Jeng, “BSIM: Berkeley short-channel IGFET model for MOS transistors,” *IEEE Journal of Solid-State Circuits*, vol. 22, no. 4, pp. 558-566, August 1987.
- [3] M. C. Jeng, P. K. Ko, and C. Hu, “A deep submicron MOSFET model for analog/digital circuit simulations,” *Tech. Dig. of IEDM*, pp. 114-117, San Francisco, December 1988.
- [4] K. Y. Toh, P. K. Ko, and R. G. Meyer, “An engineering model for short-channel MOS devices,” *IEEE Journal of Solid-State Circuits*, vol. 23, no. 4. pp. 950-958, August 1988.
- [5] J. H. Huang, Z. H. Liu, M. C. Jeng, P. K. Ko, and C. Hu, “A physical model for MOSFET output resistance,” *Tech. Dig. of IEDM*, pp. 569-572, San Francisco, December 1992.
- [6] J. H. Huang, Z. H. Liu, M. C. Jeng, K. Hui, M. Chan, P. K. Ko, and C. Hu, “BSIM3 Manual”, University of California, Berkeley, 1993.
- [7] Z.-H. Liu, C. Hu, J.-H. Huang, T.-Y. Chan, M.-C. Jeng, P. K. Ko, and Y. C. Cheng, “Threshold voltage model for deep-submicrometer MOSFET’s,” *IEEE Trans. on Electron Devices*, vol. 40, no. 1, pp. 86-95, January 1993.
- [8] Chenming Hu, Simon C. Tam, Fu-Chieh Hsu, Ping-Keung Ko; Tung-Yi Chan; K. W. Terrill, “Hot-electron induced MOSFET degradation – Model, monitor, and improvement,” *IEEE Trans. Electron Devices*, vol. ED-32, pp. 375-385, February 1985, and *IEEE Journal Solid-State Circuits*, vol. SC-20, pp. 295-305, February 1985.

10 *BSIM4 AND MOSFET MODELING FOR IC SIMULATION*

*By Weidong Liu and Chenming Hu*

- [9] T. Y. Chan, J. Chen, P. K. Ko, and C. Hu, "The impact of gate-induced drain leakage current on MOSFET scaling," Tech. Dig. of IEDM, pp. 718-721, Washington D. C., December 1987.
- [10] K. K. Hung, P. K. Ko, C. Hu, and Y.C. Cheng, "A unified model for the flicker noise in metal-oxide-semiconductor field-effect transistors," IEEE Trans. on Electron Devices, vol. 37, no. 3, pp. 654-665, March 1990.
- [11] P. Fang, K. K. Hung, P. K. Ko, and Chenming Hu, "Characterizing a single hot-electron-induced trap in submicron MOSFET using random telegraph noise," Digest of Tech. Papers of Symp. on VLSI Technology, pp. 37-38, Honolulu, Hawaii, June 1990.
- [12] K. K. Hung, P. K. Ko, C. Hu, and Y. C. Cheng, "A physics-based MOSFET noise model for circuit simulators," IEEE Trans. Electron Devices, vol. 37, no. 4, pp. 1323-1333, May 1990.
- [13] Kai Chen, H. C. Wann, J. Dunster, P. K. Ko, Chenming Hu, and M. Yoshida, "MOSFET carrier mobility model based on gate oxide thickness, threshold and gate voltages," Solid-State Electronics, pp. 1515-1518, October 1996.
- [14] Yuhua Cheng, Kai Chen, K. Imai, and Chenming Hu, "A unified MOSFET channel charge model for device modeling in circuit simulation," IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems, pp. 641-644, August 1998.
- [15] Yuhua Cheng, Mie-Chie Jeng, Zhihong Liu, Mansun Chan, J. H. Huang, Kai Chen, P. K. Ko, and Chenming Hu, "A physical and scalable I-V model in BSIM3v3 for analog/digital circuit simulation," IEEE Trans. Electron Devices, pp. 277-287, 1997.
- [16] Yuhua Cheng and Chenming Hu, "MOSFET modeling and BSIM3 user's guide," Kluwer Academic Publishers, 1999.
- [17] Weidong Liu, Xiaodong Jin, James Chen, Min-Chie Jeng, Zhihong Liu, Yuhua Cheng, Kai Chen, Mansun Chan, Kelvin Hui, Jianhui Huang, Robert Tu, Ping K. Ko, and Chenming Hu, "BSIM3v3.2 MOSFET model — Users' manual", Memorandum No. UCB/ERL M98/51. Electronics Research Laboratory, College of Engineering, University of California, Berkeley, August 21, 1998.
- [18] Weidong Liu, Xiaodong Jin, Kanyu M. Cao, and Chenming Hu, "BSIM4.0.0 MOSFET model – User's manual," Memorandum No. UCB/ERL M00/38, Electronics Research Laboratory, College of Engineering, University of California, Berkeley. August 3, 2000.
- [19] Xiaodong Jin, Jia-Jiunn Ou, Chin-Hung Chen, Weidong Liu, M. Deen, P. R. Gray, and Chenming Hu , "An effective gate resistance model for CMOS RF and noise modeling," Tech. Dig. of IEDM, pp. 961-964, San Francisco, December 1998.
- [20] Jia-Jiunn Ou, Xiaodong Jin, Chenming Hu, and P. R. Gray, "Submicron CMOS thermal noise modeling from an RF perspective," VLSI Technology Symposium, pp.151-152, 1999.

- [21] Weidong Liu, Xiaodong Jin, Ya-Chin King, and C. Hu, "An efficient and accurate compact model for thin-oxide MOSFET intrinsic capacitance considering the finite charge layer thickness," IEEE Trans. Electron Devices, pp.1070-1072, May 1999.
- [22] Kanyu M. Cao, W. C. Lee, Weidong Liu, Xiaodong Jin, Pin Su, S. K. H. Fung, Judy X. An, B. Yu, and Chenming Hu, "BSIM4 gate leakage model including source-drain partition," Tech. Dig. of IEDM, pp. 815-818, San Francisco, December 2000.
- [23] Kanyu Mark Cao, Weidong Liu, Xiaodong Jin, Karthik Vasanth, Keith Green, John Krick, Tom Vrotsos, and Chenming Hu, "Modeling of pocket implanted MOSFETs for anomalous analog behavior," Tech. Dig. of IEDM, pp. 171-174, Washington D. C., December 1999.
- [24] Mansun Chan, Pin Su, Hui Wan, C. H. Lin, Samuel K.-H. Fung, A. M. Niknejad, Chenming Hu, and P. K. Ko, "Modeling the floating-body effects of fully depleted, partially depleted, and body-grounded SOI MOSFETs," Solid State Electronics, pp. 969-978, June 2004.
- [25] Wei Jin, Weidong Liu, S. K. H. Fung, P. C. Chan, and Chenming Hu, "SOI thermal impedance extraction methodology and its significance for circuit simulation," IEEE Trans. Electron Devices, vol. 48, no. 4, pp. 730-736, April 2001.
- [26] Hui Wan, Pin Su, Samuel K. H. Fung, A. Niknejad, and Chenming Hu, "RF modeling for FDSOI MOSFET and self heating effect on RF parameter extraction," NanoTech Workshop on Compact Modeling, 2002.
- [27] Xuejue Huang, Wen-Chin Lee, Charles Kuo, D. Hisamoto, Leland Chang, J. Kedzierski, E. Anderson, H. Takeuchi, Yang-Kyu Choi, K. Asano, K. V. Subramanian, Tsu-Jae King, J. Bokor, and Chenming Hu, "Sub-50 nm FinFET: PMOS", Tech. Dig. of IEDM, pp. 67-70, Washington D. C., December 1999.
- [28] M. V. Dunga, Chung-Hsun Lin, D. D. Lu, Weize Xiong, C. R. Cleavelin, P. Patruno, Jiunn-Ren Hwang, Fu-Liang, A. M. Niknejad, and Chenming Hu "BSIM-MG: A versatile multi-gate FET model for mixed-signal design," *VLSI Technology Symposium*, pp. 60-61, Kyoto, June 2007.

**This page intentionally left blank**

## Chapter 2

# Fundamental MOSFET Physical Effects and Their Models for BSIM4

### 2.1 Introduction and Chapter Objectives

This chapter presents and analyzes the fundamental physical effects in the modern MOSFET transistor and their mathematical formulations in BSIM4. In order to make this task easy, Section 2.1 will first introduce to the readers the basic terminologies, definitions, and geometry, and the material model options of BSIM4. The model options are intended to facilitate the customization of BSIM4 for various gate and channel geometries, model card and parameter binning, and gate and channel building materials such as high- $k$  metal gate stacks and silicon/non-silicon substrates.

Following this, in the subsequent sections of this chapter, will be the detailed presentations and discussions of the fundamental physical effects found in a modern MOSFET device structure and the modeling methodology of these physical effects in BSIM4. The topics herein include the device channel surface potential and threshold voltage, poly-silicon gate depletion, bulk-charge effects, LDD (Lightly-Doped Drain/source) resistance, inversion charge layer thickness owing to the quantum mechanical effects, carrier mobility, and layout-dependent mechanical stress and well-proximity effects.

Accurate modeling of these effects is of paramount importance for circuit simulation. This is because these effects are so fundamental that they determine how a MOSFET transistor operates. This is also because they serve as the foundation for the accurate modeling of the various device charges ( $Q$ ), currents ( $I$ ), trans-conductances ( $G$ ) and trans-capacitances ( $C$ ), which are the constituents of the circuit equations that a SPICE simulator solves. Thus, this chapter also serves to prepare the

readers to understand the presentation of the chapters to follow in this book.

Robust convergence in numerical computations is prerequisite for successful circuit simulation. Fast simulation improves the turnaround time of circuit design and tapeout. The SPICE implementation of BSIM4 pays unparalleled attention to the details and optimization of numerical robustness and computational efficiency, which leads to up to ten times speedup of a BSIM4 C-code implementation over its Verilog-A implementation. Some of these implementation techniques will be exemplified and discussed in this chapter as well as the subsequent chapters.

## 2.2 Gate and Channel Geometries and Materials

### 2.2.1 Gate and Channel Lengths and Widths

BSIM4 defines several different gate and channel lengths and widths. These parameters are needed to represent advanced CMOS process implementations and circuit design practices. The first is the drawn length ( $L_{drawn}$ ) and width ( $W_{drawn}$ ) of the gate. They are specified in the instance statement cards of SPICE MOSFET elements to designate the drawn geometries. Their values are sometimes scaled by a unit-conversion (for instance, from micrometers to meters in the MKS unit system) or a size-scaling factor. An example is the shrinking from one technology node to its half node or to the next full node of process technologies. This mechanism is used to facilitate the reuse of existing circuit netlists by applying a multiplying factor to  $L_{drawn}$  and  $W_{drawn}$  to obtain  $L_{designed}$  and  $W_{designed}$ , instead of regenerating a new netlist from scratch. These are the gate geometries that designers want to use with a particular process technology.

The actual size of the fabricated or physical gate ( $L_{physical}$  and  $W_{physical}$ ) can deviate from  $L_{designed}$  and  $W_{designed}$ , either because of unintentional process variations and/or from intentional gate size reduction performed to achieve better performance as a result of a larger device driving current. To model this important process effect within SPICE, BSIM4 introduces two new model parameters  $XL$  and  $XW$  such that the physical gate length and width become

$$L_{physical} = L_{designed} + XL \quad (2.1a)$$

and

$$W_{physical} = \frac{W_{designed}}{NF} + XW \quad (2.1b)$$

where  $NF$  (default value = 1) represents the number of fingers of a multi-finger MOSFET structure, which is preferred over a wider single finger in high-speed circuits such as RF ICs.  $XL$  and  $XW$  have default values of zero. They are extracted from silicon test transistor data and typically have negative values with a magnitude of a few nanometers resulting from intentional trimming of the gate size.

Yet another set of channel length and width are their effective values. They differ from their physical values, owing to the overlap or underlap between the gate and the source and drain diffusions ( $\Delta L$ ) and the gate control over the channel edge adjacent to the shallow trench isolation ( $\Delta W$ ). BSIM4 models these dimensions as

$$\Delta L = LINT + \frac{LL}{L_{physical}^{LLN}} + \frac{LW}{W_{physical}^{LWN}} + \frac{LWL}{L_{physical}^{LLN} \cdot W_{physical}^{LWN}} \quad (2.2a)$$

and

$$\Delta W = WINT + \frac{WL}{L_{physical}^{WLN}} + \frac{WW}{W_{physical}^{WWN}} + \frac{WWL}{L_{physical}^{WLN} \cdot W_{physical}^{WWN}} \quad (2.2b)$$

$LINT$  and  $WINT$  are the delta  $L$  and delta  $W$  of large devices, those with large  $L_{physical}$  and  $W_{physical}$ .  $LINT$  can be extracted from the measured  $V_{ds}/I_{ds}$  versus  $L_{designed}$  of a few large devices. Likewise,  $WINT$  can be extracted from measured data of  $I_{ds}/V_{ds}$  versus  $W_{designed}$ . Other parameters of the above formulas are the length and width dependence model parameters that improve the model accuracy over a wide range of the gate length and width.

The effective channel length and width used for the BSIM4 DC IV models are

$$L_{eff} = L_{physical} - 2 \cdot \Delta L \quad (2.3a)$$

and

$$W_{eff} = W_{physical} - 2 \cdot \Delta W \quad (2.3b)$$

The junction diode IV and CV and the gate-induced drain/source leakage (GIDL/GISL) current models use a different effective width because of subtle corner effects at the width-ends of the channel. It is

$$W_{effJCT} = W_{physical} - 2 \cdot \Delta W_{JCT} \quad (2.4)$$

where

$$\Delta W_{JCT} = DWJ + \frac{WLC}{L_{physical}^{WLN}} + \frac{WWC}{W_{physical}^{WWN}} + \frac{WWLC}{L_{physical}^{WLN} \cdot W_{physical}^{WWN}} \quad (2.5)$$

with **DWJ** being the overlap length in the width direction. Other parameters are the length and width dependence model parameters to improve the model scalability.

The MOSFET intrinsic and overlap capacitances are determined by the region between the source-body and drain-body junctions. BSIM4 uses yet another set of effective channel length and width for these capacitance models

$$L_{effCV} = L_{physical} - 2 \cdot \Delta L_{CV} \quad (2.5a)$$

and

$$W_{effCV} = W_{physical} - 2 \cdot \Delta W_{CV} \quad (2.5b)$$

where the gate-channel overlaps are

$$\Delta L_{CV} = DLC + \frac{LLC}{L_{physical}^{LLN}} + \frac{LWC}{W_{physical}^{LWN}} + \frac{LWLC}{L_{physical}^{LLN} \cdot W_{physical}^{LWN}} \quad (2.6a)$$

and

$$\Delta W_{CV} = DWC + \frac{WLC}{L_{physical}^{WLN}} + \frac{WWC}{W_{physical}^{WWN}} + \frac{WWLC}{L_{physical}^{WLN} \cdot W_{physical}^{WWN}} \quad (2.6b)$$

with **DLC** being the overlap length and width and other parameters being the length and width dependence model parameters.

### 2.2.2 Model Card and Parameter Binning

SPICE model parameter extraction tools can be configured to generate two different types of SPICE model parameter cards from a group of test devices with varying channel lengths and widths. They are global models and binned models.

A **global model** uses a single model card, i.e., a single set of model parameters, relying only on the model equations to accurately reproduce the nuanced electrical characteristics of transistors of any combinations of  $L_{designed}$  and  $W_{designed}$ . Building a global model requires significant extraction efforts aided by specialized software tools in order to attain satisfactory accuracies. Success depends on the quality of a model, i.e., its equation formulations and the expertise applied in parameter extractions. A global model produces smoother dependence of device behaviors on device geometries than a binned model.

In contrast, a **binned model** uses multiple, perhaps dozens model parameter cards, i.e., sets of model parameters, to model different ranges of geometries. For example, one model card is used for  $L > 1 \mu\text{m}$  and  $W > 1 \mu\text{m}$ , a second model card is used for  $1 \mu\text{m} > L > 0.3 \mu\text{m}$  and  $W > 1 \mu\text{m}$ , and a third model card is used for another range. Each of these parameter cards is called a bin. Since each model card needs to represent only a small range of  $L$  and  $W$  combination, parameter extraction can be easier for a model card but many model cards need to be generated. One shortcoming is that the binned model is not a smooth function of  $L$  and  $W$ . Moreover, the size of a model library increases with the number of bins, leading to extra overhead in library maintenance as well as repetitive parameter parsing/setup and more memory usage during SPICE simulations.

In either of these two scenarios, model card selection parameters, such as **LMIN**, **LMAX**, **WMIN** and **WMAX** in the case of BSIM4, are required to clarify the range of designed gate lengths and widths, for which the model cards provide adequate accuracy. Note that for multi-finger devices, the length and width values here refer to those of one transistor finger.

In either a global or binned model, a subset of the model parameters are formulated as simple functions of  $L_{eff}$  and  $W_{eff}$ . This practice improves model accuracy and makes parameter extraction easier, particularly in the case of generating a global model. In binned models, usually only this same subset of model parameters are binned and the rest of the model parameter are not binned. This subset is known as the binnable parameters.

Take the BSIM4 **VTH0** parameter as an example. This base parameter is given three binning parameters: **LVTH0**, **WVTH0**, and **PVTH0**. They are provided to improve the accuracy of threshold voltage fitting over all combinations of channel length and width. BSIM4 and

BSIM3v3 have used the same parameter binning scheme successfully for many technology nodes. The formulation of this scheme is simple and it computes the value of the binned  $V_{TH0}$  as

$$V_{TH0\text{binned}} = V_{TH0} + \frac{LV_{TH0}}{L_{eff}} + \frac{WV_{TH0}}{W_{eff}} + \frac{PV_{TH0}}{L_{eff} \cdot W_{eff}} \quad (2.7)$$

These binning parameters default to zero. Their dimensions are chosen such that every term in Eq. (2.7) has the same unit as that of the base parameter  $V_{TH0}$ .

In the above equation, the effective channel length and width are given in the unit of micrometers. BSIM4 provides a binning parameter value scaling selector, **BINUNIT**. When it is set to 0, Eq. (2.7) is used. When it is set to 1 (the default case), Eq. (2.7) changes to

$$V_{TH0\text{binned}} = V_{TH0} + \frac{LV_{TH0} \cdot 10^{-6}}{L_{eff}} + \frac{WV_{TH0} \cdot 10^{-6}}{W_{eff}} + \frac{PV_{TH0} \cdot 10^{-12}}{L_{eff} \cdot W_{eff}} \quad (2.8)$$

This means that device channel lengths and widths used in model card extraction are given in meters. Eqs. (2.7) and (2.8) must give the same value regardless of which unit is used in extraction and circuit net-listing. The value of **BINUNIT** (in model libraries) is set by model extraction engineers.

### **2.2.3 Gate Stack and Substrate Material Model Options**

The continual scaling of CMOS technologies needs to keep the power consumption issue contained. Power consumption includes the dynamic (proportional to  $V_{dd}$  squared) and the static (due to leakage currents through the channel as well as the tunneling current between the gate and other terminals). Power supplies ( $V_{dd}$ ) have been reduced aggressively below 1 volt. However, lower power supply voltages lead to lower transistor and circuit speed and weaken the gate control over turning on and off MOSFET devices. In fact, reduced gate voltage requires thinner gate oxide thickness (about 1 nanometer or three molecular layers of  $\text{SiO}_2$ ) in order to keep transistor current and circuit speed high. One drawback of the thin oxide is that it degrades the mobilities of channel charge carriers. As counter measures, techniques such as strained silicon and non-silicon channel material have been employed or are being developed for advanced planar CMOS process technologies. Another

drawback of thin  $\text{SiO}_2$  is excessive gate tunneling currents and static leakage. The counter measure is to employ a high- $k$  gate dielectric to increase the gate insulator thickness in order to reduce the gate tunneling current and to employ a metal gate material in order to reduce or eliminate the poly-silicon gate depletion effect. This section will discuss the BSIM4 modeling of non-Si channel and high- $k$  metal gate technologies. The modeling of the mechanical stress effect will be presented later in this chapter.

Inside an  $n^+$ -poly/oxide/p-Si (NMOS) or  $p^+$ -poly/oxide/n-Si (PMOS) transistor structure, there exist two physical effects that make the electrical gate oxide thickness ( $\text{TOXE}$ ) thicker than the physical oxide thickness,  $\text{TOXP}$ . One is the poly-Si gate depletion when the gate bias is high. The other is the significant channel charge layer thickness that is a quantum mechanical effect. Both effects will be analyzed shortly. They bring about a difference,  $\text{DTOX}$ , between  $\text{TOXE}$  and  $\text{TOXP}$ .  $\text{TOXE}$  or  $\text{DTOX}$  is bias dependent. In practice, however,  $\text{TOXE}$  is usually taken as a constant and determined from measured gate capacitance  $C_{gg}$  data at  $V_g = V_{dd}$  [1], [2].

In the BSIM4 implementation, when all these three model parameters (given in the unit of meters) are specified in model card libraries,  $\text{DTOX}$  will be ignored. If  $\text{TOXE}$  is given but not  $\text{TOXP}$ , then  $\text{TOXP}$  is computed to be  $(\text{TOXE} - \text{DTOX})$ , and vice versa. The electrical gate oxide capacitance is

$$C_{oxe} = \frac{\varepsilon_{ox}}{\text{TOXE}} = \frac{8.85418 \times 10^{-12} \cdot \text{EPSROX}}{\text{TOXE}} \quad (2.9)$$

where **EPSROX** is the model parameter for the relative dielectric constant of  $\text{SiO}_2$  and has a default value of 3.9. If the gate dielectric layer is not  $\text{SiO}_2$ , say an oxynitride  $\text{SiO}_x\text{N}_y$  layer, then **EPSROX** can be explicitly specified to be different from 3.9 and an equivalent  $\text{TOXE}$  can again be extracted from  $C_{gg}$  with the given **EPSROX**.  $C_{oxe}$  is used to compute the threshold voltage, subthreshold swing factor, mobilities, bulk-charge effect coefficient  $A_{bulk}$ , effective and smoothing gate ( $V_{gsteff}$ ) and drain ( $V_{dseff}$ ) voltages, and **CAPMOD** = 0 and 1 capacitance models of BSIM4. Similarly, the physical gate oxide capacitance is

$$C_{oxp} = \frac{\varepsilon_{ox}}{\text{TOXP}} = \frac{8.85418 \times 10^{-12} \cdot \text{EPSROX}}{\text{TOXP}} \quad (2.10)$$

It is used to compute the BSIM4 effective gate oxide capacitance  $C_{oxeff}$  for the channel current and **CAPMOD = 2** capacitance models, which take into account the finite channel charge layer thickness effect.

Note that in the case of poly-Si/oxide/Si device structures, the permittivity of the silicon substrate is  $\epsilon_{sub} = 1.03594 \cdot 10^{-10}$  Farad/meter. This quantity is used to compute other BSIM4 parameter values in the sections and chapters to follow in this book.

In order to account for all material cases other than poly-Si/oxide/Si, BSIM4 provides a new-material model selector **MTRLMOD**. It has two possible settings, 0 (the default) and 1. **MTRLMOD = 0** applies to the classical poly-Si/SiO<sub>2</sub>/Si scenario. When **MTRLMOD = 1** is chosen, device structures other than poly-Si/Oxide/Si, such as a high-*k* metal gate stack, and a non-Si substrate material, such as germanium or SiGe, can be modeled. In the latter case, one needs to extract, from measured gate capacitance  $C_{gg}$  an equivalent oxide thickness (EOT) relative to SiO<sub>2</sub> with **EPSROX = 3.9**. EOT can be specified explicitly or can be substituted for **TOXE** in BSIM4 model card libraries. EOT needs to be obtained under the inversion condition with a gate bias of **VDDEOT**, a model parameter denoting the power supply voltage designated for a particular manufacturing process. With this, the electrical gate dielectric capacitance is calculated in the same way as  $C_{oxe}$  above. The same holds true for the physical gate dielectric capacitance  $C_{oxp}$  as **MTRLMOD = 0**.

Note that in the case of **MTRLMOD = 1**, **TOXP** does not need to be specified because BSIM4 provides an auxiliary code snippet to have it computed automatically. It subtracts from **TOXE** or **EOT** the contributions from poly gate depletion and channel charge layer thickness effects. In order to do so, a few model parameters of **MTRLMOD = 1** have to be supplied to specify the condition at which **TOXP** is extracted. They are the temperature (**TEMPEOT**), the effective channel length (**LEFFEOT**) and width (**WEFFEOT**), and **VDDEOT**.

For a non-Si channel material, the substrate permittivity now becomes  $\epsilon_{sub} = (8.85418 \cdot 10^{-12} \cdot \text{EPSRSUB})$  in the unit of Farad/meter, where **EPSRSUB** is the substrate relative dielectric constant of a non-Si substrate material. **EPSRSUB** has a default value of 11.7 for a silicon substrate.

There are a few other device parameters, such as the energy-band gap,  $E_g$ , the intrinsic carrier concentration  $n_i$ , and the flat-band voltage, on which MTRLMOD has effects. They will be discussed in the ensuing sections and chapters.

## 2.3 Temperature-Dependence Model Options

BSIM4 provides four options for modeling the temperature dependencies of multiple device parameters, such as the energy-band gap ( $E_g$ ), intrinsic carrier concentration ( $n_i$ ), thermal voltage, surface potential, threshold voltage, flat-band voltage, carrier velocity and mobility, parasitic resistances, junction built-in potential, junction leakage current and capacitances. Each option is turned on by specifying a model parameter selector TEMPMOD in model cards. TEMPMOD has four possible values, which are 0 (the default), 1, 2 and 3. The temperature dependencies of these parameters will be presented throughout this book. In this section, only  $E_g$  and  $n_i$  are discussed.

For the silicon substrate (MTRLMOD = 0), the energy band gap in the unit of volt at nominal (TNOM) and operating ( $T_{emp}$ ) temperatures is

$$E_{g0}(\text{TNOM}) = 1.16 - \frac{7.02 \cdot 10^{-4} \cdot \text{TNOM}^2}{300.15 + 1108} \quad (2.11)$$

and

$$E_g(T_{emp}) = 1.16 - \frac{7.02 \cdot 10^{-4} \cdot T_{emp}^2}{T_{emp} + 1108} \quad (2.12)$$

The intrinsic carrier concentration in the unit of  $\text{cm}^{-3}$  at TNOM is

$$n_i(\text{TNOM}) = 1.45 \cdot 10^{10} \cdot \left( \frac{\text{TNOM}}{300.15} \right)^{3/2} \cdot \exp \left[ q \cdot \frac{E_g(300.15) - E_{g0}(\text{TNOM})}{2 \cdot k_B \cdot \text{TNOM}} \right] \quad (2.13)$$

where  $k_B = 1.3806226 \cdot 10^{-23}$  in the unit of  $\text{J} \cdot \text{K}^{-1}$  is the Boltzmann constant.

For a non-silicon substrate (MTRLMOD = 1), a few material parameters are introduced. The energy-band gap at nominal (TNOM) and operating ( $T_{emp}$ ) temperatures is

$$E_{g0}(\text{TNOM}) = \text{BG0SUB} - \frac{\text{TBGASUB} \cdot \text{TNOM}^2}{300.15 + \text{TBGBSUB}} \quad (2.14)$$

and

$$E_g(T_{emp}) = \text{BG0SUB} - \frac{\text{TBGASUB} \cdot T_{emp}^2}{T_{emp} + \text{TBGBSUB}} \quad (2.15)$$

The intrinsic carrier concentration at TNOM is

$$n_i(\text{TNOM}) = \text{NI0SUB} \cdot \left( \frac{\text{TNOM}}{300.15} \right)^{3/2} \cdot \exp \left[ q \cdot \frac{E_g(300.15) - E_{g0}(\text{TNOM})}{2 \cdot k_B \cdot \text{TNOM}} \right] \quad (2.16)$$

Here **BG0SUB** is the energy-band gap at 0 degree Kelvin, **TBGASUB** and **TBGBSUB** are the temperature-dependence parameters of the energy-band gap, and **NI0SUB** is the intrinsic carrier concentration at 300.15 degree Kelvin. Note that 300.15 ( $= 273.15 + 27$ ) degree Kelvin is the reference ambient temperature of SPICE3. Both **TNOM** and  $T_{emp}$  default to 300.15 degree Kelvin. Commercial SPICE simulators may have different default values for these two temperatures.

## 2.4 Threshold Voltage

The threshold voltage  $V_{th}$  of a MOS transistor is a very important and useful parameter for CMOS and BSIM4. It determines how the transistor operates and behaves. Circuit designers can choose a high-performance (HP; low  $V_{th}$ ), low-power (LP; high  $V_{th}$ ) process technology (flavor) or something inbetween that differs mostly in  $V_{th}$ . For each type of these processes, the designers are also given the choices of two or three different  $V_{th}$ 's to further optimize their chip speed and power consumption. Therefore, accurate modeling of  $V_{th}$  is a must, regardless of whatever compact modeling approach is to be taken.

BSIM4 does this by formulating the process and physical effects in a comprehensive manner. This section will present and analyze the formulations and the thinking behind them.

### 2.4.1 Long Channel with Uniform Substrate Doping

Take a large NMOS transistor with uniform channel doping **NSUB** for example. When a positive and large enough gate-to-body voltage  $V_{gb}$  is applied such that the energy-band bending in the channel surface region

reaches  $2\varphi_B$ , the surface region is depleted of holes and an inversion layer of electrons starts to form with a surface electron density equal to that of the holes found in the bulk substrate. According to the charge neutrality requirement, the charge density per unit area at any location  $y$  in the surface inversion layer is

$$q_{inv}(y) = -C_{oxe} \cdot [V_{gb} - VFB - \Phi_s(y)] - q_b(y) \quad (2.17)$$

Note that the first term bracketed without the minus sign on the right-hand side denotes the gate charge density  $q_g(y)$  and the voltage/potential terms in the square brackets give the voltage drop  $V_{ox}$  across the gate dielectric/oxide layer.  $VFB$  is the flat-band voltage.  $q_b(y)$  in the above equation is the density of the body charge in the body depletion region. It is obtained by solving the Poisson equation for the surface region

$$q_b(y) = -C_{oxe} \cdot \gamma \cdot \sqrt{\Phi_s(y)} \quad (2.17a)$$

$\gamma$  is the process technology-dependent body-bias coefficient of the body charge model

$$\gamma = \frac{\sqrt{2 \cdot \epsilon_{sub} \cdot q \cdot NSUB}}{C_{oxe}} \quad (2.17b)$$

$\Phi_s(y)$  of the above equations is the surface potential relative to the bulk. It is equal to  $[\varphi_s(y) + V_{sb}]$ , where  $\varphi_s(y)$  also denotes the surface potential, but with the source terminal voltage as the reference. In the modeling and analysis of MOSFET device operation, this is an important distinction to note. When the source-body voltage  $V_{sb}$  is zero, these two surface potentials will be the same. Fig. 2.1 gives a graphical illustration.

With these, Eq. (2.17) is transformed into the following

$$q_{inv}(y) = -C_{oxe} [V_{gs} - (VFB + 2\varphi_B) - V(y) - \gamma \sqrt{2\varphi_B - V_{bs} + V(y)}] \quad (2.18)$$

Under inversion, the surface potential is taken as

$$\Phi_s(y) = \varphi_s - V_{bs} + V(y) = 2\varphi_B - V_{bs} + V(y) \quad (2.18a)$$

Here  $\varphi_s$  takes on twice the Fermi potential, i.e.,  $2\varphi_B$ .  $\varphi_s$  and  $2\varphi_B$  will be used interchangeably in the remainder of this book. A refined definition for them is given in Eq. (2.38) after the non-uniform channel doping effect is introduced.

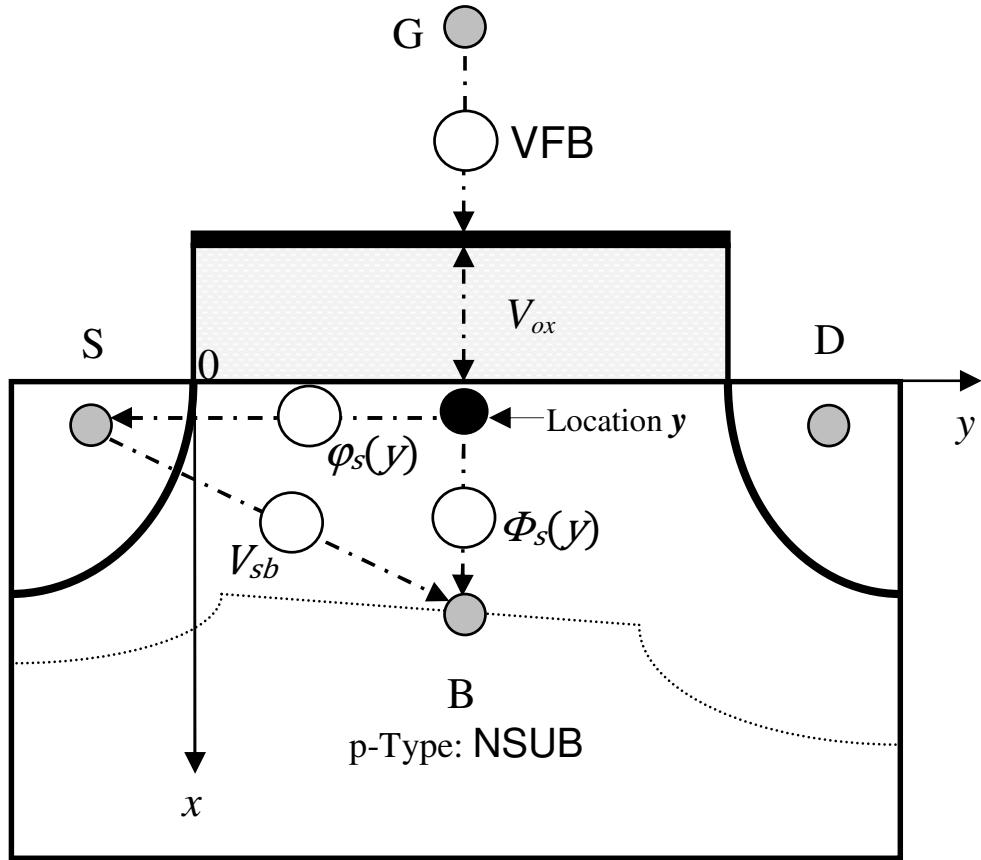


Fig. 2.1 Graphical illustration of the flat-band voltage  $V_{FB}$ , oxide voltage  $V_{ox}$  and surface potentials  $\varphi_s$  and  $\Phi_s$  in a MOSFET device structure. Open circles denote these voltages and surface potentials. Grayed circles represent the device terminals. The black dot designates the location  $y$  in the channel.

In the above equations,  $V(y)$  is the channel voltage at  $y$  with respect to the source end and ranges from 0 to the source-drain voltage  $V_{ds}$ . In order to simplify the model formulations, Taylor series expansions are performed for the square-root term of Eq. (2.18). Keeping only the first-order term yields

$$q_{inv}(y) = -C_{oxe} \cdot [V_{gs} - V_{th} - A_{bulk} \cdot V(y)] \quad (2.19)$$

where

$$V_{th} = VTH0 + \gamma \cdot (\sqrt{2\varphi_B - V_{bs}} - \sqrt{2\varphi_B}) \quad (2.19a)$$

is the threshold voltage of a long and wide-channel MOS transistor and  $VTH0$  is the long and wide-channel zero body-bias threshold voltage.  $VTH0$  is bias independent. It is extracted as a process technology model parameter

$$VTH0 = VFB + 2\varphi_B + \gamma \cdot \sqrt{2\varphi_B} \quad (2.19b)$$

$A_{bulk}$  of Eq. (2.19) is called the bulk charge coefficient. It is

$$A_{bulk} = 1 + \frac{\gamma}{2\sqrt{2\varphi_B - V_{bs}}} \quad (2.19c)$$

which is always greater than 1. The second term on the right-hand side of Eq. (2.19c) is the derivative of the bulk charge  $q_b(y)$  with respect to the channel potential  $V(y)$  at  $V_{ds} = 0$ . The bulk charge effect reduces the channel inversion charge density as a consequence of the reduced gate oxide voltage drop nearing the drain end. As will be shown later, a more accurate  $A_{bulk}$  model needs to take into account the channel length, width and gate-bias dependencies.

#### 2.4.2 Short-Channel Effect: $V_{th}$ Roll-Off and Drain Bias Effects

At a given drain-source voltage  $V_{ds}$ , the threshold voltage of a MOSFET device  $V_{th}$  decreases with decreasing channel length, a phenomenon known as  $V_{th}$  roll-off. This is caused by the increased influence of the drain voltage on the channel potential and charge density when the channel length becomes shorter. In the case of NMOS transistors, where the inverted channel carriers are electrons, the energy barrier (for electrons) will become lower at the source end. As a result, more electrons can be pulled into the channel inversion layer.

On the other hand, for a given channel length, the source-end energy barrier becomes lower as  $V_{ds}$  increases, which further reduces  $V_{th}$ . This drain bias effect is known as DIBL (drain-induced barrier lowering). The BSIM3v3 and BSIM4  $V_{th}$  roll-off and DIBL models are based upon the work published in [3] with useful modifications that improve the model accuracy and SPICE simulation robustness.

$V_{th}$  is defined as the gate-source voltage  $V_{gs}$  at which the channel surface potential  $\varphi_s$  is made equal to  $2\varphi_B$ .  $\varphi_s$  is determined by process parameters such as channel doping and gate dielectric thickness, device geometries and biases. The modeling of roll-off and DIBL effects then becomes a task of solving Poisson equation for  $\varphi_s$  in the channel depletion layer. In order to obtain a closed-form equation for compact

modeling, a quasi two-dimensional analysis is often utilized to find  $\varphi_s$  under proper boundary conditions. By applying Gauss' law, the analyses start out by relating the net electric field flux entering a finite region of the depletion layer to the charge densities in that region (often a rectangular box, known as the Gaussian box as shown in Fig. 2.2). The relation under the quasi two-dimensional approximations is

$$\varepsilon_{sub} \frac{X_{dep}}{\eta} \cdot \frac{dE_s(y)}{dy} + \varepsilon_{ox} \frac{V_{gs} - VFB - V_s(y)}{TOXE} = q \cdot NDEP \cdot X_{dep} \quad (2.20)$$

where  $X_{dep}$  is the depletion layer thickness,  $\varepsilon_{sub}$  and  $\varepsilon_{ox}$  are the permittivities of the substrate and gate dielectric layer, respectively, and NDEP is the doping concentration of the channel region (whose physical meaning is to be defined in the next subsection),  $E_s(y)$  and  $V_s(y)$  are the lateral electric field and surface potential at the interface, respectively. When the surface is depleted of mobile carriers, the depletion layer is formed and has a thickness of

$$X_{dep} = \sqrt{\frac{2\varepsilon_{sub}(\varphi_s - V_{bs})}{q \cdot NDEP \cdot 10^6}} \quad (2.21a)$$

This expression does not take into account the dependencies on the channel length and the drain bias. One may approximate this term in Eq. (2.20) with an average thickness by introducing a fitting parameter  $\eta$ . In the following, the zero-bias depletion layer thickness will be referred to frequently. It is given here

$$X_{dep0} = \sqrt{\frac{2\varepsilon_{sub}\varphi_s}{q \cdot NDEP \cdot 10^6}} \quad (2.21b)$$

The first term on the left-hand side of Eq. (2.20) represents the net field flux entering the box in the lateral direction. The second term is a familiar expression representing the net field flux entering the box from the top. Both of the lateral and vertical net fluxes are terminated inside the box by the needed amount of ionized dopants given on the right-hand side. At the bottom edge of the depletion layer, the electrical potential is fixed to the body terminal voltage and the field becomes zero.

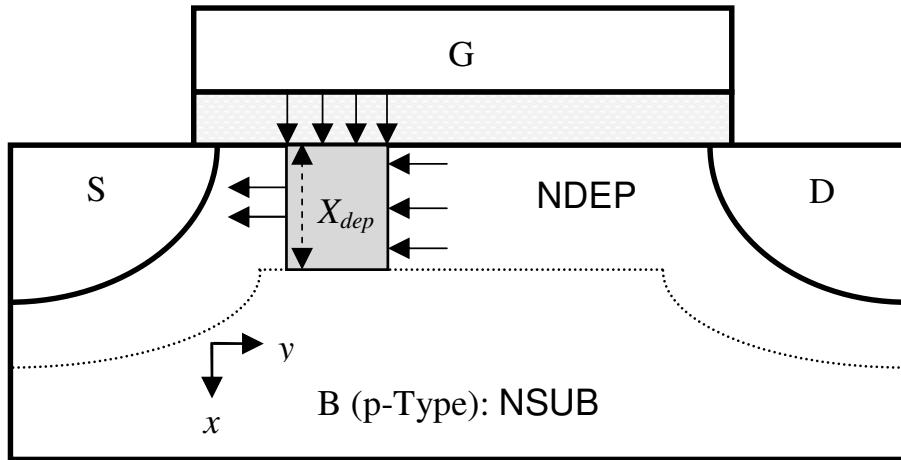


Fig. 2.2 Gaussian box in the MOSFET depletion layer as part of the quasi 2-D problem that determines the surface potential profile. The average doping concentration in the surface area is **NDEP** and that in the bulk is **NSUB**. This approximates the retrograde doping profiles found in modern CMOS process technologies, where the depletion layer hardly extends into the **NSUB** region.

From Eq. (2.20), the solution of the surface potential  $V_s(y)$  under the boundary conditions  $V_s(0) = V_{bi}$  and  $V_s(L_{eff}) = V_{bi} + V_{ds}$  is

$$V_s(y) = V_{gs} - VTH0 + \varphi_s + (V_{bi} - V_{gs} + VTH0 - \varphi_s + V_{ds}) \cdot \frac{\sinh(y/l_c)}{\sinh(L_{eff}/l_c)} + (V_{bi} - V_{gs} + VTH0 - \varphi_s) \cdot \frac{\sinh[(L_{eff}-y)/l_c]}{\sinh(L_{eff}/l_c)} \quad (2.22)$$

$VTH0$  is the long-channel zero body-bias threshold voltage.  $l_c$  is the characteristic length of MOSFET short-channel effects (SCE). It is proportional to the square root of the product of the gate oxide thickness and the depletion layer thickness. The smaller  $l_c$  is, the less the device is prone to SCE. Therefore, thinner gate oxides and higher channel dopings are desired for reducing SCE. In the surface potential model above,  $V_{bi}$  is the built-in potential of the source-body and drain-body junctions

$$V_{bi} = \frac{k_B \cdot TNOM}{q} \cdot \log\left(\frac{NSD \cdot NDEP}{n_i^2}\right) \quad (2.23)$$

where  $n_i$  is the intrinsic carrier concentration at temperature  $TNOM$ .

In order to obtain the threshold voltage reduction caused by the roll-off ( $L$  dependence) and DIBL ( $V_{ds}$  and  $L$  dependence) effects, one needs

to find the minimum value of the surface potential along the channel, which is

$$V_{smin} = V_{gs} - VTH0 + \varphi_s + \left[ 2 \cdot (V_{bi} - V_{gs} + VTH0 - \varphi_s) + V_{ds} \right] \cdot \frac{\sinh(L_{eff}/2l_c)}{\sinh(L_{eff}/l_c)} \quad (2.24)$$

Equating  $V_{smin}$  and  $2\varphi_B$  and letting  $V_{th} = V_{gs}$  yield

$$V_{th} = VTH0 - \left[ \frac{V_{bi} - \varphi_s}{\cosh(L_{eff}/2l_c) - 1} + \frac{0.5 \cdot V_{ds}}{\cosh(L_{eff}/2l_c) - 1} \right] \equiv VTH0 - \Delta V_{th} \quad (2.25)$$

The first term in the brackets represents a  $V_{th}$  reduction  $\Delta V_{th}$ (roll-off) attributed to short-channel  $V_{th}$  roll-offs, whereas the second represents the reduction owing to DIBL, denoted by  $\Delta V_{th}$ (DIBL). [Note that in a coincidence, these powerful terms were also illustrated in Professor Chih-Tang Sah's 1991 book *Fundamentals of Solid-State Electronics – Study Guide*. See the two energy-band diagrams given in Figs. B1.1 and B1.2 on pages 386 and 387, respectively, of that book.]

In BSIM4,  $\Delta V_{th}$ (roll-off) and  $\Delta V_{th}$ (DIBL) are enhanced with the body-bias dependencies and by incorporating more parameters for ease of parameter extractions and better accuracy of the channel-length dependencies. The enhanced version of  $\Delta V_{th}$ (roll-off) is

$$\Delta V_{th}(\text{roll-off}) = \left[ \frac{0.5 \cdot DVT0}{\cosh(DVT1 \cdot L_{eff}/l_{c1}) - 1} \right] \cdot (V_{bi} - \varphi_s) \quad (2.26)$$

where DVT0 and DVT1 are the model parameters for the  $V_{th}$  roll-off dependencies on the channel length. For uniform channel doping, the body-bias effects on the characteristic length are partially determined by the depletion layer thickness  $X_{dep}$ . For MOSFETs with non-uniform channel doping,  $X_{dep}$  is no longer a simple square-root function of ( $\varphi_s$  -  $V_{bs}$ ). The characteristic length is found to have a different body-bias dependence than the uniform doping case. As the reverse body bias increases, the characteristic length becomes larger and the device has stronger short-channel effects. In Eq. (2.26), the characteristic length is modeled

$$l_{c1} = \sqrt{\frac{\varepsilon_{sub} \cdot TOXE \cdot X_{dep}}{\varepsilon_{ox}}} \cdot (1 + DVT2 \cdot V_{bseff}) \quad (2.26a)$$

where  $DVT2$  is the body-bias coefficient parameter.

The enhanced version of  $\Delta V_{th}$ (DIBL) of BSIM4 is

$$\Delta V_{th}(\text{DIBL}) = \frac{0.5}{\cosh(DSUB \cdot L_{eff}/l_{c0}) - 1} \cdot (\text{ETA0} + \text{ETAB} \cdot V_{bseff}) \cdot V_{ds} \quad (2.27)$$

where  $\text{ETA0}$  is the DIBL coefficient parameter,  $\text{DSUB}$  models the length dependencies of DIBL and  $\text{ETAB}$  is a parameter to model the body-bias dependence. The characteristic length in  $\Delta V_{th}$ (DIBL) is found to be bias independent

$$l_{c0} = \sqrt{\frac{\varepsilon_{sub} \cdot TOXE \cdot X_{dep0}}{\varepsilon_{ox}}} \quad (2.27a)$$

It should be noted that when  $V_{ds}$  is large, the DIBL-induced  $V_{th}$  reduction will no longer be the linear function of  $V_{ds}$ . It is actually proportional to

$$\Delta V_{th}(\text{DIBL}) \sim (A \cdot V_{ds} + B \cdot \sqrt{V_{ds}}) \quad (2.27b)$$

This deviation comes from the assumption used in deriving Eq. (2.25) that when  $V_{ds}$  is small, the minimum surface potential  $V_{smin}$  is located at  $y = 0.5L_{eff}$  [3]. In fact, as  $V_{ds}$  increases,  $V_{smin}$  will move closer to the source end of the channel and the barrier lowering will be less affected by the drain bias. The coefficient  $B$  above is negative.

In SPICE implementation, BSIM4 uses many useful numerical techniques, including smoothing and limiting functions, elimination of potential numerical round-off and divide-by-zero errors, and use of computationally efficient mathematical functions and floating-point operations to improve the model evaluation performance and convergence robustness. Some of these techniques are applied to the  $V_{th}$  model implementation as well. They will be discussed below and throughout the book.

For old process technologies where the ratio of the channel length to the characteristic length is usually large, the  $\cosh(x)$  term in the above formulations can be approximated by

$$\frac{1}{\cosh(x)-1} \approx 2 \cdot (e^{-x} + 2 \cdot e^{-2x}) \quad (2.28)$$

This approximation is used in the BSIM3v3  $V_{th}$  model [4], [5]. The right-hand side has a finite value even at  $L_{eff} = 0$ . As will be shown shortly, the  $V_{th}$  roll-up term attributed to pocket implants will be a very large number for a very-short-channel device. Therefore, in some occasions where model parameters are not properly extracted, the computed threshold voltage using BSIM3v3 might roll up again after the roll-off at some channel length less than the minimum length LMIN, for which the model card should be used. BSIM4 pays particular attention to such details. It uses the  $\cosh(x)$  term with no approximations to eliminate the phantom deficiencies of this type.

$\cosh(x)$  is a well-behaved function for compact modeling. It has a minimum value of 1 and is exponentially symmetrical about  $x = 0$ . However, the left-hand side of Eq. (2.28) will lead to a divide-by-zero error when  $x = 0$ . To prevent this from happening, it is replaced by the following treatment in the BSIM4 SPICE implementation

$$\frac{1}{\cosh(x)-1} \approx \frac{2}{e^x + e^{-x} - 2 + MIN\_EXP} \quad (2.29)$$

where  $MIN\_EXP$  is set to the value of exponential function  $\exp(-34) = 1.713908431 \cdot 10^{-15}$ . Furthermore, in order to prevent numerical over- and under-flow of the  $\exp(x)$  function, the maximum and minimum function value is clamped at  $5.834617425 \cdot 10^{+14}$  and  $1.713908431 \cdot 10^{-15}$  when  $x$  moves outside the range of [-34, 34]. The constant 34 has proven to be an adequate and safe choice for possible  $V_{dd}$  and device geometries. This can also reduce model evaluation CPU time by avoiding unnecessary computations of expensive functions such as  $\exp(x)$  in SPICE iterations. BSIM4 also uses this technique in the computation of any power functions, which is always transformed into the less expensive  $\log(x)$  and  $\exp(x)$  functions as

$$\text{pow}(y, x) = y^x = \exp[x \cdot \log(y)] \quad (2.30)$$

Linear functions are simple to use in modeling. However, they often lead to inaccuracy or even unreasonable results when used outside the region they are intended for. Take the linear term of the characteristic length  $l_{c1}$  as example. Let it be denoted by  $f(x) = 1 + x$  where  $x = DVT2 \cdot V_{bseff}$ .  $f(x)$  does its job as expected in the reverse body-bias region (i.e.,  $x > 0$ ) but it does not otherwise. For instance,  $l_{c1}$  can become negative when  $x < -1$ . To avoid this, BSIM4 uses rational polynomials

$g(x)$  to replace  $f(x)$  beyond certain  $x_0$ , where the linear function is not acceptable. Rational polynomials, if defined properly, can have continuous derivatives and approach a physically correct asymptotic value smoothly. Continuity of derivatives is an important property for iterative algorithms. The choice of  $x_0$  should be made based upon device physics. In the case of  $f(x) = 1 + \text{DVT2} \cdot V_{bseff}$ ,  $x_0 = -0.5$  is a good choice.  $g(x)$  can have the form

$$g(x) = \frac{1+a \cdot x}{b+c \cdot x} \quad (2.31)$$

with  $a$ ,  $b$ , and  $c$  being the coefficients that are determined by the criterion that  $g(x)$  is continuous up to its first-order derivative with  $f(x)$  at  $x_0$ . Equalizing the second-order derivatives of  $g(x)$  and  $f(x)$  is actually often unnecessary. Multiple sets of values for  $a$ ,  $b$ , and  $c$  are sometimes able to meet the criterion. Choose one simple set. In the case of  $x = \text{DVT2} \cdot V_{bseff}$ , that simple set is (3, 3, 8) for  $a$ ,  $b$  and  $c$ , respectively. It is observed that as  $x$  approaches infinity,  $g(x)$  smoothly approaches its lower limit of 3/8, rather than an unphysical large negative number as  $f(x)$  otherwise would lead to.

### 2.4.3 Narrow-Width Effects

For narrow-width devices, the preceding quasi 2-D analyses along the depth and channel length direction become less accurate. This is because the effect of the fringing fields along the two channel edges (in the length direction) becomes more appreciable. The smaller the channel length is, the more significant the  $V_{th}$  change due to narrow width will be. For older CMOS process technologies where a field oxide much thicker than the gate oxide is employed for device isolation, the threshold voltage increases with decreasing channel width owing to the well-known “LOCOS bird’s beak” effect (LOCOS stands for LOCal Oxidation of Silicon). For advanced technologies where the channel is surrounded and confined by shallow-trench isolation (STI) oxide, the threshold voltage actually decreases for narrower device width.

In BSIM4, the narrow-width induced threshold voltage change is modeled in two terms for today’s device isolation technology:

$$\Delta V_{th}(\text{narrow width}) = (K3 + K3B \cdot V_{bseff}) \cdot \frac{TOXE}{W0 + W_{eff}} \cdot \varphi_s - \left[ \frac{0.5 \cdot DVT0W}{\cosh(DVT1W \cdot L_{eff} \cdot W_{eff} / l_{cw}) - 1} \right] \cdot (V_{bi} - \varphi_s) \quad (2.32)$$

where  $K3$ ,  $K3B$  and  $W0$  model the width effect that does not depend on the channel length.  $DVT0W$  and  $DVT1W$  are used to model the channel length dependencies for narrower devices. Like the analyses conducted in the previous section, the characteristic length of the narrow-width effect model is also found to have additional body-bias dependence than that of uniform channel dopings. A separate characteristic length  $l_{cw}$  is allowed for the modeling of the narrow-width effects

$$l_{cw} = \sqrt{\frac{\varepsilon_{sub} \cdot TOXE \cdot X_{dep}}{\varepsilon_{ox}}} \cdot (1 + DVT2W \cdot V_{bseff}) \quad (2.32a)$$

where  $DVT2W$  is the body-bias coefficient parameter.

#### 2.4.4 Non-Uniform Substrate Doping

In the preceding sections, uniform substrate doping was assumed in the derivation of the threshold voltage model. Low channel dopings produce large depletion layer thicknesses ( $X_{dep}$ ) and large characteristic lengths ( $l_c$ ). That leads to strong drain-induced barrier lowering. It gives rise to strong short-channel effects, which includes  $V_{th}$  roll-off (reduction) at short channel lengths, increasing with increasing drain voltage.

Increasing the channel doping density everywhere uniformly reduces electron and hole mobility and therefore the driving current. Two techniques are widely used to overcome this shortcoming. One is adding pocket (or halo) implants at the source and drain corners of a dopant type opposite to that of the source and drain. The other technique is to increase the channel doping but not in the region near the silicon/oxide interface with a “steep retrograde” doping profile. Both techniques are illustrated in Fig. 2.3. The steep retrograde doping can reduce  $X_{dep}$  and therefore the characteristic length  $l_c$ . The lightly doped region at the interface helps to increase carrier mobility through reduction of charge scattering by ionized dopants.

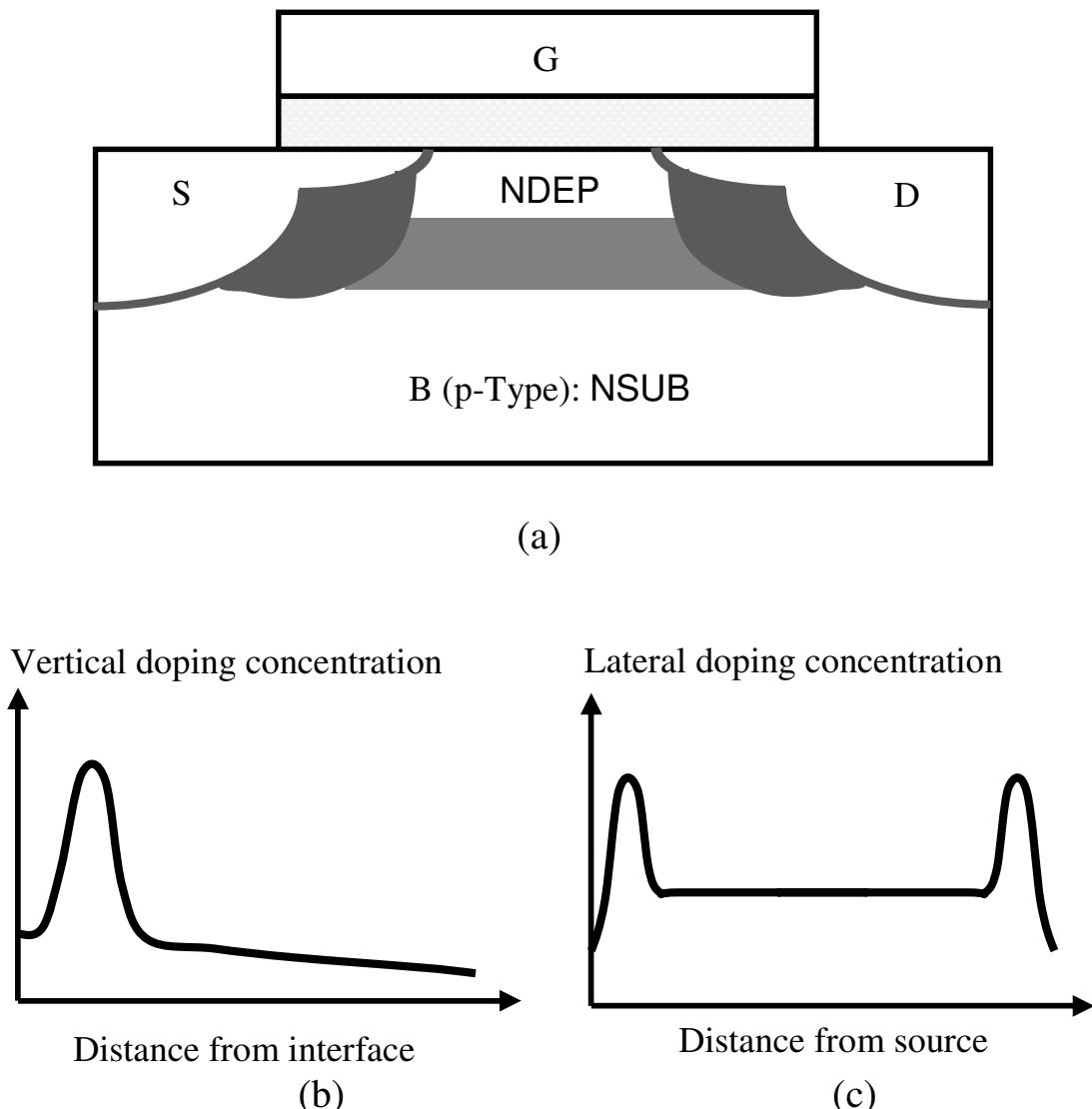


Fig. 2.3 Shown in (a) is the device structure with non-uniform channel doping: Pocket implants at the source and drain corners represented by the dark regions and steep retrograde or delta channel doping marked in gray. Both are effective for suppressing the short-channel effects. (b) and (c) are the doping profiles along the vertical and lateral directions, respectively.

#### 2.4.4.1 Non-Uniform Vertical Doping

It was shown that in the case of uniform substrate doping NSUB, the threshold voltage of a large device is

$$V_{th} = V_{TH0} + \gamma \cdot (\sqrt{\varphi_s - V_{bs}} - \sqrt{\varphi_s}) \quad (2.33)$$

This formulation needs to be revised for a non-uniform vertical doping profile such as that shown in Fig. 2.3(b). The body-bias coefficient  $\gamma$  and the inversion surface potential  $\varphi_s = 2\varphi_B$  need to be revised, too. Define an effective channel doping concentration NDEP that can produce the same  $X_{dep0}$  as that from a non-uniform doping under zero body bias.  $V_{th}$  is now re-written

$$V_{th} = VTH0 + \frac{q \cdot D_0}{C_{oxe}} + K1 \cdot \left( \sqrt{\varphi_s - \frac{q \cdot D_1}{\varepsilon_{sub}} - V_{bs}} - \sqrt{\varphi_s - \frac{q \cdot D_1}{\varepsilon_{sub}}} \right) \quad (2.34)$$

To take into account the various possible channel materials and doping concentrations, BSIM4 changes the inversion channel surface potential from the traditional  $\varphi_s = 2\varphi_B$  to

$$\varphi_s = 0.4 + \frac{k_B \cdot TNOM}{q} \cdot \log \left( \frac{NDEP}{n_i} \right) \quad (2.35)$$

In Eq. (2.34),  $D_0$  and  $D_1$  represent the 0<sup>th</sup> and 1<sup>st</sup> moments of the vertical doping profile, respectively. They are

$$D_0 = \int_0^{X_{dep0}} [N(x) - NDEP] \cdot dx + \int_{X_{dep0}}^{X_{dep}} [N(x) - NDEP] \cdot dx \quad (2.36a)$$

and

$$D_1 = \int_0^{X_{dep0}} [N(x) - NDEP] \cdot x \cdot dx + \int_{X_{dep0}}^{X_{dep}} [N(x) - NDEP] \cdot x \cdot dx \quad (2.36b)$$

where  $N(x)$  is the doping profile along the depth direction, perpendicular to the channel. The first term of Eq. (2.36a) is constant and only dependant on process conditions. Hence, it should be included in  $VTH0$  of Eq. (2.34), which is extracted from measured data. Assume a steep delta doping and hence  $X_{dep}$  becomes less sensitive to  $V_{bs}$  and is close to  $X_{dep0}$ . A Taylor series expansion of  $X_{dep}$  at  $V_{bs} = 0$  produces approximately a linear dependence of the second term of Eq. (2.36a) on  $V_{bs}$ , which, upon substitution into Eq. (2.34), gives a linear  $V_{bs}$  term, designated as  $(-K2 \cdot V_{bs})$ .

In Eq. (2.36b), the first term is also constant and much larger than the second term, which is thus neglected. Substituting this equation into Eq. (2.34) yields the new  $V_{th}$  formulation for non-uniform vertical channel doping

$$V_{th} = VTH0 + K1 \cdot (\sqrt{\varphi_s - V_{bs}} - \sqrt{\varphi_s}) - K2 \cdot V_{bs} \quad (2.37)$$

The VTH0 formulation is revised to

$$VTH0 = \varphi_s + VFB + K1 \cdot \sqrt{\varphi_s} \quad (2.37a)$$

and the inversion surface potential is now changed to

$$\varphi_s = 0.4 + \frac{k_B \cdot TNOM}{q} \cdot \log\left(\frac{NDEP}{n_i}\right) + PHIN \quad (2.38)$$

PHIN is a model parameter with a default value of zero.

In SPICE implementation, if either VTH0 or VFB is specified, the other will be computed from the one that is specified according to Eq. (2.37a). If neither is given, VFB will first be obtained as follows (and then VTH0 with Eq. (2.37a)). For a poly-Si/SiO<sub>2</sub>/Si structure, VFB is set to -1V. For a high- $k$  metal gate device, it is obtained from

$$VFB = PHIG - EASUB - 0.5 \cdot E_{g0}(TNOM) + \frac{k_B \cdot TNOM}{q} \cdot \ln\left(\frac{NSD}{n_i}\right) \quad (2.39)$$

Here, PHIG and EASUB are model parameters denoting the gate work function and the electron affinity of the substrate, and  $E_{g0}$  is the substrate energy band gap at TNOM.

In the case of K1 and K2, K1 is set to  $0.53V^{-1/2}$  if only K2 is given. K2 is set to -0.0186 if only K1 is given. If neither is specified, they are computed as follows.

Define a few process parameters. XT is the depth of the delta doping; VBX is the body bias at which the depletion width is equal to XT; VBM is the maximum applicable body voltage, and GAMMA1 and GAMMA2 are the body-bias coefficients corresponding to NDEP and NSUB. If GAMMA1, GAMMA2 and VBX are not specified, they are computed

$$GAMMA1 = \frac{\sqrt{2 \cdot \epsilon_{sub} \cdot q \cdot NDEP}}{C_{oxe}} \quad (2.40)$$

and

$$GAMMA2 = \frac{\sqrt{2 \cdot \epsilon_{sub} \cdot q \cdot NSUB}}{C_{oxe}} \quad (2.41)$$

By solving the Poisson equation,  $V_{BX}$  is obtained for a given  $XT$

$$V_{BX} = \varphi_s + \frac{q \cdot NDEP \cdot XT^2}{2 \cdot \epsilon_{sub}} \quad (2.42)$$

$K_1$  and  $K_2$  are now ready to be computed as

$$K_2 = \frac{(GAMMA1 - GAMMA2) \cdot (\sqrt{\varphi_s - V_{BX}} - \sqrt{\varphi_s})}{2 \cdot \sqrt{\varphi_s} \cdot (\sqrt{\varphi_s - V_{BM}} - \sqrt{\varphi_s}) + V_{BM}} \quad (2.43)$$

and

$$K_1 = GAMMA2 - 2 \cdot K_2 \cdot \sqrt{\varphi_s - V_{BM}} \quad (2.44)$$

Use of these process parameters is beneficial to the predictive ability of BSIM4. IC designers need to have access to the SPICE models of the next technology node for early designs even before Si data is available. Device engineers can generate these pre-Si model cards from the existing model card of the previous technology node by revising the process-related model parameters according to the process specifications or TCAD simulations. This has been a distinctive and significant use model of BSIM4. Moreover, being predictive makes statistical and parametric variability modeling easier and more accurate.

#### 2.4.4.2 Non-Uniform Lateral Doping: Pocket Implants

Suppose the doping concentration and the length (in the channel direction) of the source and drain pocket implants are  $N_{poc}$  and  $L_{poc}$ , respectively. The equivalent channel doping concentration is increased by a factor of

$$\frac{LPE0}{L_{eff}} = \frac{2L_{poc} \cdot (N_{poc} - NDEP)}{L_{eff} \cdot NDEP} \quad (2.45)$$

It produces a  $V_{th}$  roll-up prior to its roll-off as the channel length decreases. In addition, the pocket implants change the body-bias effect by a similar amount as given by Eq. (2.45). By combining these two effects, the change in the threshold voltage of BSIM4 is

$$\Delta V_{th}(\text{roll-up}) = (K1_{ox}\sqrt{\varphi_s - V_{bseff}} - K1\sqrt{\varphi_s}) \cdot \sqrt{1 + \frac{LPEB}{L_{eff}}} + K1_{ox} \cdot \left( \sqrt{1 + \frac{LPE0}{L_{eff}}} - 1 \right) \cdot \sqrt{\varphi_s} \quad (2.46)$$

where LPE0 and LPEB are model parameters. They represent the increases in the effective channel doping concentration and the body-bias effect. Both are given in the unit of meters after the pocket dopings are normalized by the channel doping NDEP.

Note that in Eq. (2.46), the original body-bias coefficient K1 has been replaced with  $K1_{ox}$ . This is intended to improve the model scalability over the gate dielectric thickness. According to the equation of the body-bias coefficient, K1 is dependent upon TOXE. The actual TOXE of a device may deviate from the target oxide thickness TOXM, at which parameters such as K1 are extracted. For this reason,  $K1_{ox}$  is given by

$$K1_{ox} = K1 \cdot \frac{TOXE}{TOXM} \quad (2.46a)$$

The same applies to the model parameter K2. Note also that K1 on the right-hand side of Eq. (2.46) is unchanged because VTH0 has the same term and they cancel out each other anyway.

According to the P-N junction energy-band diagram, the pocket implants introduces two energy barriers (potential valleys) located at the source and drain ends as shown in Fig. 2.4 [6]. In the situation where the barrier is high enough, the transistor may not be turned on even if the inversion layer has already formed in the main channel region between the valleys. The picture here is very different from that of a non-pocket-implanted device. This is because the inversion layer is not present at the source and the drain p/n junction barriers. No significant channel current flows from source to drain until the two energy barriers are reduced sufficiently by the gate voltage or by the drain voltage. Since the main channel region between the valleys is inverted and conductive, the drain voltage can have significant impact on the source energy barrier height, even at long channel lengths. In other words, the DIBL effect does not drop off rapidly with increasing channel length as in non-pocket-implanted MOSFETs. Therefore, the threshold voltage of a pocket-implanted MOSFET device is significantly affected by  $V_{ds}$ , even when the channel length is very large. This is illustrated in Fig. 2.5. This effect was first modeled and named as DITS (drain-induced threshold shifts) by

BSIM4 [6]. DITS is a very significant quantity for analog and RF circuits, where longer channel devices are employed and an accurate threshold voltage model is required. In addition, the pocket implants and DITS reduce the output resistance  $R_{out}$  of a device appreciably. This will be discussed in Chapter 3. The DITS effect needs to be distinguished from the DIBL effect that has already been presented in this section. DIBL is significant only when the channel length is small. In BSIM4, DITS has the following channel-length and  $V_{ds}$  dependences.

$$\Delta V_{th}(\text{DITS}) = \frac{n \cdot k_B \cdot T_{\text{NOM}}}{q} \cdot \log \left( \frac{L_{eff}}{L_{eff} + DVTP0 \cdot [1 + \exp(-DVTP1 \cdot V_{ds})]} \right) \quad (2.47)$$

for  $\text{TEMPMOD} = 2$  and 3.  $T_{\text{NOM}}$  will be replaced with the circuit operating temperature  $T_{emp}$  if  $\text{TEMPMOD} = 0$  or 1.  $DVTP0$  is the length-dependence parameter and  $DVTP1$  is the drain-bias dependence parameter. Eq. (2.47) will be skipped and no DITS will be computed in BSIM4 if the binned value of  $DVTP0$  is less than or equal to zero.  $DVTP0$  and  $DVTP1$  are extracted from measured threshold voltages of an array of short-to-long devices under various drain biases.

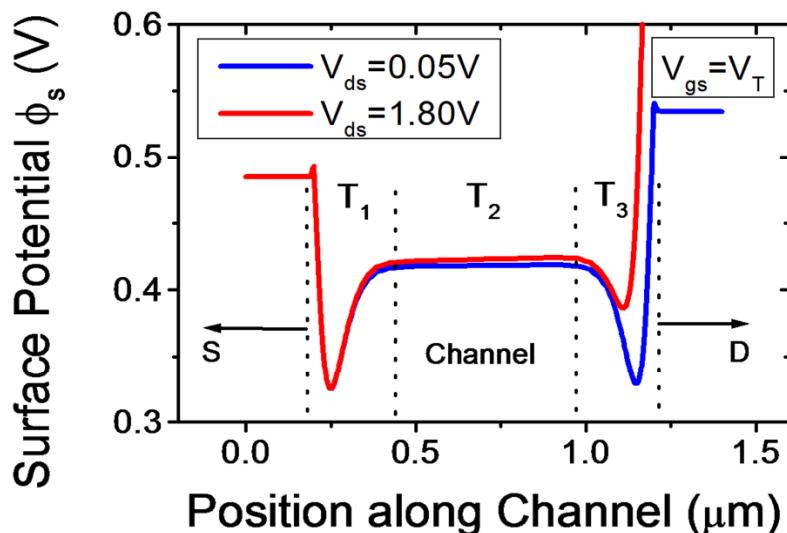


Fig. 2.4 Surface potential distribution along the channel and two potential valleys at source and drain from 2-D simulations of a pocket-implanted device. The gate bias is chosen at  $V_{th}$ . Two different drain biases are applied.

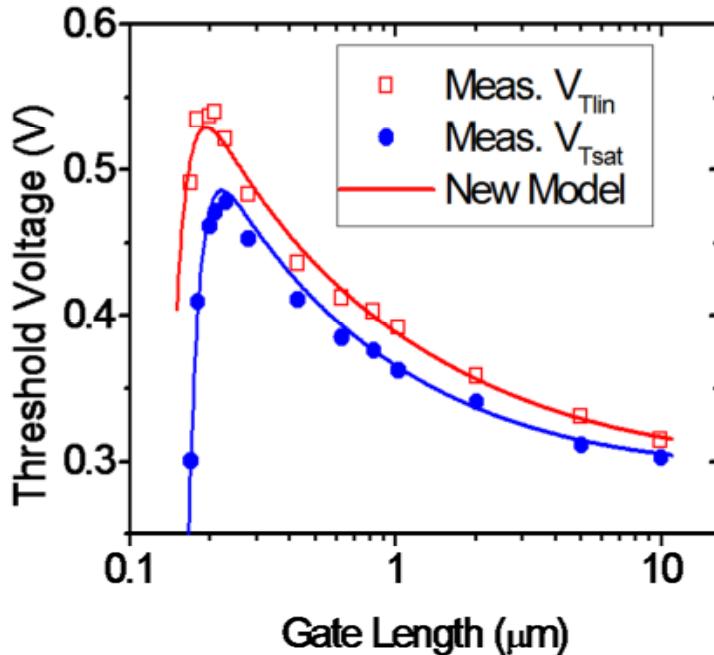


Fig. 2.5 Illustrations of DITS (drain-induced threshold shifts) of pocket-implanted devices with various channel lengths up to  $10\mu\text{m}$  under  $V_{ds} = 50\text{mV}$  and  $1.8\text{V}$ . New Model in the plot refers to the BSIM4 DITS model.

#### 2.4.5 $V_{th}$ Temperature Dependence

MOSFET threshold voltage  $V_{th}$  has a negative temperature dependence. That is, the lower the operating temperature  $T_{emp}$ , the larger the  $V_{th}$ . Recall that  $V_{th}$  is expressed as a function of the surface and Fermi potential, intrinsic carrier concentration, energy-band gap, and the built-in junction potential. They, in addition to the dopant ionization, are all temperature dependant. One possible approach of modeling the  $V_{th}$  temperature dependence is to model the temperature dependencies for all these quantities individually. Unfortunately, this would be a quite complex model, and a lot of Si measurement and model parameter extraction effort would be required for these quantities for various temperatures  $T_{emp}$ . In addition, it would be inefficient in SPICE simulation, because repeated computation of these quantities and, their temperature derivatives if required to have for self-heating modeling, would be needed.

The approach that BSIM3 and BSIM4 use is simple and accurate:  $V_{th}$  as a function of channel lengths and biases is first measured and modeled at the reference nominal temperature  $T_{NOM}$ . Then,  $V_{th}$  is adjusted as a function of  $(T_{emp} - T_{NOM})$ . Given this, an efficient yet accurate temperature dependence model for the threshold voltage is

$$\Delta V_{th}(T_{emp}) = \left( KT1 + \frac{KT1L}{L_{eff}} + KT2 \cdot V_{bseff} \right) \cdot \left( \frac{T_{emp}}{T_{NOM}} - 1 \right) \quad (2.48)$$

This delta term is added to the threshold voltage at the nominal temperature  $T_{NOM}$ .  $KT1$ ,  $KT1L$  and  $KT2$  are model parameters. The first term on the right-hand side is expected to be negative.

#### 2.4.6 BSIM4 $V_{th}$ Equation

By combining all the  $V_{th}$  terms developed, the complete  $V_{th}$  model equation of BSIM4 is

$$\begin{aligned} V_{th} = & V_{TH0} + \left( K1_{ox} \cdot \sqrt{\varphi_s - V_{bseff}} - K1 \cdot \sqrt{\varphi_s} \right) \cdot \sqrt{1 + \frac{L_{PEB}}{L_{eff}}} - K2_{ox} \cdot \\ & V_{bseff} + K1_{ox} \cdot \left( \sqrt{1 + \frac{L_{PE0}}{L_{eff}}} - 1 \right) \cdot \sqrt{\varphi_s} + \left( K3 + K3B \cdot V_{bseff} \right) \cdot \\ & \frac{TOXE}{W_0 + W_{eff}} \cdot \varphi_s - \\ & 0.5 \cdot \left[ \frac{DVT0W}{\cosh(DVT1W \cdot L_{eff} \cdot W_{eff} / l_{cw}) - 1} + \frac{DVT0}{\cosh(DVT1 \cdot L_{eff} / l_{c1}) - 1} \right] \cdot (V_{bi} - \varphi_s) - \\ & \frac{0.5}{\cosh(DSUB \cdot L_{eff} / l_{c0}) - 1} \cdot (\text{ETA0} + \text{ETAB} \cdot V_{bseff}) \cdot V_{ds} - \frac{n \cdot k_B \cdot T}{q} \cdot \\ & \log \left( \frac{L_{eff}}{L_{eff} + DVTP0 \cdot [1 + \exp(-DVTP1 \cdot V_{ds})]} \right) + \left( KT1 + \frac{KT1L}{L_{eff}} + KT2 \cdot V_{bseff} \right) \cdot \\ & \left( \frac{T_{emp}}{T_{NOM}} - 1 \right) \quad (2.49) \end{aligned}$$

Note that the temperature  $T$  of the drain-induced threshold voltage shift term is replaced with  $T_{emp}$  for TEMPMOD = 0 or 1 and with  $T_{NOM}$  for TEMPMOD = 2 or 3.

The body-bias voltage  $V_{bs}$  in the above equations has been replaced by the effective body bias voltage  $V_{bseff}$ . The same has been applied to other model formulations in the remainder of this chapter, especially to the intrinsic DC and capacitance models, unless otherwise noted (for instance, the junction diode IV and CV models will still use the actual body voltage). For a forward junction bias, when  $V_{bs} > (0.95 \cdot \varphi_s)$ ,  $V_{bseff}$  is pinned at  $(0.95 \cdot \varphi_s)$ . For a reverse operation, when  $V_{bs} < V_{bsc}$ ,  $V_{bseff}$  is clamped to  $V_{bsc}$ . This is illustrated in Fig. 2.6.

Equation (2.50) provides limiting of  $V_{bseff}^*$  at  $V_{bsc}$  for the reverse operation.

$$V_{bseff}^* = V_{bsc} + \frac{1}{2} \cdot \left[ V_{bs} - V_{bsc} - \delta_1 + \sqrt{(V_{bs} - V_{bsc} - \delta_1)^2 - 4 \cdot V_{bsc} \cdot \delta_1} \right] \quad (2.50)$$

with  $\delta_1 = 0.001\text{V}$ . The effective body bias  $V_{bseff}$  that is also limited at  $(0.95 \cdot \varphi_s)$  in the junction forward bias scenario is

$$V_{bseff} = 0.95 \cdot \varphi_s - \frac{1}{2} \cdot \left[ 0.95 \cdot \varphi_s - V_{bseff}^* - \delta_1 + \sqrt{(0.95 \cdot \varphi_s - V_{bseff}^* - \delta_1)^2 + 4 \cdot \delta_1 \cdot 0.95 \cdot \varphi_s} \right] \quad (2.51)$$

Here  $V_{bsc}$  is computed from the following equation

$$V_{bsc} = 0.9 \cdot \left( \varphi_s - \frac{\kappa_1^2}{4 \cdot \kappa_2^2} \right) \quad (2.52)$$

If the calculated  $V_{bsc}$  is greater than  $-3\text{V}$ , it is set to  $-3\text{V}$ . If it is less than  $-30\text{V}$ , it is set to  $-30\text{V}$ . If the body bias parameter  $\kappa_2$  is greater than or equal to 0,  $V_{bsc}$  will not be computed but it will be set to  $-30\text{V}$  directly. In any possible situation where  $V_{bsc}$  should become greater than  $\text{VBM}$  (the maximum body-bias voltage, which defaults to  $-3\text{V}$ ), it will be set to  $\text{VBM}$ .

The effective body-bias voltage makes model formulations behave more robustly and SPICE simulation converge faster.

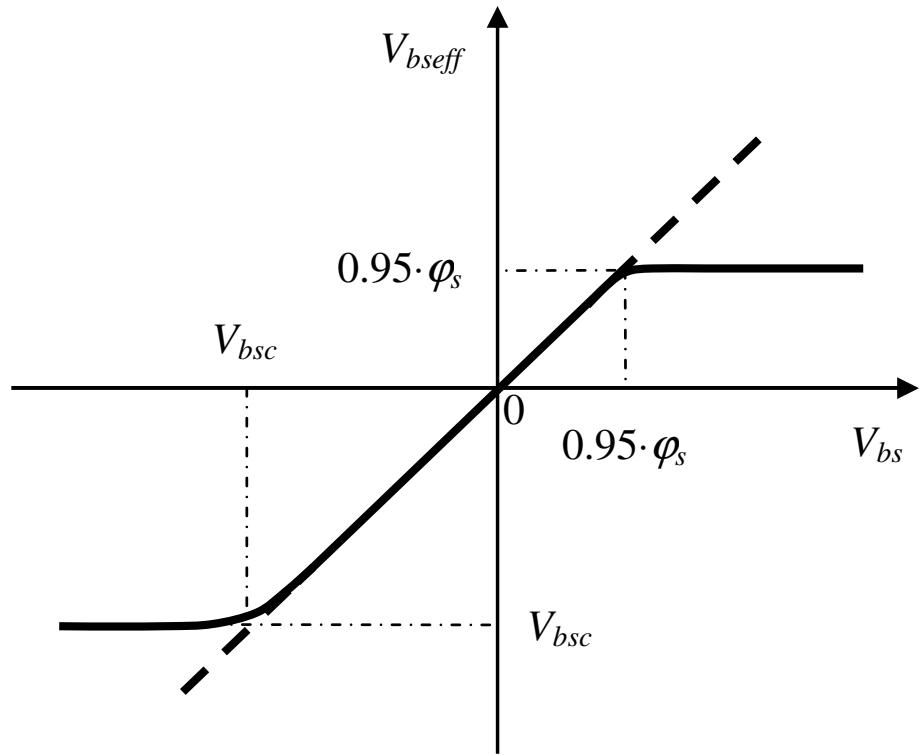


Fig. 2.6 Effective body-bias voltage  $V_{bseff}$  versus  $V_{bs}$ .

## 2.5 Poly-Silicon Gate Depletion

Although the poly-silicon gate is heavily doped to reduce the gate resistance, depletion of carriers (electrons for the  $n^+$ -poly gate and holes for the  $p^+$ -poly gate) can take place at the gate oxide interface because of the high gate oxide electric fields under inversion operation conditions. The depletion-layer thickness  $X_{poly}$  shown in Fig. 2.7 is not negligible compared to the gate oxide thickness. This reduces the gate capacitance. Its impact on the channel current can be understood as the result of the voltage drop  $V_{poly}$  across the depletion layer reducing the effective gate bias  $V_{gse} = (V_{gs} - V_{poly})$ . (In the most advance technologies, this effect is mitigated with the metal-gate stack technologies.)

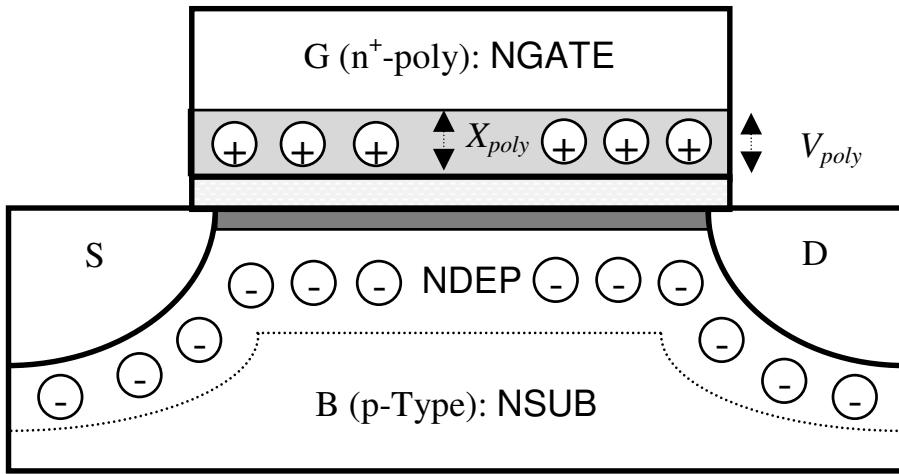


Fig. 2.7 Illustration of poly-silicon gate depletion and the voltage drop  $V_{poly}$  that reduces the effective gate drive.

Assume a uniformly doped gate and therefore an electric field, increasing from zero across  $X_{poly}$  to its maximum ( $E_m$ ) value at the interface.  $V_{poly}$  can be obtained by integrating the field over the poly depletion region

$$V_{poly} = 0.5 \cdot X_{poly} \cdot E_m \quad (2.53)$$

It is known from the previous discussions that like the derivation of the substrate depletion width, the poly gate depletion width is

$$X_{poly} = \sqrt{\frac{2\epsilon_{gate} \cdot V_{poly}}{q \cdot NGATE \cdot 10^6}} \quad (2.54)$$

where NGATE is the poly-gate doping concentration in the unit of  $\text{cm}^{-3}$ .  $\epsilon_{gate}$  is the permittivity of the gate. If the material model selection parameter MTRLMOD is zero, it selects poly-silicon gate. Hence,  $\epsilon_{gate} = 1.03594 \cdot 10^{-10}$  Farad/meter. Otherwise,  $\epsilon_{gate} = 8.85418 \cdot 10^{-12} \cdot \text{EPSRGATE}$ , in the unit of Farad/meter, is the relative dielectric constant of the gate electrode material.

Substituting Eq. (2.54) into Eq. (2.53) yields an expression of  $E_m$  of the  $V_{poly}$  expression

$$E_m = \sqrt{\frac{2q \cdot NGATE \cdot 10^6 \cdot V_{poly}}{\epsilon_{gate}}} \quad (2.55)$$

The voltage drop  $V_{poly}$  across the poly-silicon gate depletion layer is

$$V_{poly} = V_{gs} - \text{VFB} - \varphi_s - V_{ox} \quad (2.56)$$

where  $V_{ox}$  is the voltage across the gate oxide. At the gate-oxide interface, the electric displacement must be continuous. This requires

$$V_{ox} = \frac{\varepsilon_{gate} \cdot E_m \cdot \text{TOXE}}{\varepsilon_{ox}} \quad (2.57)$$

Substituting Eqs. (2.57) and (2.55) into Eq. (2.56) and solving for  $V_{poly}$  give the effective gate bias  $V_{gse} = (V_{gs} - V_{poly})$  of the BSIM4 model

$$V_{gse} = \text{VFB} + \varphi_s + a \cdot (b - 1) \quad (2.58)$$

with

$$a = \frac{q \varepsilon_{gate} \cdot \text{NGATE} \cdot 10^6 \cdot \text{TOXE}^2}{\varepsilon_{ox}^2} \quad (2.59a)$$

and

$$b = \sqrt{1 + \frac{2 \cdot (V_{gs} - \text{VFB} - \varphi_s)}{a}} \quad (2.59b)$$

In the BSIM4 code implementation, the poly-silicon gate energy-band bending,  $V_{poly}$ , is clamped at the silicon energy-band gap 1.12eV when the gate voltage is very large. The clamping is made smooth by using a smoothing function in the form of

$$V_{poly} = 1.12 - \frac{1}{2} \left[ 1.12 - V_{poly} - \delta + \sqrt{(1.12 - V_{poly} - \delta)^2 + 4\delta \cdot 1.12} \right] \quad (2.60)$$

where  $\delta = 0.05$ .  $\delta$  can be any small number between 0.01 and 0.08. It determines how rapidly  $V_{poly}$  can approach that clamping energy gap 1.12. Eq. (2.60) is one of those smoothing functions that BSIM4 uses. Choice of a good smoothing function is crucial for accuracy and numerical robustness.

When the gate-source voltage  $V_{gs}$  is in the vicinity of  $(\text{VFB} + \varphi_s)$ , the variable  $b$  of Eq. (2.59b) can be very close to 1. The term  $(b - 1)$  of Eq. (2.58) can therefore be a meaningless number, a consequence of computer floating-point round-off errors. A round-off error often results from a subtraction of two operands of approximately the same numeric values, such as  $(b - 1)$ , or any arithmetic operations on two very small

numeric numbers. Numerical round-off errors can be disastrous to the computation of trans-conductances and capacitances, one result of which is the deviation of computed circuit operating point from reality. Models with their parameters and variables coded in single-precision data type are more prone to round-off errors or accuracy loss than those coded in double precisions.

In order to remedy a potential round-off error of Eq. (2.58) in the BSIM4 code execution, direct computation of the term  $(b - 1)$  is avoided by applying the following useful technique

$$a \cdot (b - 1) \equiv a \cdot \frac{(b-1) \cdot (b+1)}{b+1} = \frac{2 \cdot (V_{gs} - V_{FB} - \varphi_s)}{b+1} \quad (2.61)$$

In addition, an effective gate-drain bias  $V_{gde}$  is computed by using the same formulation as  $V_{gse}$  but with  $V_{gs}$  replaced by  $V_{gd}$ . Note that  $V_{gde}$  is needed to compute those BSIM4 terms that are functions of the gate-drain bias.

The poly-silicon gate depletion model will be evaluated only when the following conditions are met (take the computation of  $V_{gse}$  as an example)

$$\left\{ \begin{array}{l} 1 \times 10^{18} < \text{NGATE} < 1 \times 10^{25} \\ V_{gs} > (V_{FB} + \varphi_s) \\ \varepsilon_{gate} \neq 0 \end{array} \right. \quad (2.62)$$

Otherwise all gate depletion terms are ignored.

With the  $V_{th}$  and  $V_{gse}$  models established, an effective gate overdrive voltage  $V_{gsteff}$  model is readily developed. It is presented in the BSIM4 unified channel charge model to be given in Chapter 3.

## 2.6 Bulk-Charge Effects

Repeated below is the inversion charge density  $q_{inv}(y)$ , Eq. (2.19).

$$q_{inv}(y) = -C_{oxe} \cdot [V_{gs} - V_{th} - A_{bulk} \cdot V(y)] \quad (2.19)$$

A bulk-charge coefficient  $A_{bulk}$  was obtained, from the Taylor series expansion of the bulk charge

$$A_{bulk} = 1 + \frac{\gamma}{2 \cdot \sqrt{2\varphi_B - V_{bseff}}} \quad (2.63)$$

The bulk-charge effect results from the widening of the space depletion region. Under non-zero  $V_{ds}$ , the threshold voltage and the inversion charges will both vary when moving along the channel from source to the drain. Eq. (2.63) is simply a first-order model. The second term on the right-hand side is obtained as the derivative of the bulk charge  $q_b(y)$  with respect to the channel potential  $V(y)$  at  $V_{ds} = 0$ , which is mathematically equal to the derivative of the bulk charge  $q_b(y)$  with respect to the body potential  $V_b$  at  $V(y) = 0$ . Furthermore, it assumes a long channel and a uniform channel doping.

For better accuracy,  $A_{bulk}$  needs to take into account the effects of non-uniform vertical and lateral doping, geometry, and gate and body biases. BSIM4 uses an  $A_{bulk}$  model similar to that of BSIM3v3 but incorporates many enhancements suggested by numerous users.

$$A_{bulk} = \left\{ 1 + \left( \frac{K1_{ox}}{2\sqrt{\varphi_s - V_{bseff}}} \cdot \sqrt{1 + \frac{LPEB}{L_{eff}}} + K2_{ox} - \frac{K3B \cdot TOXE \cdot \varphi_s}{W_{eff} + W_0} \right) \cdot \left[ \frac{A0 \cdot L_{eff}}{L_{eff} + 2\sqrt{XJ \cdot X_{dep}}} \cdot \left( 1 - AGS \cdot V_{gsteff} \cdot \left( \frac{L_{eff}}{L_{eff} + 2\sqrt{XJ \cdot X_{dep}}} \right)^2 \right) + \frac{B0}{W_{eff} + B1} \right] \right\} \cdot \frac{1}{1 + KETA \cdot V_{bseff}} \quad (2.64)$$

All the model parameters are described in the Parameter Table at the end of this chapter. For the BSIM4 capacitance model that will be presented in Chapter 5, a simplified bulk-charge coefficient  $A_{bulk0}$  that has no  $V_{gsteff}$  dependence (by setting  $AGS = 0$ ) is used. Note that  $A_{bulk}$  increases when the channel becomes longer.

## 2.7 LDD Resistances

There are three components of parasitic resistances in the MOSFET source and drain regions. They are associated with the metal/source and metal/drain contacts, the deep-junction diffusions, and the shallow-junction source and drain extensions. The latter is often referred to as LDD diffusion resistances. These three components are illustrated in Fig. 2.8.

In a typical IC design flow, the contact resistance  $R_{sc}$  and  $R_{dc}$  are usually extracted and supplied by an interconnect RC extraction software tool. Some SPICE simulators support  $R_{sc}$  and  $R_{dc}$  as instance parameters that are specified in MOSFET element cards. They can be used as a rough estimate to achieve better LVS (layout versus schematic) consistency. For this consideration, BSIM4 does not provide a contact resistance model.

The diffusion resistances  $R_{s,diff}$  and  $R_{d,diff}$  are determined by the resistivity and geometry of the deep diffusion regions and current spreading from the channel into diffusion regions. Traditionally, they are modeled as the product of the number of squares (NRS and NRD, which are instance parameters) and the sheet resistance RSH (a process-dependent model parameter). For advanced CMOS technologies and modern analog/RF IC designs, this treatment turns out to be inadequate. BSIM4 provides a versatile and comprehensive layout-dependent diffusion resistance model, which will be presented in Chapter 8. NRD and NRS and other geometrical instance parameters are specific to the source and drain layout of each individual transistor. These instance parameters are usually extracted by an LPE (layout parasitics extraction) software program.

The LDD resistances  $R_s(V)$  and  $R_d(V)$  significantly affect the high speed/frequency performance for advanced technologies. They can be comparable to the channel resistance in the linear region of operation. The LDD region can either be in accumulation or in depletion depending upon the gate and source/drain biases. Therefore, the LDD resistances, through both charge density and carrier mobility, have strong bias dependencies.

LDD resistance model parameter extractions can be carried out with a group of devices having varying channel lengths and widths and minimal source and drain diffusion lengths.

BSIM4 provides two topological options for the modeling of the  $R_s(V)$  and  $R_d(V)$  effects. They are selected via the model selector parameter RDMSMOD. The default is RDMSMOD = 0 (*Internal*  $R_{ds}(V)$ ), which combines  $R_s(V)$  and  $R_d(V)$  into one term  $R_{ds}(V)$ .  $R_{ds}(V)$  is then integrated into the channel current model formulation (refer to Chapter 3 for details). This option originated from BSIM3v3. It works well for digital IC simulations and speeds up the circuit matrix solving time by avoiding the creation of the two source and drain internal nodes. It is a good choice if the diffusion resistances are small.

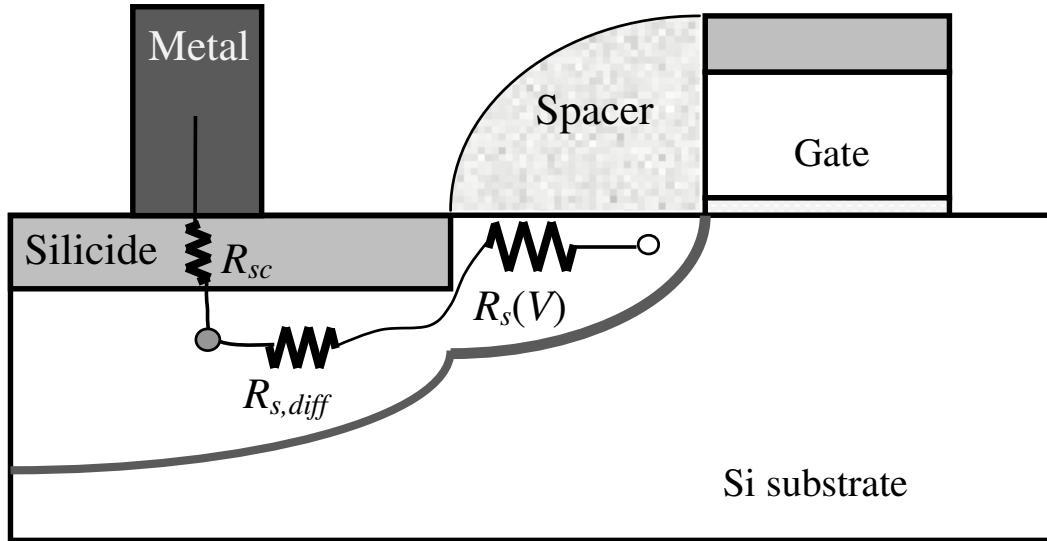


Fig. 2.8 MOSFET source and drain parasitic resistances. Only the source side is shown for ease of illustration.

In the second option (**RDSMOD** = 1: *External*  $R_s(V)$  and  $R_d(V)$ ), the configuration is similar to that of Fig. 2.8. In this case, the LDD and the diffusion resistances are also lumped, but they are situated outside the basic (intrinsic) MOSFET as shown in Fig. 2.9. Thus, two internal source and drain nodes are required. This choice is often taken in analog/RF IC simulations for accurate S/Y-parameter and noise modeling.

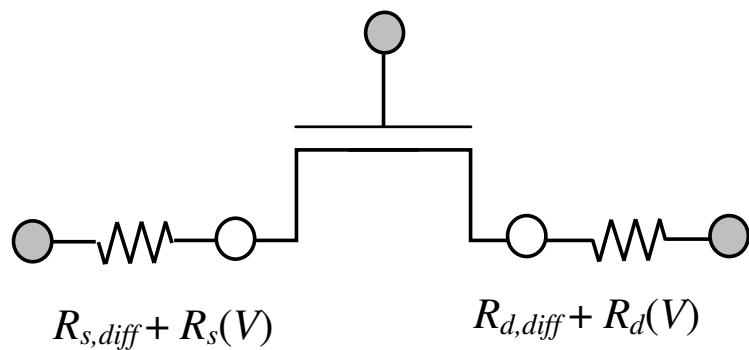


Fig. 2.9 The BSIM4 topology of **RDSMOD** = 1. The LDD bias-dependence resistances are lumped with the source and drain diffusion resistances and placed between the internal and external nodes of the source and the drain.

The formulations of LDD resistances are presented in the following. In the case of **RDSMOD** = 0, the lumped LDD resistance  $R_{ds}(V)$  is given by

$$R_{ds}(V) = \frac{rdswmin + rdsw \cdot \left[ PRWB \cdot (\sqrt{\varphi_s - V_{bseff}} - \sqrt{\varphi_s}) + \frac{1}{1+PRWG \cdot V_{gsteff}} \right]}{(W_{eff}JCT \cdot 10^6)^{WR}} \quad (2.65)$$

with

$$rdswmin = RDSWMIN + PRT \cdot \left( \frac{T_{emp}}{TNOM} - 1 \right) \quad (2.65a)$$

and

$$rdsw = RDSW + PRT \cdot \left( \frac{T_{emp}}{TNOM} - 1 \right) \quad (2.65b)$$

when TEMPMOD = 0. For other TEMPMOD settings (1, 2 or 3),  $rdswmin$  and  $rdsw$  are given

$$rdswmin = RDSWMIN \cdot [1 + PRT \cdot (T_{emp} - TNOM)] \quad (2.65c)$$

and

$$rdsw = RDSW \cdot [1 + PRT \cdot (T_{emp} - TNOM)] \quad (2.65d)$$

In the above equations, **RDSWMIN** is the zero-bias resistance. **RDSW** can be extracted under zero body bias and high gate voltages. They have the unit of ohm  $\cdot \mu\text{m}$ . **PRWB** and **PRWG** are the body and gate bias-dependence coefficients.  $W_{eff}JCT$  is the width of the source and drain diffusions. The constant  $10^6$  is a meter-to-micrometer unit conversion factor. For more explanations of these parameters, please refer to the Parameter Table at the end of this chapter.

In the case of **RDSMOD** = 1,  $R_s(V)$  and  $R_d(V)$  are modeled separately. This supports asymmetry by providing a separate set of model parameters for the source and the drain.

$$R_d(V) = \frac{rdwmin + rdw \cdot \left[ -PRWB \cdot V_{bd} + \frac{1}{1+PRWG \cdot (V_{gd} - V_{fbsd})} \right]}{NF \cdot (W_{eff}JCT \cdot 10^6)^{WR}} \quad (2.66)$$

where

$$rdwmin = RDWMIN + PRT \cdot \left( \frac{T_{emp}}{TNOM} - 1 \right) \quad (2.66a)$$

and

$$rdw = RDW + PRT \cdot \left( \frac{T_{emp}}{TNOM} - 1 \right) \quad (2.66b)$$

when TEMPMOD = 0. For other TEMPMOD settings (1, 2 or 3),  $rdwmin$  and  $rdw$  are

$$rdwmin = \text{RDWMIN} \cdot [1 + \text{PRT} \cdot (T_{emp} - \text{TNOM})] \quad (2.66c)$$

and

$$rdw = \text{RDW} \cdot [1 + \text{PRT} \cdot (T_{emp} - \text{TNOM})] \quad (2.66d)$$

$R_s(V)$  and its temperature dependencies are formulated in a similar way.

$$R_s(V) = \frac{rswmin + rsw \cdot \left[ -\text{PRWB} \cdot V_{bs} + \frac{1}{1 + \text{PRWG} \cdot (V_{gs} - V_{fbsd})} \right]}{\text{NF} \cdot (W_{effJCT} \cdot 10^6)^{\text{WR}}} \quad (2.67)$$

where

$$rswmin = \text{RSWMIN} + \text{PRT} \cdot \left( \frac{T_{emp}}{\text{TNOM}} - 1 \right) \quad (2.67a)$$

and

$$rsw = \text{RSW} + \text{PRT} \cdot \left( \frac{T_{emp}}{\text{TNOM}} - 1 \right) \quad (2.67b)$$

when TEMPMOD = 0. For other TEMPMOD settings (1, 2 or 3),  $rswmin$  and  $rsw$  are given

$$rswmin = \text{RSWMIN} \cdot [1 + \text{PRT} \cdot (T_{emp} - \text{TNOM})] \quad (2.67c)$$

and

$$rsw = \text{RSW} \cdot [1 + \text{PRT} \cdot (T_{emp} - \text{TNOM})] \quad (2.67d)$$

For RDSMOD = 1, these formulations and parameters are similar to those of Eq. (2.65). However, the bias-dependence terms are improved. For example, the resistance is now formulated to be inversely proportional to the difference of the gate voltage and the gate-source/drain flat-band voltage  $V_{fbsd}$ . This is a more physical model because the mobile charge in the diffusion region is a linear function of  $(V_g - V_{fbsd})$ .  $V_{fbsd}$  is given as follows.

In the case of MTRLMOD = 0 (a traditional poly-Si/SiO<sub>2</sub>/Si device structure),

$$V_{fbsd} = \frac{k_B \cdot \text{TNOM}}{q} \cdot \ln \left( \frac{\text{NGATE}}{\text{NSD}} \right) \quad (2.68)$$

When **NGATE** is equal to or less than zero,  $V_{fbsd}$  will be set to zero. **NGATE** and **NSD** are the gate and source/drain doping concentrations, respectively. If different device materials, such as high- $k$  metal gates, are employed (**MTRLMOD** = 1), the flat-band voltage between the gate and the source/drain diffusion is

$$V_{fbsd} = \text{PHIG} - \text{EASUB} - 0.5 \cdot E_{g0}(\text{TNOM}) + \frac{k_B \cdot \text{TNOM}}{q} \cdot \ln\left(\frac{\text{NSD}}{n_i(\text{TNOM})}\right) \quad (2.69)$$

**PHIG** and **EASUB** are model parameters denoting the gate work function and the electron affinity of the substrate.  $E_{g0}$  is the substrate energy band gap at **TNOM**.

## 2.8 Finite Charge Thickness

MOSFET models for IC simulation before BSIM3v3.2 and BSIM4 ignore the finite thickness of an inversion or accumulation charge layer. They assume that the charges are all located at the interface with zero thickness. In reality, the energy-band diagram of an NMOS, for example, clearly indicates the presence of a quantum well between the gate oxide and the semiconductor conduction band at the interface. The solution of the Schrodinger equation for such a quantum well indicated that the electron density is nearly zero at the interface and peaks at some distance below the interface before the density falls back to zero deep below the interface. The average depth of the inversion charge is called the charge centroid or charge thickness. Traditionally, for the electrostatic analyses, all the inversion charges were assumed to be in a sheet.

The finite charge-thickness model starts with the DC charge thickness  $X_{DC}$ .  $X_{DC}$  is defined as the integral  $\int_0^\infty \rho(x) \cdot x \cdot dx / \int_0^\infty \rho(x) \cdot dx$ , where  $\rho(x)$  is the charge density as a function of depth. This finite charge thickness introduces a capacitance in series with  $C_{oxp}$  determined by the physical gate dielectric thickness **TOXP**. This results in an effective gate oxide capacitance that becomes bias dependent

$$C_{oxeff} = \frac{C_{oxp} \cdot C_{cen}}{C_{oxp} + C_{cen}} \quad (2.70)$$

Here  $C_{cen} = \epsilon_{sub}/X_{DC}$  is the charge-thickness capacitance.

The BSIM4 inversion charge layer thickness in the unit of centimeters is derived from quantum mechanical analyses (refer to Chapter 5 for details)

$$X_{DC} = \frac{1.9 \times 10^{-9} \cdot ADOS}{1 + \left[ \frac{V_{gsteff} + 4 \cdot (VTH0 - VFB - \varphi_s)}{2 \cdot TOXP \cdot 10^8} \right]^{0.7 \cdot BDOS}} \quad (2.71)$$

ADOS and BDOS are model parameters to account for the density of states. The second term in the denominator has the unit of MV/cm. For N<sup>+</sup> or P<sup>+</sup> poly-silicon gates, (VFB +  $\varphi_s$ ) is approximately zero, but it is kept for other gate stacks such as high- $k$  metal gates. Fig. 2.10 shows the comparison of this inversion charge thickness model against the numerical quantum analysis for various gate oxide thicknesses (TOXP) and channel doping concentrations (NDEP).

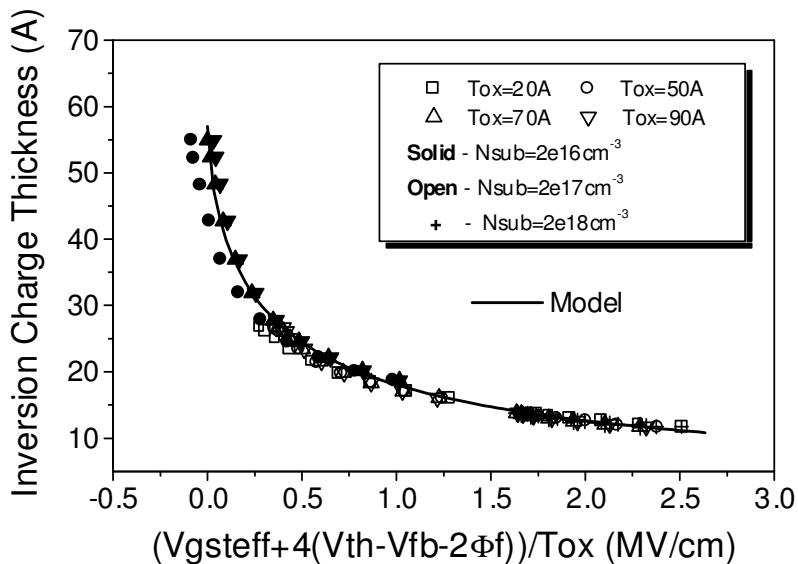


Fig. 2.10 The inversion charge thickness model agrees well with numerical quantum mechanical simulations for various channel doping concentrations and gate oxide thicknesses.

## 2.9 Effective Mobility

MOSFET channel carriers have limited mobility because of various scattering events while they are traversing the channel to the drain end. Three scattering mechanisms, each limiting its corresponding mobility component, have been identified to be inherent in the MOS system.

Coulombic scattering can be caused by trapped or fixed charges at the interface and/or inside the gate insulator. It is also caused by the ionized dopant atoms in the channel region, which are the primary Coulombic scattering centers for the SiO<sub>2</sub>/Si system. It is the dominant mechanism when the inversion carrier density is low. As the number of inversion carriers increases, the Coulombic scattering events decrease because of the increased screening effect of the mobile charge carriers. This mechanism is negligible in traditional, production-quality MOSFETs in the on-state ( $V_g > V_{th}$ ); it is important in the most advanced MOSFET transistors because of very high channel dopings and/or the use of high- $k$  metal gate stacks.

Phonon scattering refers to the scattering of carriers by thermal phonons. This scattering mechanism is important in the intermediate range of the transverse electric field.

Surface scattering is the dominant mechanism in the high transverse field range and is usually attributed to the roughness of the silicon/dielectric interface.

The combined phonon and surface scattering mechanisms result in an effective inversion layer carrier mobility  $\mu_{eff}$ . It is a strong function of the transverse electric field in the channel surface region as shown in Fig. 2.11. By defining an effective (transverse) electric field  $E_{eff}$ , the average value of this field across the thickness of the inversion charge layer,  $\mu_{eff}$  follows a universal mobility-field relationship

$$\mu_{eff} = \frac{U_0}{1 + \left(\frac{E_{eff}}{E_0}\right)^v} \quad (2.72)$$

$U_0$  and  $E_0$  are the low-field mobility and the critical vertical electric field. The power  $v$  ranges from 1 to 2 and may be slightly process dependent. The effective mobility plotted as a function of the effective field follows a single curve regardless of the body doping concentration and profile, the gate and body biases, and the gate oxide thickness for a given operating temperature.  $E_{eff}$  is found to be [7]

$$E_{eff} = \frac{1}{\varepsilon_{sub}} \cdot (q_b + 0.5 \cdot q_{inv}) \quad (2.73)$$

It is the average of the field at the surface and that at the bottom edge of the inversion layer. It can be formulated explicitly in terms of device parameters and the gate bias [8] as

$$E_{eff} \approx \frac{V_{gs} + V_{th}}{6 \cdot \text{TOXE}} \quad (2.74)$$

Fig. 2.11 shows that the effective mobility indeed is a function of only  $E_{eff}$  when  $E_{eff}$  is large, which is the regime where Coulombic scattering is not important. At small  $E_{eff}$  or low  $V_g$ , the measured mobility falls off the model curve because Coulombic scattering becomes important. See the discussion of Eq. (2.79) for how Coulombic scattering is included in the BSIM4 model.

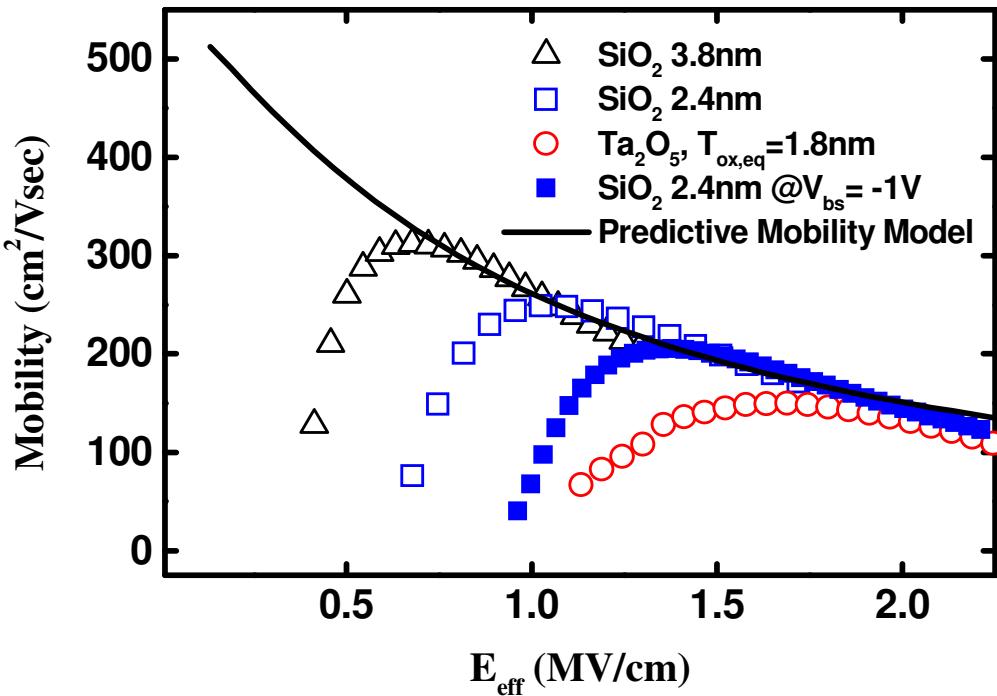


Fig. 2.11 The electric field dependencies of the effective mobility for various gate dielectric materials and thicknesses. The line represents the result using the BSIM4 MOBMOD = 2 mobility model, which is accurate for inversion region where Coulombic scattering can be neglected.

BSIM4 provides four different mobility models, each of which is selected by specifying the model parameter **MOBMOD**. In the case of poly-Si/SiO<sub>2</sub>/Si structures (**MTRLMOD** = 0, the default),  $E_{eff}$  is given

$$E_{eff} = \frac{V_{gsteff} + 2 \cdot V_{th}}{\text{TOXE}} \quad (2.75\text{a})$$

whereas when a different structure is used (**MTRLMOD** = 1),

$$E_{eff} = \frac{V_{gsteff} + 2 \cdot V_{th} + 2 \cdot (\text{PHIG} - \text{EASUB} - 0.5 \cdot E_{g0} + 0.45)}{\text{TOXE}} \quad (2.75\text{b})$$

**PHIG** and **EASUB** are model parameters denoting the gate work function and the electron affinity of the substrate, and  $E_{g0}$  is the substrate energy band gap at TNOM. The term within the parentheses in the numerator represents the offset to the flat-band voltage of the poly/SiO<sub>2</sub>/Si device structures.  $\mu_{eff}$  of **MOBMOD** = 0 is

$$\mu_{eff} = \frac{\mu_0(T, L)}{1 + (UA + UC \cdot V_{bseff}) \cdot E_{eff} + UB \cdot E_{eff}^2 + UD \cdot \left( \frac{V_{th} \cdot \text{TOXE}}{V_{gsteff} + 2 |V_{th}|} \right)^2} \quad (2.76)$$

**UA**, **UB**, **UC** and **UD** are model parameters. The  $E_{eff}$  terms are obtained from the Taylor series expansion of the  $E_{eff}$  term of Eq. (2.72). The **UD** term in the denominator takes into account the Coulombic scattering. Mobility has strong temperature dependence because of the change of phonon scattering. This is modeled through the temperature dependence of the low-field mobility **U0**. The other cause of the mobility temperature dependence is the temperature dependence of the effective electric field due to the change of inversion carrier densities, which can be modeled through the temperature dependence model of  $V_{th}$ . In the above equation,  $\mu_0(T, L)$  is the low-field mobility at operating temperature  $T_{emp}$ . It also includes the length dependence of pocket implant dopings as

$$\mu_0(T, L) = U0 \cdot \left( \frac{T_{emp}}{\text{TNOM}} \right)^{\text{UTE}} \cdot \left[ 1 - UP \cdot \exp \left( - \frac{L_{eff}}{LP} \right) \right] \quad (2.76\text{a})$$

**UTE**, **UP** and **LP** are also model parameters. The effective mobility has strong temperature dependence, proportional to  $T_{emp}^{-(0.8 \sim 1.5)}$  with  $T_{emp}$  given in degrees Kelvin. **UTE** has a default value of -1.5. The lower  $T_{emp}$  is, the larger the mobility is. Moreover, as the channel length becomes shorter, the channel doping concentration becomes higher because of the pocket implants at the source and drain corners. As a result, mobility

decreases accordingly. Note that BSIM4 will issue a fatal error message if  $\mu_0(T, L)$  becomes less than or equal to zero, a condition that leads to SPICE simulation failures. MOBMOD = 0 is the most frequently used model in production. It will be shown shortly that  $\mu_0(T, L)$  will need to incorporate mechanical stress effects and well-proximity effects as well.

MOBMOD = 1 has a similar formulation

$$\mu_{eff} = \frac{\mu_0(T, L)}{1 + (UA \cdot E_{eff} + UB \cdot E_{eff}^2) \cdot (1 + UC \cdot V_{bseff}) + UD \cdot \left( \frac{V_{th} \cdot TOXE}{V_{gsteff} + 2|V_{th}|} \right)^2} \quad (2.77)$$

MOBMOD = 2 preserves the predictiveness and universality of Eq. (2.72). It is intended for the poly-Si/SiO<sub>2</sub>/Si structure, for which the universal mobility has been well studied. It includes the body-bias and Coulombic scattering effects.

$$\mu_{eff} = \frac{\mu_0(T, L)}{1 + (UA + UC \cdot V_{bseff}) \cdot \left[ \frac{V_{gsteff} + C_{n,p} \cdot (V_{TH0} - V_{FB} - \varphi_s)}{TOXE} \right]^{EU} + UD \cdot \left( \frac{V_{th} \cdot TOXE}{V_{gsteff} + 2|V_{th}|} \right)^2} \quad (2.78)$$

The coefficient  $C_{n,p}$  is experimentally found to be 2 for NMOS and 2.5 for PMOS.

The carrier mobility in high- $k$  metal gate stacks is significantly lower than the poly-Si gate and SiO<sub>2</sub> stacks. Surface roughness scattering as well as remote soft-phonon (i.e., low-energy optical phonon) scattering are among the causes. These scattering mechanisms can still be modeled well with the above models with proper choice of parameter values. However, Coulombic scattering from fixed charges of high- $k$  dielectrics, which is gate bias dependent, cannot be represented well by the above model. BSIM4 provides a predictive, effective mobility model MOBMOD = 3, with accurate Coulombic scattering model for the most advanced gate stacks [2]:

$$\mu_{eff} = \frac{\mu_0(T, L)}{1 + (UA + UC \cdot V_{bseff}) \cdot \left[ \frac{V_{gsteff} + C_{n,p} \cdot (V_{TH0} - V_{FB} - \varphi_s)}{6 \cdot TOXE} \cdot 10^{-8} \right]^{EU} + UD / \left[ 0.5 \cdot (1 + V_{gsteff} / V_{gs\_on})^{UCS} \right]} \quad (2.79)$$

In Eqs. (2.78) and (2.79), EU and UCS are model parameters, both having a default value of 1.67 for NMOS and 1 for PMOS.  $V_{gs\_on}$  is equal to  $V_{gsteff}$  taken at  $V_{ds} = V_{bs} = 0$  and under the effective gate bias (because of the poly-Si gate depletion effect)  $V_{gse} = V_{th}$ . The last term in the denominator in Eq. (2.79) models the Coulombic scattering. The inversion charge dependence is taken into account through  $V_{gsteff}$ .

The mobility model parameters U0, UA, UB, UC, UD and UCS are temperature dependent. BSIM4 provides several temperature-dependence models. They are selected with the model selector parameter TEMPMOD. A brief description of them is given below.

When TEMPMOD = 0 is selected,

$$UA(T_{emp}) = UA + UA1 \cdot \left( \frac{T_{emp}}{TNOM} - 1 \right) \quad (2.80a)$$

$$UB(T_{emp}) = UB + UB1 \cdot \left( \frac{T_{emp}}{TNOM} - 1 \right) \quad (2.80b)$$

$$UC(T_{emp}) = UC + UC1 \cdot \left( \frac{T_{emp}}{TNOM} - 1 \right) \quad (2.80c)$$

$$UD(T_{emp}) = UD + UD1 \cdot \left( \frac{T_{emp}}{TNOM} - 1 \right) \quad (2.80d)$$

When TEMPMOD = 1 or 2 is selected,

$$UA(T_{emp}) = UA \cdot [1 + UA1 \cdot (T_{emp} - TNOM)] \quad (2.81a)$$

$$UB(T_{emp}) = UB \cdot [1 + UB1 \cdot (T_{emp} - TNOM)] \quad (2.81b)$$

$$UC(T_{emp}) = UC \cdot [1 + UC1 \cdot (T_{emp} - TNOM)] \quad (2.81c)$$

$$UD(T_{emp}) = UD \cdot [1 + UD1 \cdot (T_{emp} - TNOM)] \quad (2.81d)$$

When TEMPMOD = 3 is selected,

$$UA(T_{emp}) = UA \cdot \left( \frac{T_{emp}}{TNOM} \right)^{UA1} \quad (2.82a)$$

$$UB(T_{emp}) = UB \cdot \left( \frac{T_{emp}}{TNOM} \right)^{UB1} \quad (2.82b)$$

$$UC(T_{emp}) = UC \cdot \left( \frac{T_{emp}}{TNOM} \right)^{UC1} \quad (2.82c)$$

$$UD(T_{emp}) = UD \cdot \left( \frac{T_{emp}}{TNOM} \right)^{UD1} \quad (2.82d)$$

Note that the same temperature dependence model of U0 and UCS is used for all TEMPMOD settings

$$U0(T_{emp}) = U0 \cdot \left( \frac{T_{emp}}{TNOM} \right)^{UTE} \quad (2.83)$$

$$UCS(T_{emp}) = UCS \cdot \left( \frac{T_{emp}}{TNOM} \right)^{UCSTE} \quad (2.84)$$

## 2.10 Layout-Dependent Effects: Mechanical Stress and Proximity Effects

This section discusses the modeling of the layout-dependent effects of mechanical stress and proximity on the MOSFET intrinsic characteristics. BSIM4 also provides another layout-dependence model, which is associated the parasitic effects brought about by those such as the source and drain diffusions and junctions. That parasitics model will be presented in Chapter 8. In this sub-section, the BSIM4 model of the STI/LOD mechanical stress effect will be presented first, followed by the model of the WPE effects.

The Si channel region of the state-of-the-art MOS transistors is surrounded by various other materials that have different coefficients of thermal expansion than Si. These materials are deposited at elevated temperatures and unavoidably exert mechanical stress on the channel. The stress can be tensile or compressive force per unit area usually measured in megapascals (MPa) or gigapascals (GPa). Stress causes compression or expansion of the silicon channel. The technical term for such material deformation is strain. For example, the strain in the channel length direction is defined as the ratio of the length change to the length with no stress exerted.

Strain leads to a change in the inter-atomic spacing and, in turn, causes the splitting or warping (bending or twisting) of the energy bands. As a result, the carrier effective mass and inter-band scattering rate can change. Therefore, the carrier mobility either increases or decreases. In other words, strain can be good to carrier mobility and transistor driving current or bad to them. Strain may be created unintentionally or intentionally. In either case, it is difficult to model their effects well. Sometimes, bad strains arise as the unintended consequence of

fabrication steps including STI (shallow-trench isolation) and nitride capping over the transistor. At other times these process and other steps are carefully engineered or new process steps are introduced to create good strains from either or both compressive and tensile stress to improve carrier mobility and transistor performance. The engineered strain technology is called the strained-silicon technology, which is an important technology differentiator to have. Some of the favorable/good strains (tensile or compressive) for NMOS and PMOS are illustrated in Fig. 2.12.

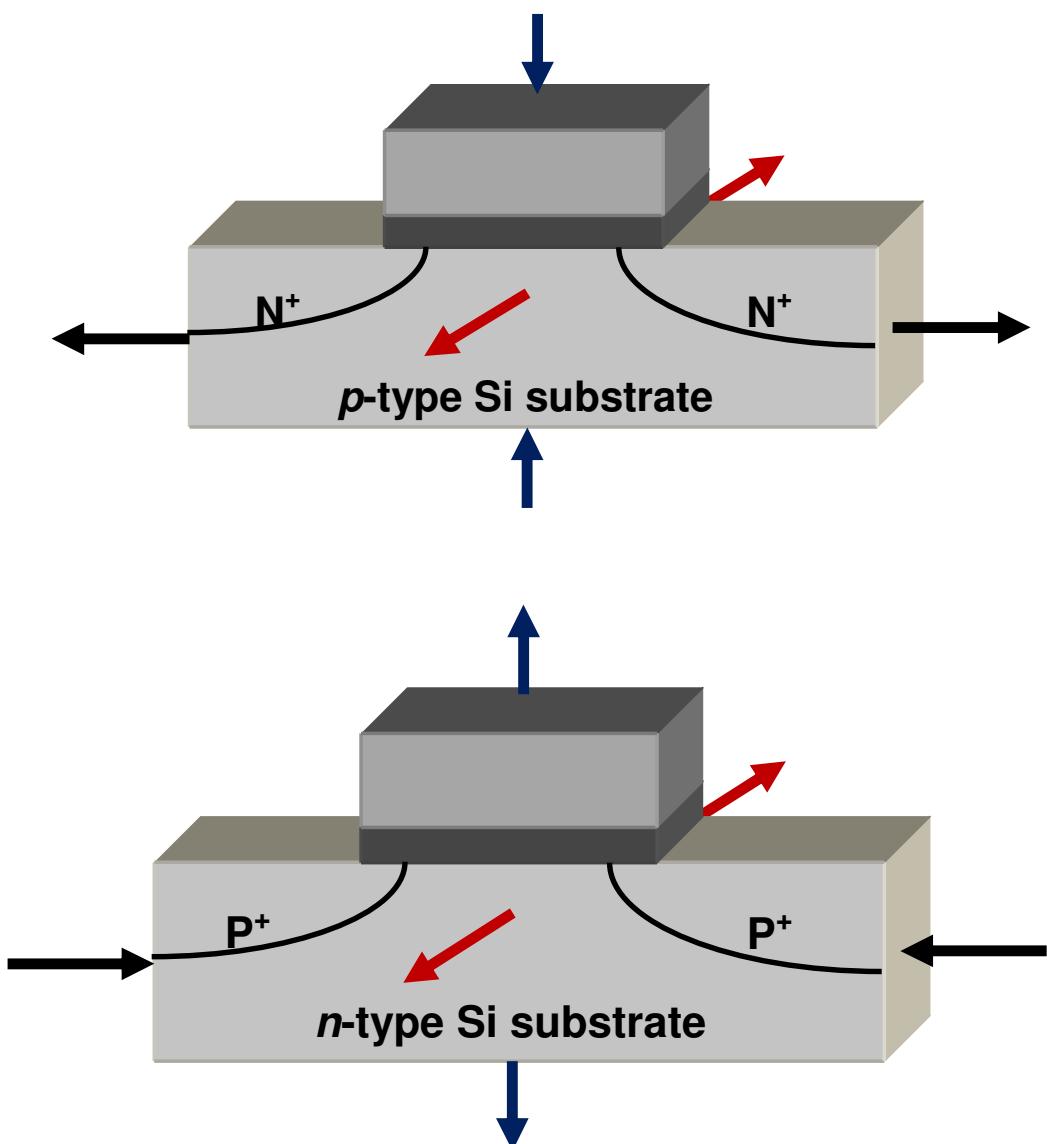


Fig. 2.12 Desirable compressive and/or tensile stress orientations in silicon NMOS (on the top) and PMOS (at the bottom) transistors on a 100 cut wafer with the current flowing in the 110 direction of the crystal.

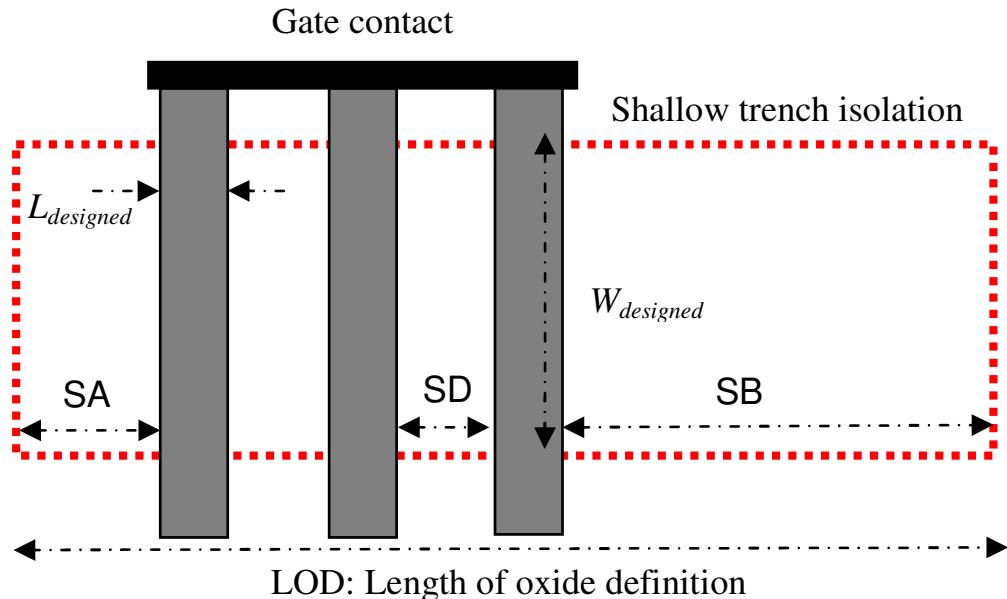


Fig. 2.13 Schematic illustration of the layout dependence of the STI/LOD stress effects.

Various strain engineering techniques have been invented. Examples are the dual contact etch stopper layer (CESL), stress memorization technique (SMT), and STI with and without  $\text{Si}_3\text{N}_4$  liner to achieve desired stressors for NMOS and PMOS. Other techniques include the embedded SiGe as source and drain diffusions for PMOS only. Strain engineering, measurement and characterization, and SPICE modeling and parameter extraction are sophisticated tasks because the strain has a very complex dependence on process and layout implementations. It depends on the size, location and orientation of the transistor and its surrounding materials. Fab and design companies have their own customized and unique ways of modeling and extraction of the mechanical stress effects on top of BSIM4.

BSIM4 provides a layout-dependent STI/LOD stress effect model. The amount of the stress that is exerted on the transistor channel region depends upon how far the channel is located from the edges of STI and how many fingers this device may have (as shown in Fig. 2.13). One major factor is the so-called the length of oxide definition or LOD that defines the active region of a die where that device is built.

There are five BSIM4 parameters that describe the LOD layout dependence and need to be preprocessed prior to the BSIM4 IV/CV model evaluation and SPICE circuit matrix loading. They are the low-field mobility  $\mu_0(T, L)$ ,  $VTH0$ ,  $\text{ETA}0$ ,  $K2$  as well as the electron and

hole saturation velocity VSAT (the velocity-field relationship will be presented in Chapter 3). Consider  $\mu_0(T, L)$  first.

$$\mu_0(T, L)_{\text{LOD}} = \mu_0(T, L) \cdot \frac{1 + \rho_\mu}{1 + \rho_{\mu-\text{ref}}} \quad (2.85)$$

$\rho_\mu$  is the percentage change of the low-field mobility  $U_0$  relative to an *infinitely* large MOSFET that has negligible mechanical stress effect.

$$\rho_\mu = \frac{KU0}{KU0\_Geo\_Temp} \cdot Inv\_SAB_{eff} \quad (2.86)$$

$KU0$  is a model parameter that specifies how mobility changes with SA, SB, and SD.  $KU0\_Geo\_Temp$  accounts for the dependencies of  $KU0$  on channel geometries and temperatures.

$$KU0\_Geo\_Temp = \left[ 1 + \frac{LKU0}{(L_{\text{physical}})^{\text{LLODKU0}}} + \frac{WKU0}{(W_{\text{physical}} + W_{\text{LOD}})^{\text{WLODKU0}}} + \frac{PKU0}{(L_{\text{physical}})^{\text{LLODKU0}} \cdot (W_{\text{physical}} + W_{\text{LOD}})^{\text{WLODKU0}}} \right] \cdot \left[ 1 + TKU0 \cdot \left( \frac{T_{\text{emp}}}{T_{\text{NOM}}} - 1 \right) \right] \quad (2.87)$$

This formulation has traces of the familiar BSIM4 methods of geometrical binning and scaling and temperature de-rating. These new parameters are further described in the Parameter Table at the end of this chapter.

$\rho_\mu$  is inversely proportional to SA, SB, and SD.  $Inv\_SAB_{eff}$  of Eq. (2.86), taking into account both single-finger ( $NF = 1$ ) and multi-finger ( $NF \geq 2$ ) device structures, is

$$Inv\_SAB_{eff} = \frac{1}{NF} \cdot \left[ \sum_{i=0}^{NF-1} \frac{1}{SA + 0.5 \cdot L_{\text{designed}} + i \cdot (SD + L_{\text{designed}})} + \sum_{i=0}^{NF-1} \frac{1}{SB + 0.5 \cdot L_{\text{designed}} + i \cdot (SD + L_{\text{designed}})} \right] \quad (2.88)$$

Using an *infinitely* large device as the reference is not realistic. Thus, one thinks of introducing a denominator into Eq. (2.85) to de-rate the  $\rho_\mu$

term. This is a useful practice and helps parameter extraction.  $\rho_{\mu\text{-ref}}$  is modeled with

$$\rho_{\mu\text{-ref}} = \frac{KU0}{KU0\_Geo\_Temp} \cdot Inv\_SAB_{ref} \quad (2.89)$$

Where

$$Inv\_SAB_{ref} = \frac{1}{SAREF + 0.5 \cdot L_{desinged}} + \frac{1}{SBREF + 0.5 \cdot L_{desinged}} \quad (2.90)$$

SAREF and SBREF are the references for SA and SB, respectively. Both are model parameters and set to  $1\mu\text{m}$  as their default values.

The LOD stress effect on the saturation velocity VSAT is modeled similarly as for mobility.

$$VSAT_{LOD} = VSAT \cdot \frac{1 + KVSAT \cdot \rho_\mu}{1 + KVSAT \cdot \rho_{\mu\text{-ref}}} \quad (2.91)$$

where KVSAT is the coefficient parameter.

The LOD stress effect on the threshold voltage  $V_{th}$  is modeled by considering the following  $V_{th}$  parameters of BSIM4.

$$VTH0_{LOD} = VTH0 + \frac{KVTH0}{KVTH0\_Geo\_Temp} \cdot (Inv\_SAB_{eff} - Inv\_SAB_{ref}) \quad (2.92)$$

$$ETA0_{LOD} = ETA0 + \frac{STETA0}{KVTH0\_Geo\_Temp^{LODETA0}} \cdot (Inv\_SAB_{eff} - Inv\_SAB_{ref}) \quad (2.93)$$

and

$$K2_{LOD} = K2 + \frac{STK2}{KVTH0\_Geo\_Temp^{LODK2}} \cdot (Inv\_SAB_{eff} - Inv\_SAB_{ref}) \quad (2.94)$$

In the above equations, the formulation of the denominator  $KVTH0\_Geo\_Temp$  is

$$KVTH0\_Geo\_Temp = \left[ 1 + \frac{LKVT0}{(L_{physical})^{LLODVTH}} + \frac{WKVT0}{(W_{physical}+WLOD)^{WLLODVTH}} + \frac{PKVT0}{(L_{physical})^{LLODVTH} \cdot (W_{physical}+WLOD)^{WLLODVTH}} \right] \quad (2.95)$$

The new model parameters in the above equations are explained in the Parameter Table.

The LOD stress model evaluation will be turned on if both **SA** and **SB** are greater than zero for  $NF = 1$  and if **SA**, **SB**, and **SD** are all greater than zero for  $NF \geq 2$ .

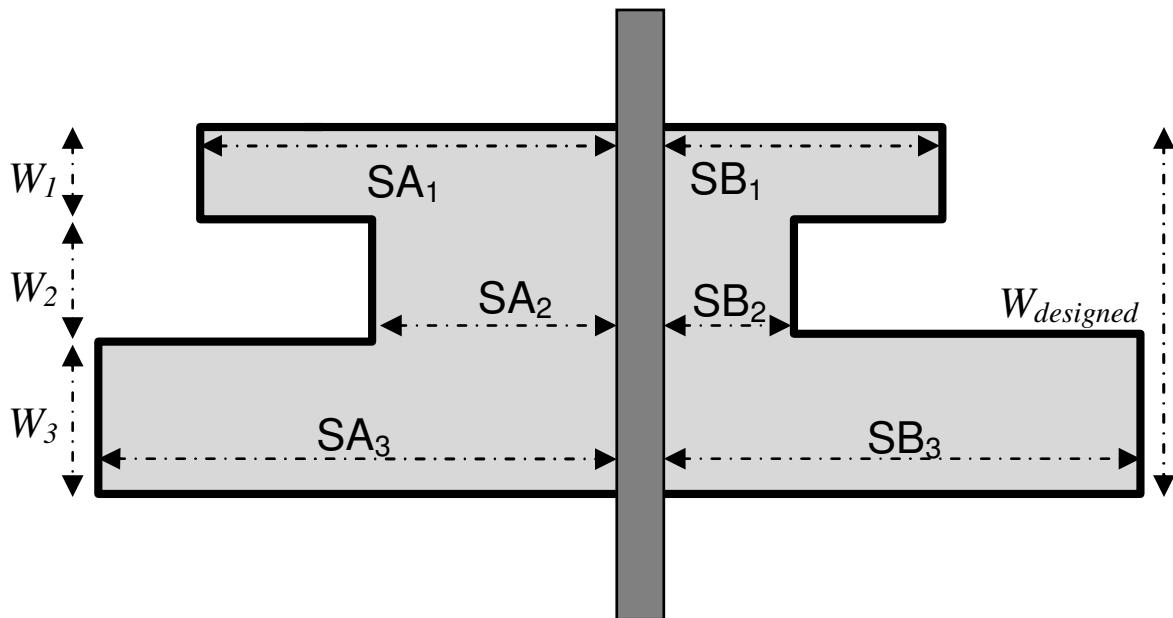


Fig. 2.14 A MOSFET transistor with irregular LODs. An equivalent  $\mathbf{SA}_e$  and  $\mathbf{SB}_e$  can be obtained from Eq. (2.96a) and Eq. (2.96b) to compute  $Inv\_SABeff$ .

Equation (2.88) can be generalized to obtain an equivalent  $\mathbf{SA}_e$  and  $\mathbf{SB}_e$  for a device with irregular LODs as shown in Fig. 2.14, where each of the  $n$  segments in the width direction has its own  $\mathbf{SA}_i$  and  $\mathbf{SB}_i$ . This way, no additional instance parameters  $\mathbf{SA}_i$  and  $\mathbf{SB}_i$  need to be introduced.

$$\frac{1}{\mathbf{SA}_e + 0.5 \cdot L_{designed}} = \sum_{i=1}^n \frac{W_i}{W_{designed}} \cdot \frac{1}{\mathbf{SA}_i + 0.5 \cdot L_{desinged}} \quad (2.96a)$$

and

$$\frac{1}{\mathbf{SB}_e + 0.5 \cdot L_{designed}} = \sum_{i=1}^n \frac{W_i}{W_{designed}} \cdot \frac{1}{\mathbf{SB}_i + 0.5 \cdot L_{desinged}} \quad (2.96b)$$

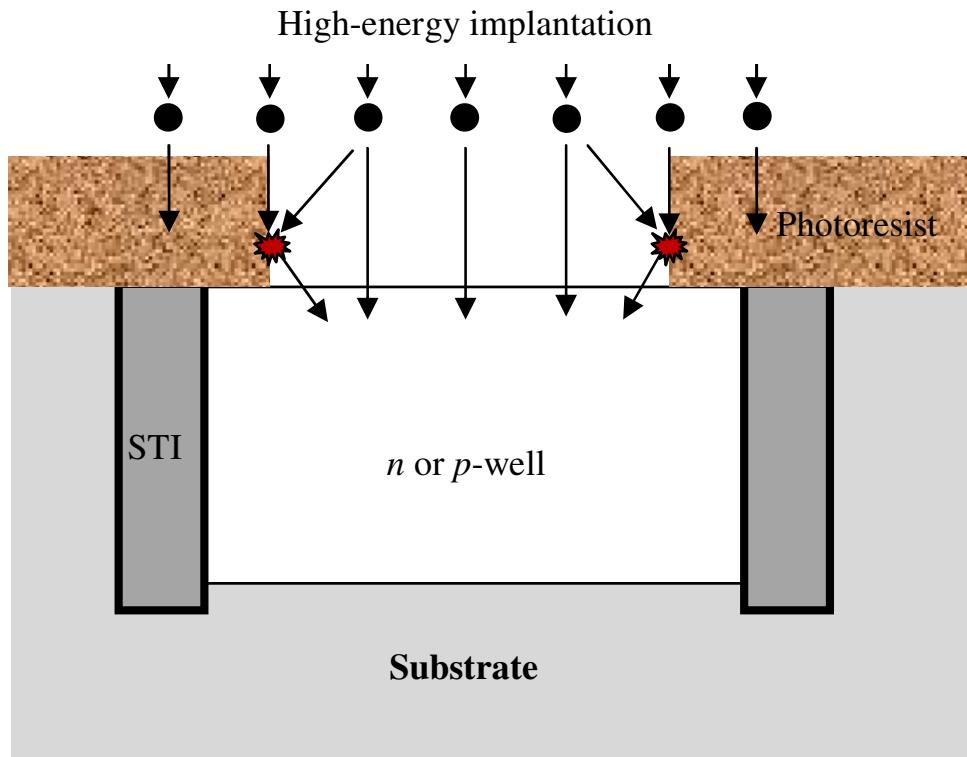


Fig. 2.15 Schematic illustration of the WPE effects. Some of the high-energy dopant ions are scattered by the photoresist into the edge region of the well, thus changing the intended channel doping concentrations and hence parameters of the devices that are located in the edge region.

BSIM4 models a second layout-dependent effect known as the well-proximity effects (WPE). In realizing CMOS well doping profiles, high-energy ion implantation is the process technique to use. Some of the dopant ions that are being implanted are scattered (deflected) by the photoresist wall before entering the silicon wafer. This is as shown in Fig. 2.15. These scattered ions make the doping concentration higher near the edge of the well. This means the device that is located near the edge of the well will have a higher threshold voltage and a lower carrier mobility.

Define **SC** as the distance from the gate edge to the edge of the well (**SC** not shown in Fig. 2.15). The low-field mobility is now modified

$$\mu_0(T, L)_{\text{WPE}} = \mu_0(T, L) \cdot (1 + K_{\mu 0} \cdot \text{WE} \cdot SC_{eff}) \quad (2.97)$$

The long-channel zero-body bias threshold voltage is now

$$V_{TH0\text{WPE}} = V_{TH0} + K_{VTH0} \cdot \text{WE} \cdot SC_{eff} \quad (2.98)$$

In addition, the body-bias coefficient  $K2$  for non-uniform vertical doping is changed to

$$K2_{WPE} = K2 + K2WE \cdot SC_{eff} \quad (2.99)$$

where  $KU0WE$ ,  $KVTH0WE$ , and  $K2WE$  are binnable coefficient parameters to account for the WPE effect. For details, refer to the Parameter Table at the end of this chapter. In the above equation, the quantity  $SC_{eff}$  represents the photoresist-scattered dopant

$$SC_{eff} = SCA + WEB \cdot SCB + WEC \cdot SCC \quad (2.100)$$

$SCA$ ,  $SCB$ , and  $SCC$  are instance parameters. Model parameters  $WEB$  and  $WEC$  are the coefficients of  $SCB$  and  $SCC$ , respectively.

The well-proximity effect model will be turned on when the model selector  $WPEMOD$  is set to 1. In the case where  $SCA$ ,  $SCB$ , and  $SCC$  are not specified, if either  $SC$  is not given or it is negative or both, no WPE model evaluation will be performed. However, if  $SC$  is both given and greater than zero, then  $SCA$ ,  $SCB$ , and  $SCC$  are computed as follows

$$SCA = SCREF^2 \cdot \frac{1}{SC \cdot (SC + W_{designed})} \quad (2.101)$$

$$SCB = \frac{1}{W_{designed}} \cdot \left\{ (0.1 \cdot SC + 0.01 \cdot SCREF) \cdot \exp \left( -\frac{10 \cdot SC}{SCREF} \right) - [0.1 \cdot (SC + W_{designed}) + 0.01 \cdot SCREF] \cdot \exp \left( -\frac{10 \cdot (SC + W_{designed})}{SCREF} \right) \right\} \quad (2.102)$$

and

$$SCC = \frac{1}{W_{designed}} \cdot \left\{ (0.05 \cdot SC + 0.0025 \cdot SCREF) \cdot \exp \left( -\frac{20 \cdot SC}{SCREF} \right) - [0.05 \cdot (SC + W_{designed}) + 0.0025 \cdot SCREF] \cdot \exp \left( -\frac{20 \cdot (SC + W_{designed})}{SCREF} \right) \right\} \quad (2.103)$$

Note that these expressions for computing  $SCA$ ,  $SCB$ , and  $SCC$  are particularly useful for layout parasitics extraction (LPE) tools to attain good pre- and post-layout simulation accuracy correlations.

The layout-dependent effects of the state-of-the-art CMOS technologies are much more complex for numerous process flavors and recipes. The nature of this subject is truly a 3-D problem. The number of

stressors is large. Devices can no longer be deemed solely and uniquely as a function of their own geometries (e.g., lengths and widths), but of their orientations and details of surroundings as well. Restricted design rules and regularized layout patterns can solve the problem only partially. Custom models built with extensions upon standard models such as BSIM4 have been the choice solutions. Here, a typical approach is to employ SPICE sub-circuit based macro-modeling using BSIM4 as a core primitive with layout-dependence device parameters, such as  $V_{th}$ , carrier mobility, body-bias coefficient, saturation velocity, drain-induced barrier lowering (DIBL), and subthreshold swing parameter, corrected with the layout-dependent effects. In this approach, tens of layout-dependence instance parameters are introduced, and thousands of lines of layout-dependence model formulations are coded in a SPICE input deck in the ASCII format. Consequently, more simulation overhead associated with parameter parsing and searching, expression evaluation, and device isomorphic matching results. The turnaround time in circuit design and verification becomes longer. More efficient layout-dependence modeling solutions continue to be pursued in SPICE modeling and simulation as well as layout parasitics extraction. This is an important problem for Design For Manufacturing (DFM).

## 2.11 Chapter Summary

This chapter presents and analyzes the fundamental physical effects of a modern MOSFET transistor and their SPICE modeling methodology and BSIM4 implementations. This chapter provides the stage for the chapters to be presented in this book.

## 2.12 Parameter Table

Name (type)	Description and default	Can be binned?	Note
LEVEL (Global; integer)	The SPICE3 device model level selector.  Default = 14 (BSIM4) together with the keyword NMOS or PMOS specified in the model cards; dimensionless.	No	-

<b>VERSION</b> (Global; string)	BSIM4 model version number. The latest version is BSIM4.7.0.  Default = “4.7.0”; dimensionless.	No	-
<b>BINUNIT</b> (Global; integer)	Model parameter geometrical binning model selector.  Default = 1; dimensionless. The other optional value is 0.	No	-
<b>MTRLMOD</b> (Global; integer)	Gate stack and substrate material model selector.  Default = 0 (Poly Si/oxide/Si MOSFETs); dimensionless. The other optional value is 1 (High- <i>k</i> and metal gates).	No	-
<b>TEMPMOD</b> (Global; integer)	Temperature-dependence model selector.  Default = 0; dimensionless. The other optional values are 1, 2, and 3.	No	-
<b>RDSMOD</b> (Global; integer)	Bias-dependent LDD source and drain resistance model and topology selector.  Default = 0 (Resistances are integrated into the channel current model formulation; dimensionless. The other optional value is 1(the resistances are lumped into the source and drain diffusion resistances and connected between the internal and external source and drain nodes).	No	-
<b>MOBMOD</b> (Global; integer)	Mobility model selector.  Default = 0. The other optional values are 1, 2, and 3.	No	-
<b>WPEMOD</b> (Global; integer)	The well-proximity effect model.  Default = 0; dimensionless. The other optional value is 1.	No	-
<b>L</b> (Local; double)	Drawn length of the gate.  Default = $5.0 \times 10^{-6}$ in [m].	No	A fatal error message will be issued if ( <i>L</i> + <i>XL</i> ) is less than or equal to <i>XGL</i> .

W (Local; double)	Drawn width of the gate.  Default = $5.0 \times 10^{-6}$ in [m].	No	-
SA (Local; double)	Distance of one gate edge to one LOD edge.  Default = 0.0 in [m].	No	-
SB (Local; double)	Distance of the other gate edge to the other LOD edge.  Default = 0.0 in [m].	No	-
SD (Local; double)	Spacing between the edges of two adjacent gates of a multi-finger device.  Default = $2.0 \cdot \text{DMCG}$ in [m].	No	Refer to Chapter 8 for the definition of the parameter DMCG.
SCA (Local; double)	Integral of the first-order photoresist-scattered dopant distribution function.  Default = 0.0; dimensionless.	No	A warning message will be issued if it is less than zero and it will be reset to zero when WPEMOD is set to 1.
SCB (Local; double)	Integral of the second-order photoresist-scattered dopant distribution function.  Default = 0.0; dimensionless.	No	A warning message will be issued if it is less than zero and it will be reset to zero when WPEMOD is set to 1.
SCC (Local; double)	Integral of the third-order photoresist-scattered dopant distribution function.  Default = 0.0; dimensionless.	No	A warning message will be issued if it is less than zero and it will be reset to zero when WPEMOD is set to 1.

SC (Local; double)	Distance from the gate edge to the edge of the well.  Default = 0.0 in [m].	No	A warning message will be issued if it is less than zero and it will be reset to zero when WPEMOD is set to 1.
TOXE (Global; double)	(Equivalent) electrical gate oxide thickness.  Default = $30.0 \times 10^{-10}$ in [m].	No	If it is not positive, a fatal error message will be issued. When MTRLMOD = 1, it is EOT.
TOXM (Global; double)	Target (equivalent) electrical gate oxide thickness at which the BSIM4 model parameters are extracted.  Default = TOXE in [m].	No	If it is not positive, a fatal error message will be issued.
TOXP (Global; double)	(Equivalent) physical gate oxide thickness.  Default = TOXE in [m].	No	If it is not positive, a fatal error message will be issued. When MTRLMOD = 1, it is computed by BSIM4 automatically.
TOXREF (Global; double)	The target gate oxide/dielectric layer thickness for the gate direct-tunneling model [refer to the chapter on the gate-direct tunneling current model].  Default = $30.0 \times 10^{-10}$ in [m].	No	If it is not positive, a fatal error message will be issued.

70 BSIM4 AND MOSFET MODELING FOR IC SIMULATION  
By Weidong Liu and Chenming Hu

DTOX (Global; double)	The difference between TOXE and TOXP.  Default = 0.0 in [m].	No	If it is not equal to the difference between TOXE and TOXP, it will be ignored by BSIM4.
EOT (Global; double)	Equivalent oxide thickness of the gate dielectric layer for MTRLMOD = 1.  Default = $15.0 \times 10^{-10}$ in [m].	No	If it is not positive, a fatal error message will be issued.
VDDEOT (Global; double)	The power supply voltage for the gate voltage at which EOT is extracted from measured gate capacitance when MTRLMOD = 1.  Default = 1.5 for NMOS and -1.5 for PMOS in [V].	No	If it is not positive, a fatal error message will be issued.
TEMPEOT (Global; double)	The temperature at which TOXP is extracted from EOT when MTRLMOD = 1.  Default = 300.15 in [Kelvin].	No	-
LEFFEOT (Global; double)	The effective channel length at which TOXP is extracted from EOT when MTRLMOD = 1.  Default = $1.0 \times 10^{-6}$ in [m].	No	-
WEFFEOT (Global; double)	The effective channel width at which TOXP is extracted from EOT when MTRLMOD = 1.  Default = $10.0 \times 10^{-6}$ in [m].	No	-
EPSROX (Global; double)	Relative dielectric constant of the gate dielectric layer for MTRLMOD = 0.  Default = 3.9; dimensionless.	No	If it is less than 0.0, a fatal error message will be issued.
EPSRSUB (Global; double)	Relative dielectric constant of the substrate for MTRLMOD = 1.  Default = 11.7; dimensionless.	No	-

<b>EPSRGATE</b> (Global; double)	Relative dielectric constant of the gate. If <b>MTRLMOD</b> = 0, the material is poly silicon. If it is zero, a metal gate is implied and no gate depletion will take place and be modeled.  Default = 11.7; dimensionless.	No	If it is less than 0.0, a fatal error message will be issued.
<b>XL</b> (Global; double)	Offset between designed and actual physical gate lengths.  Default = 0.0 in [m].	No	If it makes $L_{physical}$ less than <b>XGL</b> (refer to Chapter 6), a fatal error message will be issued.
<b>XW</b> (Global; double)	Offset between designed and actual physical gate widths.  Default = 0.0 in [m].	No	-
<b>LINT</b> (Global; double)	Gate-source or -drain overlap length.  Default = 0.0 in [m].	No	-
<b>DLC</b> (Global; double)	Gate-source or -drain overlap length for capacitance models.  Default = <b>LINT</b> in [m].	No	-
<b>WINT</b> (Global; double)	Gate and channel edge overlap length in the width direction.  Default = 0.0 in [m].	No	-
<b>DWC</b> (Global; double)	Gate and channel edge overlap length in the width direction for capacitance models.  Default = <b>WINT</b> in [m].	No	-
<b>DWJ</b> (Global; double)	Gate and channel edge overlap length in the width direction for junction diode models.  Default = <b>DWC</b> in [m].	No	-
<b>LL, LW, LWL, LLN, and LWN</b> (Global; double)	Length and width dependence parameters of <b>LINT</b> .  Default = 0.0. Their dimensions are given such that the overlap length $\Delta L$ is in meters.	No	-

72 BSIM4 AND MOSFET MODELING FOR IC SIMULATION  
By Weidong Liu and Chenming Hu

LLC, LWC, and LWLC (Global; double)	Length and width dependence parameters of DLC for capacitance models.  Default = LL, LW, and LWL, respectively.	No	-
WL, WW, WWL, WLN, and WWN (Global; double)	Length and width dependence parameters of WINT.  Default = 0.0. Their dimensions are given such that the overlap length $\Delta W$ is in meters.	No	-
WLC, WWC, and WWLC (Global; double)	Length and width dependence parameters of DWC.  Default = WL, WW, and WWL, respectively.	No	-
TNOM (Global; double)	Nominal temperature at which model parameters are extracted.  Default = 300.15 in [Kelvin].	No	-
BG0SUB (Global; double)	Band gap of the substrate at 0 Kelvin for MTRLMOD = 1.  Default = 1.16 in [eV].	No	-
TBGASUB (Global; double)	Temperature dependence parameter of the band gap of the substrate for MTRLMOD = 1.  Default = $7.02 \times 10^{-4}$ in [eV/Kelvin].	No	-
TBGBSUB (Global; double)	Temperature dependence parameter of the band gap of the substrate for MTRLMOD = 1.  Default = 1108.0 in [Kelvin].	No	-
NI0SUB (Global; double)	Intrinsic carrier concentration of the substrate at 300.15 degree Kelvin for MTRLMOD = 1.  Default = $1.45 \times 10^{10}$ in [ $\text{cm}^{-3}$ ].	No	A fatal error message will be issued if it is not positive.
NSD (Global; double)	Doping concentrations of the source and drain diffusion regions.  Default = $1.0 \times 10^{20}$ in [ $\text{cm}^{-3}$ ].	Yes	-

NDEP (Global; double)	Channel doping concentrations.  Default = $1.7 \times 10^{17}$ in [cm <sup>-3</sup> ].	Yes	If its binned value is not positive, a fatal error message will be issued.
NGATE (Global; double)	The doping concentration of poly-silicon gate  Default = 0.0 in [cm <sup>-3</sup> ].	Yes	If its binned value is less than zero, a fatal error message will be issued.
NSUB (Global; double)	Substrate doping concentrations.  Default = $6.0 \times 10^{16}$ in [cm <sup>-3</sup> ].	Yes	If its binned value is not positive, a fatal error message will be issued.
VTH0 or VTHO (Global; double)	Long-channel zero body bias threshold voltage.  Default = Computed if not given, in unit of [V].	Yes	-
VFB (Global; double)	Flat-band voltage.  Default = Computed if not given, in unit of [V].	No	-
DVT0 (Global; double)	$V_{th}$ roll-off coefficient parameter.  Default = 2.2; dimensionless.	Yes	If its binned value is less than 0, a warning message will be issued.
DVT1 (Global; double)	$V_{th}$ roll-off length-dependence parameter.  Default = 0.53; dimensionless.	Yes	If its binned value is less than 0, a fatal error message will be issued.
DVT2 (Global; double)	Body-bias coefficient parameter of the characteristic length of the $V_{th}$ roll-off model.  Default = -0.032 in [V <sup>-1</sup> ].	Yes	-

74 BSIM4 AND MOSFET MODELING FOR IC SIMULATION  
By Weidong Liu and Chenming Hu

K1 (Global; double)	First-order body bias coefficient of the $V_{th}$ model.  Default = 0.53 in [ $V^{-1/2}$ ].	Yes	If it is not given but K2 is, it will be set to 0.53. If neither is given, it will be computed.
K2 (Global; double)	Second-order body bias coefficient of the $V_{th}$ model for non-uniform vertical doping.  Default = -0.0186; dimensionless.	Yes	If it is not given but K1 is, it will be set to -0.0186. If neither is given, it will be computed.
XT (Global; double)	The depth of the delta doping.  Default = $1.55 \times 10^{-7}$ in [m].	Yes	-
VBM (Global; double)	The maximum body voltage applicable to a given process technology.  Default = -3.0 in [V].	Yes	-
GAMMA1 (Global; double)	First-order body bias coefficient of the $V_{th}$ model.  Default = Computed from NDEP if not specified, in unit of [ $V^{-1/2}$ ].	Yes	-
GAMMA2 (Global; double)	First-order body bias coefficient of the $V_{th}$ model.  Default = Computed from NSUB if not specified, in unit of [ $V^{-1/2}$ ].	Yes	-
VBX (Global; double)	The body voltage at which the depletion width is equal to XT.  Default = Computed if not specified, in unit of [V].	Yes	-
K3 (Global; double)	Narrow-width effect coefficient parameter for the $V_{th}$ model.  Default = 80.0; dimensionless.	Yes	-
K3B (Global; double)	Narrow-width effect coefficient parameter for the $V_{th}$ model.  Default = 0.0 in [ $V^{-1}$ ].	Yes	-

PHIN (Global; double)	Surface potential offset parameter owing to non-uniform vertical doping.  Default = 0.0 in [V].	Yes	A fatal error message will be issued if it produces a negative surface potential.
W0 (Global; double)	Narrow-width effect model parameter for the widest device for the $V_{th}$ model.  Default = $2.5 \times 10^{-6}$ in [m].	Yes	A fatal error message will be issued if its binned value has the opposite value $W_{eff}$ because of divide-by-zero error.
DVT0W (Global; double)	Narrow-width and short-channel coefficient parameter for the $V_{th}$ model.  Default = 0.0; dimensionless.	Yes	-
DVT1W (Global; double)	Narrow-width and short-channel dependence parameter for the $V_{th}$ model.  Default = $5.3 \times 10^{-6}$ ; dimensionless.	Yes	If its binned value is less than 0, a fatal error message will be issued.
DVT2W (Global; double)	Body-bias coefficient parameter of the width-dependence characteristic length for the $V_{th}$ model.  Default = -0.032 in [ $V^{-1}$ ].	Yes	-
DSUB (Global; double)	Length-dependence parameter for the $V_{th}$ DIBL effects.  Default = DROUT = 0.56; dimensionless.	Yes	If its binned value is negative, a fatal error message will be issued.
ETA0 (Global; double)	$V_{ds}$ -dependence coefficient for the $V_{th}$ DIBL model.  Default = 0.08; dimensionless.	Yes	If its binned value is negative, a warning message will be issued.

76 BSIM4 AND MOSFET MODELING FOR IC SIMULATION  
By Weidong Liu and Chenming Hu

ETAB (Global; double)	Body-bias coefficient for the $V_{th}$ DIBL model.  Default = $-0.07$ ; dimensionless.	Yes	If its binned value is negative, a warning message will be issued.
LPE0 (Global; double)	$V_{th}$ roll-up parameter owing to pocket implants. It models the increase in the equivalent channel doping.  Default = $1.74 \times 10^{-7}$ in [m].	Yes	If its binned value is less than $L_{eff}$ , a fatal error message will be issued.
LPEB (Global; double)	$V_{th}$ roll-up parameter owing to pocket implants. It models the increase in the body-bias effect because of the increase in the equivalent channel doping.  Default = 0.0 in [m].	Yes	If its binned value is less than $L_{eff}$ , a fatal error message will be issued.
DVTP0 (Global; double)	Length-dependence parameter of DITS (drain-induced threshold shifts) of pocket-implanted devices.  Default = 0.0 in [m].	Yes	If its binned value is not positive, no DITS will be modeled.
DVTP1 (Global; double)	Drain bias dependence parameter of DITS (drain-induced threshold shifts) of pocket-implanted devices.  Default = 0.0 in [ $V^{-1}$ ].	Yes	-
XJ (Global; double)	Source and drain junction depth.  Default = $1.5 \times 10^{-7}$ in [m].	Yes	If its binned value is negative, a fatal error message will be issued.
KT1 (Global; double)	Temperature-dependence coefficient parameter of the threshold voltage model.  Default = $-0.11$ in [V].	Yes	-
KT1L (Global; double)	Channel-length dependence parameter of the threshold voltage temperature-dependence model.  Default = 0.0 in [V · m].	Yes	-

KT2 (Global; double)	Body bias coefficient parameter for the threshold voltage temperature-dependence model.  Default = 0.022; dimensionless.	Yes	-
A0 (Global; double)	Length-dependence parameter of the $A_{bulk}$ model.  Default = 1.0; dimensionless.	Yes	-
B0 (Global; double)	The first width-dependence parameter of the $A_{bulk}$ model.  Default = 0.0 in [m].	Yes	-
B1 (Global; double)	The second width-dependence parameter of the $A_{bulk}$ model.  Default = 0.0 in [m].	Yes	If the sum its binned value and $W_{eff}$ becomes zero, a fatal error message will be issued.
KETA (Global; double)	Body-bias coefficient of the $A_{bulk}$ model.  Default = -0.047 in [ $V^{-1}$ ].	Yes	-
AGS (Global; double)	Gate bias dependence parameter of the $A_{bulk}$ model.  Default = 0.0 in [ $V^{-1}$ ].	Yes	-
RDSW (Global; double)	Zero-bias source and drain LDD resistance component per unit width for RDSMOD=0.  Default = 200.0 in [ $\text{ohm} \cdot (\mu\text{m})^{WR}$ ].	Yes	A warning message will be issued if its binned value is less than zero. It will then be reset to zero.
AGS (Global; double)	Gate bias dependence parameter of the $A_{bulk}$ model.  Default = 0.0 in [ $V^{-1}$ ].	Yes	-

RDSW (Global; double)	Zero-bias source and drain LDD resistance component per unit width for RDSMOD = 0.  Default = 200.0 in [ohm · ( $\mu\text{m}$ ) <sup>WR</sup> ].	Yes	A warning message will be issued if its binned value is less than zero. It will then be reset to zero.
WR (Global; double)	The power of the width dependence of source and drain LDD resistances.  Default = 1.0; dimensionless.	Yes	-
RDSWMIN (Global; double)	Source and drain LDD resistance component per unit width at zero body bias and high gate voltages for RDSMOD = 0.  Default = 0.0 in [ohm · ( $\mu\text{m}$ ) <sup>WR</sup> ].	No	A warning message will be issued if it is less than zero. It will then be reset to zero.
PRWG (Global; double)	Gate bias dependence parameter of the LDD source and drain resistances.  Default = 1.0 in [V <sup>-1</sup> ].	Yes	A warning message will be issued if its binned value is less than zero. It will then be reset to zero.
PRWB (Global; double)	Body bias dependence parameter of the LDD source and drain resistances.  Default = 0.0 in [V <sup>-0.5</sup> ].	Yes	-
RDW (Global; double)	Zero-bias drain LDD resistance component per unit width for RDSMOD = 1.  Default = 100.0 in [ohm · ( $\mu\text{m}$ ) <sup>WR</sup> ].	Yes	A warning message will be issued if its temperature de-rated value at $T_{emp}$ is negative. Its temperature de-rated value will then be reset to zero.

<b>RDWMIN</b> (Global; double)	Drain LDD resistance component per unit width at zero body bias and high gate voltages for RDMSMOD = 1.  Default = 0.0 in [ohm · ( $\mu\text{m}$ ) <sup>WR</sup> ].	No	A warning message will be issued if its temperature de-rated value at $T_{emp}$ is negative. Its temperature de-rated value will then be reset to zero.
<b>RSW</b> (Global; double)	Zero-bias source LDD resistance component per unit width for RDMSMOD = 1.  Default = 100.0 in [ohm · ( $\mu\text{m}$ ) <sup>WR</sup> ].	Yes	A warning message will be issued if its temperature de-rated value at $T_{emp}$ is negative. Its temperature de-rated value will then be reset to zero.
<b>RSWMIN</b> (Global; double)	Source LDD resistance component per unit width at zero body bias and high gate voltages for RDMSMOD = 1.  Default = 0.0 in [ohm · ( $\mu\text{m}$ ) <sup>WR</sup> ].	No	A warning message will be issued if its temperature de-rated value at $T_{emp}$ is negative. Its temperature de-rated value will then be reset to zero.
<b>PRT</b> (Global; double)	Temperature dependence coefficient of the source and drain LDD resistances.  Default = 0.0 in [ohm · ( $\mu\text{m}$ ) <sup>WR</sup> ] for TEMPMOD = 0 and 0.0 in [1/Kelvin] for other TEMPMOD settings (= 1, 2 or 3).	Yes	-
<b>ADOS</b> (Global; double)	Inversion charge thickness parameter to account the density of states, used in both the DC channel current and the CAPMOD = 2 charge and capacitance models.  Default = 1.0; dimensionless.	No	-

BDOS (Global; double)	Inversion charge thickness power parameter to account the density of states, used in both the DC channel current and the CAPMOD = 2 charge and capacitance models.  Default = 1.0; dimensionless.	No	-
PHIG (Global; double)	Work function of (metal) gates.  Default = 4.05 in [V].	No	-
EASUB (Global; double)	Electron affinity of substrates.  Default = 4.05 in [V].	No	A fatal error message will be issued if it is negative.
U0 (Global; double)	Low-field mobility.  Default = 0.067 for NMOS and 0.025 for PMOS in [ $\text{m}^2/(\text{V} \cdot \text{s})$ ].	Yes	A fatal error message will be issued if its temperature de-rated value at $T_{emp}$ is not positive.
UP (Global; double)	Coefficient of the channel-length dependence of the low-field mobility U0.  Default = 0.0; dimensionless.	Yes	-
LP (Global; double)	Exponent of the channel-length dependence of the low-field mobility U0.  Default = $1.0 \times 10^{-8}$ in [m].	Yes	-
UA (Global; double)	Gate bias dependence parameter for the effective mobility model.  Default = $1.0 \times 10^{-15}$ for MOBMOD = 2 and $1.0 \times 10^{-9}$ for all others in [m/V].	Yes	-
UB (Global; double)	Gate bias dependence parameter for the effective mobility model.  Default = $1.0 \times 10^{-19}$ in [(m/V) <sup>2</sup> ].	Yes	-
UC (Global; double)	Body bias dependence parameter for the effective mobility model.  Default = -0.0465 for MOBMOD = 1 and - $0.0465 \times 10^{-9}$ for all others in [m/V <sup>2</sup> ] for MOBMOD = 0; V <sup>-1</sup> for MOBMOD = 1; and m <sup>EU</sup> /V <sup>EU+1</sup> for MOBMOD = 2 and 3].	Yes	-

UD (Global; double)	Coulomb scattering dependence parameter for the effective mobility model.  Default = 0.0 in [ $\text{m}^{-2}$ for MOBMOD = 0, 1, and 2; dimensionless for MOBMOD = 3].	Yes	-
EU (Global; double)	Power of the field dependence for the Coulomb scattering of the effective mobility model MOBMOD = 2 and 3.  Default = 1.67 for NMOS and 1.0 for PMOS; dimensionless.	Yes	A warning message will be issued if its binned value is negative. It will be reset to 0.0.
UCS (Global; double)	Power of the gate bias dependence for the Coulomb scattering of the effective mobility model MOBMOD = 3.  Default = 1.67 for NMOS and 1.0 for PMOS; dimensionless.	Yes	A warning message will be issued if its binned value is negative. It will be reset to 0.0.
UTE (Global; double)	Power of low-field mobility U0 temperature dependence.  Default = -1.5; dimensionless.	Yes	-
UCSTE (Global; double)	Power of temperature dependence of UCS.  Default = $-4.775 \times 10^{-3}$ ; dimensionless.	Yes	-
UA1 (Global; double)	Temperature dependence coefficient of the parameter UA.  Default = $1.0 \times 10^{-9}$ in [The same unit as UA for TEMPMOD = 0; 1/Kelvin for TEMPMOD = 1 and 2; and dimensionless for TEMPMOD = 3].	Yes	-
UB1 (Global; double)	Temperature dependence coefficient of the parameter UB.  Default = $-1.0 \times 10^{-18}$ in [The same unit as UB for TEMPMOD = 0; 1/Kelvin for TEMPMOD = 1 and 2; and dimensionless for TEMPMOD = 3].	Yes	-
UC1 (Global; double)	Temperature dependence coefficient of the parameter UC.	Yes	-

	Default = $-0.056$ for MOBMOD = 1 and $-0.056 \times 10^{-9}$ for all others in [The same unit as UC for TEMPMOD = 0; 1/Kelvin for TEMPMOD = 1 and 2; and dimensionless for TEMPMOD = 3].		
UD1 (Global; double)	Temperature dependence coefficient of the parameter UD.  Default = 0.0 in [The same unit as UD for TEMPMOD = 0; 1/Kelvin for TEMPMOD = 1 and 2; and dimensionless for TEMPMOD = 3].	Yes	-
SAREF (Global; double)	Reference spacing in the channel length direction between one edge of the OD and one edge of the gate.  Default = $1.0 \times 10^{-6}$ in [m].	No	A fatal error message will be issued if it is not positive when the LOD model is turned on.
SBREF (Global; double)	Reference spacing in the channel length direction between the other edge of the OD and the other edge of the gate.  Default = $1.0 \times 10^{-6}$ in [m].	No	A fatal error message will be issued if it is not positive when the LOD model is turned on.
KU0 (Global; double)	Mobility enhancement/reduction coefficient because of LOD.  Default = 0.0 in [m].	No	-
TKU0 (Global; double)	Temperature dependence parameter for KU0 of the LOD model.  Default = 0.0 in [1/Kelvin].	No	-
LKU0 (Global; double)	Length dependence parameter of KU0.  Default = 0.0 in [ $m^{LLODKU0}$ ].	No	-
WKU0 (Global; double)	Width dependence parameter of KU0.  Default = 0.0 in [ $m^{WLODKU0}$ ].	No	-
PKU0 (Global; double)	Width and length cross-term dependence parameter of KU0.  Default = 0.0 in [ $m^{LLODKU0+WLODKU0}$ ].	No	-

LLODKU0 (Global; double)	The power of the length dependence of KU0. Default = 0.0; dimensionless.	No	-
WLODKU0 (Global; double)	The power of the width dependence of KU0. Default = 0.0; dimensionless.	No	-
WLOD (Global; double)	Channel width offset parameter for LOD. Default = 0.0 in [m].	No	-
KVSAT (Global; double)	Saturation velocity enhancement/reduction coefficient because of LOD.  Default = 0.0; dimensionless	No	A warning message will be issued and it will be reset to -1.0 if less than -1.0. A warning message will be issued and it will be reset to 1.0 if greater than 1.0.
KVTH0 (Global; double)	Threshold voltage increase/decrease parameter because of LOD.  Default = 0.0 in [V · m].	No	-
LKVTH0 (Global; double)	Length dependence parameter of KVTH0.  Default = 0.0 in [ $m^{LLODVTH}$ ].	No	-
WKVTH0 (Global; double)	Width dependence parameter of KVTH0.  Default = 0.0 in [ $m^{WLODVTH}$ ].	No	-
PKVTH0 (Global; double)	Width and length cross-term dependence parameter of KVTH0.  Default = 0.0 in [ $m^{LLODVTH+WLODVTH}$ ].	No	-
LLODVTH (Global; double)	The power of the length dependence of KVTH0. Default = 0.0; dimensionless.	No	-
WLODVTH (Global; double)	The power of the width dependence of KVTH0. Default = 0.0; dimensionless.	No	-

STK2 (Global; double)	K2 increase/decrease parameter because of LOD.  Default = 0.0 in [m].	No	-
LODK2 (Global; double)	The power of the geometry dependence of STK2 because of LOD.  Default = 1.0; dimensionless.	No	A warning message will be issued when it is not positive and when the LOD model evaluation is turned on.
STETA0 (Global; double)	ETA0 increase/decrease parameter because of LOD.  Default = 0.0 in [m].	No	-
LODETA0 (Global; double)	The power of the geometry dependence of STETA0 because of LOD.  Default = 1.0; dimensionless.	No	A warning message will be issued when it is not positive and when the LOD model evaluation is turned on.
SCREF (Global; double)	The reference value of SC of the WPE model.  Default = $1 \times 10^{-6}$ in [m].	No	A warning message will be issued if it is less than or equal to zero when the WPE model is turned. It is then reset to $1.0 \times 10^{-6}$ m.
KU0WE (Global; double)	Coefficient for the low-field mobility U0 because of WPE.  Default = 0.0; dimensionless.	Yes	-
KVTH0WE (Global; double)	Coefficient for VTH0 because of WPE.  Default = 0.0 in [V].	Yes	-

K2WE (Global; double)	Coefficient for K2 because of WPE.  Default = 0.0; dimensionless.	Yes	-
WEB (Global; double)	Coefficient of SCB because of WPE.  Default = 0.0; dimensionless.	No	-
WEC (Global; double)	Coefficient of SCC because of WPE.  Default = 0.0; dimensionless.	No	-

## References

- [1] Weidong Liu, Xiaodong Jin, Kanyu M. Cao, and Chenming Hu, “BSIM4.0.0 MOSFET Model – User’s Manual,” *Memorandum No. UCB/ERL M00/38*, Electronics Research Laboratory, College of Engineering, University of California, Berkeley, August 3, 2000.
- [2] Tanvir Hasan Morshed, Wenwei (Morgan) Yang, Mohan V. Dunga, Xuemei (Jane) Xi, Jin He, Weidong Liu, Kanyu M. Cao, Xiaodong Jin, Jeff J. Ou, Mansun Chan, Ali M. Niknejad, and Chenming Hu, “BSIM4.6.4 User’s Manual,” University of California, Berkeley, <http://www-device.eecs.berkeley.edu/~bsim3/bsim4.html>, August 2009.
- [3] Z.-H. Liu, C. Hu, J.-H. Huang, T.-Y. Chan, M.-C. Jeng, P. K. Ko, and Y. C. Cheng, “Threshold voltage model for deep-submicrometer MOSFET’s,” *IEEE Trans. on Electron Devices*, vol. 40, no. 1, pp. 86-95, January 1993.
- [4] Weidong Liu and Chenming Hu, “BSIM3v3 MOSFET Model” – Silicon and Beyond: Advanced Device Models and Circuit Simulators,” edited by Michael S. Shur and Tor A. Fjeldly, pp. 1-31, ISBN: 981-02-4280-8, World Scientific, 2000.
- [5] Weidong Liu, Xiaodong Jin, James Chen, Min-Chie Jeng, Zhihong Liu, Yuhua Cheng, Kai Chen, Mansun Chan, Kelvin Hui, Jianhui Huang, Robert Tu, Ping K. Ko, and Chenming Hu, “BSIM3v3.2 MOSFET model — Users’ manual,” *Memorandum No. UCB/ERL M98/51*. Electronics Research Laboratory, College of Engineering, University of California, Berkeley, August 21, 1998.
- [6] Kanyu Mark Cao, Weidong Liu, Xiaodong Jin, Karthik Vasanth, Keith Green, John Krick, Tom Vrotsos, and Chenming Hu, “Modeling of pocket implanted MOSFETs for anomalous analog behavior”, *Tech. Dig. of IEDM*, pp. 171-174, 1999.
- [7] Anant G. Sabnis, and James T. Clemens, “Characterization of the electron mobility in the inverted <100> Si surface,” *Tech. Dig. of IEDM*, pp. 18-21, 1979.
- [8] Mong-Song Liang, Jeong Yeol Choi, Ping-Keung Ko, and Chenming Hu, “Inversion-layer capacitance and mobility of very think gate-oxide MOSFET’s,” *IEEE Trans. Electron Devices*, vol. ED-33, no. 3, pp. 409-413, March 1986.

**This page intentionally left blank**

## Chapter 3

# Channel DC Current and Output Resistance

### 3.1 Introduction and Chapter Objectives

This chapter and the remaining chapters of this book are a continuation of the BSIM4 MOSFET model presented in Chapter 2. This chapter is focused on the channel current model of BSIM4. It will briefly review the channel current theory and formulation, followed by the BSIM4 unified channel charge model for the subthreshold and linear regions of operation. Velocity saturation will then be introduced and discussed. Several physical mechanisms in combination determine the channel current saturation and output resistance  $R_{out}$  characteristics of modern MOSFET devices. In addition, carrier velocity overshoot and source-end velocity limit are believed to play increasingly more important roles for sub-50nm devices. The BSIM4 models for velocity overshoot and source-end velocity limit will also be presented and discussed in this chapter.

The BSIM4 channel charge and current formulation is the cornerstone of BSIM4, which has been used for production from the 130-nanometer down to 20-nanometer CMOS technology nodes. The goal of compact modeling is to capture device physics and the effects of fabrication processes with precise and yet computationally efficient mathematical equations. These equations are required to be simple, accurate, and numerically robust. They must also be flexible for various process technologies and circuit analyses.

### 3.2 Channel Current Theory

The channel current density  $J_{ch}(y)$  of an  $n$ -channel MOSFET is

$$J_{ch}(y) = -q \cdot n(x, y) \cdot \mu_n(x, y) \cdot \frac{d\varphi_{fn}(y)}{dy} \quad (3.1)$$

In this equation,  $n(x, y)$  is the channel electron density,  $\mu_n(x, y)$  is the electron mobility, and  $\varphi_{fn}(y)$  is the channel electron quasi-Fermi potential at the location  $y$ . It includes both drift and diffusion carrier transport. The difference in  $\varphi_{fn}(y)$  between source and drain results in channel carriers drifting and diffusing along the channel from source to drain. The channel current  $I_{ch}(y)$  at this location is obtained by integrating Eq. (3.1) over the channel width  $W_{eff}$  and the depth  $x$ . By assuming that it is uniform along the width direction,  $I_{ch}(y)$  is given by

$$I_{ch}(y) = -W_{eff} \cdot \int_0^{\infty} \left[ q \cdot n(x, y) \cdot \mu_n(x, y) \cdot \frac{d\varphi_{fn}(y)}{dy} \right] \cdot dx \quad (3.2)$$

Further assuming  $\mu_n$  is independent of  $x$ , we get

$$I_{ch}(y) = -W_{eff} \cdot q_{ch}(y) \cdot \mu_n(y) \cdot \frac{d\varphi_{fn}(y)}{dy} \quad (3.3)$$

where

$$q_{ch}(y) = \int_0^{\infty} [q \cdot n(x, y)] \cdot dx \quad (3.4)$$

$\varphi_{fn}(y)$  is equal to the source-to-body voltage  $V_{sb}$  at the source end of the channel and the drain-to-body voltage  $V_{db}$  at the drain side. In the inversion range,

$$\varphi_{fn}(y) \approx V(y) + V_{sb} \quad (3.5)$$

which permits the integration of Eq. (3.3) after multiplying both sides by  $dy$

$$I_{ch} = -\frac{W_{eff}}{L_{eff}} \cdot \int_0^{V_{ds}} [q_{ch}(y) \cdot \mu_n(y)] \cdot dV(y) \quad (3.6)$$

There are several approaches to obtain the solutions of Eqs. (3.3) and (3.4). The Pao-Sah model [1] provides a general treatment but it can only be solved numerically. The charge-sheet approach [2] assumes zero inversion-layer thickness that has no voltage drop across it and a depletion region free of mobile carriers. A rigorous charge-sheet model requires numerical iteration in solving for the surface potential. The BSIM4 model provides an efficient approximate channel charge model

for both the inversion and the subthreshold operation to speed up circuit simulation.

### 3.3 Single Continuous Channel Charge Model

BSIM4, similar to BSIM3v3, uses a single, continuous channel charge formulation to ensure the continuity and smoothness of the channel charge and current model and their first- and higher-order derivatives. Applications of this model, from the 130nm technology node to the state-of-the-art technology today, have proven its deliveries of excellent accuracy and also robust numerical convergence in SPICE simulation.

In the inversion region, the channel charge density is

$$q_{inv}(y) = -C_{oxeff} \cdot [V_{gse} - V_{th} - A_{bulk} \cdot V(y)] \quad (3.7)$$

As given in Chapter 2,  $C_{oxeff}$  is the electrical gate capacitance with the inversion charge layer thickness considered (owing to the quantum mechanics effects),  $V_{gse}$  is the effective gate bias containing the poly-Si gate depletion effect,  $V_{th}$  is the threshold voltage, and  $A_{bulk}$  is the bulk charge effect coefficient.  $V(y)$  is the channel potential at  $y$ . Eq. (3.7) can be transformed into

$$q_{inv}(y) = q_{inv0} \cdot \left[ 1 - \frac{A_{bulk} \cdot V(y)}{V_{gse} - V_{th}} \right] \quad (3.8)$$

where  $q_{inv0} = -C_{oxeff} \cdot (V_{gse} - V_{th})$  is the inversion channel charge density (per unit area) under zero drain biases.

In the subthreshold bias range, diffusion dominates. According to the diffusion theory, the channel charge density along the channel is

$$q_{subVth}(y) = -C_{dep0} \cdot \nu_t \cdot \exp\left(\frac{V_{gse} - V_{th} - V_{off}^*}{n \cdot \nu_t}\right) \cdot \exp\left[-\frac{A_{bulk} \cdot V(y)}{n \cdot \nu_t}\right] \quad (3.9)$$

where the thermal voltage is defined by

$$\nu_t = \frac{k_B \cdot T_{emp}}{q} \quad (3.9a)$$

$C_{dep0}$  is the depletion layer capacitance per unit area under zero body bias

$$C_{dep0} = \sqrt{\frac{q \cdot NDEP \cdot 10^6 \cdot \epsilon_{sub}}{2 \varphi_s}} \quad (3.9b)$$

In Eq. (3.9),  $n$  is the subthreshold swing parameter. It has complex geometry and bias dependencies. A detailed discussion will be given later in this section.  $V_{off}^*$  is the offset voltage between the threshold voltage  $V_{th}$  that determines the strong inversion charge (and current) in Eq. (3.7) and the “ $V_{th}$ ” that determines the subthreshold charge and current.  $V_{off}^*$ , in conjunction with the  $n$  factor, determines the channel current.  $V_{off}^*$  is temperature dependent and varies with the channel length (because of non-uniform lateral pocket doping profile).

$$V_{off}^* = \left( V_{OFF} + \frac{V_{OFFL}}{L_{eff}} \right) \cdot [1 + TV_{OFF} \cdot (T_{emp} - TNOM)] \quad (3.9c)$$

By applying Taylor series expansion and keeping the first two terms, Eq. (3.9) is approximated as

$$q_{subVth}(y) \approx q_{subVth0} \cdot \left[ 1 - \frac{A_{bulk} \cdot V(y)}{n \cdot v_t} \right] \quad (3.10)$$

with

$$q_{subVth0} = -C_{dep0} \cdot v_t \cdot \exp\left(\frac{V_{gse} - V_{th} - V_{off}^*}{n \cdot v_t}\right) \quad (3.10a)$$

An examination of Eqs. (3.8) and (3.10) suggests that one can unify them into a single continuous channel charge model that is valid from subthreshold to strong inversion

$$q_{ch}(y) = q_{cho} \cdot \left[ 1 - \frac{A_{bulk}}{V_{gsteff} + 2 \cdot v_t} \cdot V(y) \right] \quad (3.11)$$

in which the constant coefficient 2 is substituted for the  $n$  factor in Eq.(3.10) where  $n$  has very little effect on accuracy.

$q_{cho}$  is the channel charge density under  $V_{ds} = 0$  and is taken as

$$q_{cho} = \frac{q_{inv0} \cdot q_{subVth0}}{q_{inv0} + q_{subVth0}} \quad (3.12)$$

Upon substituting the formulations of  $q_{inv0}$  and  $q_{subVth0}$  and after a few mathematical simplifications,  $q_{cho}$  can now be expressed in a concise form

$$q_{cho} = -C_{oxeff} \cdot V_{gsteff} \quad (3.13)$$

where the effective gate bias  $V_{gsteff}$  is

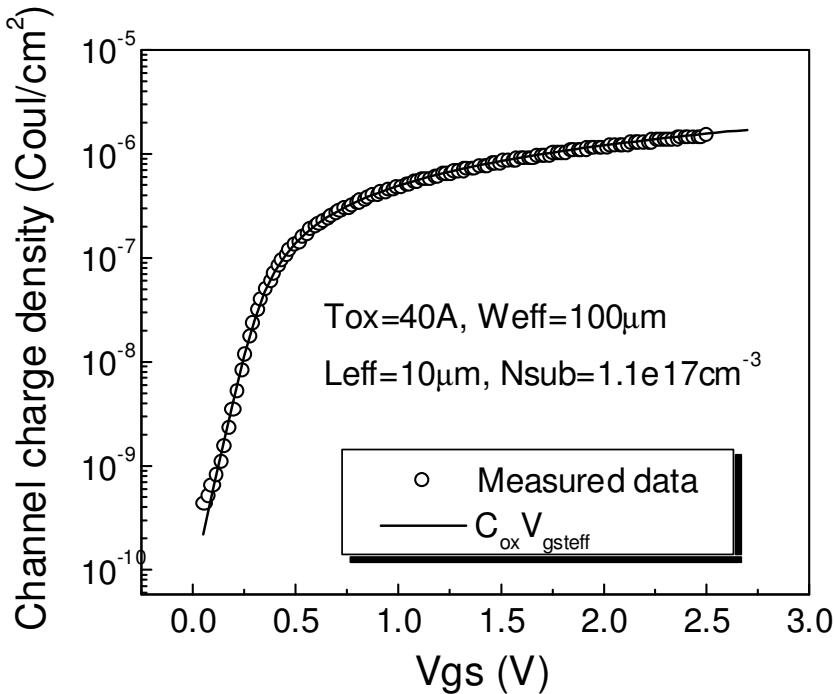
$$V_{gsteff} = \frac{n \cdot v_t \ln \left\{ 1 + \exp \left[ \frac{m^* (V_{gse} - V_{th})}{n \cdot v_t} \right] \right\}}{m^* + n \cdot \frac{C_{oxeff}}{C_{dep0}} \cdot \exp \left[ - \frac{(1 - m^*) (V_{gse} - V_{th}) - V_{off}^*}{n \cdot v_t} \right]} \quad (3.14)$$

In this formulation, a model parameter **MINV** is incorporated into the  $m^*$  term to further improve the model accuracy in the moderate inversion region for  $G_m$ ,  $G_m/I_d$ , and  $G_m^2/I_d$ , which are critical for analog and RF IC design.

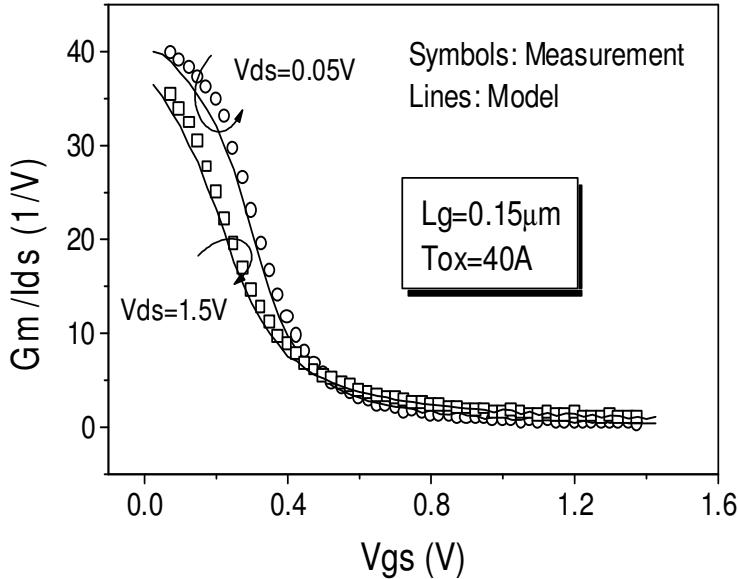
$$m^* = 0.5 + \frac{\arctan(\text{MINV})}{\pi} \quad (3.15)$$

$m^*$  is in the range of 0 and 1 and has a typical value around 0.5. Here, an *arctan* function is utilized for easy parameter extraction and optimization of **MINV**.

Note that with  $V_{gsteff}$  defined in Eq. (3.14), the unified channel charge model  $q_{ch0}$  of Eq. (3.13) transitions smoothly to  $q_{inv0} = -C_{oxeff} \cdot (V_{gse} - V_{th})$  in the inversion region and become  $q_{subVth0}$  of Eq. (3.10a) in the sub-threshold region. The accuracy of this model is good for many generations of technologies and exemplified in Fig. 3.1 [3].



(a)



(b)

Fig. 3.1 Comparison of the unified channel charge model and the circuit gain versus bias current parameter  $G_m/I_{ds}$  against measured data. (a) Channel charge density versus the gate voltage. (b)  $G_m/I_{ds}$  versus the gate voltage.

In the remainder of this subsection, the subthreshold swing parameter  $n$  will be discussed. This parameter determines how rapidly the MOS transistors can be turned on and off when sweeping the gate voltage. The smaller  $n$  is, the steeper the  $\log(I_d)$  versus  $V_{gs}$  transfer curve will be. This parameter is closely linked to the subthreshold swing  $S$ -factor of MOSFETs. The  $S$ -factor describes the amount of the gate voltage change that is required in order to increase or decrease the subthreshold channel current by one order of magnitude. It is given in the unit of mV/decade, typically around 90mV/decade for bulk CMOS and 70mv/decade for SOI and FinFETs or multi-gate transistors. The theoretical ideal minimum  $S$ -factor for MOSFETs is  $(k_B T_{emp}/q)\log_e 10 \cong 60\text{mV/decade}$  at room temperature.  $n$  is defined such that  $S = n \cdot 60\text{mV/decade}$ . A small  $S$  or small  $n$  is the result of a good device design that provides a stronger gate control over the channel and a steeper turn on/off, which helps to reduce the subthreshold channel leakage current.

According to MOS device physics, the steepness of the transfer curve reflects the ratio of the delta surface potential to a delta gate voltage. This ratio is determined by a capacitor voltage divider circuit. The

capacitances involved are the surface depletion capacitance  $C_{dep}$ , the interface capacitance CIT representing the effect of charging and discharging of interface states, as well as the coupling capacitance CDSC between drain/source and the channel [4]. Therefore,  $n$  depends on bias, geometry, and operating temperature.

The  $n$  parameter of BSIM4 is formulated by

$$n = 1 + \text{NFACTOR} \cdot \frac{C_{dep}}{C_{oxe}} + \frac{\text{CIT}}{C_{oxe}} + \frac{\text{CDSC} + \text{CDSCB} \cdot V_{bseff} + \text{CDSCD} \cdot V_{ds}}{C_{oxe}} \cdot \frac{0.5}{\cosh(DVT1 \cdot L_{eff}/l_{c1}) - 1} \quad (3.16)$$

where CDSC and CDSCB as well as CDSCD are the model parameters for the channel-length dependent drain/source to channel coupling capacitance and its body and drain bias dependence coefficients. The term on the far-right hand side represents the channel-length dependencies of  $n$ , which is the same term used in the  $V_{th}$  roll-off model [refer to Chapter 2].  $C_{dep}$  in the above formulation is the surface depletion layer capacitance and given by

$$C_{dep} = \sqrt{\frac{\varepsilon_{sub} \cdot q \cdot NDEP \cdot 10^6}{2 \cdot (\varphi_s - V_{bseff})}} \quad (3.16a)$$

$n$  usually has a value between 1 and 2.

### 3.4 Channel Current in Subthreshold and Linear Operations

Substitute the BSIM4 unified channel charge density  $q_{ch}(y)$  equation (3.11) into Eq. (3.3). The BSIM4 channel current model is obtained

$$I_{ch}(y) = -W_{eff} \cdot q_{ch0} \cdot \left[ 1 - \frac{A_{bulk}}{V_{gsteff} + 2 \cdot v_t} \cdot V(y) \right] \cdot \mu_n(y) \cdot \frac{d\varphi_{fn}(y)}{dy} \quad (3.17)$$

where  $q_{ch0}$  is given in Eq. (3.13)

$$q_{ch0} = -C_{oxeff} \cdot V_{gsteff} \quad (3.13)$$

and  $\mu_n(y)$  is the channel carrier mobility, which considers the longitudinal electric field dependence and carrier velocity saturation

$$\mu_n(y) = \frac{\mu_{eff}}{1 + \frac{E(y)}{E_{sat}}} \quad (3.18)$$

$\mu_{eff}$  is the effective mobility that was discussed in Chapter 2,  $E_{sat}$  is the longitudinal critical electric field beyond which the carrier velocity will start to saturate (more discussions of  $E_{sat}$  and saturation electric field will be given in Section 3.5), and  $E(y)$  is the longitudinal electric field at  $y$ , defined as

$$E(y) = \frac{dV(y)}{dy} \quad (3.19)$$

Substituting Eqs. (3.18) and (3.19) into Eq. (3.17), multiplying both sides of Eq. (3.17) by  $dy$  and integrating over the channel length yield the channel current  $I_{ch0}$  of BSIM4 for the subthreshold and linear regions of operation

$$I_{ch0} = \frac{W_{eff} \cdot C_{oxeff} \cdot V_{gsteff}}{L_{eff} \cdot \left(1 + \frac{V_{ds}}{E_{sat} \cdot L_{eff}}\right)} \cdot \mu_{eff} \cdot V_{ds} \cdot \left[1 - \frac{A_{bulk}}{2(V_{gsteff} + 2 \cdot v_t)} \cdot V_{ds}\right] \quad (3.20)$$

Note that Eq. (3.5) was used in deriving  $I_{ch0}$

$$\varphi_{fn}(y) \approx V(y) + V_{sb} \quad (3.5)$$

No source and drain resistance effects have been taken into account so far. The channel current that has just been obtained is called the *intrinsic channel current*. The bias-dependent source and drain resistances  $R_{ds}(V)$  (i.e., the LDD resistances only, not the source/drain diffusion resistances), at the user's discretion by specifying **RDSMOD** = 1, can be taken care of by connecting them between the internal and external source nodes and between the internal and external drain nodes, respectively. The alternative is the *extrinsic* case when **RDSMOD** = 0 (the default setting) is selected by the user. In this case, the LDD source and drain resistances are lumped into the internal source and drain nodes [refer to Chapter 2 for more]. The *extrinsic* case is modeled as follows.

By applying Ohm's law and assuming  $R_{ds}(V)$  is small relative to the channel resistance  $R_{ch}(V)$ , the channel current can be approximated as

$$I_{ch} = \frac{V_{ds}}{R_{ch}(V) + R_{ds}(V)} \quad (3.21)$$

where  $R_{ch}(V)$  is equal to  $V_{ds}/I_{ch0}$  at zero  $R_{ds}(V)$ . Thus

$$I_{ch} = \frac{I_{ch0}}{1 + \frac{R_{ds}(V) \cdot I_{ch0}}{V_{ds}}} \quad (3.22)$$

When  $R_{ds}(V)$  becomes zero, Eq. (3.22) reduces to Eq. (3.20). If  $R_{ds}(V)$  is significant or the power supply is large (greater than 5 volts for high-voltage applications for instance), Eq. (3.22) will no longer be accurate. In either case it is advisable to use **RDSMOD** = 1 and/or a more accurate LDD resistance model. Setting **RDSMOD** = 1 is often necessary for advanced technology nodes because  $R_{ds}(V)$  is comparable to the channel resistance in the linear region ( $V_{gs} \geq V_{th}$ ) and for the analog and RF modeling to give accurate input  $Y$  and S-parameter computation.

### 3.5 Velocity Saturation and Velocity Overshoot

The channel carrier velocity and the channel current start to saturate when the source-drain voltage  $V_{ds}$  exceeds its saturation voltage  $V_{dsat}$ , i.e., when the longitudinal electric field  $E(y)$  in the channel is higher than the critical electric field  $E_{sat}$ . Recall that the BSIM4 carrier velocity and electric field relationship that was used in deriving Eq. (3.20) is

$$v = \mu_n(y) \cdot E(y) = \frac{\mu_{eff}}{1 + \frac{E(y)}{E_{sat}}} \cdot E(y) \quad (3.23)$$

which states that when  $E(y)$  is much smaller than  $E_{sat}$ , the velocity is linearly proportional to the field and when  $E(y)$  is much greater than  $E_{sat}$ , the velocity saturates at **VSAT**. In order to make Eq. (3.23) fit the measured dependence of velocity on the field,  $E_{sat}$  is chosen to be [4]

$$E_{sat} = \frac{2 \cdot VSAT}{\mu_{eff}} \quad (3.24)$$

where **VSAT** is the saturation velocity model parameter that is extracted from measurement. Carrier velocity has strong temperature dependencies: The lower the temperature, the higher the velocity. In BSIM4, when **TEMPMOD** = 0 (the default case) takes effect, it is modeled by

$$VSAT(T_{emp}) = VSAT(TNOM) - AT \cdot \left( \frac{T_{emp}}{TNOM} - 1 \right) \quad (3.25a)$$

**VSAT(TNOM)** means the saturation velocity extracted from measurement under the nominal temperature **TNOM**. When **TEMPMOD** is set to other values such as 1, 2 or 3, it is

$$\text{VSAT}(T_{emp}) = \text{VSAT}(\text{TNOM}) \cdot [1 - \text{AT} \cdot (T_{emp} - \text{TNOM})] \quad (3.25\text{b})$$

For very short-channel devices (such as sub-100nm), when the longitudinal electric field and its gradient are both sufficiently high, the carrier velocity can actually exceed VSAT, a phenomenon known as the velocity overshoot. Velocity saturation is caused by energetic electrons losing their energy through the generation of optical phonons. However, it takes a finite time for the optical phonons to generate. During this time the velocity can continue to increase. In the context of the energy-transport theory, velocity overshoot takes place when the carrier energy relaxation time becomes longer than their momentum relaxation time. It can also be said that velocity overshoot results from the *non-local* field effects, which is to say the carrier velocity is not only determined by the field strength itself, but also determined by the gradient of the field. The overshooting velocity can be expressed as

$$v_{vo} = v \cdot \left[ 1 + \frac{\text{LAMBDA}}{E(y)} \cdot \frac{\partial E(y)}{\partial y} \right] \quad (3.26)$$

where LAMBDA is the velocity overshooting coefficient. In BSIM4, the velocity overshooting effect is modeled by way of multiplying  $E_{sat}$  with a factor

$$E_{sat} = E_{sat\_original} \cdot \left[ 1 + \frac{\text{LAMBDA}}{\mu_{eff} \cdot L_{eff}} \cdot \frac{\left( \frac{V_{ds} - V_{dseff}}{E_{sat\_original} \cdot \text{LitL}} \right)^2 - 1}{\left( \frac{V_{ds} - V_{dseff}}{E_{sat\_original} \cdot \text{LitL}} \right)^2 + 1} \right] \quad (3.27)$$

$E_{sat\_original}$  is the  $E_{sat}$  in Eq. (3.24).  $V_{dseff}$  and  $\text{LitL}$  will be explained later in this sub-section. The evaluation of the BSIM4 velocity overshoot model will be performed only when LAMBDA is given and its  $L$ - and  $W$ -binned value is greater than zero (refer to chapter 2 for parameter binning).

It is now possible to derive the source-drain saturation voltage  $V_{dsat}$ . In the case where no LDD resistances are included the channel current formulation (the *intrinsic* case),  $V_{dsat}$  is

$$V_{dsat} = \frac{E_{sat} \cdot L_{eff} \cdot (V_{gsteff} + 2 \cdot v_t)}{A_{bulk} \cdot E_{sat} \cdot L_{eff} + V_{gsteff} + 2 \cdot v_t} \quad (3.28)$$

simply by equating  $I_{ch0}$  of Eq. (3.20) with  $V_{ds}$  replaced by  $V_{dsat}$ , and

$$I_{ch,sat} = -W_{eff} \cdot q_{ch}(V(y) = V_{dsat}) \cdot \text{VSAT} \quad (3.29)$$

It is apparent that for a long-channel device in the saturation region, Eq. (3.28) gives a familiar saturation voltage

$$V_{dsat} \approx \frac{V_{gs} - V_{th}}{A_{bulk}} \quad (3.30)$$

In the *extrinsic* case (where the LDD resistance effect is modeled within the channel current equation), equating  $I_{ch}$  of Eq. (3.22) and

$$I_{ch,sat} = -W_{eff} \cdot q_{ch}(V(y) = V_{dsat}) \cdot \text{VSAT} \cdot \lambda \quad (3.31)$$

yields, after lengthy mathematical manipulations,

$$V_{dsat} = -\frac{b + \sqrt{b^2 - 4 \cdot a \cdot c}}{2 \cdot a} \quad (3.32)$$

where

$$a = \text{VSAT} \cdot W_{eff} \cdot C_{oxeff} \cdot A_{bulk}^2 \cdot R_{ds}(V) - A_{bulk} \cdot (1 - 1/\lambda) \quad (3.33)$$

$$b = (V_{gsteff} + 2 \cdot v_t) \cdot (1 - 2/\lambda) - A_{bulk} \cdot E_{sat} \cdot L_{eff} - 3 \cdot \text{VSAT} \cdot W_{eff} \cdot C_{oxeff} \cdot A_{bulk} \cdot R_{ds}(V) \cdot (V_{gsteff} + 2 \cdot v_t) \quad (3.34)$$

and

$$c = (V_{gsteff} + 2 \cdot v_t) \cdot E_{sat} \cdot L_{eff} + 2 \cdot \text{VSAT} \cdot W_{eff} \cdot C_{oxeff} \cdot R_{ds}(V) \quad (3.35)$$

In Eq. (3.31), the factor  $\lambda$  is incorporated mainly for PMOS devices. Besides their lower mobilities, holes have a more gradual velocity-field transition than electrons from the linear region to the saturation region. A useful formulation of  $\lambda$  is

$$\lambda = A1 \cdot V_{gsteff} - A2 \quad (3.36)$$

$A1$  and  $A2$  are model parameters. They are extracted from the sharpness of the transition between the linear and the saturation regions from measured  $I_{ds}$ - $V_{ds}$  characteristics of PMOSFETs.

The channel current model of Eq. (3.22) is valid for  $V_{ds} < V_{dsat}$ . It can be extended into the saturation region by replacing  $V_{ds}$  with an effective source-drain voltage  $V_{dseff}$ , which is equal to  $V_{dsat}$  in the saturation region and  $V_{ds}$  in the linear region with a smooth transition in between with the assistance of the model parameter **DELTA** as illustrated in Fig. 3.2 [3].

$$V_{dseff} = V_{dsat} - \frac{1}{2} \cdot \left[ \frac{(V_{dsat} - V_{ds} - \text{DELTA}) + \sqrt{(V_{dsat} - V_{ds} - \text{DELTA})^2 + 4 \cdot \text{DELTA} \cdot V_{dsat}}}{\sqrt{(V_{dsat} - V_{ds} - \text{DELTA})^2 + 4 \cdot \text{DELTA} \cdot V_{dsat}}} \right] \quad (3.37)$$

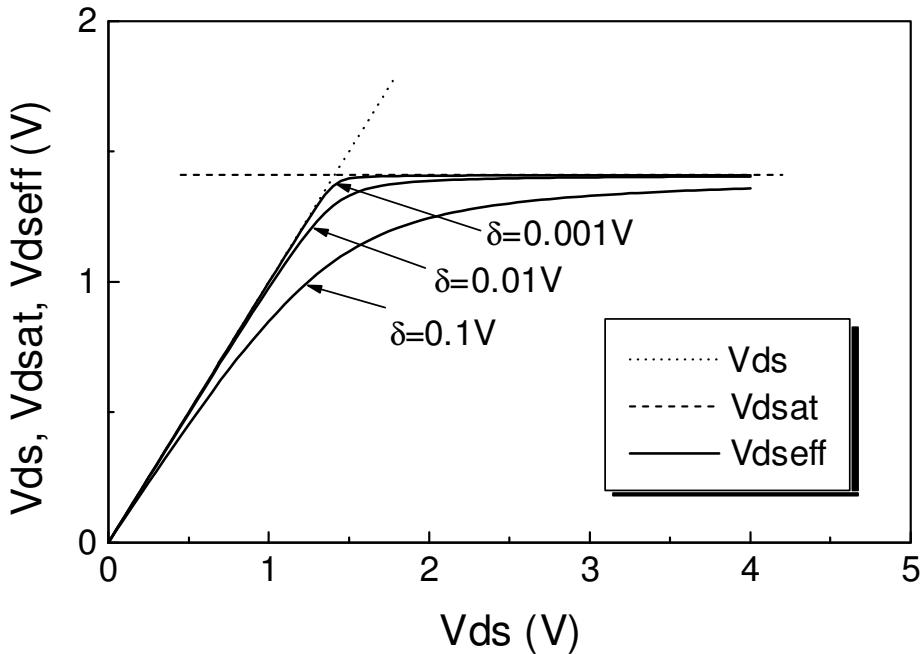


Fig. 3.2 The effective source-drain voltage  $V_{dseff}$  (solid lines) versus  $V_{ds}$  for various  $\text{DELTA}$ .

### 3.6 Output Resistance in Saturation Region

The MOSFET channel current continues to increase when  $V_{ds}$  increases beyond its saturation voltage  $V_{dsat}$ . That also leads to a finite output resistance  $R_{out}$ . In fact,  $R_{out}$  can even decrease as the drain voltage becomes sufficiently high. There are several physical mechanisms at work here.

The output resistance is defined as the inverse of the partial derivative of the channel current with respect to the drain voltage  $V_d$  (not  $V_{ds}$  although the resulting partial derivative can have the same numeric value coincidentally, for instance, when the channel current formulation does not contain  $V_{dg}$  and/or  $V_{db}$ ).

The channel current in saturation can be generally expressed in the form of an integral

$$I_{ch} = I_{ch\_sat}(V_{dsat}) \cdot \left[ 1 + \int_{V_{dsat}}^{V_{ds}} \frac{1}{V_A} \cdot dV_d \right] \quad (3.38)$$

where  $V_A$  is called the Early voltage and defined as

$$V_A = I_{ch\_sat}(V_{dsat}) \cdot \left[ \frac{\partial I_{ch}}{\partial V_d} \right]^{-1} \quad (3.39)$$

It is schematically illustrated in Fig. 3.3. The smaller the  $V_A$  is, the smaller the  $R_{out}$  is. There are four primary physical mechanisms that need to be taken into account in the modeling of the Early voltage. They are channel-length modulation (CLM), drain-induced barrier lowering (DIBL), drain-induced threshold voltage shifts (DITS) due to the pocket implant, and substrate current induced body effects (SCBE). They are all modeled by the BSIM4 model. Fig. 3.4 depicts the respective regions in which each of them plays their respective role.

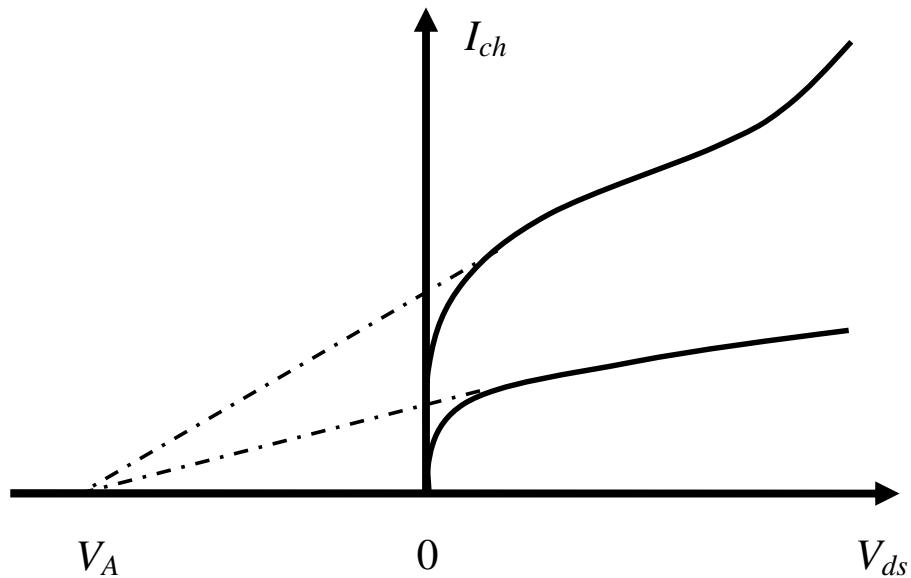


Fig. 3.3 The Early voltage and its effects on MOSFET output characteristics.

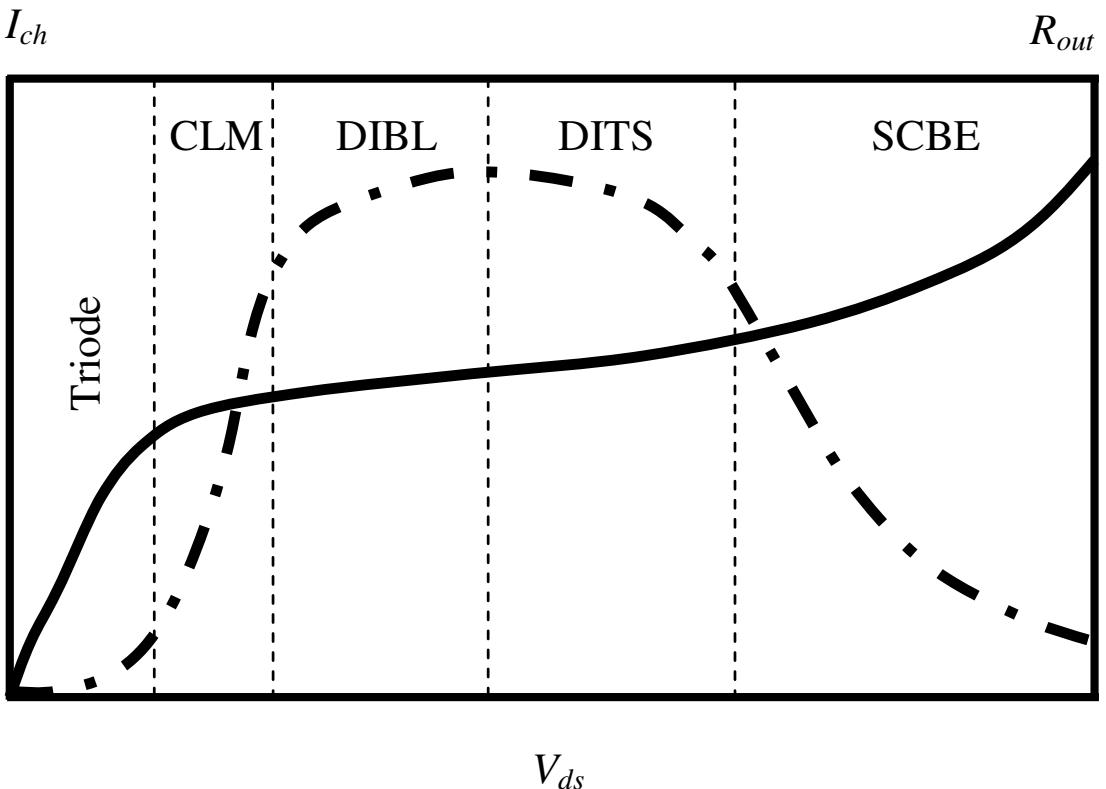


Fig. 3.4 MOSFET output current and output resistance characteristics and the physical mechanisms that determine the Early voltage in various  $V_{ds}$  regions. The solid line represents the channel current whereas the dash-dotted line denotes the output channel resistance.

### 3.6.1 CLM: Channel Length Modulation

As the name suggests, when  $V_{ds} > V_{dsat}$ , **CLM** makes the effective channel length shorter (refer to Chapter 2 for the definitions of the channel length). **CLM** happens because of the lengthening of the channel pinch-off region, or more accurately, velocity-saturation region, at the drain end of the channel. Fig. 3.5 sketches the distribution of the channel potential, electric field, carrier velocity and channel charge density along the channel when **CLM** takes place.

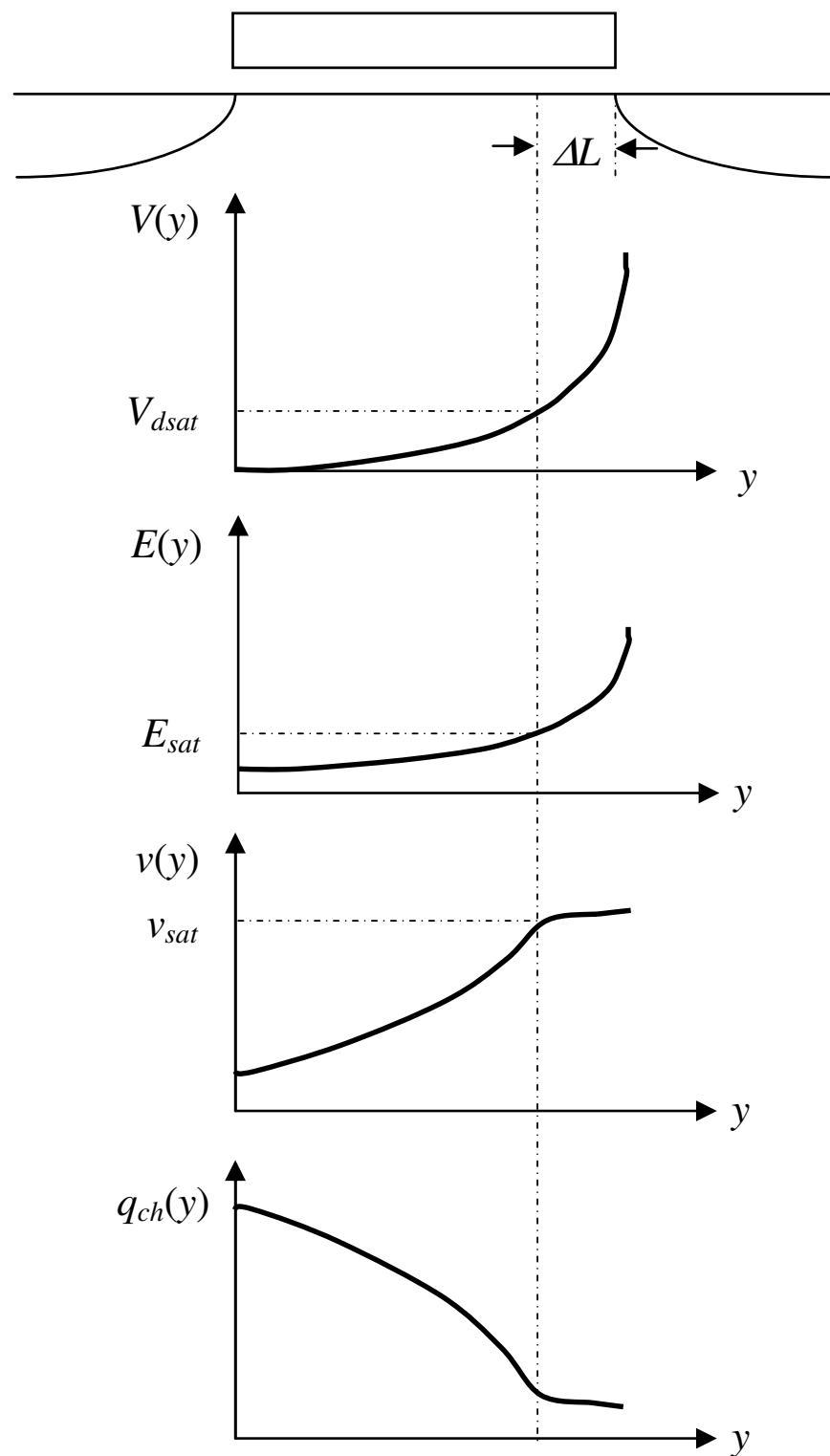


Fig. 3.5 MOSFET channel potential  $V(y)$ , electric field  $E(y)$ , carrier velocity  $v(y)$  and channel charge density  $q_{ch}(y)$  distributions along the channel when *CLM* takes place.

The channel length reduction,  $\Delta L$ , is a complex function of the electric potential and field distributions. Classical quasi-2D analyses give a simple  $\Delta L$

$$\Delta L = LitL \cdot \sinh^{-1} \left( \frac{V_{ds} - V_{dsat}}{E_{sat} \cdot LitL} \right) \quad (3.40)$$

$LitL$  is known as the characteristic drain-field length

$$LitL = \sqrt{\frac{\varepsilon_{sub} \cdot TOXE \cdot XJ}{EPSROX}} \quad (3.40a)$$

The Early voltage of  $CLM$  is

$$V_{ACLM} = I_{ch\_sat}(V_{dsat}) \cdot \left[ \frac{\partial I_{ch}}{\partial L} \cdot \frac{\partial L}{\partial V_d} \right]^{-1} \quad (3.41)$$

The formulation of  $V_{ACLM}$  of BSIM4 is based upon BSIM3v3 [5 - 8] and has incorporated numerous enhancements to the basic quasi-2D theory by the BSIM team and many users for better bias and geometry scalability and the pocket implant effects. The channel current is given

$$\begin{aligned} I_{ch} &= I_{ch\_sat}(V_{dsat}) \cdot \left[ 1 + \int_{V_{dsat}}^{V_{ds}} \frac{1}{V_A} \cdot dV_d \right] \\ &= I_{ch\_sat}(V_{dsat}) \cdot \left[ 1 + \frac{1}{C_{clm}} \cdot \ln \left( \frac{V_{A\_sat} + V_{ACLM}}{V_{A\_sat}} \right) \right] \end{aligned} \quad (3.42)$$

with

$$V_{ACLM} = C_{clm} \cdot (V_{ds} - V_{dseff}) \quad (3.43)$$

The bias and geometry dependencies of  $C_{clm}$  are

$$C_{clm} = \frac{F_{pocket}}{PCLM} \cdot \left( 1 + PVAG \cdot \frac{V_{gsteff}}{E_{sat} \cdot L_{eff}} \right) \cdot \left( 1 + \frac{R_{ds}(V) \cdot I_{cho}}{V_{dseff}} \right) \cdot \left( L_{eff} + \frac{V_{dsat}}{E_{sat}} \right) \cdot \frac{1}{LitL} \quad (3.44)$$

$V_{A\_sat}$  is the Early voltage at  $V_{ds} = V_{dsat}$ . It facilitates a smooth transition of Eq. (3.42) between the linear and saturation regions.

$$V_{A\_sat} = \frac{E_{sat} \cdot L_{eff} + V_{dsat} + 2 \cdot VSAT \cdot W_{eff} \cdot C_{oxeff} \cdot R_{ds}(V) \cdot V_{gsteff} \cdot \left[ 1 - \frac{A_{bulk} V_{dsat}}{2 \cdot (V_{gsteff} + 2 \cdot v_t)} \right]}{VSAT \cdot W_{eff} \cdot C_{oxeff} \cdot R_{ds}(V) \cdot V_{gsteff} - 1 + 2/\lambda} \quad (3.45)$$

In the above equations,  $PCLM$  and  $PVAG$  are model parameters to be extracted from measured MOSFET data. In Eq. (3.44),  $F_{pocket}$  is called

the  $R_{out}$  degradation factor for devices with pocket implants relative to those with uniformly doped channel (see Chapter 2 for more detailed discussions).  $F_{pocket}$  has a strong dependence on the channel length for given gate and drain biases. It is modeled by

$$F_{pocket} = \frac{1}{1 + FPROUT \cdot \frac{\sqrt{L_{eff}}}{V_{gsteff} + 2 \cdot v_t}} \quad (3.46)$$

where **FPROUT** is a model parameter also. Fig. 3.6 shows the good agreement of Eq. (3.44) with measured Si data [9].

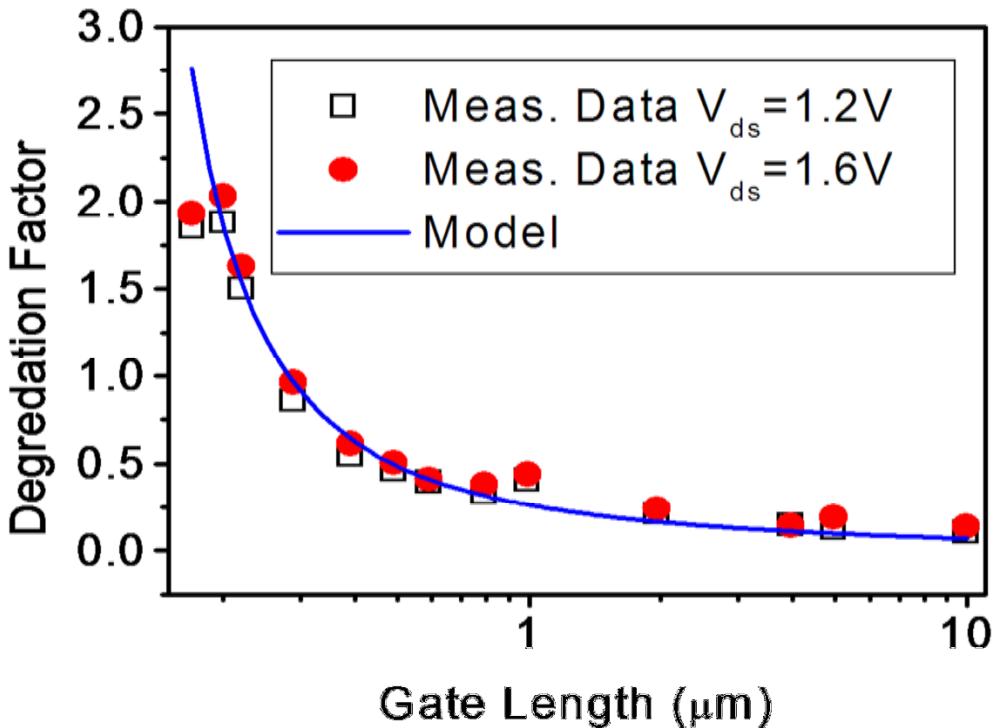


Fig. 3.6 The  $R_{out}$  degradation factor  $F_{pocket}$  for devices with pocket implants relative to those with uniform lateral channel dopings.

### 3.6.2 DIBL: Drain-Induced Barrier Lowering

The energy barrier at the source end of the channel can be lowered by the drain bias via the drain to channel capacitive coupling. The Early voltage due to the **DIBL** effect is

$$V_{ADIBL} = I_{ch\_sat}(V_{dsat}) \cdot \left[ \frac{\partial I_{ch}}{\partial V_{th}} \cdot \frac{\partial V_{th}}{\partial V_d} \right]^{-1} \quad (3.47)$$

The formulation of  $V_{ADIBL}$  of BSIM4 is also based upon BSIM3v3 with additional enhancements. The channel current is thus given as

$$I_{ch} = I_{ch\_sat}(V_{dsat}) \cdot \left[ 1 + \int_{V_{dsat}}^{V_{ds}} \frac{1}{V_{ADIBL}} \cdot dV_d \right] \quad (3.48)$$

with

$$V_{ADIBL} = \frac{V_{gsteff} + 2 \cdot v_t}{\theta_{rout} \cdot (1 + PDIBLB \cdot V_{bseff})} \cdot \left( 1 - \frac{A_{bulk} \cdot V_{dsat}}{A_{bulk} \cdot V_{dsat} + V_{gsteff} + 2 \cdot v_t} \right) \cdot \left( 1 + PVAG \cdot \frac{V_{gsteff}}{E_{sat} \cdot L_{eff}} \right) \quad (3.49)$$

with PDIBLB and PVAG being model parameters.  $\theta_{rout}$  has a similar channel-length dependence as that of the  $V_{th}$  DIBL model (refer to Chapter 2). The smaller the channel length is, the larger the DIBL effect becomes. However,  $\theta_{rout}$  uses a different set of model parameters to improve the model flexibility. Somewhat different channel length dependences should be expected because the gate voltage here is larger than  $V_{th}$  while the gate voltage used in the  $V_{th}$  DIBL model is, of course,  $V_{th}$ , and this changes the channel charge and the drain-channel capacitance.

$$\theta_{rout} = \frac{PDIBLC1}{2 \cdot \cosh\left(\frac{DROUT \cdot L_{eff}}{l_{c0}}\right) - 2} + PDIBLC2 \quad (3.50)$$

where PDIBLC1, PDIBLC2, and DROUT are model parameters. PDIBLC2 is usually quite small and comes into play for long-channel devices. The short-channel characteristic length  $l_{c0}$  given in Chapter 2 is repeated here for convenience of reference

$$l_{c0} = \sqrt{\frac{\epsilon_{sub} \cdot TOXE \cdot X_{dep0}}{\epsilon_{ox}}} \quad (3.50a)$$

### 3.6.3 DITS: Drain-Induced Threshold Voltage Shift Due to Non-Uniform Doping

Modern MOS transistors use pocket implants to improve the  $V_{th}$  roll-off and short-channel effects. As discussed in Chapter 2, two energy barriers are introduced by the pocket implantation, one each at the source end and

the drain end. As the drain bias increases, the energy barrier on the drain side will be lowered first before the one on the source side. This phenomenon gives long-channel devices a new effect from the drain bias in the form of so-called threshold voltage shifts [9].

To model this effect, the channel current in the saturation region can be written as

$$I_{ch} = I_{ch\_sat}(V_{dsat}) \cdot \left[ 1 + \int_{V_{dsat}}^{V_{ds}} \frac{1}{V_{ADITS}} \cdot dV_d \right] \quad (3.51)$$

where the Early voltage  $V_{ADITS}$  is

$$V_{ADITS} = \frac{F_{pocket}}{\text{PDITS}} \cdot \left[ 1 + (1 + \text{PDITSL} \cdot L_{eff}) \cdot \exp(\text{PDITSD} \cdot V_{ds}) \right] \quad (3.52)$$

The model parameters **PDITS**, **PDITSL**, and **PDITSD** are introduced as the coefficients of the channel-length and drain bias dependencies.  $F_{pocket}$  is the  $R_{out}$  degradation factor introduced in Eq. (3.46).

### 3.6.4 SCBE: Substrate Current Induced Body-Bias Effect

When the drain bias is high and the channel electric field near the drain is in the range of MV/cm, channel carriers will gain a few electron-volt (eV) kinetic energies from the accelerating voltages applied between the drain and source. Some of them will become “hot” (1eV of energy can heat up an electron from room temperature to approximately 12,000 degrees in Kelvin ( $300\text{K} \times (1\text{eV}/0.025\text{eV})$ ) to generate electron and hole pairs through the collision with the valence electrons of silicon, a physical process known as impact ionization or inter-band (across the silicon energy gap) generation of electron-hole pairs. The generated holes (in an NMOSFET) flow out of the body of the transistor, which is an easily measurable substrate current called the impact ionization current ( $I_{ii}$  or interchangeably  $I_{sub}$ ). In this case, electrons generated from this process will flow out of the drain terminal and contribute to the drain current. As  $I_{ii}$  flows through the Si body resistance to the body terminal, it creates a positive voltage drop across the body and source junction (assuming this junction is zero biased externally).

The consequence is that the threshold voltage is now reduced because of the elevated body potential, resulting in an additional (and much larger) increase in the channel current, an effect termed the substrate current induced body bias effect or *SCBE*. It is further analyzed and modeled in [10] and in Chapter 4. The interesting fact here is how it changes the Early voltage for the  $R_{out}$  modeling. By following the approach established in the previous sub-sections, the channel current can be written as

$$I_{ch} = I_{ch\_sat}(V_{dsat}) \cdot \left[ 1 + \int_{V_{dsat}}^{V_{ds}} \frac{1}{V_{ASCBE}} \cdot dV_d \right] \quad (3.53)$$

where the Early voltage  $V_{ASCBE}$  associated with *SCBE* is modeled in BSIM4 as

$$\frac{1}{V_{ASCBE}} = \frac{PSCBE2}{L_{eff}} \cdot \exp\left(-\frac{PSCBE1 \cdot LitL}{V_{ds} - V_{dseff}}\right) \quad (3.54)$$

$LitL$  is the characteristic drain-field length as defined in Eq. (3.40a), and **PSCBE1** and **PSCBE2** are model parameters.

### 3.6.5 Channel Current Model for All Regions of Operation

The BSIM4 channel current model, which is valid for all regions of operation and considers all the effects that have been discussed above are written

$$I_{ch} = \frac{I_{ch0} \cdot NF}{1 + \frac{R_{ds}(V) \cdot I_{ch0}}{V_{dseff}}} \cdot \left[ 1 + \frac{1}{C_{clm}} \cdot \ln\left(\frac{V_A}{V_{A\_sat}}\right) \right] \cdot \left( 1 + \frac{V_{ds} - V_{dseff}}{V_{ADIBL}} \right) \cdot \\ \left( 1 + \frac{V_{ds} - V_{dseff}}{V_{ADI}} \right) \cdot \left( 1 + \frac{V_{ds} - V_{dseff}}{V_{ASCBE}} \right) \quad (3.55)$$

**NF** is an instance parameter with a default of 1. It designates the number of fingers of a multi-finger MOS transistor.

It is noted that although the charge carrier type, the device terminal voltage polarities, and the channel current flow direction are all presented for the case of NMOS transistors and forward operation mode ( $V_{ds} \geq 0$ ), the model formulations are also developed for the reverse mode

of operation and PMOS as well. This is the practice for compact modeling and accomplished by performing the swapping of the device type, operation mode, and current polarity in SPICE implementation and model parameter extraction tools. The same also applies to the BSIM4 charge/capacitance models. Refer to Chapter 10 for a detailed implementation methodology.

### 3.7 Source-End Velocity Limit

The channel current is determined by the product of the total channel charges and the velocity at which these charges traverse the channel to the drain terminal. In other words, the faster these charges could move *through the channel*, the higher the drain current would be.

In this chapter we have thus far assumed that the carrier velocity is determined by the electric field profile in the channel. However, the channel electric field in devices with very short-channel lengths (< 20nm) can rise so suddenly and become so large that, with velocity overshoot, some other effects will limit the charge carrier velocity. Indeed the carrier velocity and therefore the current are limited by the maximum possible speed at which the charge carriers can enter the channel region from the source – the thermal velocity. This phenomenon has been known as the thermionic emission limited current in the context of metal/semiconductor and p/n junctions (See C. T. Sah's *Fundamentals of Solid-State Electronics* (FSSE), World Scientific Publishing Co., pp. 474 – 497, 1991) [11]. In CMOS, this is conveniently termed the source-end velocity limit or *SEVL*, in contrast to the velocity saturation or velocity overshoot taking place at the drain end.

BSIM4 provides an efficient approach to modeling the *SEVL* effects on the channel current. According to the *SEVL* theory, the maximum possible source end velocity is [12]

$$v_{SEVL} = \frac{1-r_{bs}}{1+r_{bs}} \cdot VTL \quad (3.56)$$

where the model parameter *VTL* denotes the carrier thermal injection speed at the source with a default value of  $2 \times 10^7$  cm/second, and  $r_{bs}$  represents the backward scattering coefficient

$$r_{bs} = \frac{L_{eff}}{XN \cdot L_{eff} + LC} \quad (3.57)$$

XN and LC are the scattering effect parameters, defaulting to 3 and  $5 \times 10^{-9}$  m, respectively, which gives a  $v_{SEVL}$  approximately half of VTL, i.e.,  $1 \times 10^7$  cm/second.

The BSIM4 channel current model with *SEVL* taken into account becomes

$$I_{ch\_final} = \frac{I_{ch}}{\sqrt[6]{1 + \left(\frac{v_{s,eff}}{v_{SEVL}}\right)^6}} \quad (3.58)$$

where  $I_{ch}$  is given by Eq. (3.55). The velocity  $v_{s,eff}$  is computed by

$$v_{s,eff} = \frac{I_{ch}}{W_{eff} \cdot C_{oxeff} \cdot V_{gsteff}} \quad (3.59)$$

an effective velocity at which the channel charges travel at the source end of the channel.

Although BSIM4 has offered the velocity overshoot model and the *SEVL* model since the year 2003, most users have found that the simpler velocity-saturation model of BSIM4 continues to match the measured device data well even for the advanced MOSFET technologies. The reason is that the velocity saturation model limits the velocity to VSAT while the *SEVL* model limits the velocity to  $v_{SEVL}$ . However, the two limiting velocities can coincidentally have about the same value from measured silicon data and parameter extraction. At very short-channel lengths where *SEVL* is expected to be significant, these two models are nearly indistinguishable. For long-channel lengths, of course, the velocity saturation model is the correct model to use.

### 3.8 Chapter Summary

This chapter presented the MOSFET DC channel current theory first. It then focused on the model derivation and analyses of the BSIM4 unified channel charge, subthreshold and linear channel current, velocity saturation, and output resistance ( $R_{out}$ ) models. A unified, smooth and accurate BSIM4 DC channel current model was obtained, which considers the velocity-field effects, parasitic source/drain resistances, pocket implants, and various physical mechanisms for the accurate modeling of  $R_{out}$ . They include velocity saturation, channel-length modulation (CLM), drain-induced barrier lowering (DIBL), drain-

induced threshold shifts (*DITS*), substrate-induced body effects (*SCBE*), velocity overshoot and source-end velocity limit (*SEVL*). The BSIM4 channel charge and current model is the cornerstone of BSIM4, which has been used for production from the 130-nanometer down to 20-nanometer CMOS technology nodes.

### 3.9 Parameter Table

Name (type)	Description and default	Can be binned?	Note
VOFF (Global; double)	Offset voltage of the effective gate drive.  Default = -0.08 in [V].	Yes	-
VOFFL (Global; double)	Length dependence parameter for VOFF.  Default = 0.0 in [V·m].	No	-
TVOFF (Global; double)	Temperature dependence parameter for VOFF.  Default = 0.0 in [1/Kelvin].	Yes	-
MINV (Global; double)	Parameter for the DC channel current in the moderate inversion region.  Default = 0.0; dimensionless.	Yes	-
NFACTOR (Global; double)	Coefficient of the depletion capacitance dependence of the subthreshold slope parameter $n$ .  Default = 1.0; dimensionless.	Yes	A warning message will be issued if its binned value becomes negative.
CDSC (Global; double)	Drain-source and the channel coupling capacitance for the subthreshold slope parameter $n$ .  Default = $2.4 \times 10^{-4}$ in [Farad/m <sup>2</sup> ].	Yes	A warning message will be issued if its binned value becomes negative.
CDSCD (Global; double)	Drain-bias dependence coefficient of the drain-source and the channel coupling capacitance CDSC for the subthreshold slope parameter $n$ .  Default = 0.0 in [Farad/m <sup>2</sup> /V].	Yes	A warning message will be issued if its binned value becomes negative.

## 110 BSIM4 AND MOSFET MODELING FOR IC SIMULATION

By Weidong Liu and Chenming Hu

CDSCB (Global; double)	Body-bias dependence coefficient of the drain-source and the channel coupling capacitance CDSC for the subthreshold slope parameter $n$ .  Default = 0.0 in [Farad/m <sup>2</sup> /V].	Yes	-
CIT (Global; double)	Unit-area capacitance owing to the interface states for the subthreshold slope parameter $n$ .  Default = 0.0 in [Farad/m <sup>2</sup> ].	Yes	-
VSAT (Global; double)	Channel carrier saturation velocity extracted at TNOM.  Default = $8.0 \times 10^4$ in [m/s].	Yes	A fatal error message will be issued if its value after temperature de-rating is not positive; a warning message will be issued if its value after temperature de-rating is less than $1 \times 10^3$ .
AT (Global; double)	Temperature-dependence parameter for the saturation velocity VSAT.  Default = $3.3 \times 10^4$ in [m/s] for TEMPMOD = 0 and $3.3 \times 10^4$ in [1/Kelvin] for other TEMPMOD settings (i.e., 1, 2, and 3).	Yes	-
A1 (Global; double)	PMOS non-saturation effect parameter.  Default = 0.0 in [V <sup>-1</sup> ].	Yes	Reset to zero if A2 is greater than 1.0.
A2 (Global; double)	PMOS non-saturation effect parameter.  Default = 0.0; dimensionless.	Yes	A warning message will be issued if its binned value is less than 0.1 and will be reset to 0.1. A warning message will be issued if its binned value is greater than 1.0 and it will be reset to 1.0.

DELTA (Global; double)	Parameter of the effective source-drain voltage $V_{dseff}$ used in the DC current model. It models how rapid carrier velocity saturates when $V_{ds}$ near the saturation voltage $V_{dsat}$ .  Default = 0.01; dimensionless.	Yes	A fatal error message will be issued if its binned value is negative.
PCLM (Global; double)	Parameter to model the channel-length modulation effects on the channel output resistance degradation.  Default = 1.3; dimensionless.	Yes	A fatal error message will be issued if its binned value is not positive.
PVAG (Global; double)	Gate bias dependence parameter to model the channel-length modulation effects on the channel output resistance degradation.  Default = 0.0; dimensionless.	Yes	-
PDIBLB (Global; double)	Body bias dependence parameter to model the drain-induced barrier lowering effects on the channel output resistance degradation.  Default = 0.0 in [V <sup>-1</sup> ].	Yes	-
PDIBL1 (Global; double)	Channel-length dependence parameter to model the drain-induced barrier lowering effects on the channel output resistance degradation.  Default = 0.39; dimensionless.	Yes	A warning message will be issued if its binned value is less than zero.
PDIBL2 (Global; double)	Channel-length dependence parameter to model the drain-induced barrier lowering effects on the channel output resistance degradation.  Default = 0.0086; dimensionless.	Yes	A warning message will be issued if its binned value is less than zero.
DROUT (Global; double)	Channel-length dependence coefficient parameter to model the drain-induced barrier lowering effects on the channel output resistance degradation.  Default = 0.56; dimensionless.	Yes	A fatal error message will be issued if its binned value is less than zero.

112 BSIM4 AND MOSFET MODELING FOR IC SIMULATION  
By Weidong Liu and Chenming Hu

FPOUT (Global; double)	Coefficient for the channel output resistance degradation due to pocket implants.  Default = 0.0 in [V/m <sup>-1/2</sup> ].	Yes	A fatal error message will be issued if its binned value is negative.
PDITS (Global; double)	$R_{out}$ degradation coefficient parameter due to drain-induced threshold voltage shifts caused by pocket implants.  Default = 0.0 in [V <sup>-1</sup> ].	Yes	A fatal error message will be issued if its binned value is negative.
PDITSL (Global; double)	Length-dependence parameter for drain-induced threshold voltage shifts caused by pocket implants.  Default = 0.0 in [m <sup>-1</sup> ].	No	A fatal error message will be issued if its value is negative.
PDITSD (Global; double)	Drain bias dependence parameter for drain-induced threshold voltage shifts caused by pocket implants.  Default = 0.0 in [V <sup>-1</sup> ].	Yes	-
PSCBE1 (Global; double)	Drain bias dependence parameter for the substrate current induced body effect.  Default = $4.24 \times 10^8$ in [V/m].	Yes	-
PSCBE2 (Global; double)	Length-dependence parameter for the substrate current induced body effect.  Default = $1.0 \times 10^{-5}$ in [m/V].	Yes	A warning message will be issued if its binned value is not positive.
LAMBDA (Global; double)	Parameter to model the velocity overshoot effect on the channel critical electric field $E_{sar}$ .  Default = 0.0 in [m <sup>3</sup> /(V · s)].	Yes	Velocity overshoot model evaluation will be turned on if this parameter is given and greater than zero. A warning message will also be issued if its value is greater than $1 \times 10^{-9}$ (too large!).

VTL (Global; double)	Electron and hole thermal injection speed at source of the source-end velocity limit model.  Default = $2.0 \times 10^5$ in [m/s].	Yes	A warning message will be issued when its binned value is less than $6 \times 10^4$ (too small), if it is given and its binned value is greater than zero.
XN (Global; double)	Channel-length dependence coefficient parameter for the backward scattering of the source-end velocity limit model.  Default = 3.0; dimensionless.	Yes	A warning message will be issued when its binned value is less than 3 (too small), if VTL is given and the binned value of VTL is greater than zero. It will then be reset to 3.
LC (Global; double)	Critical channel length for the backward scattering of the source-end velocity limit model.  Default = $5.0 \times 10^{-9}$ in [m].	No	A warning message will be issued when its value is less than 0 (too small), if VTL is given and the binned value of VTL is greater than zero. It will then be reset to zero.

## References

- [1] H. C. Pao, and C. T. Sah, “Effects of diffusion current on characteristics of metal-oxide (insulator)-semiconductor transistors,” *Solid-State Electronics*, vol. 9, pp. 927-937, 1966.
- [2] J. R. Brews, “A charge sheet model of the MOSFET,” *Solid-State Electronics*, vol. 21, pp. 345-355, 1978.
- [3] Weidong Liu, and Chenming Hu, “BSIM3v3 MOSFET Model” – Silicon and Beyond: Advanced Device Models and Circuit Simulators, edited by Michael S. Shur and Tor A. Fjeldly, pp. 1-31, ISBN: 981-02-4280-8, World Scientific, 2000.
- [4] Chenming Calvin Hu, “Modern Semiconductor Devices for Integrated Circuits,” Chapter 6, pp. 195 – 257, Pearson Prentice Hall, 2010.

## 114 BSIM4 AND MOSFET MODELING FOR IC SIMULATION

By Weidong Liu and Chenming Hu

- [5] J. H. Huang, Z. H. Liu, M. C. Jeng, K. Hui, M. Chan, P. K. Ko, and C. Hu, "BSIM3 Version 2.0 User's Manual," University of California, Berkeley, March 1994.
- [6] J. H. Huang, Z. H. Liu, M. C. Jeng, P. Ko, and C. Hu, "A physical model MOSFET output resistance," Tech. Dig. of IEDM, pp. 569-572, San Francisco, December 1992.
- [7] Y. Cheng, M. Jeng, Z. Liu, J. H. Hang, M. Chan, K. Chen, P. Ko, and C. Hu, "A physical and scalable I-V model in BSIM3v3 for analog/digital circuit simulation," IEEE Tran. Electron Devices, vol. 44, pp. 277-287, 1997.
- [8] Weidong Liu, Xiaodong Jin, James Chen, Min-Chie Jeng, Zhihong Liu, Yuhua Cheng, Kai Chen, Mansun Chan, Kelvin Hui, Jianhui Huang, Robert Tu, Ping K. Lo, and Chenming Hu, "BSIM3v3.2 MOSFET MODEL — Users' Manual," Memorandum No. UCB/ERL M98/51. Electronics Research Laboratory, College of Engineering, the University of California at Berkeley, August 1998.
- [9] Kanyu Mark Cao, Weidong Liu, Xiaodong Jin, Karthik Vasanth, Keith Green, John Krick, Tom Vrotsos, and Chenming Hu, "Modeling of pocket implanted MOSFETs for anomalous analog behavior," Tech. Dig. of IEDM, pp. 171-174, Washington D. C., December 1999.
- [10] C. Hu, S. Tam, F.C. Hsu, P.K. Ko, T.Y. Chan, and K.W. Kyle, "Hot-Electron Induced MOSFET Degradation - Model, Monitor, Improvement," IEEE Trans. Electron Devices, vol. 32, pp. 375-385, 1985.
- [11] Chih-Tang Sah, "Fundamentals of Solid-State Electronics (FSSE)," World Scientific Publishing Co., pp. 474-497, 1991.
- [12] Tanvir H. Morshed, Wenwei (Morgan) Yang, Mohan V. Dunga, Xuemei (Jane) Xi, Jin He, Weidong Liu, Kanyu M. Cao, Xiaodong Jin, Jeff J. Ou, Mansun Chan, Ali M. Niknejad, and Chenming Hu, "BSIM4.6.4 User's Manual," The University of California at Berkeley, <http://www-device.eecs.berkeley.edu/~bsim3/bsim4.html>. 2009.

## Chapter 4

# Gate Direct-Tunneling and Body Currents

### 4.1 Introduction and Chapter Objectives

Like the MOSFET channel current presented in Chapter 3, MOSFET gate direct-tunneling current and body current are also DC currents. Unlike the channel current, these current components are unwanted and constitute circuit leakage current and may even interfere with the operations of some sensitive circuits.

Accurate modeling of these currents is indispensable for the 90-nanometer technology node and below for various CMOS ICs from logic and memory to analog and RF applications. BSIM4 provides accurate models for the gate leakage and body currents. Developed from the research by the UC Berkeley BSIM team and refined with the collective efforts within the Compact Model Council, the BSIM4 gate direct-tunneling and body current models are now production turnkey models.

This chapter is devoted to these BSIM4 models. It begins with the gate direct-tunneling current model, followed by the models of the impact ionization ( $I_{ii}$ ), gate-induced source ( $I_{GISL}$ ) and gate-induced drain leakage ( $I_{GIDL}$ ) currents. The latter three currents and the gate-to-body direct-tunneling current are the components of the body terminal current. The other two contributors to the body current are the source/body and drain/body junction diode currents. They will not be presented until Chapter 9 after the source/drain layout-dependence model has been presented in Chapter 8.

This chapter will conclude with a topological representation of the BSIM4 branch and terminal DC currents and a parameter table for all the current models presented in this chapter.

## 4.2 Gate Direct-Tunneling Current Theory and Model

The gate dielectric direct-tunneling leakage current became an important topic when the gate oxide thickness was reduced to 50Å for the 130nm technology node. The physical mechanisms that are responsible for these leakage currents are the direct tunneling or band-to-band tunneling of charge carriers (electrons and holes) between the silicon substrate and the poly-silicon gate when the gate dielectric field strength is high enough to induce such inter-band tunneling [1]. The gate direct-tunneling is undesirable as it adds static leakage power dissipation during MOSFET operations, in addition to the sub-threshold channel current. In extreme cases, the tunneling itself may even interfere with the circuit operations and functionality.

The physical mechanisms of tunneling will now be discussed for various operation regimes. This is followed by the analyses and derivations of the BSIM4 tunneling model. Particular attention is paid to the drain bias dependencies of the gate-and-channel tunneling current. The methodology for partitioning the gate-to-channel tunneling current between the source and drain terminals is then presented. Finally, the model is verified with experimental data and 2-D TCAD simulations.

### 4.2.1 Tunneling Mechanisms and Current Components

As the gate oxide thickness decreases, the energy barrier presented by the gate oxide becomes easier to penetrate by the charge carriers. The probability for the charge carriers to tunnel between the silicon substrate to the poly-silicon gate increases exponentially with decreasing oxide thickness and increasing oxide voltage. The electrons and holes are able to tunnel directly from one side of the oxide to the other without having to “fly” over the top of the conduction band of the gate oxide [2], [3], [4], and [5]. Tunneling is illustrated with the arrows in Fig. 4.1 [6].

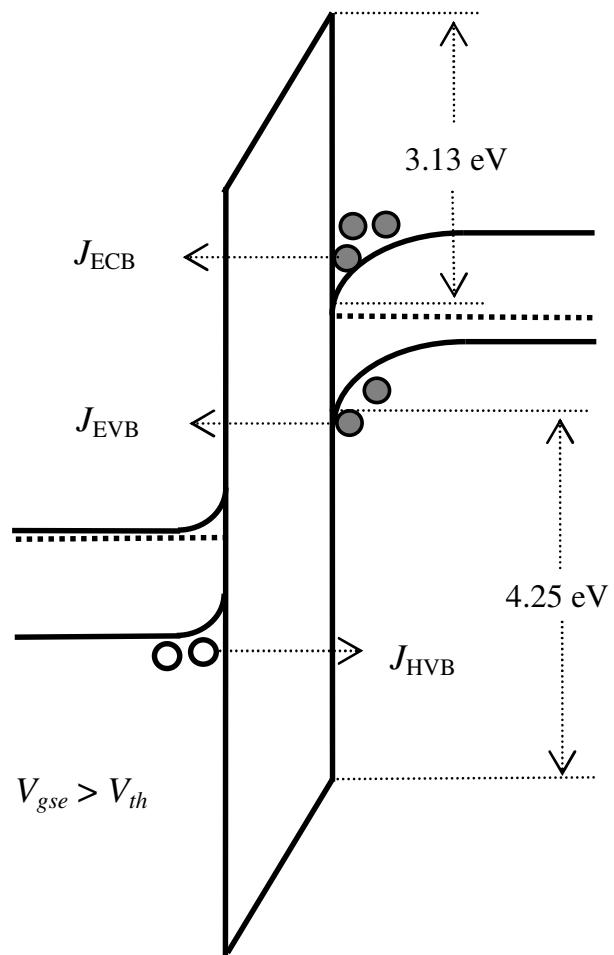


Fig. 4.1 Energy-band diagram illustrating the three electron and hole direct-tunneling mechanisms in an  $n^+$ -poly-gate NMOSFET with a thin gate dielectric layer. Here, the transistor is biased in the strong inversion region.

As shown in the figure, three physical mechanisms of gate-direct tunneling leakage of a MOS device have been identified: Conduction band electron (ECB) tunneling from the p-Si-body conduction band on the right side to the  $n^+$ -poly-Si-gate conduction band on the left side, valence band electron (EVB) tunneling from the p-Si-body valence band to the  $n^+$ -poly-Si-gate conduction band, and valence band hole (HVB) tunneling from the  $n^+$ -poly-Si-gate valence band to the p-Si-body valence band. Each mechanism is dominant in a particular region of operation of NMOS and PMOS transistors. In the case of an NMOS transistor that operates in the strong inversion region (as illustrated in Fig. 4.1), the dominant mechanism of tunneling is ECB. Electrons from the conduction

band of the inversion layer in the silicon body tunneling into the poly-Si gate, creating the gate-channel tunneling current ( $I_{gc}$ ). Under the same bias condition, electrons from the valence band (EVB) of the silicon substrate can also tunnel into the poly-silicon gate, creating the gate-body tunneling current  $I_{gb}$ . The magnitude of  $I_{gb}$  is, however, much smaller than  $I_{gc}$  as shown in Fig. 4.2 because the tunneling barrier of EVB, 4.25 eV, is higher than the tunneling barrier of ECB, 3.13 eV. Finally, holes from the valence band (HVB) of the poly-silicon gate can tunnel into the silicon substrate as well, producing another component of  $I_{gb}$ , which is even smaller under this bias condition since the effective mass of holes is significantly larger than that of electrons. Fig. 4.2 illustrates the measured and modeled contributions of the three tunneling mechanisms for the MOSFET gate leakage currents [5].

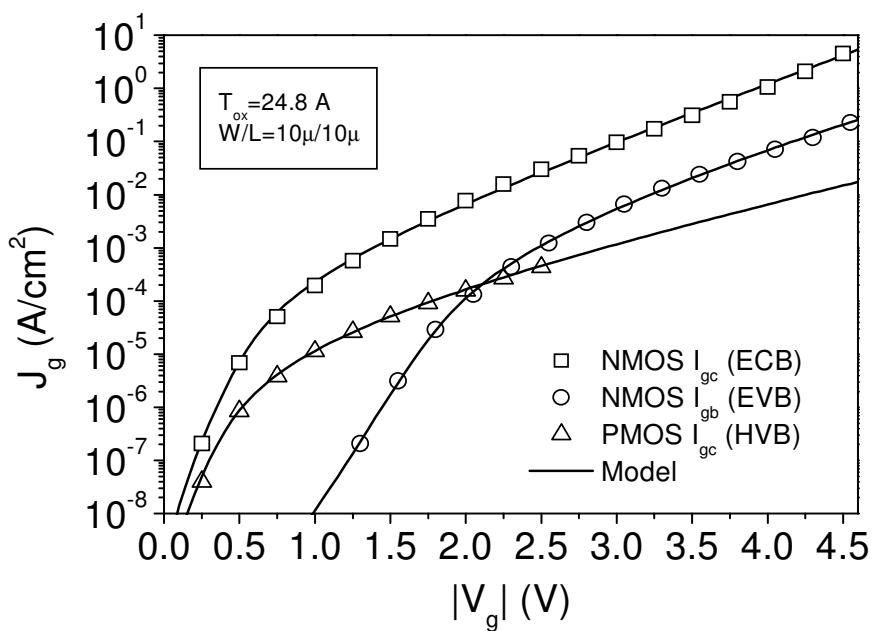


Fig. 4.2 Measured and modeled gate-channel ( $I_{gc}$ ) and gate-body ( $I_{gb}$ ) tunneling currents of NMOS and PMOS transistors when  $V_{ds} = 0$ . The tunneling mechanisms responsible for each current component are explicitly noted in the figure for each transistor type. Both NMOS and PMOS are biased in the inversion region. The model is the gate direct-tunneling current model of BSIM4, which will be presented in the text shortly.

Table 4.1 The dominant physical mechanisms responsible for the direct-tunneling current components found in NMOS and PMOS for various bias regimes

MOSFET	Inversion		Accumulation
	$I_{gc}$	$I_{gb}$	$I_{gb}$
NMOS	ECB	EVB	ECB
PMOS	HVB	EVB	ECB

Table 4.1 shows the dominant tunneling mechanism for the tunneling current components for NMOS and PMOS devices under various operating conditions. Interested readers are encouraged to verify this table for themselves using the energy band diagrams similar to Fig. 4.1.

In contrast to the gate-channel tunneling current, which dominates in long-channel transistors with thin gate oxide, the gate-source/drain tunneling currents ( $I_{gs}$  and  $I_{gd}$ , respectively) in the gate/source and gate/drain overlap regions are not negligible for short-channel transistors. ECB and HVB are responsible for these overlap tunneling components for NMOS and PMOS, respectively, in the short-channel transistors.

In SPICE implementation, it is necessary to partition the gate-channel tunneling current into two components because the channel is not a node in compact model. One component flows out of or from the source terminal ( $I_{gcs}$ ) and the other, the drain terminal ( $I_{gcd}$ ). The source and drain partitioning will be further discussed shortly. First we note that there are five direct-tunneling current components modeled. They are illustrated in Fig. 4.3. They are turned on or off by two global model flags **IGCMOD** and **IGBMOD**. That is, when **IGBMOD** is set to 1, the current  $I_{gb}$  is modeled. **IGCMOD** has two possible settings, 1 or 2 for the remaining four components to be turned on and modeled. **IGCMOD** = 1 uses the zero-bias threshold voltage  $VTH0$ , whereas **IGCMOD** = 2 uses the full equation of the BSIM4 threshold voltage  $V_{th}$ , which is more accurate but consume more computational time.

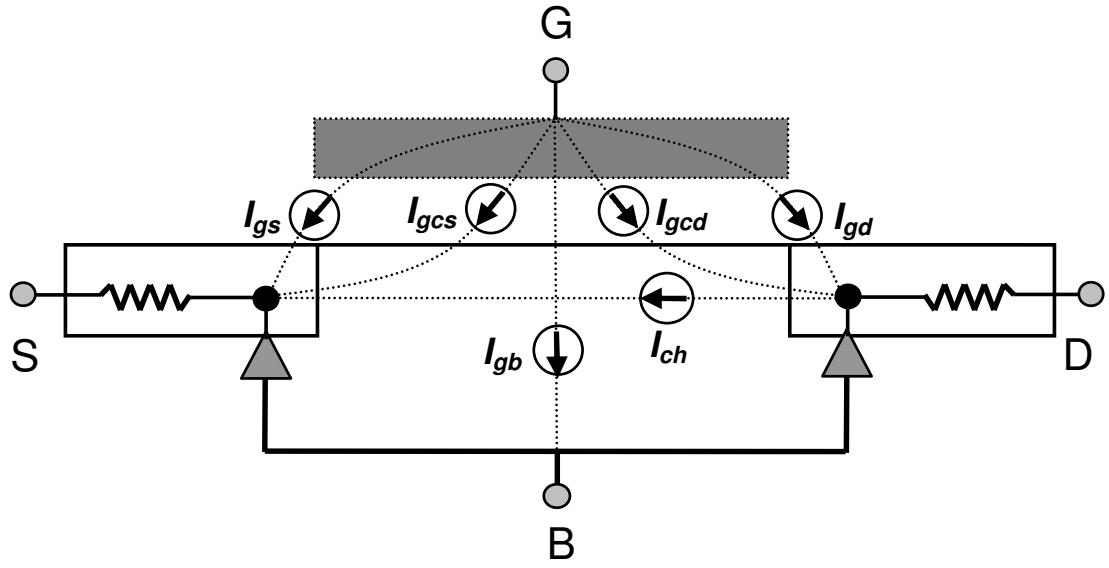


Fig. 4.3 The gate direct-tunneling current components modeled by BSIM4. Depending upon the biases,  $I_{gs}$ ,  $I_{gd}$ , and  $I_{gb}$  can be positive or negative.

#### 4.2.2 Gate Oxide Voltage

The gate direct-tunneling current is a strong function of the voltage drop across the gate oxide  $V_{ox}$ .  $V_{ox}$  is more thoroughly derived in Chapter 5. For now,  $V_{ox}$  is briefly presented.

$$V_{ox} = V_{oxacc} + V_{oxdepinv} \quad (4.1)$$

where  $V_{oxacc} = V_{fbzb} - V_{fbeff}$  is the gate oxide voltage for the accumulation region.  $V_{oxdepinv} = K1_{ox} \sqrt{\varphi_s(V) - V_{bseff}} + V_{gsteff}$  is that for the depletion and inversion regions. As will be shown in Chapter 5, the combined  $V_{ox}$

$$V_{ox} = V_{fbzb} - V_{fbeff} + K1_{ox} \sqrt{\varphi_s(V) - V_{bseff}} + V_{gsteff} \quad (4.1a)$$

and its derivatives are continuous and smooth over all bias regions. These terms were explained in the preceding chapters. Note also that in Eq. (4.1a), the surface potential is bias dependent and is not equal to  $2\cdot\varphi_B$ . The square-root term is given by

$$\sqrt{\varphi_s(V) - V_{bseff}} = \frac{K1_{ox}}{2} \cdot \left[ -1 + \sqrt{1 + \frac{4}{K1_{ox}^2} (V_{gs\_eff} - V_{fbeff} - V_{gsteff} - V_{bseff})} \right] \quad (4.1b)$$

### 4.2.3 Gate-Body Tunneling Current $I_{gb}$

In the depletion and inversion regions,  $I_{gb}$  flows from the gate to the body in an NMOSFET and from the body to the gate in a PMOSFET, with EVB dominant in both, and is modeled by

$$I_{gb\_depinv} = W_{eff} \cdot L_{eff} \cdot A \cdot \left( \frac{TOXREF}{TOXE} \right)^{NTOX} \cdot \left( \frac{V_{gb} \cdot V_{dens}}{TOXE^2} \right) \cdot \exp(-B \cdot TOXE \cdot (AIGBINV - BIGBINV \cdot V_{oxdepinv}) \cdot (1 + CIGBINV \cdot V_{oxdepinv})) \quad (4.2)$$

where  $A = q^3 / 8\pi\hbar\phi_b \approx 3.75956 \times 10^{-7}$  is given in the unit of  $[A \cdot V^{-2}]$ ,  $B = 8\pi\sqrt{2m_{ox}\phi_b^{3/2}} / 3qh \approx 9.82222 \times 10^{11}$  in the unit of  $(g/F \cdot s^2)^{0.5}$  with  $m_{ox} = 0.3m_0$  being the effective carrier mass in the oxide and  $\phi_b = 4.25$  eV being the p-Si-body valence electron tunneling barrier height. (See Fig. 4.1 pathway  $J_{EVB}$ . It is different from  $\varphi_B$ , the Fermi potential, defined in Chapter 2.) Note that  $\pi = 3.14159$ , the Planck constant  $\hbar = 6.62617 \times 10^{-34}$  J-s, the elementary charge  $q = 1.60218 \times 10^{-19}$  C, and the electron mass  $m_0 = 9.1095 \times 10^{-31}$  in the unit of kilograms are employed to compute  $A$  and  $B$ .

$TOXE$  is again the electrical gate oxide thickness.  $TOXREF$  denotes the reference or nominal oxide thickness of a particular manufacturing process. It is emphasized that a slight deviation of  $TOXE$  from  $TOXREF$  could lead to a significant change in the tunneling current because of the exponential dependence on the gate oxide thickness. This is modeled by a power function of the ratio of the two oxide thicknesses with the power  $NTOX$  being a fitting parameter.  $AIGBINV$ ,  $BIGBINV$ , and  $CIGBINV$  are global parameters to model the oxide voltage dependence.

In Eq. (4.2),  $V_{dens}$ , representing the density of states and carriers that overlap in energy and therefore are available for tunneling, is given by

$$V_{dens} = \text{NIGBINV} \cdot v_t \cdot \log\left(1 + \exp\left(\frac{V_{oxdepinv} - \text{EIGBINV}}{\text{NIGBINV} \cdot v_t}\right)\right) \quad (4.2a)$$

where the thermal voltage  $v_t$  is defined as

$$v_t = \frac{k_B T_{emp}}{q} \quad (4.2b)$$

when the temperature model selector **TEMPMOD** is equal to 0 or 1; while

$$v_t = \frac{k_B \cdot \text{TNOM}}{q} \quad (4.2c)$$

when **TEMPMOD** = 2. The Boltzmann constant  $k_B = 1.38066 \times 10^{-23}$  J/K.  $T_{emp}$ , specified in a SPICE net-list, is the temperature at which the device or circuit operates and **TNOM** is the nominal temperature at which the device is characterized for the BSIM4 model parameter extraction.

In Eq. (4.2a), **NIGBINV** approximates the energy level, above which and below which the density varies linearly and exponentially with  $V_{oxdepinv}$ , respectively. **EIGBINV** determines the rapidness of the exponential changes and the transitions around **EIGBINV**. Eq. (4.2a) is analogous to the effective gate bias  $V_{gsteff}$  in its mathematic functional form. Both **NIGBINV** and **EIGBINV** are global model parameters and extracted from the measured  $I_{gb\_inv}$ .

In the accumulation region,  $I_{gb}$  is primarily the result of ECB. (See Fig. 4.1 for NMOS transistors.) It flows from the p-Si-body to the n<sup>+</sup>-poly-Si-gate for the NMOS transistors, and from the gate to the body for PMOS transistors.

$$\begin{aligned} I_{gb\_acc} &= W_{eff} \cdot L_{eff} \cdot A \cdot \left(\frac{\text{TOXREF}}{\text{TOXE}}\right)^{\text{NTOX}} \cdot \left(\frac{V_{gb} \cdot V_{dens}}{\text{TOXE}^2}\right) \\ &\cdot \exp(-B \cdot \text{TOXE} \cdot (\text{AIGBACC} - \text{BIGBACC} \cdot V_{oxacc}) \cdot (1 + \text{CIGBACC} \cdot V_{oxacc})) \end{aligned} \quad (4.3)$$

where  $A = q^3/8\pi h\phi_b = 4.97232 \times 10^{-7}$  in the unit of  $[\text{A}\cdot\text{V}^{-2}]$  and  $B = 8\pi\sqrt{2m_{ox}}\phi_b^{3/2}/3qh = 7.45669 \times 10^{11} (\text{g/F}\cdot\text{s}^2)^{0.5}$  with  $m_{ox} = 0.4m_0$  being the effective carrier mass in oxide.  $\phi_b = 3.13$  eV is the conduction-band electron tunneling barrier height. AIGBACC, BIGBACC, and CIGBACC are global model parameters. The overlap of densities of carriers and states is modeled by

$$V_{dens} = \text{NIGBACC} \cdot v_t \cdot \log\left(1 + \exp\left(-\frac{V_{gb} - V_{fbzb}}{\text{NIGBACC} \cdot v_t}\right)\right) \quad (4.4)$$

where an analogous interpretation as given for NIGBINV can be drawn for the global model parameter NIGBACC.

The single and continuous equation of the gate-body tunneling current is the sum of Eqs. (4.2) and (4.3)

$$I_{gb} = I_{gb\_acc} + I_{gb\_depinv} \quad (4.5)$$

Equation (4.5) is valid for all regions of operation.

#### 4.2.4 Gate-Source/Drain Tunneling Through Overlap Regions

The tunneling currents between the gate and the source/drain high-impurity-concentration diffusions through the overlap regions, namely,  $I_{gs}$  and  $I_{gd}$ , respectively are determined by ECB for NMOS and HVB for PMOS transistors. They are modeled as

$$I_{gs} = W_{eff} \cdot \text{DLCIG} \cdot A \cdot T_{oxRatioEdge} \cdot v_{gs} \cdot v_{gs-fb} \cdot \exp[-B \cdot \text{TOXE} \cdot \text{POXEDGE} \cdot (\text{AIGSD} - \text{BIGSD} \cdot v_{gs-fb}) \cdot (1 + \text{CIGSD} \cdot v_{gs-fb})] \quad (4.6)$$

and

$$I_{gd} = W_{eff} \cdot \text{DLCIG} \cdot A \cdot T_{oxRatioEdge} \cdot v_{gd} \cdot v_{gd-fb} \cdot \exp[-B \cdot \text{TOXE} \cdot \text{POXEDGE} \cdot (\text{AIGSD} - \text{BIGSD} \cdot v_{gd-fb}) \cdot (1 + \text{CIGSD} \cdot v_{gd-fb})] \quad (4.7)$$

respectively. All the gate-source/drain voltages denoted by a lower-case  $v$  are the physical gate-to-source/drain voltages, without regard to the source-drain voltage being positive or negative, i.e., not requiring source and drain voltage swapping in SPICE implementations. For the sake of simplicity, these voltages do not take into account the poly-silicon gate depletion effects. Note also that in Eqs. (4.6) and (4.7),  $A=q^3/8\pi\hbar\phi_b=4.97232\times10^{-7}$  [A·V<sup>-2</sup>] for NMOS and  $3.42537\times10^{-7}$  [A·V<sup>-2</sup>] for PMOS and  $B=8\pi\sqrt{2m_{ox}}\phi_b^{3/2}/3qh=7.45669\times10^{11}$  (g/F·s<sup>2</sup>)<sup>0.5</sup> for NMOS and  $1.16645\times10^{12}$  (g/F·s<sup>2</sup>)<sup>0.5</sup> for PMOS, with  $m_{ox}=0.4m_0$  and  $\phi_b=3.13$  eV.

Note also that in these two equations, the global parameter **DLCIG** represents the length of the gate and source/drain overlap regions for the gate-direct tunneling current modeling; it defaults to **LINT**, the overlap length for the channel current model (refer to Chapter 2). The gate oxide thickness near the overlap regions may actually differ from the thickness of the oxide above the channel region as a result of such process steps as gate sidewall oxide deposition and annealing. Hence, a parameter **POXEDGE** is introduced to permit the oxide thickness to differ from **TOXE** and to be used in the modeling of  $I_{gs}$  and  $I_{gd}$ . **POXEDGE** is dimensionless and defaults to 1. With this parameter, the variable  $T_{oxRatioEdge}$  term is now transformed into

$$T_{oxRatioEdge} = \left( \frac{\text{TOXREF}}{\text{TOXE} \cdot \text{POXEDGE}} \right)^{\text{NTOX}} \cdot \frac{1}{(\text{TOXE} \cdot \text{POXEDGE})^2} \quad (4.8)$$

In Eqs. (4.6) and (4.7), the effective gate-source and gate-drain voltages,  $v_{gs-fb}$  and  $v_{gd-fb}$ , respectively, take into account the effects of the flat-band voltage associated with the source and drain diffusions. In other words, when the diffusion regions are under a flat-band bias condition or when  $v_{gs}$  and  $v_{gd}$  are zero, the tunneling through the overlaps is approximately zero. The effective gate voltages in this case are defined by

$$v_{gs-fb} = \sqrt{[v_{gs} - V_{fbsd}(T)]^2 + 1.0 \times 10^{-4}} \quad (4.9)$$

and

$$v_{gd-fb} = \sqrt{[v_{gd} - V_{fbsd}(T)]^2 + 1.0 \times 10^{-4}} \quad (4.10)$$

which are both approximately zero when  $v_{gs}$  and  $v_{gd}$  are close to  $V_{fbsd}$ . Note, in particular, that  $v_{gs-fb}$  and  $v_{gd-fb}$  are symmetrical about  $V_{fbsd}(T)$ , a useful function form such that  $I_{gs}$  and  $I_{gd}$  are symmetrical with respect to  $V_{gs-fb}$  and  $V_{gd-fb}$ , respectively, and change their current flow directions as the polarity of  $v_{gs}$  and  $v_{gd}$  changes. The source and drain flat-band voltage  $V_{fbsd}(T)$  is expressed by

$$V_{fbsd}(T) = V_{fbsd} + VFBSDOFF \cdot [1 + TVFBSDOFF \cdot (T_{emp} - TNOM)] \quad (4.11)$$

in which the first term on the right side is the base term of the flat-band voltage. It is given

$$V_{fbsd} = \frac{k_B \cdot TNOM}{q} \cdot \log\left(\frac{NGATE}{NSD}\right) \quad (4.12)$$

when the poly-silicon gate and source/drain doping concentrations, NGATE and NSD, are both finite (non zero). Otherwise,  $V_{fbsd} = 0$ . The source and drain flat-band voltage offset term VFBSDOFF that has an additional temperature dependence, accounted for on the righ-hand side of Eq. (4.11), is used because the  $V_{fbsd}$  term of Eq. (4.12) does not provide sufficient temperature-dependence accuracy and extraction flexibility. This is because the parameter NGATE has been reserved for modeling the poly-silicon gate depletion effects. Having that additional temperature term helps to minimize the coupling between the modeling of the poly-silicon gate depletion effects and that of  $V_{fbsd}(T)$ .

#### 4.2.5 Gate-Channel Tunneling Current

##### 4.2.5.1 $I_{gc0}$ : The $V_{ds} = 0$ Bias Scenario

The gate-channel tunneling current  $I_{gc}$  in the inversion bias regime is a strong function of the source and drain bias. It is attributed to ECB for NMOS and HVB for PMOS.  $I_{gc0}$ , denoting  $I_{gc}$  at  $V_{ds} = 0$ , is

$$I_{gc0} = W_{eff} \cdot L_{eff} \cdot A \cdot \left( \frac{\text{TOXREF}}{\text{TOXE}} \right)^{\text{NTOX}} \cdot \left( \frac{V_{gse} \cdot V_{dens}}{\text{TOXE}^2} \right) \cdot \exp(-B \cdot \text{TOXE} \cdot (\text{AIGC} - \text{BIGC} \cdot V_{oxdepinv}) \cdot (1 + \text{CIGC} \cdot V_{oxdepinv})) \quad (4.13)$$

where the density of states and carriers available for tunneling is represented with

$$V_{dens} = \text{NIGC} \cdot v_t \cdot \log \left( 1 + \exp \left( \frac{V_{gse} - \text{VTH0}}{\text{NIGC} \cdot v_t} \right) \right) \quad (4.14)$$

When **IGCMOD** = 1, a zero-bias, long-channel threshold voltage global model parameter **VTH0** is used to simplify the  $I_{gc0}$  model formulation; an optional  $V_{dens}$  is expressed as

$$V_{dens} = \text{NIGC} \cdot v_t \cdot \log \left( 1 + \exp \left( \frac{V_{gse} - V_{th}}{\text{NIGC} \cdot v_t} \right) \right) \quad (4.15)$$

if **IGCMOD** = 2 is selected, the full  $V_{th}$  model equation is utilized to account for the body-bias effects on the gate-channel tunneling current, which is more accurate but requires more computation.

Note that in Eq. (4.13),  $A = q^3 / 8\pi h \phi_b = 4.97232 \times 10^{-7}$  [A·V<sup>-2</sup>] for NMOS and  $3.42537 \times 10^{-7}$  given in the unit of [A·V<sup>-2</sup>] for PMOS, and  $B = 8\pi \sqrt{2m_{ox}} \phi_b^{3/2} / 3qh = 7.45669 \times 10^{11}$  (g/F·s<sup>2</sup>)<sup>0.5</sup> for NMOS and  $1.16645 \times 10^{12}$  (g/F·s<sup>2</sup>)<sup>0.5</sup> for PMOS with  $m_{ox} = 0.4m_0$  and  $\phi_b = 3.13$  eV for ECB in the case of NMOS and  $m_{ox} = 0.326m_0$  and  $\phi_b = 4.25$  eV for HVB in the case of PMOS.

With  $I_{gc0}$  modeled, one is ready to present the source and drain partitioning of the gate-channel tunneling current when a non-zero  $V_{ds}$  bias is present.

#### 4.2.5.2 $I_{gcs}$ and $I_{gcd}$ Partitioning: The Non-Zero $V_{ds}$ Scenario

In order to carry out the partitioning of the gate-channel tunneling current  $I_{gc}$  under non-zero  $V_{ds}$ , the current continuity equation is solved for the voltage along the channel as a function of  $y$  ( $y = 0$  at the source and  $y = L_{eff}$  at the drain as illustrated in Fig. 4.4). Assuming the transistor is biased in the linear region, the channel current at position  $y$  is given by

$$I_{ch}(y) = W_{eff} \mu_{eff} C_{oxe} \cdot [V_{gst} - V(y)] \cdot \frac{dV(y)}{dy} \quad (4.16)$$

where  $V_{gst} = V_{gs} - V_{th}$ . Note that the bulk-charge effect accounted for by the  $A_{bulk}$  coefficient as  $C_{oxe} \cdot [V_{gst} - A_{bulk} \cdot V(y)]$  given in Chapter 3 is modeled with  $A_{bulk} = 1$  for convenience. Having a full term  $A_{bulk}$  will not complicate the mathematical derivation of the  $V_{ds}$  dependence model for the gate-channel direct tunneling. In fact, it is found that not incorporating  $A_{bulk}$  here would not lead to much noticeable accuracy loss for numerous advanced process technologies. However, the benefit here is the decoupling from the channel current model for ease of the tunneling model parameter extraction as well as less computation cost.

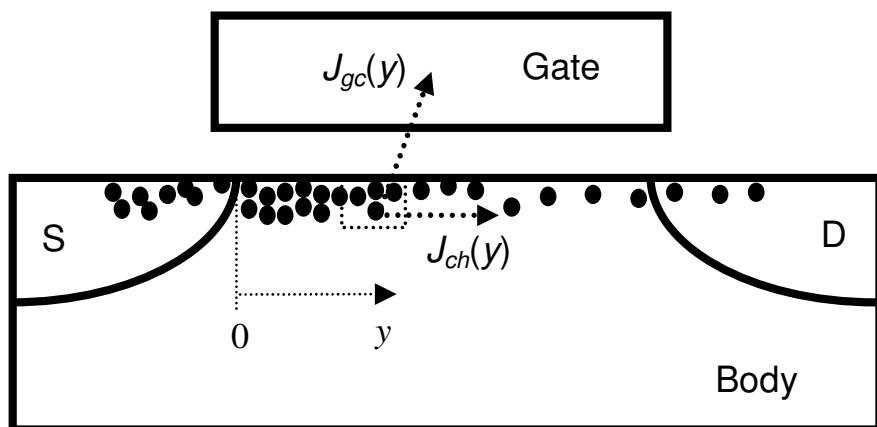


Fig. 4.4 The position-dependent gate-channel tunneling current needs to be taken into account in the current continuity equation for position  $y$  in the channel region, a method that one would start with to develop the source and drain partitioning of the gate-channel tunneling current. In this figure,  $J_{gc}(y)$  and  $J_{ch}(y)$  stand for the gate-channel and channel current densities at  $y$ , respectively.

Taking the gate-channel direct-tunneling current density  $J_{gc}(y)$  into consideration, the one-dimensional current continuity equation at any position  $y$  along the channel is [5]

$$\frac{dI_{ch}(y)}{dy} + W_{eff} \cdot J_{gc}(y) \equiv 0 \quad (4.17)$$

Note that this formulation was also discovered by Sah in 1961 [7] on the effect of surface channel and recombination on the p/n junction diode currents.

The position dependence of  $J_{gc}(y)$  is formulated with approximations such that Eq. (4.17) can be solved in a closed form as follows. The gate oxide field dependence of the tunneling mechanisms suggests

$$J_{gc}(y) \approx A \cdot E_{ox}^2(y) \cdot \exp\left(-\frac{B}{E_{ox}(y)}\right) \quad (4.18)$$

where the coefficient  $A$  given in the unit of  $[A \cdot V^{-2}]$  and exponent  $B$  in  $[V \cdot cm^{-1}]$  are material constants, and  $E_{ox}(y)$ , the electrical field strength in the gate oxide at  $y$ , is equal to

$$E_{ox}(y) = \frac{V_{ox}(0) - V(y)}{\text{TOXE}} \quad (4.19)$$

with  $V_{ox}(0) = V_{ox}(y)$  at  $y = 0$  and approximately equal to  $V_{gs}$ . This is because at the source side of the channel of an n<sup>+</sup>-poly NMOS and a p<sup>+</sup>-poly PMOS,  $V_{FB} \approx -\varphi_s$  and, thus,  $V_{gse} = V_{FB} + \varphi_s + V_{ox}(0) \approx V_{ox}(0)$ . Inserting Eq. (4.19) into Eq. (4.18), one obtains

$$J_{gc}(y) \approx A \cdot E_{ox}^2(y) \cdot \exp\left(-\frac{B \cdot \text{TOXE}}{V_{ox}(0)} \cdot \left[1 - \frac{V(y)}{V_{ox}(0)}\right]^{-1}\right) \quad (4.20)$$

Assuming that, in the linear region,  $E_{ox}(y)$  and  $V(y)$  along the channel do not differ appreciably from their respective values at the source side of the channel and applying Taylor series expansions to the second term in the exponent around  $V_{ox}(0)$  and retaining only the first-order term in the expansion, a more compact version of the  $y$  dependence of  $J_{gc}(y)$  is obtained

$$J_{gc}(y) \approx A \cdot E_{ox}^2(0) \cdot \exp\left(-\frac{B \cdot \text{TOXE}}{V_{ox}(0)} \cdot \left[1 + \frac{V(y)}{V_{ox}(0)}\right]\right) \quad (4.21)$$

An even more concise form is

$$J_{gc}(y) \approx J_{gc0} \cdot \exp(-B^* \cdot V(y)) \quad (4.22)$$

with

$$J_{gc0} = A \cdot E_{ox}^2(0) \cdot \exp\left(-\frac{B \cdot \text{TOXE}}{V_{ox}(0)}\right) \quad (4.23)$$

being the gate-channel tunneling current density at the source and

$$B^* = \frac{B \cdot \text{TOXE}}{V_{ox}(0)^2} \quad (4.24)$$

In order to prevent any potential divide-by-zero errors in SPICE simulation, a numerically robust version for  $B^*$  is implemented as

$$B^* \approx \frac{B \cdot \text{TOXE}}{V_{gse}^2} \approx \frac{B \cdot \text{TOXE}}{(V_{gsteff} + 1 \times 10^{-20})^2} \quad (4.25)$$

Substituting Eqs. (4.16) and (4.22) into the current continuity equation Eq. (4.17) yields a non-linear second-order differential equation of  $V(y)$

$$[V_{gst} - V(y)] \cdot \frac{d^2V(y)}{dy^2} - \left[ \frac{dV(y)}{dy} \right]^2 + \frac{J_{gc0}}{\mu_{eff} \cdot C_{oxe}} \cdot \exp(-B^* \cdot V(y)) = 0 \quad (4.26)$$

Equation (4.26), when  $V_{ds} < V_{gst}$ , can be readily solved if there is no tunneling, namely  $J_{gc0} = 0$ . The solution is

$$V_0(y) = V_{gst} - \sqrt{V_{gst}^2 - (2V_{gst} - V_{ds}) \cdot V_{ds} \cdot \frac{y}{L_{eff}}} \quad (4.27)$$

which satisfies the boundary conditions that  $V_0(y) = 0$  at  $y = 0$ , and  $V_0(y) = V_{ds}$  at  $y = L_{eff}$ . Again, by applying the Taylor expansion to Eq. (4.27) about  $y = 0$ , one obtains

$$V_0(y) = \eta \cdot y \quad (4.28)$$

with  $\eta$ , in the unit of  $[V \cdot m^{-1}]$ , being

$$\eta = \frac{1}{2L_{eff}} \cdot \frac{V_{ds} \cdot (2V_{gst} - V_{ds})}{V_{gst}} \quad (4.29)$$

An approximate analytical solution  $V(y)$  of Eq. (4.26), in the presence of the gate-channel tunneling, is developed by superposing on the channel potential  $V_0(y)$  the small changes  $V_\delta(y)$  induced by the tunneling process. It is

$$V(y) = V_0(y) + V_\delta(y) \quad (4.30)$$

Equation (4.26) now becomes a reduced, linear differential equation of  $V_\delta(y)$  of second order that makes an analytical solution possible.

$$(V_{gst} - \eta \cdot y) \cdot \frac{d^2 V_\delta(y)}{dy^2} - 2\eta \cdot \frac{dV_\delta(y)}{dy} + \frac{J_{gc0}}{\mu_{eff} \cdot C_{oxe}} \cdot \exp(-\eta B^* \cdot y) = 0 \quad (4.31)$$

In deriving the last equation, the  $\left(\frac{dV_0(y)}{dy}\right)^2$  and  $\left(\frac{dV_\delta(y)}{dy}\right)^2$  terms are neglected. There exists no analytical proof that all these assumptions are reasonable. Nevertheless, the difference between the analytical  $V(y)$  solution of Eq. (4.31) and the numerical solution of Eq. (4.26) are within a few percents over wide bias ranges and parameter values.

The analytical solution of Eq. (4.31),  $V_\delta(y)$ , is

$$V_\delta(y) = -\frac{J_{gc0}}{B^{*2} \eta^2 \mu_{eff} C_{oxe} \cdot (V_{gst} - \eta \cdot y)} \cdot \left\{ \left[ \exp(-B^* \eta \cdot y) - 1 \right] + \frac{y}{L_{eff}} \cdot \left[ 1 - \exp(-B^* \eta L_{eff}) \right] \right\} \quad (4.32)$$

The boundary conditions are  $V_\delta(0) = 0$  and  $V_\delta(L_{eff}) = 0$ . Note that the gate-channel tunneling process results in finite changes, either positive or negative or both depending on the bias conditions, to the potential within the channel region. It induces no changes at all to the source-drain voltage drop  $V_{ds}$  because it is fixed by the bias voltage.  $V_\delta(y)$ , in the case of n-channel MOS transistors operating in the linear region, for instance,

is greater than zero along the entire channel. This leads to the tunneling current flowing from the gate into the channel region and then splitting into source and drain.

To formulate the partitioning mathematically, the first-order derivative of  $V_\delta(y)$  with respect to  $y$  is given

$$\frac{dV_\delta(y)}{dy} = -\frac{J_{gc0}}{B^{*2} \eta^2 L_{eff} \mu_{eff} C_{oxe} \cdot (V_{gst} - \eta \cdot y)^2} \cdot \left\{ \begin{array}{l} V_{gst} \cdot [\exp(-B^* \eta L_{eff}) - 1] + \eta L_{eff} \cdot \exp(-B^* \eta \cdot y) \\ \cdot [B^*(V_{gst} - \eta \cdot y) - 1] + \eta L_{eff} \end{array} \right\} \quad (4.33)$$

With the bias and channel position dependencies of  $V_0(y)$  and  $V_\delta(y)$  already developed from the channel and gate tunneling current continuity equation, one is ready to proceed to derive the gate to source and gate to drain tunneling currents, symbolized by  $I_{gcs}$  and  $I_{gcd}$ . The approach is illustrated in Fig. 4.5, where  $I_{gcs}$ ,  $I_{gcd}$ , channel current  $I_{ch}$ , and total source ( $I_s$ ) and drain ( $I_d$ ) terminal current components are shown.

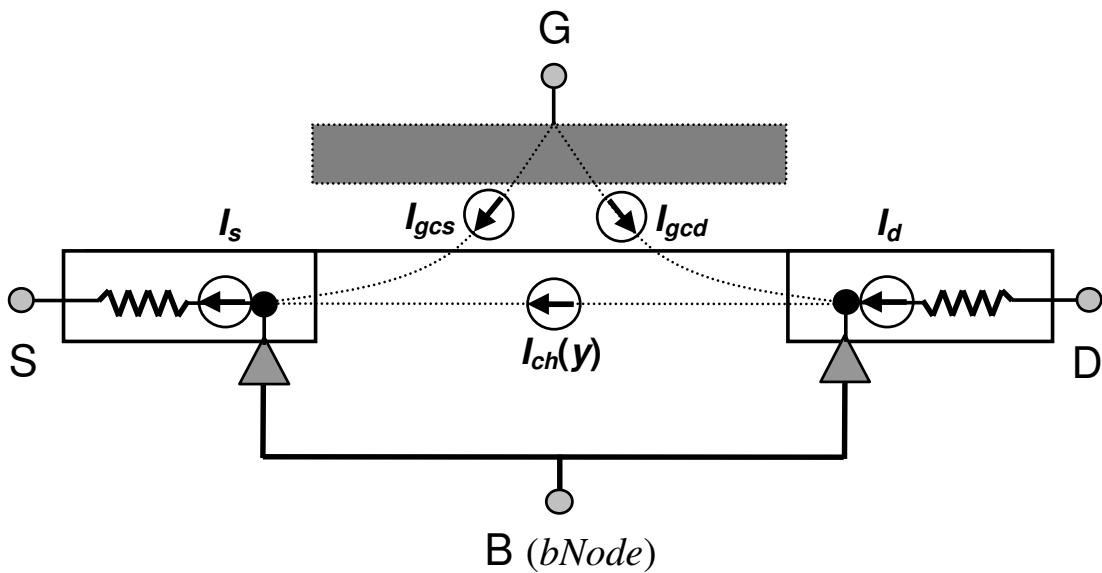


Fig. 4.5 Schematic gate-channel tunneling current model topology.

At the internal source node, KCL states

$$I_{gcs} = I_s - I_{ch}(y = 0) \quad (4.34)$$

$I_s$  is computed from Eq. (4.16) with  $y = 0$

$$I_s = W_{eff} \mu_{eff} C_{oxe} \cdot V_{gst} \cdot \frac{dV(y)}{dy} \Big|_{y=0} \quad (4.35)$$

$I_{ch}(y)$  at the source end is similarly derived with the gradient of  $V_0(y)$  at  $y = 0$ .

$$I_{ch}(y = 0) = W_{eff} \mu_{eff} C_{oxe} \cdot V_{gst} \cdot \frac{dV_0(y)}{dy} \Big|_{y=0} \quad (4.36)$$

Hence,  $I_{gcs}$  becomes

$$I_{gcs} = W_{eff} \mu_{eff} C_{oxe} \cdot V_{gst} \cdot \frac{dV_\delta(y)}{dy} \Big|_{y=0} \quad (4.37)$$

upon substituting Eqs. (4.35) and (4.36) into Eq. (4.34) and noting the definition  $V_\delta(y) = V(y) - V_0(y)$  according to Eq. (4.30). Entering  $y = 0$  into Eq. (4.33) permits rewriting  $I_{gcs}$  of Eq. (4.37)

$$I_{gcs} = \frac{W_{eff} L_{eff} \cdot J_{gc0}}{\eta^2 B^{*2} L_{eff}^2} \cdot [\exp(-\eta B^* L_{eff}) + \eta B^* L_{eff} - 1] \quad (4.38)$$

Similarly, at the internal drain node, KCL requires

$$I_{gcd} = I_{ch}(y = L_{eff}) - I_d \quad (4.39)$$

$I_d$  is obtained from Eqs. (4.16) and (4.28) and using the channel potential gradient with  $y = L_{eff}$

$$I_d = W_{eff} \mu_{eff} C_{oxe} \cdot (V_{gst} - \eta L_{eff}) \cdot \frac{dV(y)}{dy} \Big|_{y=L_{eff}} \quad (4.40)$$

$I_{ch}(y)$  at the drain end, using the gradient of  $V_0(y)$ , becomes

$$I_{ch}(y = L_{eff}) = W_{eff} \mu_{eff} C_{oxe} \cdot (V_{gst} - \eta L_{eff}) \cdot \frac{dV_0(y)}{dy} \Big|_{y=L_{eff}} \quad (4.41)$$

$I_{gcd}$  of Eq. (4.39) becomes

$$I_{gcd} = -W_{eff} \mu_{eff} C_{oxe} \cdot (V_{gst} - \eta L_{eff}) \cdot \frac{dV_\delta(y)}{dy} \Big|_{y=L_{eff}} \quad (4.42)$$

where  $V_\delta(y) = V(y) - V_0(y)$  of Eq. (4.30) has been used. Equation (4.42) becomes

$$I_{gcd} = \frac{W_{eff} L_{eff} \cdot J_{gc0}}{\eta^2 B^{*2} L_{eff}^2} \cdot \left[ -\exp(-\eta B^* L_{eff}) \cdot (\eta B^* L_{eff} + 1) + 1 \right] \quad (4.43)$$

once the derivative of  $V_\delta(y)$  with respect to  $y$  at  $y = L_{eff}$  is obtained by entering  $y = L_{eff}$  into Eq. (4.33).

Adding  $I_{gcs}$  of Eq. (4.38) and  $I_{gcd}$  of Eq. (4.43) gives the total gate-channel tunneling current

$$I_{gc}(V_{ds}) = I_{gcs} + I_{gcd} = \frac{W_{eff} L_{eff} \cdot J_{gc0}}{\eta B^* L_{eff}} \cdot [1 - \exp(-\eta B^* L_{eff})] \quad (4.44)$$

$B^*$  in  $[V^{-1}]$  and  $\eta$  in  $[V \cdot m^{-1}]$  can be rewritten for convenience as

$$B^* \approx \frac{B \cdot \text{TOXE}}{(V_{gsteff} + 1 \times 10^{-20})^2} \quad (4.25)$$

and

$$\eta = \frac{1}{2L_{eff}} \cdot \frac{V_{ds} \cdot (2V_{gst} - V_{ds})}{V_{gst}} \quad (4.29)$$

A quick check on Eqs. (4.38), (4.43), and (4.44) reveals that they do not hold true for  $V_{ds} = 0$  because according to Eq. (4.29), a divide-by-zero error would result as  $\eta$  approaches zero when  $V_{ds} = 0$ .

In order to eliminate this numerical instability, let the variable  $P_{igcd}$  be

$$P_{igcd} = B^* \cdot \left[ 1 - \frac{V_{dseff}}{2 \cdot (V_{gsteff} + 1 \times 10^{-20})} \right] \quad (4.45)$$

and Eq. (4.29) be approximated with smoothing functions  $V_{gsteff}$  and  $V_{dseff}$

$$\eta \approx \frac{1}{2L_{eff}} \cdot \frac{V_{dseff} \cdot [2(V_{gsteff} + 1 \times 10^{-20}) - V_{dseff}]}{V_{gsteff} + 1 \times 10^{-20}} \quad (4.46)$$

such that

$$\eta B^* L_{eff} = P_{igcd} \cdot V_{dseff} \quad (4.47)$$

Thus, Eqs. (4.38), (4.43), and (4.44) can now be transformed into

$$I_{gcs} = I_{gc}(V_{ds} = 0) \cdot \frac{P_{igcd} \cdot V_{dseff} + \exp(-P_{igcd} \cdot V_{dseff}) - 1 + 1 \times 10^{-4}}{P_{igcd}^2 \cdot V_{dseff}^2 + 2 \times 10^{-4}} \quad (4.48)$$

$$I_{gcd} = I_{gc}(V_{ds} = 0) \cdot \frac{1 - (P_{igcd} \cdot V_{dseff} + 1) \cdot \exp(-P_{igcd} \cdot V_{dseff}) + 1 \times 10^{-4}}{P_{igcd}^2 \cdot V_{dseff}^2 + 2 \times 10^{-4}} \quad (4.49)$$

and

$$\begin{aligned} I_{gc}(V_{ds}) &= I_{gcs} + I_{gcd} \\ &= I_{gc}(V_{ds} = 0) \cdot \frac{P_{igcd} \cdot V_{dseff} \cdot [1 - \exp(-P_{igcd} \cdot V_{dseff})] + 2 \times 10^{-4}}{P_{igcd}^2 \cdot V_{dseff}^2 + 2 \times 10^{-4}} \end{aligned} \quad (4.50)$$

where

$$I_{gc}(V_{ds} = 0) = W_{eff} L_{eff} \cdot J_{gc0} \quad (4.51)$$

Recall from Eq. (4.23) that

$$J_{gc0} = A \cdot E_{ox}(0)^2 \cdot \exp\left(-\frac{B \cdot \text{TOXE}}{V_{ox}(0)}\right) \quad (4.23)$$

In the BSIM4 implementation, the gate-channel tunneling current  $I_{gc}(V_{ds} = 0)$  under zero  $V_{ds}$  in Eqs. (4.48), (4.49), and (4.50) is actually computed by using Eq. (4.13), which is

$$\begin{aligned} I_{gc}(V_{ds} = 0) &= I_{gc0} = W_{eff} \cdot L_{eff} \cdot A \cdot \left(\frac{\text{TOXREF}}{\text{TOXE}}\right)^{\text{NTOX}} \cdot \left(\frac{V_{gse} \cdot V_{dens}}{\text{TOXE}^2}\right) \\ &\cdot \exp\left(-B \cdot \text{TOXE} \cdot (\text{AIGC} - \text{BIGC} \cdot V_{oxdepinv}) \cdot (1 + \text{CIGC} \cdot V_{oxdepinv})\right) \end{aligned} \quad (4.13)$$

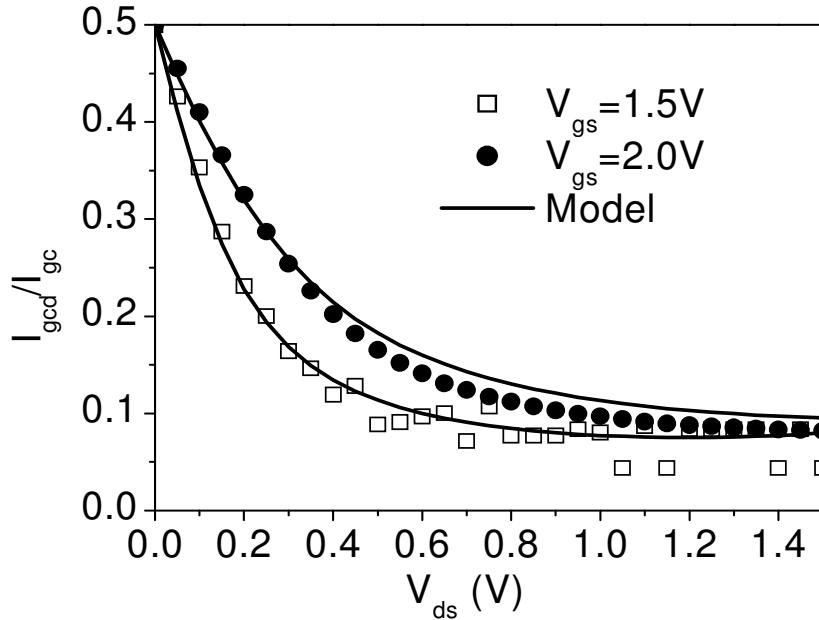
It should be noted that the two small constants ( $1 \times 10^{-4}$  and  $2 \times 10^{-4}$ ) that are introduced to avoid potential divide-by-zero errors in the above derivations help to realize that

$$I_{gcs} = I_{gcd} = 0.5 \cdot I_{gc}(V_{ds} = 0) \quad (4.52)$$

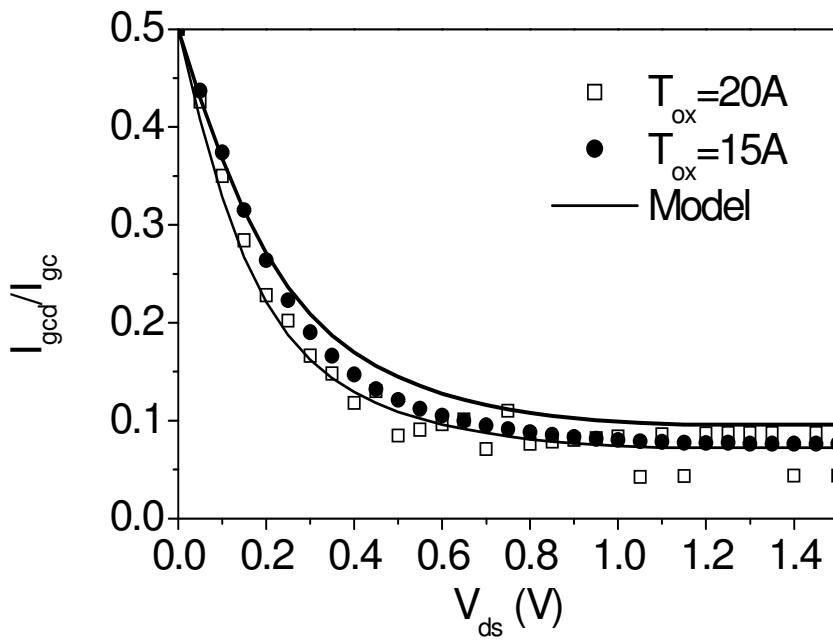
when  $V_{ds} = 0$ . By symmetry, the gate-channel tunneling current is partitioned equally between source and drain at  $V_{ds} = 0$ . Fig. 4.6 illustrates the agreement of the ratio of  $I_{gcd}$  to  $I_{gc}$  for various gate biases and the gate oxide thicknesses between the model and 2-D TCAD numerical simulations [5].

A global model parameter **PIGCD** replaces the variable  $P_{igcd}$  of Eq. (4.45) when **PIGCD** is specified in a BSIM4 model card library. This improves the model evaluation CPU runtime at the cost of slight accuracy loss.

All the gate tunneling current components derived above need to be multiplied by the multi-finger number **NF** to obtain the total value of each current component if **NF** is greater than 1.



(a)



(b)

Fig. 4.6 The ratio of  $I_{gcd}$  and  $I_{gc}(V_{ds})$  versus  $V_{ds}$  for (a) different  $V_{gs}$  and (b) gate oxide thicknesses (physical values  $\text{TOXP}$  are shown). Symbols designate the 2-D TCAD simulations and lines are given by the BSIM4 gate-channel tunneling current partitioning model. Note that at  $V_{ds} = 0$ , the ratio is 0.5, meaning equal partitioning of  $I_{gc}$  between the source and drain terminals.

#### 4.2.6 Characterization and Parameter Extraction

Equation (4.39) is the KCL branch current equation for the drain node when the gate-channel current is present

$$I_{gcd} = I_{ch}(y = L_{eff}) - I_d \quad (4.39)$$

here the other branch currents such as the impact ionization, GIDL, and junction diode currents are left out for the convenience of the discussions below. Rearranging Eq. (4.39) such that

$$I_d = I_{ch}(y = L_{eff}) - I_{gcd} \quad (4.53)$$

It implies that the drain terminal current  $I_d$  can become less than the channel current  $I_{ch}$  by the amount of  $I_{gcd}$ .

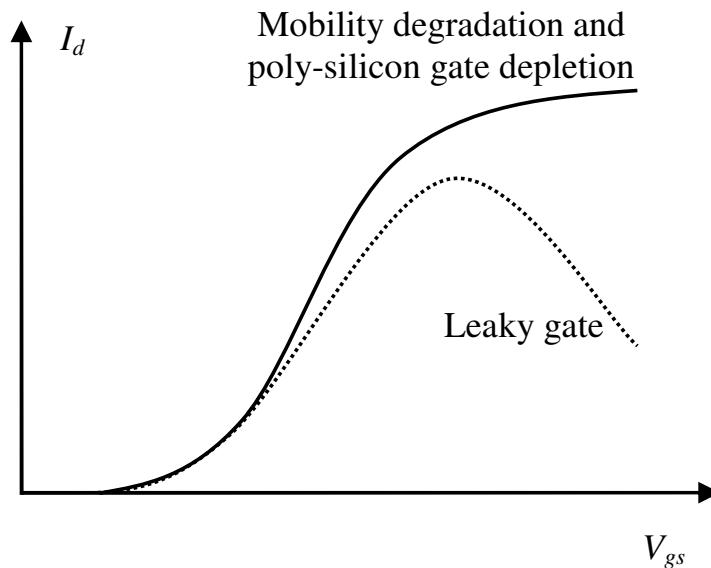


Fig. 4.7 Linear-scale drain current in the strong inversion region of operation when  $V_{ds}$  is small. In the case of the solid curve, the leveling off of the current is caused primarily by the carrier mobility degradation; the increase in the effective gate oxide thickness due to the poly-silicon gate depletion is sometimes a secondary factor. In the case illustrated by the dashed line, the large and rapid drop of the current is attributable to a large gate direct-tunneling leakage current.

Figure 4.7 illustrates the observed drain terminal current when a significant gate-channel tunneling is present in comparison with the case where the tunneling is absent. The large and rapid drop in the current results from significant amount of channel carriers (e.g., electrons in the case of NMOS) being pulled out of the channel region and tunneling into the gate before they can reach the drain terminal (refer to Fig. 4.4 again).

Attention must be paid to extracting the gate tunneling current parameter from the drain terminal current data. Otherwise, the  $V_{gs}$ -dependence model parameters of the mobility model could be overestimated and lead to a negative trans-conductance  $G_m$ . That would be both inaccurate and also harmful to SPICE simulation convergence.

In the case where transistors are very leaky due to gate tunneling (not due to the sub-threshold leakage), the channel current and gate tunneling current model parameters need to be extracted sequentially and then optimized simultaneously. This can be done in a four-step procedure. The first step is to extract the parameters for the tunneling currents  $I_{gc0}$  (from shorted source and drain) and  $I_{gb}$ ; The second is to extract the parameters of the  $P_{igcd}$  expression or simply PIGCD itself from the gate-channel tunneling current as a function of drain and gate voltages. The third step is the extraction of the  $I_{ch}$  current parameters including the mobility parameters. The last step, if desired, would be to perform a global optimization of the tunneling and  $I_{ch}$  parameters simultaneously. This practice should be applied to the characterization and extractions for the CV model as well.

### 4.3 Body Currents

In addition to the gate-body tunneling current  $I_{gb}$  that was discussed in the previous section, the impact ionization and gate-induced source/drain leakage currents (denoted by  $I_{ii}$  and  $I_{GISL}/I_{GIDL}$ ) contribute to the body terminal current. In the following, these body current components are discussed with respect to their physical mechanisms and BSIM4 models. The body current components due to the source/drain-body junction diodes will be presented in Chapter 9.

### 4.3.1 Impact Ionization

When they transit from the source to the drain, channel carriers can gain large kinetic energy (a few eVs) in the high electric-field region near the drain end of the channel. Statistically, energetic carriers can undergo ionizing collisions with the valence electrons of the silicon atoms in the silicon lattice.

An ionizing collision generates an electron and hole pair. One might say that these electrons are hot and they have a temperatures  $T_e$  that is larger than the lattice (silicon) temperature. Using NMOSFET for example, the generated electrons flow into the drain terminal together with the original channel current. On the other hand, the generated holes flow into the substrate forming the impact-ionization body current  $I_{ii}$  [8]. This process is illustrated in Fig. 4.8.

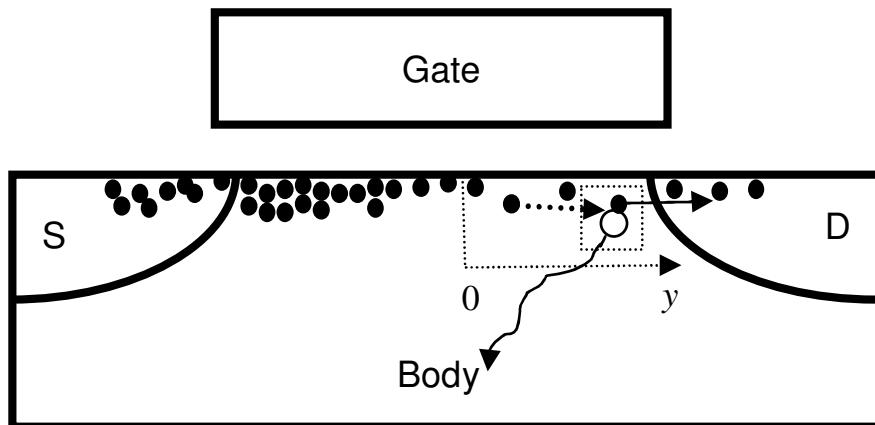


Fig. 4.8 Impact ionization and induced body current: The electron and hole pair generation event is as marked in the box. The generated holes, under the influence of the electric field, flow into the silicon substrate.

$I_{ii}$  is equal to the product of the channel current (the number of channel carriers passing through the channel per second) and the integral of the electrons or holes generated by impact ionization with each channel carrier [8, 9]

$$I_{ii} = I_{ch} \cdot A_i \cdot \int_{y=0}^{y=\Delta L} \exp\left(-\frac{B_i}{E_{ch}(y)}\right) dy \quad (4.54)$$

$A_i$  and  $B_i$  are two material constants.  $E_{ch}(y)$  is the longitudinal electric field along the current path.  $\Delta L$  is the length of the velocity saturation region of the channel and  $y = 0$  starts at the onset point of the velocity saturation.

$E_{ch}(y)$  can be obtained from a pseudo-2D solution of Poisson and Gaussian equations.

$$E_{ch}(y) = E_{sat} \cdot \cosh\left(\frac{y}{LitL}\right) \quad (4.55)$$

It has an exponential dependence on  $y$  and can be approximated by

$$E_{ch}(y) \approx \sqrt{E_{sat}^2 + \left[ \frac{V_{ch}(y) - V_{dsat}}{LitL} \right]^2} \quad (4.55a)$$

where  $LitL$ , the characteristic drain-field length of the  $E_{ch}(y)$  dependence on  $y$ , is given by

$$LitL = \sqrt{3.0 \cdot XJ \cdot TOXE} \quad (4.55b)$$

The constant 3.0 is the ratio of the permittivity of Si and  $\text{SiO}_2$ . This constant is adjusted when a non- $\text{SiO}_2$  or high- $k$  gate dielectric is employed.

Replacing  $dy$  of Eq. (4.54) with  $dE_{ch}(y)$ ,  $I_{ii}$  now becomes

$$I_{ii} = I_{ch} \cdot A_i \cdot LitL \cdot \int_{E_{sat}}^{E_{ch}(y=\Delta L)} \frac{\exp\left(-\frac{B_i}{E_{ch}(y)}\right)}{\sqrt{E_{ch}(y)^2 - E_{sat}^2}} dE_{ch}(y) \quad (4.56)$$

This leads to

$$I_{ii} \approx \frac{I_{ch} \cdot A_i \cdot LitL \cdot E_{ch}(y = \Delta L)}{B_i} \cdot \exp\left(-\frac{B_i}{E_{ch}(y = \Delta L)}\right) \quad (4.56a)$$

The electric field at the drain end of the channel  $E_{ch}(y = \Delta L)$  is

$$E_{ch}(y = \Delta L) \approx \sqrt{E_{sat}^2 + \left( \frac{V_{ds} - V_{dsat}}{LitL} \right)^2} \quad (4.57)$$

by noting from Eq. (4.55a) that at  $y = \Delta L$ ,  $V_{ch}(y) = V_{ds}$ . Equation (4.57) can also be simplified to

$$E_{ch}(y = \Delta L) \approx \frac{V_{ds} - V_{dsat}}{LitL} \quad (4.58)$$

by dropping the  $E_{sat}^2$  term as it is much smaller than  $\frac{V_{ds} - V_{dsat}}{LitL}$ . This

leads Eq. (4.56a) to

$$I_{ii} \approx \frac{A_i}{B_i} \cdot I_{ch} \cdot (V_{ds} - V_{dsat}) \cdot \exp\left(-\frac{B_i \cdot LitL}{V_{ds} - V_{dsat}}\right) \quad (4.59)$$

Equation (4.59) serves as the theoretical foundation for the BSIM4 impact ionization current  $I_{ii\text{-BSIM4}}$  model, which is implemented in BSIM4 as

$$\begin{aligned} I_{ii\text{-BSIM4}} = & \left( \text{ALPHA1} + \frac{\text{ALPHA0}}{L_{eff}} \right) \cdot I_{ch\text{-no-SCBE}} \\ & \cdot (V_{ds} - V_{dseff}) \cdot \exp\left(-\frac{\text{BETA0}}{V_{ds} - V_{dseff}}\right) \end{aligned} \quad (4.60)$$

where, benefiting from the  $V_{dseff}$  formulation given in Chapter 3,  $(V_{ds} - V_{dseff})$  is a non-negative quantity. Note that if

$$\frac{\text{BETA0}}{V_{ds} - V_{dseff}} > \text{EXP\_THRESHOLD}$$

with  $\text{EXP\_THRESHOLD}$  chosen to be 34, the exponential term of Eq. (4.60) is replaced with a constant  $\text{MIN\_EXP} = 3.720075976 \times 10^{-44}$  for both numerical robustness and a tradeoff between necessary accuracy and program execution efficiency. It is found that  $\text{EXP\_THRESHOLD} = 34$  is a good choice for the modeling of MOSFET devices.

Introducing the  $\text{ALPHA0}/L_{\text{eff}}$  term into Eq. (4.60) is found to provide better accuracy of  $I_{\text{ii}}$  over a wide range of channel lengths. One reason for such a term can be that there have been several approximations and simplifications in deriving Eq. (4.59), including those made in the pseudo-2D Gaussian analyses that lead to a linear dependence of  $E_{ch}(y)$  on  $(V_{ds} - V_{dsat})$ . It is worth noting that the impact ionization current model of BSIMSOI4 is more sophisticated in order to accurately model the body current and, hence, the floating body and history effects. It is very useful for the modeling of MOS aging due to hot-carrier injection.

The magnitude of the impact ionization body terminal current itself is usually quite small relative to the drain terminal current. In fact, it can sometimes be negligible for logic circuit designs. For memory and some precision analog circuit designs, incorporation of an accurate  $I_{\text{ii}}$  modeling is a must. For instance, this is true for a circuit that involves source followers, where the body of the transistor is connected to the source. The source is then connected to the source of the second MOSFET in series. In this case, because of the finite resistance/impedance presented by the second transistor, the presence of the impact ionization current of the first transistor modifies the body terminal voltage of its own.

However, impact ionization causes an increase in the drain current many times larger than  $I_{\text{ii}}$  due to the *SCBE* effect (refer to Chapter 3 for details), namely the substrate-current induced body effect. *SCBE* leads to an increase in the saturation channel current and a decrease of the output resistance ( $R_{\text{out}}$ ). The body of a MOSFET transistor has a finite resistance. When  $I_{\text{ii}}$  is present, the internal body node potential inside the device increases and consequently the threshold voltage decreases. This effect is modeled in BSIM4 in the following form

$$I_{ch} = I_{ch-no-SCBE} \cdot \left( 1 + \frac{V_{ds} - V_{dseff}}{V_{ASCBE}} \right) \quad (4.61)$$

$I_{ch-no-SCBE}$  includes the number of fingers  $NF$  and all the physical mechanisms of the  $R_{\text{out}}$  model (including velocity saturation, channel-length modulation *CLM*, drain-induced barrier lowering *DIBL*, pocket

implant effects *DITS*, velocity overshoot and source-end velocity limit *SEVL* as analyzed in Chapter 3) except the substrate current induced body effect. The same  $I_{ch\text{-}no\text{-}SCBE}$  is used in  $I_{ii\text{-}BSIM4}$  of Eq. (4.60).

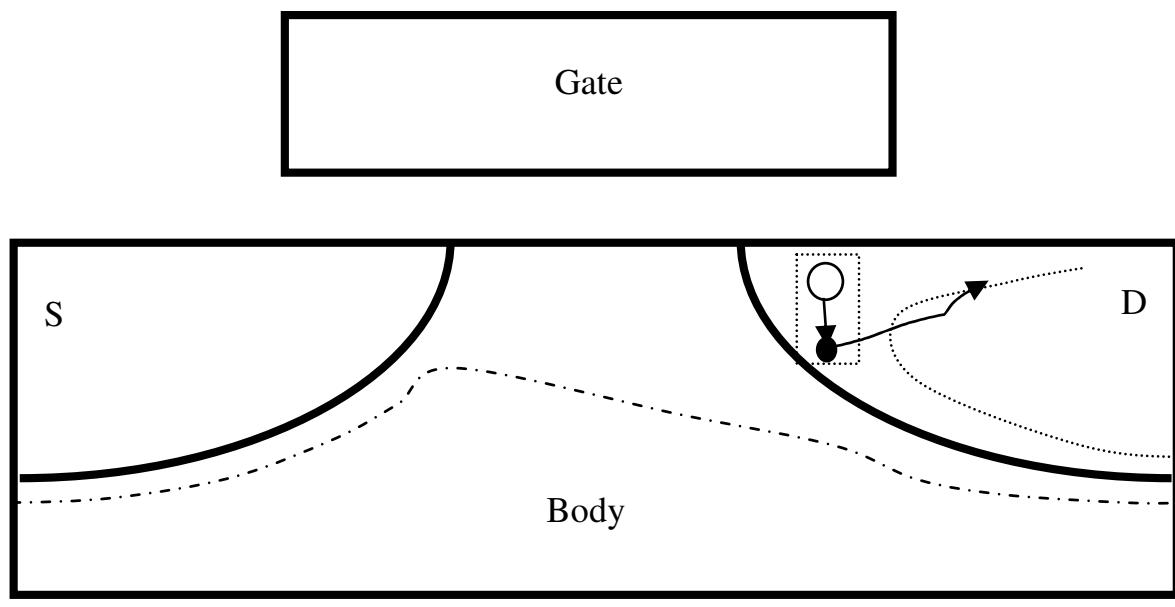
To comprehend the above discussions, one can formulate the total contribution of impact ionization to the drain terminal current as

$$I_d = I_{ch} + I_{ii} = I_{ch\text{-}no\text{-}SCBE} \cdot \left( 1 + \frac{V_{ds} - V_{dseff}}{V_{ASCBE}} \right) + I_{ii} \quad (4.62)$$

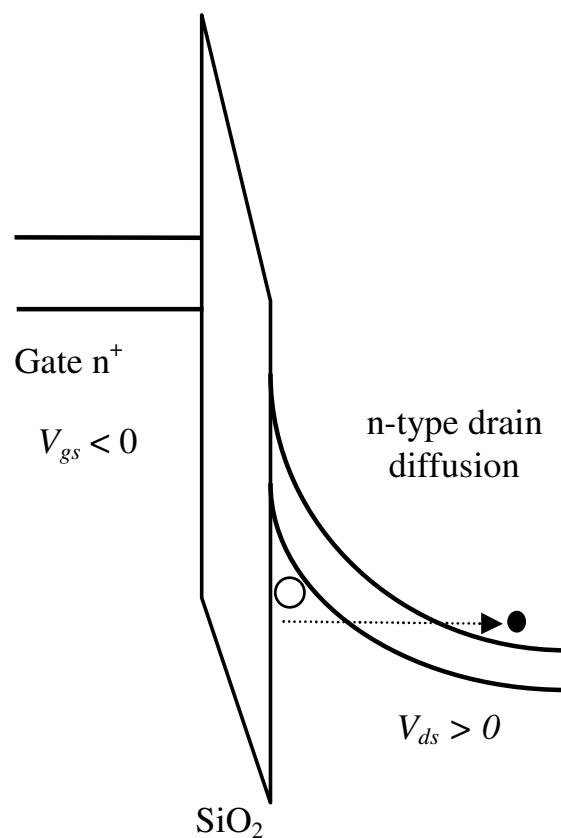
Note that a separate set of model parameters are employed for the *SCBE* and  $I_{ii}$  modeling to improve the flexibility of parameter extraction. To turn off the modeling and program execution of  $I_{ii\text{-}BSIM4}$  (Eq. (4.60)), just set both the global model parameters **ALPHA0** and **ALPHA1** to zero in model cards.

### 4.3.2 Gate-Induced Source and Drain Leakage

Take an n-channel MOSFET as example. When the gate voltage is negative and the drain voltage is positive, the surface of the n-type drain diffusion region becomes depleted of electrons and, hence, a depletion region results. This depletion region becomes wider for a larger negative gate-to-drain voltage  $v_{gd}$ . Moreover, (see Fig. 4.9 (a)), the energy band of the depletion region bends up at the interface with the gate oxide. When the amount of bending exceeds the energy-band gap, the electrons of the valence band can tunnel into the conduction band where there are available receiving energy states (see Fig. 4.9 (b)). This process is known as band-to-band tunneling. Upon finishing the tunneling process, these electrons flow out from the drain terminal, contributing to drain terminal current. It becomes larger in magnitude as  $v_g$  becomes more negative. As a result, it is more difficult to turn off the transistor (Fig. 4.9 (c)). This physical phenomenon and its related leakage current are referred to as the gate-induced drain leakage current, or simply  $I_{GIDL}$  [10]. The same takes place also in the source diffusion region, which is symbolized by  $I_{GISL}$ . Both are important leakage mechanisms for low-power/portable and memory devices.



(a)



(b)

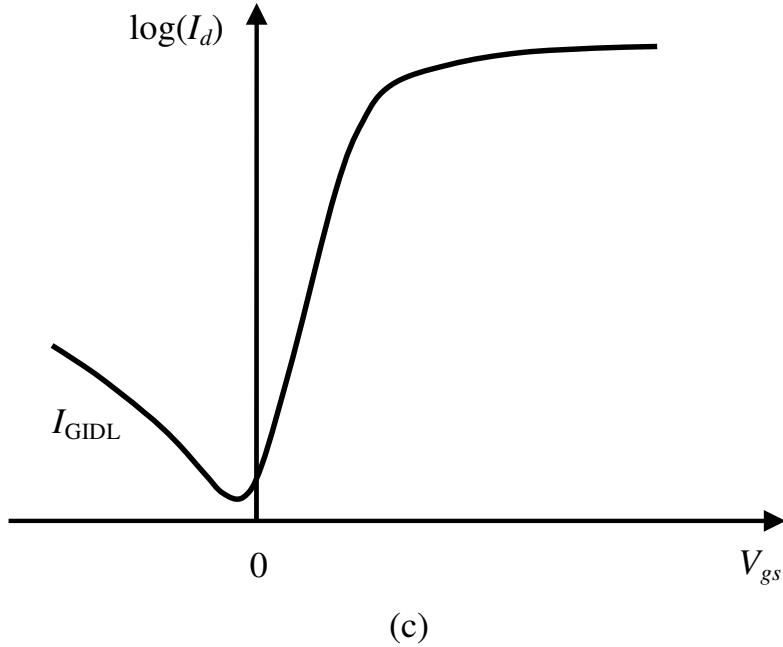


Fig. 4.9 Gate-induced drain leakage current  $I_{GIDL}$  in the case of NMOS. The deep depletion region induced by  $V_{gd}$  and  $V_b$  in the drain diffusion (a); negative  $V_{gd}$  induced energy-band bending in the overlap region that triggers the band-to-band carrier tunneling (b); and a sketch of the  $\log(I_d)$ - $V_{gs}$  illustrating the behavior of the GIDL leakage current.

BSIM4 models  $I_{GIDL}$  and  $I_{GISL}$  as

$$I_{GIDL} = AGIDL \cdot W_{effJCT} \cdot NF \cdot \frac{v_{gd\_eff} + EGIDL}{3 \cdot TOXE} \cdot \exp\left(\frac{3 \cdot TOXE \cdot BGIDL}{v_{gd\_eff} + EGIDL}\right) \cdot \frac{v_{bd}^3}{CGIDL - v_{bd}^3} \quad (4.63)$$

and

$$I_{GISL} = AGISL \cdot W_{effJCT} \cdot NF \cdot \frac{v_{gs\_eff} + EGISL}{3 \cdot TOXE} \cdot \exp\left(\frac{3 \cdot TOXE \cdot BGISL}{v_{gs\_eff} + EGISL}\right) \cdot \frac{v_{bs}^3}{CGISL - v_{bs}^3} \quad (4.64)$$

The GIDL and GISL parameters are different because CMOS processes can possibly have deliberately different source and drain diffusion doping profiles. In the model formulation, all the terminal voltages are physical terminal voltages, which implies that  $I_{\text{GIDL}}$  and  $I_{\text{GISL}}$  should not be interchanged or swapped with each other when the transistor changes its operation from forward to reverse mode, and vice versa. Note that in Eq. (4.64),  $v_{gs\_eff}$  and  $v_{gd\_eff}$  include the poly-silicon gate depletion effects. The model parameter  $\text{EGIDL}$  represents the band bending needed for the onset of a band-to-band tunneling. Other parameters are given in the Parameter Table at the end of this Chapter. It should be pointed out that Eqs. (4.63) and (4.64) are implemented in BSIM4 such that  $I_{\text{GIDL}}$  and  $I_{\text{GISL}}$  both take on non-negative values. In the case of NMOS, they flow from the drain and source terminal into the body. For PMOS transistors, they flow in the opposite directions, from the body to the source and drain. Fig. 4.10 shows the voltage dependencies of the measured and modeled  $I_{\text{GIDL}}$ .

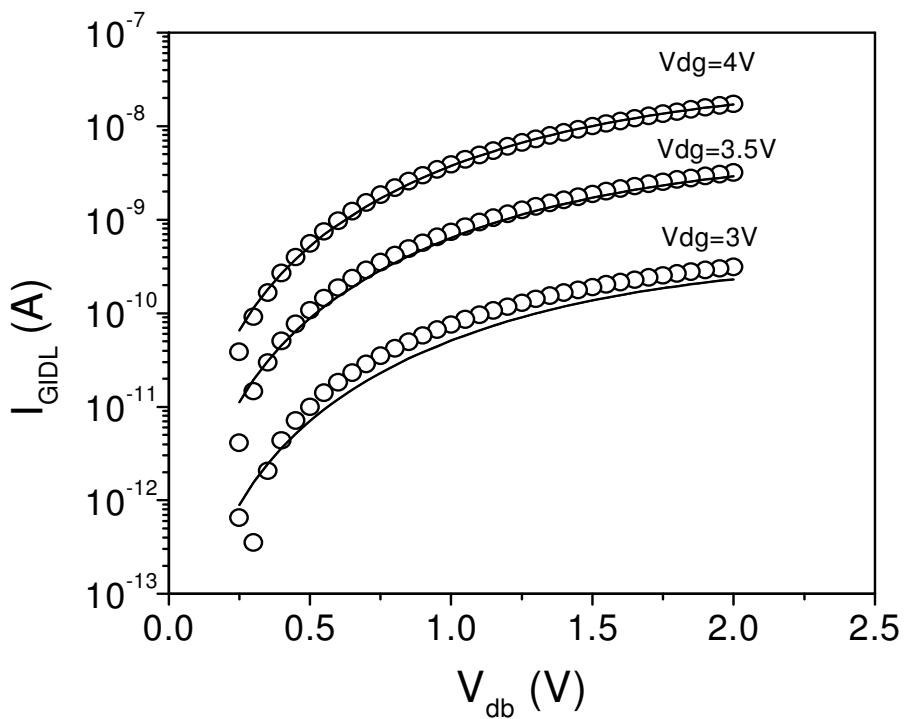


Fig. 4.10 Comparisons of measured (symbols) and modeled (lines)  $I_{\text{GIDL}}$ .  $\text{AGIDL} = 7.16 \times 10^{-13}$  mho,  $\text{BGIDL} = 2.3 \times 10^9$  V/m,  $\text{CGIDL} = 8.16 \text{ V}^3$ , and  $\text{EGIDL} = 0.8$  V.

## 4.4 Summary of BSIM4 Branch and Terminal DC Currents

Figure 4.11 shows the topology of all constitutive DC branch current components considered in BSIM4. The total current that flows into/out each device terminal is obtained as follows. This provides a good guidance for MOS transistor characterization and model extraction. It also lends a useful reference in loading the BSIM4 model into circuit simulation matrices for SPICE to solve (refer to Chapter 10 for the detailed implementation methodology).

The sum of the DC (including tunneling) currents associated with the gate terminal is

$$I_g = -(I_{gs} + I_{gd} + I_{gcs} + I_{gcd} + I_{gb}) \quad (4.65)$$

All the DC currents that flow into/out the body terminal are

$$I_b = -I_{js} - I_{jd} + I_{GISL} + I_{GIDL} + I_{ii} + I_{gb} \quad (4.66)$$

For the internal source and drain nodes, one may desire to write the KCL branch current equations instead. Thus, at the internal drain node, this equation is

$$I_{Rd} + I_{gd} + I_{gcd} + I_{jd} - I_{GIDL} - I_{ii} - I_{ch} \equiv 0 \quad (4.67)$$

where  $I_{Rd}$  (and  $I_{Rs}$  similarly below) is the current source contributed by the drain (source in the case of  $I_{Rs}$ ) resistance. At the internal source node, a similar KCL equation holds

$$-I_{Rs} + I_{gs} + I_{gcs} + I_{js} - I_{GISL} + I_{ch} \equiv 0 \quad (4.68)$$

In the above derivations, all inflow current components take a positive sign and any outflow current takes the opposite. Note that the

polarities of several current components are determined by the polarities of the terminal voltages in their model equations. These components include  $I_{js}$ ,  $I_{jd}$ ,  $I_{gs}$ ,  $I_{gd}$ , and  $I_{gb}$ . The junction diode DC current modeling will be presented in Chapter 9.

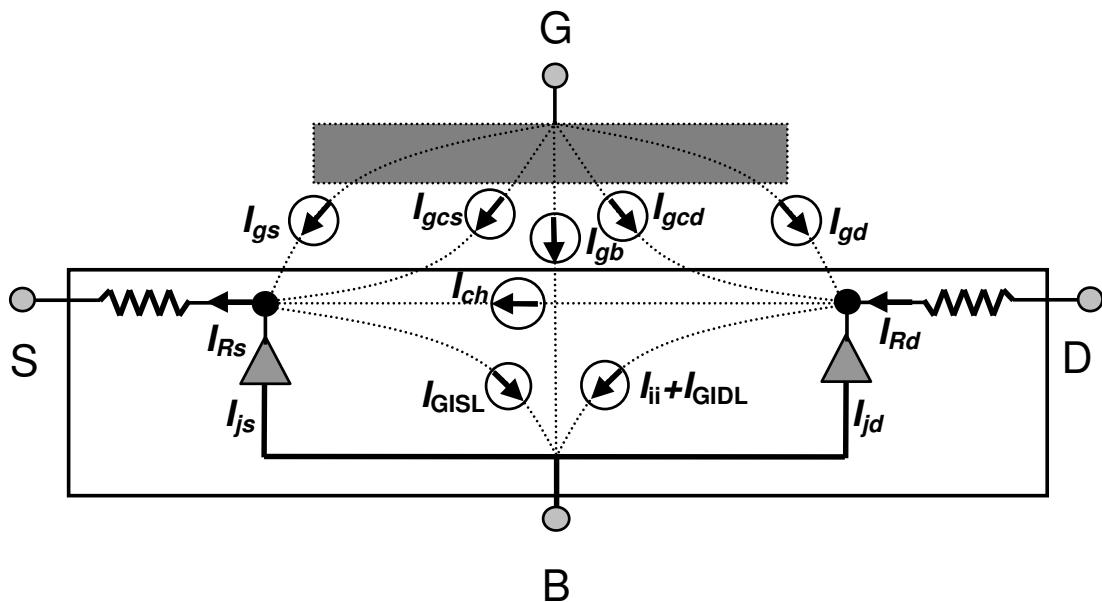


Fig. 4.11 A complete topology of BSIM4 branch DC current components. Grayed circles represent the external nodes, whereas the darkened dots denote the internal source and drain nodes.

## 4.5 Chapter Summary

This chapter presented and discussed in detail the physical mechanisms of the gate direct-tunneling, impact ionization, and gate-induced drain leakage currents and the models that have been developed for BSIM4. The gate-channel tunneling current partition was derived from the current continuity equation, which is a very useful equation for semiconductor device modeling. All these current components must be accurately modeled in a general purpose compact model. These DC current models together with those in Chapters 2 and 3 have served the CMOS technologies from the 130nm down to today's 20nm nodes with excellent accuracy.

## 4.6 Parameter Table

Name (type)	Description and default	Can be binned?	Note
IGCMOD (Global; integer)	<p>Model selector for <math>I_{gs}</math>, <math>I_{gd}</math>, <math>I_{gcs}</math> and <math>I_{gcd}</math>.</p> <p>Default = 0; dimensionless: No <math>I_{gs}</math>, <math>I_{gd}</math>, <math>I_{gcs}</math> and <math>I_{gcd}</math> to be computed. Optional values are 1 (long-channel zero-bias VTH0 to be used in the gate-and-channel tunneling current model computation) and 2 (the full-equation <math>V_{th}</math> model to be used in the gate-and-channel tunneling current model computation).</p>	No	IGCMOD = 1 and 2 turns on $I_{gs}$ , $I_{gd}$ , $I_{gcs}$ and $I_{gcd}$ .
IGBMOD (Global; integer)	<p>Model selector for <math>I_{gb}</math>.</p> <p>Default = 0; dimensionless: No <math>I_{gb}</math> to be computed.</p>	No	IGBMOD = 1 turns on $I_{gb}$ .
NF (Local; integer)	<p>The number of fingers that a multi-finger device structure has.</p> <p>Default = 1; dimensionless.</p>	No	Reset to 1 if $\leq 1$ with a warning to be issued.
AIGBACC (Global; double)	<p>Parameter for <math>I_{gb}</math> in accumulation.</p> <p>Default = 0.43 in [<math>(\text{Fs}^2/\text{g})^{0.5} \text{ m}^{-1}</math>].</p>	Yes	-
BIGBACC (Global; double)	<p>Parameter for <math>I_{gb}</math> in accumulation.</p> <p>Default = 0.054 in [<math>(\text{Fs}^2/\text{g})^{0.5} (\text{mV})^{-1}</math>].</p>	Yes	-
CIGBACC (Global; double)	<p>Parameter for <math>I_{gb}</math> in accumulation.</p> <p>Default = 0.075 in [<math>\text{V}^{-1}</math>].</p>	Yes	-
NIGBACC (Global; double)	<p>Parameter for <math>I_{gb}</math> in accumulation.</p> <p>Default = 1.0; dimensionless.</p>	Yes	Fatal errors to be issued if its binned value $\leq 0.0$ .
AIGBINV (Global; double)	<p>Parameter for <math>I_{gb}</math> in inversion.</p> <p>Default = 0.35 in [<math>(\text{Fs}^2/\text{g})^{0.5} \text{ m}^{-1}</math>].</p>	Yes	-
BIGBINV (Global; double)	<p>Parameter for <math>I_{gb}</math> in inversion.</p> <p>Default = 0.03 in [<math>(\text{Fs}^2/\text{g})^{0.5} (\text{mV})^{-1}</math>].</p>	Yes	-
CIGBINV (Global; double)	<p>Parameter for <math>I_{gb}</math> in inversion.</p> <p>Default = 0.006 in [<math>\text{V}^{-1}</math>].</p>	Yes	-

150 BSIM4 AND MOSFET MODELING FOR IC SIMULATION  
By Weidong Liu and Chenming Hu

EIGBINV (Global; double)	Parameter for $I_{gb}$ in inversion.  Default = 1.1 in [V].	Yes	-
NIGBINV (Global; double)	Parameter for $I_{gb}$ in inversion.  Default = 3.0; dimensionless.	Yes	Fatal errors to be issued if its binned value $\leq 0.0$ .
AIGC (Global; double)	Parameter for the $I_{gcs}$ and $I_{gcd}$ model.  Default = 0.43 for NMOS; 0.31 for PMOS in [ $(\text{Fs}^2/\text{g})^{0.5} \text{m}^{-1}$ ].	Yes	-
BIGC (Global; double)	Parameter the $I_{gcs}$ and $I_{gcd}$ model.  Default = 0.054 for NMOS; 0.024 for PMOS in [ $(\text{Fs}^2/\text{g})^{0.5} (\text{mV})^{-1}$ ].	Yes	-
CIGC (Global; double)	Parameter for the $I_{gcs}$ and $I_{gcd}$ model.  Default = 0.075 for NMOS; 0.03 for PMOS in [ $\text{V}^{-1}$ ].	Yes	-
AIGSD (Global; double)	Parameter for the $I_{gs}$ and $I_{gd}$ model.  Default = 0.43 for NMOS; 0.31 for PMOS in [ $(\text{Fs}^2/\text{g})^{0.5} \text{m}^{-1}$ ].	Yes	-
BIGSD (Global; double)	Parameter for the $I_{gs}$ and $I_{gd}$ model.  Default = 0.054 for NMOS; 0.024 for PMOS in [ $(\text{Fs}^2/\text{g})^{0.5} (\text{mV})^{-1}$ ].	Yes	-
CIGSD (Global; double)	Parameter for the $I_{gs}$ and $I_{gd}$ model.  Default = 0.075 for NMOS; 0.03 for PMOS in [ $\text{V}^{-1}$ ].	Yes	-
VFBSDOFF (Global; double)	The offset flat-band voltage for the source and drain diffusion regions.  Default = 0.0 in [V].	Yes	-
TVFBSDOFF (Global; double)	The temperature-dependence coefficient for VFBDOFF.  Default = 0.0 in [ $\text{K}^{-1}$ ].	Yes	-
DLCIG (Global; double)	Source and drain overlap length for the $I_{gs}$ and $I_{gd}$ models.  Default = LINT in [meter].	No	-

<b>NIGC</b> (Global; double)	Parameter for the $I_{gs}$ , $I_{gd}$ , $I_{gcs}$ , and $I_{gcd}$ models. Default = 1.0; dimensionless.	Yes	Fatal errors to be issued if its binned value $\leq 0.0$ .
<b>POXEDGE</b> (Global; double)	Factor for the gate oxide thickness of the S/D overlap regions.  Default = 1.0; dimensionless.	Yes	Fatal errors to be issued if its binned value $\leq 0.0$ .
<b>PIGCD</b> (Global; double)	Gate-and-channel tunneling current $I_{gc}$ partitioning parameter.  Default = 1.0; dimensionless.	Yes	Fatal errors to be issued if its binned value $\leq 0.0$ .
<b>NTOX</b> (Global; double)	Exponent for the ratio of the nominal and electrical gate oxide thickness.  Default = 1.0; dimensionless.	Yes	-
<b>TOXREF</b> (Global; double)	Nominal gate oxide thickness.  Default = $3.0 \times 10^{-9}$ in [meter].	No	Fatal errors to be issued if $\leq 0.0$ .
<b>ALPHA0</b> (Global; double)	$I_{ii}$ channel-length dependence coefficient.  Default = 0.0 in [meter/V].	Yes	-
<b>ALPHA1</b> (Global; double)	$I_{ii}$ coefficient parameter.  Default = 0.0 in [ $V^{-1}$ ].	Yes	-
<b>BETA0</b> (Global; double)	$I_{ii}$ exponent coefficient parameter to represent a critical field strength required to trigger impact ionization.  Default = 0.0 in [V].	Yes	-
<b>AGIDL</b> (Global; double)	$I_{GIDL}$ coefficient parameter.  Default = 0.0 in [mho].	Yes	If not positive, $I_{GIDL}$ will not be computed.
<b>AGISL</b> (Global; double)	$I_{GISL}$ coefficient parameter.  Default = AGIDL if AGIDL is given; otherwise is set to 0.0 in [mho].	Yes	If not positive, $I_{GISL}$ will not be computed.
<b>BGIDL</b> (Global; double)	$I_{GIDL}$ exponential parameter.  Default = $2.3 \times 10^9$ in [V/m].	Yes	If not positive, $I_{GIDL}$ will not be computed.

BGISL (Global; double)	$I_{GISL}$ exponential parameter. Default = BGIDL if BGIDL is given; otherwise it is set to $2.3 \times 10^9$ in [V/m].	Yes	If not positive, $I_{GISL}$ will not be computed.
CGIDL (Global; double)	$I_{GIDL}$ body bias dependence parameter. Default = 0.5 in [V <sup>3</sup> ].	Yes	If not positive, $I_{GIDL}$ will not be computed.
CGISL (Global; double)	$I_{GISL}$ body bias dependence parameter. Default = CGISL if CGISL is given; otherwise it is set to 0.5 in [V <sup>3</sup> ].	Yes	If not positive, $I_{GISL}$ will not be computed.
EGIDL (Global; double)	$I_{GIDL}$ parameter representing the band bending needed for the onset of a band-to-band tunneling.  Default = 0.8 in [V].	Yes	-
EGISL (Global; double)	$I_{GISL}$ parameter representing the band bending needed for the onset of a band-to-band tunneling.  Default = EGISL if EGISL is given; otherwise, it is set to 0.8 in [V].	Yes	-

## References

- [1] Chenming Calvin Hu, “Modern Semiconductor Devices for Integrated Circuits,” Pearson Prentice Hall, Chapter 7, pp. 259 – 289, 2010.
- [2] Wen-Chin Lee, Tsu-Jae King, and Chenming Hu, “Evidence of hole direct tunneling through ultrathin gate oxide using P poly-SiGe gate,” *IEEE EDL*, vol. 20, no. 6, pp. 268–271, June, 1999.
- [3] Pin Su, Samuel K. H. Fung, Weidong Liu, and Chenming Hu, “Studying the impact of gate tunneling on dynamic behaviors of partially-depleted SOI CMOS using BSIMPD,” *Proceedings of the International Symposium on Quality Electronic Design (ISQED)*, pp. 487–491, 2002.
- [4] Yee-Chia Yeo, Tsu-Jae King, and Chenming Hu, “Direct tunneling leakage current and scalability of alternative gate dielectrics,” *Applied Physics Letters*, 81(11), pp. 2091–2093, 2002.
- [5] K. M. Cao, W.-C. Lee, W. Liu, X. Jin, P. Su, S. K. H. Fung, J. X. An, B. Yu, and C. Hu, “BSIM4 gate leakage model including source-drain partition,” *Tech. Dig. of IEDM*, pp. 815 – 818, San Francisco, December 2000.

- [6] The authors here use 4.25 eV for the SiO<sub>2</sub>/Si hole energy barrier height, instead of the frequently used 4.9 eV. The smaller value was first extrapolated from the optical spectra of SiO<sub>2</sub> film by Sah and described in Chapter 17.9 on pages 564 to 565 of the handbook, *Properties of Silicon*, INSPECT, The Institution of Electrical Engineers, London and New York, 1988. It was later used by Sah in *Fundamentals of Solid-State Electronics (FSSE)*, 1991, on page 356, and its *Study Guide, FSSE-SG*, 1993, World Scientific Publishing Co. See FSSE-SG, Appendix B and figures therein, such as Fig. B2.3 on page 404.
- [7] Chih-Tang Sah, “A new semiconductor tetrode, the surface-potential controlled transistor,” *Proc. IRE*, vol. 49(11), pp. 1623-1634, November 1961. “Effect of surface recombination and channel on p-n junction and transistor characteristics,” *IEEE Trans. Electron Devices*, vol. 9, no. 1, pp. 94-108, January 1962.
- [8] C. Hu, S. Tam, F. C. Hsu, P. K. Ko, T. Y. Chan, and K. W. Kyle, “Hot-electron induced MOSFET degradation – Model, monitor and improvement,” *IEEE Trans. Electron Devices*, vol. 32, pp. 375–385, 1985.
- [9] Yuhua Cheng, and Chenming Hu, “MOSFET Modeling & BSIM3 User’s Guide,” Kluwer Academic Publishers, 1999.
- [10] T. Y. Chan, J. Chen, P. K. Ko, and C. Hu, “The impact of gate-induced drain leakage current on MOSFET scaling,” *Tech. Dig. of IEDM*, pp. 718 -721, 1987.

**This page intentionally left blank**

## Chapter 5

# Charge and Capacitance Models

### 5.1 Introduction and Chapter Objectives

This chapter presents the BSIM4 charge and capacitance models. MOSFET transistors have intrinsic capacitances, which limit the MOSFET operation, and parasitic capacitances which further impact the MOSFET performance. Therefore, when time-varying voltages are applied to the terminal nodes, charging and discharging events take place inside the MOSFETs, leading to capacitive currents flowing into the terminal nodes. These capacitive currents are sometimes referred to as AC currents in solid-state devices, such as this book. They are the well-known displacement currents from the time rate of change of the electric field given in (distributed) electromagnetic (EM) theories. Our lumped circuit definition comes directly from the general, distributed definition or the three-dimensional charge conservation law in the EM theory, given by  $\oint\oint\mathbf{J}\cdot d\mathbf{S} + (\partial/\partial t)\iiint\rho dv = 0$ . The integrand  $\mathbf{J}$  is the areal density of the conduction current vector of the surface integral with the traditional calculus-defined outward normal vector  $d\mathbf{S}$  representing the surface areal element  $d\mathbf{S}$  (such as  $d\mathbf{S} = \mathbf{i}_z dS = \mathbf{i}_z dx dy$  in the Cartesian coordinate). The integrand  $\rho$  is volume density of the charge enclosed (or stored) in the volume element  $dv=dx dy dz$ , enclosed by the surface of the volume integral. In vacuum electronic devices, such as vacuum tubes, cathode ray display tubes, and magnetrons, the electrical conduction current is due to the motion of electrons moving ballistically

(without collision at low electron densities when electron-electron scattering can be neglected) following Newton's law of motion in an electric field from the Coulomb force. In contrast, in solid-state electronic devices, electric conduction currents arise from the drift and diffusion motions of the charge carriers, which encounter many random, collisions (both scattering by the vibrating host atomic ions, impurities, defects, and generation-recombination-trapping at trapping centers). The charge carriers are the electrons and holes in semiconductors and electrons in metals. Of these two conduction current components in MOSFETs, the drift current is dominant in the inversion range and diffusion current is dominant in the subthreshold range. They were treated in Chapter 3 and labeled by  $I_{DC}$ . In this chapter, we shall continue to use the symbol,  $I_{DC}$ , and the results of  $I_{DC}$  obtained in Chapter 3, with the extension that  $I_{DC}$  can be time varying if the voltage (or electric field) applied is time varying; for example, one of the voltages applied to the gate, drain, source or body terminals, is time varying; but the response of the current  $I_{DC}$  to these time-varying excitations is assumed instantaneous, without delay. The delay arises from the time it takes to distribute (via diffusion and drift) the charges in the volume of the transistor, represented by the capacitive current and the volume integral of the volume charge density. In our lumped element representation of the transistor by multiple terminal nodes (i.e. terminal contacts) the current flowing inward or into each of the terminal nodes is taken positive for the convenience of not carrying the negative sign from the surface integral. Thus, the (conductive) current flowing into a terminal or surface node, denoted by  $i$ , through a metal wire in contact to the terminal node  $i$  (or metal pad) is the sum of the conductive (drift and diffusion) current flowing into this  $i$ -th node,  $I_{DC}$ , plus the capacitive current from time rate of change of the charge stored in the transistor,  $dQ_i/dt$ , which depends on the voltages applied to all the terminals, with reference to this  $i$ -th node,  $Q_i(V_j)$  which is given by the volume integral above:  $i_i(t) = I_{DC} + dQ_i(t)/dt$ .

There is a good application reason, aside from its fundamental rigorous basis, to express the capacitive current as  $dQ(V)/dt$  (a charge-based model) rather than  $C(V) \cdot (dV/dt)$  (a capacitance-based model).

Any theoretical analyses using the capacitance-based model can easily cause charge non-conservation unless the 16 components of  $C_{ij}(V_j)$ , where  $i$  and  $j$  are the four nodes, gate, drain, source and body, satisfy a set of self-consistency requirements [1 – 2]. In a charge-based model, on the other hand,  $Q_i(V_j)$  can be arbitrary functions of multiple device terminal bias voltages such as  $V_{gs}$ ,  $V_{ds}$  and  $V_{bs}$  without causing unphysical device terminal charge build-up as long as the four  $Q_i$  components add to zero. And only nine of the 16 components of the capacitance are independent. No charge can be built up or the four  $Q_i$  must add up to zero under any arbitrary bias condition. This is an important property of a compact model, known as charge conservation, which is needed to correctly describe the transient and frequency response of the transistor.

In high-speed digital and analog circuits, the capacitive current can have significant effects on the response speed and functionality of the circuits. In the frequency domain, capacitive currents change both the magnitude and the phase of the terminal currents. Therefore, it is very important that circuit simulations use physically accurate and numerically efficient charge and capacitance models.

Like its DC models, the BSIM4 charge and capacitance models have been used in industry production for many technology nodes down to 20nm. This chapter will first present the device physics and SPICE simulation of MOSFET charge and capacitance theory. This is then followed by presenting the BSIM4 intrinsic charge and capacitance models, with particular emphasis on its **CAPMOD = 2** model. The gate-to-source/drain overlap capacitances play an important role and is discussed subsequently. For ease of reference, the model parameters of the BSIM4 charge and capacitance model are given in a parameter table at the end of this chapter. The BSIM4 p/n junction diode charge and capacitance models will be presented in Chapter 9.

## 5.2 MOSFET Capacitance Theory

In the various operation regions of an MOSFET, a charge  $Q_i$  can be defined at terminal node  $i$ , which depends on the voltage  $V_j$  applied to the

terminals  $j$ . Here,  $i$  and  $j$  designate the drain ( $d$ ), gate ( $g$ ), source ( $s$ ) or body ( $b$ ) terminal nodes.  $Q_i$ , in general, is a function of the terminal node voltage  $V_j$  applied to terminal node  $j$  relative to a common reference, which is arbitrary and understood but left unspecified, for simpler notation, such as in the application with a metal box, to shield noise, as an indirect ground. (This is well known for more than fifty years, for example, in the textbook by Ernst A. Guillemin (Father of circuit theory and network analysis and synthesis and Webster Professor of Electrical Communication at MIT), *Theory of Linear Physical Systems*, John Wiley & Sons, New York, P.586, 1963.) The capacitive current at node  $i$  induced (contributed) by the voltage change at terminal  $j$  is  $(\partial Q_i / \partial t) = (\partial Q_i / \partial V_j) \cdot (dV_j / dt)$ . When all significant device physical effects are taken into account mathematically,  $Q_i$  is usually a complex nonlinear function of the voltages applied to all terminal nodes. The total capacitive current at the terminal node  $i$  is  $dQ_i/dt$ , i.e., the sum of all the capacitive currents at node  $i$  caused by the voltage change at every terminal  $j$ .

Thus, a MOSFET device with four terminals has sixteen partial derivatives of the terminal charge with respect to the same or another terminal voltage. These partial derivatives  $\partial Q_i / \partial V_j$  are known as the MOSFET trans-nodal capacitances  $C_{ij}$  (or simply trans-capacitances, consistent with the commonly known transconductances) between node  $i$  and node  $j$ . Each derivative signifies how  $Q_i$  and, therefore, the capacitive current of terminal  $i$  respond to a change in  $V_j$ . In general,  $C_{ij}$  is non-reciprocal, or  $C_{ij} \neq C_{ji}$ , because  $\partial Q_i / \partial V_j$  is not equal to  $\partial Q_j / \partial V_i$  in nonlinear and active devices, such as a transistor. But they are equal in passive devices, such as a 2-terminal nonlinear resistance that has only drift current, or a 2-terminal nonlinear capacitance that has only displacement current.

In obtaining these partial derivatives, a single subscript is used for the terminal voltages,  $V_j$ , with the understanding that a common reference node is always implied for all terminal voltages, as demanded by the physics of electricity, with electrons as the elementary charge-

carrying particle, according to the Coulomb electric force law. This basis for the simplification of notation in the circuit representation may sometimes be forgotten. Thus, it is important to remember the fundamentals, that is, the definition of  $C_{gd}$  is the delta  $Q_g$  resulting from a delta drain voltage with all the other nodal voltages of the device held constant in addition,  $\partial V_{ds} = \partial V_{dg} = \partial V_{db}$  are all equal to  $\partial V_d$ .

$C_{ij}$  is numerically positive if  $i = j$  and it is usually negative when  $i \neq j$ . However, the known exceptions are  $C_{ds}$ ,  $C_{sd}$ , and also  $C_{gd}$ , owing to the short-channel and drain-induced barrier lowering (DIBL) effects. Clearly, modeling capacitances for devices with more than two nodes requires special attention. As long as one adopts the charge based modeling approach, one needs to only focus on finding functions that accurately describe nodal charges in terms of device terminal voltages.

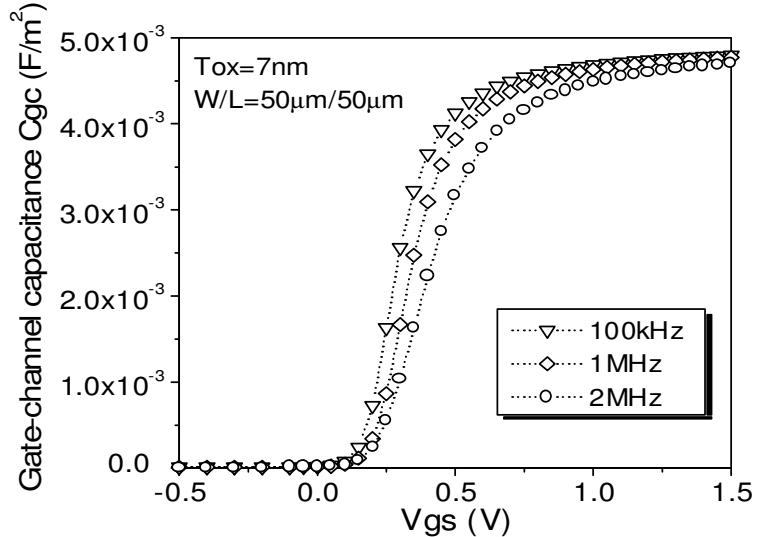
From the “measurement” or presentation point of view, MOSFET capacitances are usually plotted or presented as

$$C_{ij,measured} = \delta_{ij} \cdot C_{ij} = \delta_{ij} \cdot \frac{\partial Q_i}{\partial V_j} \quad (5.1)$$

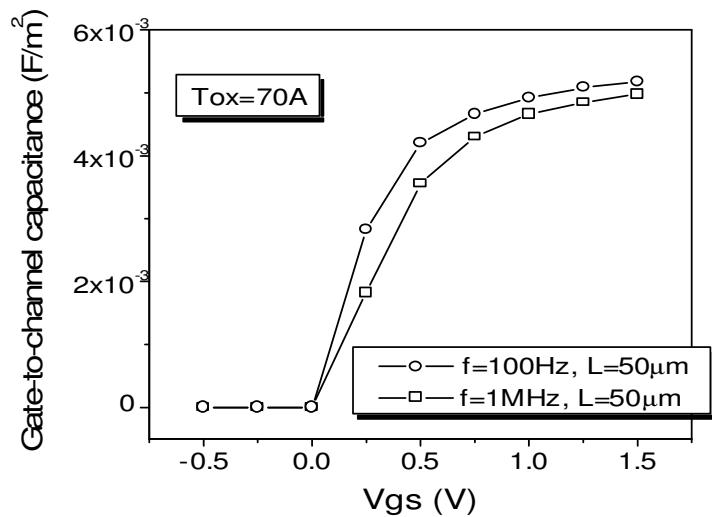
with the Kronecker delta function  $\delta_{ij}$  defined as usual for this application, that is  $\delta_{ij} = 1$  if  $i = j$  and  $-1$  if  $i \neq j$ . This avoids the spectacle of a negative capacitance although the partial derivatives  $C_{ij}$ 's themselves are indeed negative if  $i \neq j$ . Again,  $C_{ds}$  and  $C_{sd}$  are the possible exceptions and the total charge accumulated at node  $i$  from changes in the nodal voltage  $V_j$  is given by the integral

$$Q_i = \int_0^{V_j} C_{ij} dV_j \quad (5.2)$$

where  $C_{ij}$  is a bias dependent nonlinear capacitance.



(a)



(b)

Fig. 5.1 The gate-to-channel capacitance  $C_{gc}$  versus the gate voltage (a) and the SPICE-simulated BSIM4  $C_{gc}$  (b) at different frequencies.  $V_{ds} = V_{bs} = 0$  in both cases. Note the capacitance frequency dependence for a large device, which makes the quasi-static assumptions in charge and capacitance modeling inaccurate or invalid (Refer to Chapter 6 for the detailed analysis of non-quasi-static effects).

Take the gate-to-channel capacitance  $C_{gc}$  of Fig. 5.1 as example.  $C_{gc}$  is the capacitance that is equal to the sum of  $C_{gs}$  and  $C_{gd}$ . The channel charge can be obtained by integrating this capacitance  $C_{gc}$  over the gate voltage minus the body charge. A close-form integration of  $C_{ij}$  over a voltage is not always possible. The alternative is the sum of each individual  $C_{ij}(V) \cdot \Delta V_j$ . Here, the voltage step  $\Delta V_j$  should be made very small to ensure good accuracy. This is useful for validating the formulations of a MOSFET charge and capacitance model.

As can also be observed from Fig. 5.1, capacitances are frequency dependent, especially for large-size devices. This is attributed to the non-quasi-static (NQS) effects associated with the distributed-RC nature of the MOSFET channel region [3]. The longer the channel is, the more prominent the gate-channel RC time constant can be. Consequently, NQS becomes more significant at a given frequency. The NQS model and SPICE implementation will be presented in Chapters 6 and 10.

Nevertheless, a large device, such as 100 $\mu\text{m}$ /100 $\mu\text{m}$  in the gate length and width, is often used in charge and capacitance measurements to improve the measurement accuracy owing to the limited precision of measurement equipment. However, as the measurement frequency increases, the slope of the  $C_{gc}$  versus  $V_{gs}$  curve becomes more gradual as shown in Fig. 5.1(a) and (b). Therefore, the frequency employed in capacitance measurement has to be very low, often in the kilo hertz range in practice.

MOSFET charges and capacitances have some important general characteristics. Consider a MOSFET transistor with independent time-varying voltage sources connected to its four terminals, as shown in Fig. 5.2. Suppose the sources are all small signals such that they can be deemed as minute perturbations to the quiescent, operating point (OP) of the transistor (OP not shown in the figure). Assume also that during a very short period of time, only one of the four sources is changing, say  $v_g(t)$ , while the rest is held unchanged for the moment (only one at a time). The capacitive components of the four terminal currents can be expressed

$$i_d(t) = \frac{dQ_d}{dt} = \frac{d}{dt} \cdot \left[ Q_{d,OP} + \left. \frac{\partial Q_d}{\partial V_g} \right|_{OP} \cdot v_g(t) \right] = \frac{d}{dt} \cdot [C_{dg} \cdot v_g(t)] \quad (5.3a)$$

$$i_g(t) = \frac{dQ_g}{dt} = \frac{d}{dt} \cdot \left[ Q_{g,OP} + \left. \frac{\partial Q_g}{\partial V_g} \right|_{OP} \cdot v_g(t) \right] = \frac{d}{dt} \cdot [C_{gg} \cdot v_g(t)] \quad (5.3b)$$

$$i_s(t) = \frac{dQ_s}{dt} = \frac{d}{dt} \cdot \left[ Q_{s,OP} + \left. \frac{\partial Q_s}{\partial V_g} \right|_{OP} \cdot v_g(t) \right] = \frac{d}{dt} \cdot [C_{sg} \cdot v_g(t)] \quad (5.3c)$$

and

$$i_b(t) = \frac{dQ_b}{dt} = \frac{d}{dt} \cdot \left[ Q_{b,OP} + \left. \frac{\partial Q_b}{\partial V_g} \right|_{OP} \cdot v_g(t) \right] = \frac{d}{dt} \cdot [C_{bg} \cdot v_g(t)] \quad (5.3d)$$

where  $V_g$  in the denominator and similarly other such terminal voltages (not shown) in capital  $V$  are the OP voltages of the quasi-static charge model. The word “OP” to the right of the vertical bars represents the quiescent condition under which the partial charge derivatives are obtained. These derivatives are trans-nodal capacitances with respect to the gate node.

Several useful facts are observed from Eqs. (5.3a) through (5.3d) and Fig. 5.2. The sum of the terminal charges on the two electrodes of a linear capacitor is zero, which is required by the charge neutrality law or Guass’ law. This also applies to MOS transistors, that is, the sum of the charges stored at all of its nodes is zero. This comes directly from the EM theory and Coulomb law. Thus, for MOSFETs.

$$\sum_i Q_i \equiv 0 \quad (i = d, g, s, \text{ and } b) \quad (5.4)$$

Differentiating Eq. (5.4) with respect to any given  $V_j$  results in Eq. (5.5), without loss of generality, which is in agreement with the fact that the sum of Eqs. (5.3a) though (5.3d) must be zero.

$$\sum_i C_{ij} \equiv 0 \quad (i = d, g, s, \text{ and } b \text{ for any given } j) \quad (5.5)$$

This equation states that the sum of any column of the following trans-capacitance matrix is zero.

$$\begin{bmatrix} C_{dd} & \cdots & C_{db} \\ \vdots & \ddots & \vdots \\ C_{bd} & \cdots & C_{bb} \end{bmatrix}$$

In the case where the voltage sources at all the terminals of the transistor change by the same amount at the same time, the transistor is still in the same state as before the change and there will be no change in any of the node charges. Therefore, this leads to a similar statement that the sum of any row of the capacitance matrix is also zero.

These are useful properties. For instance, only three out of the four terminal charges (such as  $Q_d$ ,  $Q_s$ , and  $Q_b$ ) need to be modeled (with the understanding that the device system is electrically neutral). The fourth one ( $Q_g$ ) can be obtained from Eq. (5.4) without a loss of any information.

Similarly, only nine out of the sixteen trans-capacitances are independent because of the relationship provided by Eq. (5.5).

These two powerful properties are fully preserved and utilized in the BSIM4 SPICE implementation to provide simulation efficiency and modeling accuracy. [Having SPICE compute all the nodal charges and capacitances is runtime inefficient and can lead to violation of Eqs. (5.4) and (5.5) owing to limited machine precision and possible computational floating-point errors]. In fact, these properties are also utilized to debug and ensure whether the loading of a model into a circuit matrix has been performed correctly.

As stated before, the trans-nodal capacitances of a MOS transistor are non-reciprocal because transistor is active network, thus,  $C_{ij} \neq C_{ji}$ . In contrast, a two-terminal linear capacitor is reciprocal. For example  $C_{dg} \neq C_{gd}$ , because the influence of  $V_g$  on  $Q_d$  is not the same as that of  $V_d$  on  $Q_g$ . In general, the trans-capacitances of active semiconductor devices such as MOS and BJT transistors are non-reciprocal.

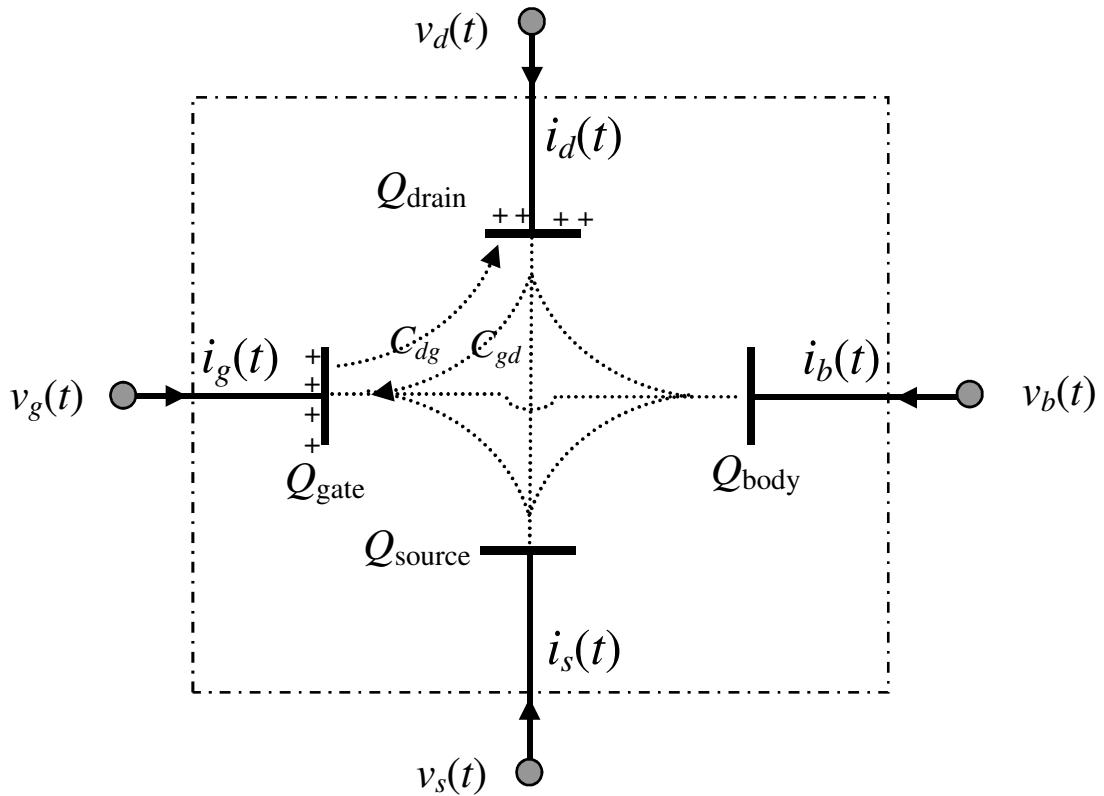


Fig. 5.2 A holistic view of MOSFET trans-nodal capacitive coupling. The two dotted arrows between the gate and drain nodes signify the non-reciprocal nature of the trans-nodal capacitances,  $C_{gd}$  and  $C_{dg}$  as examples. The charge neutrality law applies to the system enclosed within the dash-dotted box. The sums of all the charge and terminal capacitive currents are and must be zero. The charge at each node can be positive or negative depending upon the bias conditions, although positive signs are used here to represent the charges at the gate and drain nodes.

Note that the non-reciprocity nature of trans-nodal capacitances is not at all a hindrance, when it comes to the SPICE implementation and execution. However, it does impose a challenge in the construction of the MOSFET small-signal equivalent circuit or in hand calculations of small-signal circuit parameters, because there are now two different trans-capacitances  $C_{gd}$  and  $C_{dg}$  connecting the gate and drain nodes. For example, which one should be used to derive the circuit gain and an analytical expression for the Miller effect?

This is mitigated in circuit analysis as illustrated by the small-signal equivalent circuit in Fig. 5.3. Here, the trans-capacitances  $C_{dg}$  and  $C_{sg}$  are transformed, in the frequency domain, into an equivalent voltage-controlled current source  $j\omega \cdot C_m \cdot v_{gs}$ , represented by the rhombic symbol. This leaves only one capacitance between the gate and the

source, and the gate and the drain. This equivalence can be verified by using Fig. 5.2.  $C_m = (C_{gd} - C_{dg})$ . This equivalent-circuit analysis can be also applied to the intrinsic body-drain and body-source capacitances and the intrinsic gate-body and body-gate capacitances. It is also applicable to the time-domain transient analysis.

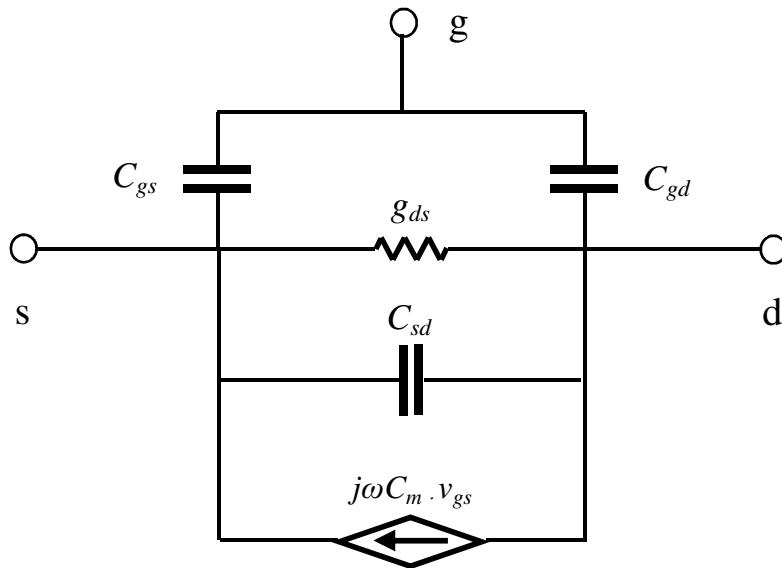


Fig. 5.3 Partial MOSFET small-signal equivalent circuit in the frequency domain. The effects of the drain-to-gate and source-to-gate trans-capacitances  $C_{dg}$  and  $C_{sg}$  are represented by the voltage-controlled capacitive current source connected between the source and drain nodes. The body node and its associated conductances, trans-capacitances and low-frequency-d.c. transconductance current source are omitted for simplicity. See Chapter 3 for these missing parts.

In MOSFET charge and capacitance modeling, channel charge partitioning is a particularly interesting problem. This was first investigated by John E. Mayer of RCA Laboratories (“MOS models and circuit simulation,” RCA Review 32(3), 42-63, March 1971.) although the analyses of transistor charges were well-known for more than a decade prior to Mayer, from the extensive investigations reported by J. J. Sparks and R. Beaufoy of British Telecommunication Research, “The Junction Transistor as a Charge Controlled Device,” Proc. IRE, 45(12), 1740-1742, December 1957; and by E. O. Johnson and Albert Rose of RCA Laboratories, “Simple General Analysis of Amplifier Devices with Emitter, Control, and Collector Functions,” Proc. IRE, 47(3), 407-418, March 1959.

For the purpose of our discussion in this chapter, we ignore the presence of the body (depletion) charge. It is not difficult to derive the

potential at every point in the channel between the source and the drain. From this potential distribution, one can relate the channel charge to the terminal voltages. On account of electrical neutrality, and ignoring the body depletion charge,  $Q_g$  is simply  $-Q_{ch}$ . The channel charge is supplied from the source and collected by the drain. The remaining step in the MOSFET charge modeling is to partition or allocate  $Q_{ch}$  into two components,  $Q_s$  and  $Q_d$ , with  $Q_{ch} = Q_s + Q_d$  so that the source and drain capacitive current components can be computed from  $dQ_s/dt$  and  $dQ_d/dt$ . At  $V_{ds} = 0$  and zero currents, the channel charge is obviously partitioned equally between the source and the drain. This is known as the 50/50 charge partition scheme. In this bias scenario, the source and the drain are electrically symmetrical and claim the same amount of the channel charge. The challenge is then how to partition the channel charge for any non-zero source-drain bias  $V_{ds}$ . There are three commonly used schemes: 50/50 again, 40/60 and 0/100. The 50/50 partition scheme still assumes equal partition even for applied voltages, from  $V_{ds} > 0$  (the linear region) through  $V_{ds} = V_{dsat}$  and beyond (the saturation region). In the 40/60 partition scheme,  $Q_d$  drops gradually from 50% of  $Q_{ch}$  at  $V_{ds} = 0$  to 40% of  $Q_{ch}$  at  $V_{ds} = V_{dsat}$  and stays at 40% of  $Q_{ch}$  beyond  $V_{dsat}$ .  $Q_s$  increases gradually from 50% of  $Q_{ch}$  to 60% of  $Q_{ch}$  and stays at 60% of  $Q_{ch}$  beyond  $V_{dsat}$ . Similarly, for the 0/100 partition scheme,  $Q_d$  drops from 50% of  $Q_{ch}$  to 0% of  $Q_{ch}$  at  $V_{dsat}$  and stays at 0 beyond  $V_{dsat}$ , while  $Q_s$  rises from 50% of  $Q_{ch}$  to 100% of  $Q_{ch}$  at  $V_{dsat}$  and stays at 100% of  $Q_{ch}$  beyond  $V_{dsat}$ .

The BSIM4 charge and capacitance models support all three schemes and users select one of them with the parameter **XPART**. For example, in the saturation region of operation, the **CAPMOD = 0** model gives

$$Q_{ch} = -\frac{2}{3}C_{oxe}W_{effCV}L_{effCV} \cdot NF \cdot (V_{gs} - V_{th}) \quad (5.6)$$

where **NF** is the number of device fingers. Set **XPART = 0** for 40/60, 0.5 for 50/50 and 1 for 0/100.

The 50/50 partition scheme is the simplest scheme. There is no good justification for this model except that it is simple and the other two partitioning schemes are not perfect or rigorous either.

The 40/60 partitioning has attractive theoretical backing. In this sense it is superior to the 50/50 model. However, the theoretical backing is only valid under the quasi-static assumption, i.e., when the voltages change at a time scale much slower than the transistor channel transit time. In this case, the channel charges are divided into the source and drain nodes by assuming that the channel charge density attributed to

source and drain decreases and increases, respectively, as a linear function of the location along the channel from the source to the drain [1]. Thus, one obtains  $Q_s$  and  $Q_d$  by performing the following integration

$$Q_s = W_{effCV} \cdot \int_0^{L_{effCV}} q_{ch}(y) \cdot \left(1 - \frac{y}{L_{effCV}}\right) \cdot dy \quad (5.7)$$

and

$$Q_d = W_{effCV} \cdot \int_0^{L_{effCV}} q_{ch}(y) \cdot \frac{y}{L_{effCV}} \cdot dy \quad (5.8)$$

where  $y$  is the location along the channel.  $q_{ch}(y)$  is the channel charge density per unit area, which gives Eq. (5.6) when  $V_{ds} = 0$  in the case of **CAPMOD** = 0.

The quasi-static assumption is acceptable for most applications other than RF. However, the 40/60 and 50/50 partition models cause unrealistic drain current spikes (e.g., current flowing against a voltage drop) in fast transient operations.

The 0/100 partitioning scheme is designed to eliminate these unrealistic current spikes by arbitrarily attributing all channel charges to the source terminal [2]. However, it pays a price because it underestimates the capacitive drain current (setting it to zero) and overestimates the source capacitive current at all transient operation speeds.

The ultimate solution to these problems is to abandon the quasi-static assumption and adopt the more accurate but computationally expensive non-quasi-static model, which will be presented in Chapter 6. In the remainder of this chapter, the BSIM4 quasi-static charge and capacitance models will be detailed. These models find good use even down to the 20nm technology node today.

### 5.3 Intrinsic Charge and Capacitance Models

BSIM4 provides three intrinsic charge and capacitance models to be described here in detail. They are selected by the model parameter **CAPMOD**. All of these models support those three charge partitioning schemes discussed above (by setting the model parameter **XPART** to 0 for 40/60, 0.5 for 50/50 and 1 for 0/100). These three charge and

capacitance models are inherited from BSIM3v3 with modifications [4-8]. The **CAPMOD = 0** model is a piece-wise, long-channel model with simple mathematical formulations. It is rarely used in advanced process technologies and circuit designs. However, it is still kept in BSIM4 to provide intuitive insights into the MOS operation and helps to aid quick analysis for hand calculations.

The **CAPMOD = 1** model is a continuous model by using a single and smooth charge formulation for all regions of operation. This is more accurate and also improves SPICE simulation convergence robustness. This model considers short-channel and bulk-charge effects. It has been the mainstream model from the quarter micron to the 90nm technology node.

The third model, **CAPMOD = 2**, is the default model. It is built upon **CAPMOD = 1** and incorporates the charge-thickness model (CTM) into all regions of operation [4 – 6]. This is the model for sub-90nm down to the present 20nm node, where the finite thicknesses of the channel and body charges become increasingly comparable to the gate dielectric thickness.

In the next section, only the **CAPMOD = 2** model will be presented and analyzed. The CTM model will be discussed first.

### 5.3.1 Charge-Thickness Model (CTM)

Traditional MOSFET charge and capacitance models ignore the finite thickness (up to 3nm in the inversion region) of a charge layer (Refer to Chapter 2 also). They assume that the inversion and accumulation charges are all located at the interface with no distributions vertical (normal) to the channel. In reality, the energy band diagram of an NFET, for example, clearly indicates the presence of a quantum well between the gate oxide and the semiconductor conduction band at the interface.

The solution of the Schrodinger equation for such a quantum well requires that the electron density is nearly zero at the interface and hence results in a peak at some distance below the interface before the density falls to zero deep below and far away from the interface. The weighted average depth of the inversion charges is called the charge centroid or simply the charge thickness. For capacitance modeling, it is as if all the inversion charges are now in a sheet that is located not at the interface, but a distance beneath the oxide interface which is equal to the charge thickness.

Ignoring the depth of the inversion charge leads to inaccuracies. The inaccuracy is increasingly more pronounced in advanced technologies where the equivalent oxide/gate dielectric thickness is now scaled down to 1 nanometer. Similarly, the majority carrier or electron accumulation layer in NFET also has a finite thickness.

The finite charge-thickness model was first introduced into BSIM3v3.2. It was developed from the 1-D self-consistent numerical solution of the Schrodinger and Poisson equations using Fermi-Dirac statistics for high carrier concentration [4–6, 9]. It starts with the DC charge thickness  $X_{DC}$ .  $X_{DC}$  is defined as the integral defined by  $\int_0^\infty \rho(x) \cdot x \cdot dx / \int_0^\infty \rho(x) \cdot dx$ , where  $\rho(x)$  is the charge density as a function of depth. This finite charge thickness is a strong function of  $V_{gs}$  and introduces a capacitance in series with the gate dielectric capacitance  $C_{oxp}$ , which is determined by the physical gate oxide or gate dielectric thickness  $TOXP$  as discussed in Chapter 2 and as shown in Fig. 5.4. The series connection results in a reduced effective gate oxide capacitance

$$C_{oxeff} = \frac{C_{oxp} \cdot C_{cen}}{C_{oxp} + C_{cen}} \quad (5.9)$$

where  $C_{cen} = \epsilon_{sub}/X_{DC}$  is the charge thickness capacitance and  $\epsilon_{sub}$  is the dielectric constant of the material of the channel region.  $C_{cen}$  represents the parallel capacitances  $C_{acc}$ ,  $C_{dep}$  and  $C_{inv}$  in Fig. 5.4. Each of the three parallel capacitances varies in relative importance depending on the voltage range of operation.

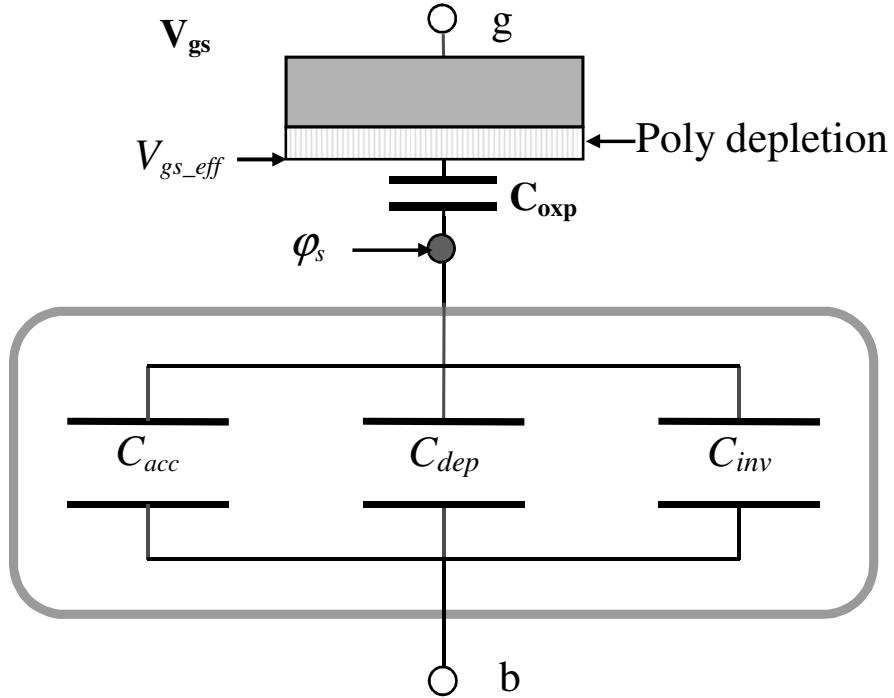


Fig. 5.4 The BSIM4 charge-thickness capacitance is in series with the gate oxide/dielectric capacitance.  $V_{gs\_eff}$  denotes the reduced gate voltage owing to the poly-silicon gate depletion effect (if any).  $C_{acc}$ ,  $C_{dep}$  and  $C_{inv}$  represent the charge-thickness capacitances due to the finite charge thickness in the accumulation, depletion and inversion regions of operation, respectively.

It turns out that from the accumulation to depletion region, a universal model for the finite charge thickness is accurately given by

$$X_{DC} = \frac{1}{3} L_{Debye} \cdot \exp \left[ ACDE \cdot \left( \frac{NDEP}{2 \times 10^{16}} \right)^{-0.25} \frac{V_{gs\_eff} - V_{bseffCV} - V_{fbzb}}{TOXP \cdot 10^8} \right] \quad (5.10)$$

where  $L_{Debye} = \sqrt{\epsilon_{sub} k_B \cdot TNOM / (q^2 \cdot NDEP \cdot 10^6)}$  is the Debye length in the unit of centimeters. The voltage dependence term of the exponential function is given in the unit of mega volts per centimeter.  $\epsilon_{sub}$  is the permittivity of the substrate. The CAPMOD = 1 and 2 models

use an effective body-source voltage  $V_{bseffCV}$ , which is the same effective body-source voltage  $V_{bseff}$  used in the DC models. It incorporates forward bias clamping at  $0.95\varphi_s$  and reverse bias limiting at VBSC. (Refer to Chapter 2 for details). Note that the physical gate dielectric/oxide thickness, instead of the electrical oxide thickness, is used in Eq. (5.10). The model parameter ACDE is a fitting parameter with a typical value of 0.25 for most process technologies. Fig. 5.5(a) shows that Eq. (5.10) has good accuracy compared to numerical quantum mechanical analyses for various channel doping concentrations, gate voltages, and oxide thicknesses.

Eq. (5.10) cannot be implemented as is in SPICE because the exponential term needs to be curbed numerically. Although a very large accumulation/depletion charge layer thickness is physically correct as shown in Fig. 5.5(a), it is not numerically robust when  $V_{gs\_eff}$  is much larger than the sum of  $V_{bseffCV}$  and  $V_{fbzb}$  of the exponential term. One remedy is to clamp it with an upper bound  $L_{Debye}/3$  when this happens. This is an important method of clamping useful for SPICE modeling in addition to those introduced in Chapter 2. For this consideration,  $X_{DC}$  of Eq. (5.10) is now replaced with a typical BSIM4 smoothing function for SPICE implementation

$$X_{DC} = \frac{1}{3}L_{Debye} - 0.5 \cdot \left( T_0 + \sqrt{T_0^2 + 4 \cdot 10^{-3} \cdot \text{TOXP} \cdot \frac{L_{Debye}}{3}} \right) \quad (5.11)$$

where  $T_0 = (L_{Debye}/3) - X_{DC} - 10^{-3} \cdot \text{TOXP}$  with  $X_{DC}$  given by Eq. (5.10). Then, Eq. (5.11) reduces to Eq. (5.10) in the accumulation and depletion ranges of operation. That is when  $V_{gs\_eff}$  is greater than the sum of  $V_{bseffCV}$  and  $V_{fbzb}$ .

In BSIM4, there are two flat-band voltages that need to be distinguished. One is the model parameter VFB that can be extracted from measured data and specified by the user or can be computed from the long-channel zero-bias threshold voltage VTH0 if it is not given. If specified, it can be used to compute VTH0 when VTH0 is not specified in model cards. It is also employed to determine the electric-field boundary conditions associated with such terms as  $(VTH0 - VFB - \varphi_s)$  for the mobility models or the inversion charge thickness model  $X_{DC}$  to be presented shortly. One lesson the authors learned about MOSFET SPICE modeling is that a constant flat-band voltage makes accurate

charge and capacitance modeling difficult, if not impossible. This is true for both  $V_{th}$  and surface potential-based MOSFET models. It makes the V-shaped gate capacitance ( $C_{gg}$ ) versus  $V_g$  curve too narrow and too shallow as the gate length becomes shorter. To prevent this from happening for charge and capacitance modeling [Eq. (5.10) and below], BSIM4 uses the other flat-band voltage that is derived from and is consistent with the zero-bias threshold voltage. It is

$$V_{fbzb} = V_{th}|_{V_{ds}=V_{bs}=0} - \varphi_s - K_1 \cdot \sqrt{\varphi_s} \quad (5.12)$$

where  $V_{th}$  is the threshold voltage given in Chapter 2. It is the complete threshold voltage that includes all the known bias, geometry and process effects. This zero-bias flat-band voltage is also used in the gate direct-tunneling current model presented in Chapter 4 for the same reason.

The charge thickness model for the accumulation to depletion region of operation is given in Eqs. (5.10) and (5.11). The inversion charge layer thickness in the unit of centimeters is obtained similarly from quantum mechanical analyses

$$X_{DC} = \frac{1.9 \times 10^{-9} \cdot ADOS}{1 + \left[ \frac{V_{gsteff} CV^{-4} \cdot (VTH0 - VFB - \varphi_s)}{2 \cdot TOXP \cdot 10^8} \right]^{0.7 \cdot BDOS}} \quad (5.13)$$

The surface potential is given in Chapter 2 and is repeated below for quick reference.

$$\varphi_s = \frac{k \cdot TNOM}{q} \cdot \log \left( \frac{NDEP}{n_i} \right) + PHIN + 0.4 \quad (5.14)$$

The second term in the denominator of Eq. (5.13) is given in the unit of MV/cm. For N<sup>+</sup> or P<sup>+</sup> poly-silicon gates, (VFB +  $\varphi_s$ ) is approximately zero, but in the very beginning of the BSIM4 development (back to 1998), this term was intentionally kept for the then-emerging high- $k$  metal gate technology. The physics-based philosophy and practice make BSIM4 useful and predictive for more than half a dozen CMOS technology nodes to date. Fig. 5.5(b) shows the comparison of this inversion charge thickness model against the numerical quantum analysis for various gate oxide thicknesses (TOXP) and channel doping concentrations (NDEP).

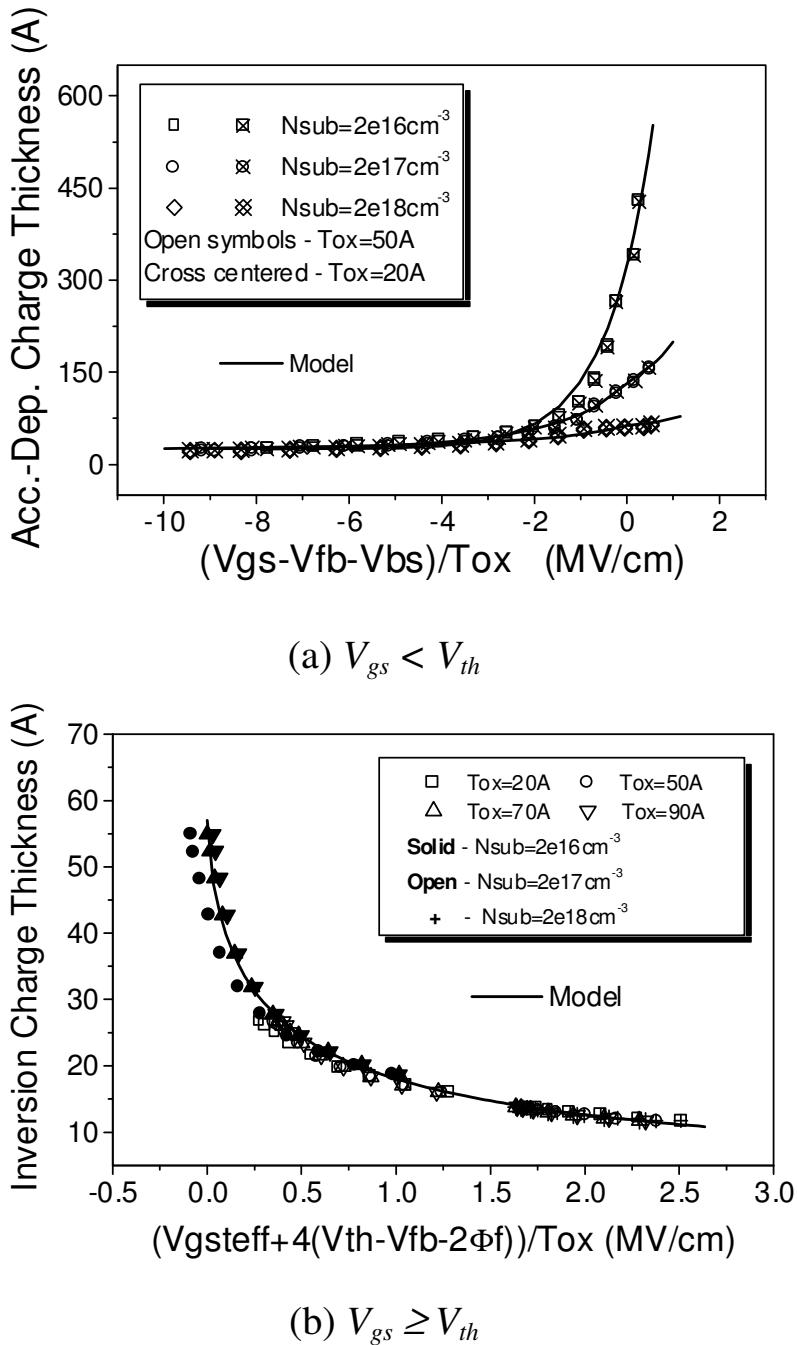


Fig. 5.5 The charge-thickness model agrees with numerical quantum mechanical simulations for various channel doping concentrations and gate oxide thicknesses, throughout all regions of operation: (a) Accumulation and depletion regions and (b) inversion region [5].

The effective gate voltage  $V_{gsteff_{CV}}$  used in CAPMOD = 1 and 2 has two versions: The simple version is turned on by setting the model selector CVCHARGEMOD = 0 (the default setting)

$$V_{gsteffCV} = n \cdot \text{NOFF} \cdot \frac{k_B T_{emp}}{q} \cdot \log \left[ 1 + \exp \left( \frac{V_{gs\_eff} - V_{th} - \text{VOFFCV}}{n \cdot \text{NOFF} \cdot \frac{k_B T_{emp}}{q}} \right) \right] \quad (5.15)$$

where  $n$  is the sub-threshold swing (slope) factor which is dependent on drain and body biases.  $\text{NOFF}$  and  $\text{VOFFCV}$  are fitting parameters with default values of 1 and zero, respectively. These two parameters are introduced to improve the model flexibility for different technologies by adjusting the sub-threshold slope and location with respect to the gate voltage [For details and insights, refer to Chapter 2]. If a non-zero  $\text{CVCHARGEMOD}$  is explicitly specified in model cards, the more sophisticated version of  $V_{gsteffCV}$  will be turned on

$$V_{gsteffCV} = \frac{n \cdot \frac{k_B T_{emp}}{q} \cdot \log \left[ 1 + \exp \left( m_{CV}^* \frac{V_{gs\_eff} - V_{th}}{n \cdot \frac{k_B T_{emp}}{q}} \right) \right]}{m_{CV}^* + n \cdot \frac{k_B T_{emp}}{q} \cdot \frac{C_{oxe}}{C_{dep0}} \cdot \exp \left( \left[ -\frac{(1-m_{CV}^*) \cdot (V_{gs\_eff} - V_{th}) - V_{offL}}{n \cdot \frac{k_B T_{emp}}{q}} \right] \right)} \quad (5.16)$$

This is the same  $V_{gsteff}$  function used for the BSIM4 DC models except for the separate  $m_{CV}^*$  and  $V_{offL}$  terms for the  $\text{CAPMOD} = 1$  and 2 charge and capacitance models for better accuracy and extraction flexibility. Refer to the detailed discussions about the development of  $V_{gsteff}$  in Chapter 3. It is worthwhile to note that  $V_{gsteffCV}$  is introduced with emphases on a single-piece and continuous model formulation and with the dedicated care to the modeling of moderate inversion region.  $n$  and  $C_{dep0}$  are two key factors determining the sub-threshold slope.  $C_{dep0}$  is the depletion layer capacitance per unit area under zero body biases

$$C_{dep0} = \sqrt{\frac{q \cdot \text{NDEP} \cdot 10^6 \cdot \varepsilon_{sub}}{2 \varphi_s}} \quad (5.17)$$

The constant  $10^6$  is a  $\text{cm}^2$ -to- $\text{m}^2$  unit conversion coefficient. The  $m_{CV}^*$  term is very useful to improve the accuracy for the moderate inversion region. It is typically around 0.5 and must fall into the range of 0 to 1, exclusive. For the sake of ease of parameter extractions, the  $m_{CV}^*$  term is also required to change slowly with the model parameter  $\text{MINVCV}$ . BSIM4 provides a handy mathematical formulation to fulfill these requirements

$$m_{CV}^* = 0.5 + \frac{\tan^{-1}(\text{MINVCV})}{\pi} \quad (5.18)$$

where the second term on the right side changes gradually from  $-0.5$  to  $0.5$  for any value in  $(-\infty, +\infty)$  of the model parameter MINVCV. The  $V_{offL}$  term is

$$V_{offL} = \text{VOFFCV} + \frac{\text{VOFFCVL}}{L_{eff}} \quad (5.19)$$

It enables the modeling of the length dependence of the offset voltage parameter VOFFCV to improve the length scalability of BSIM4 in the sub-threshold/depletion region of operation.

The Poisson equation states that the channel surface potential is a function of MOSFET terminal biases. This is especially true in the moderate inversion region of operation. Being able to do so has never been more critical for accurate simulation of advanced analog and RF circuits. Surface potential based charge-sheet models have the potential of modeling this region of operation more accurately. Unfortunately, they still ignore the finite charge layer thickness.

Consider the thickness of the unit-area inversion charge ( $q_{inv}$ ) and bulk charge ( $q_B$ ) densities in the inversion region. The bias-dependent surface potential  $\Phi_s$  can be written as

$$\Phi_s = -\frac{X_{DC} \cdot q_{inv}}{\varepsilon_{sub}} - \frac{X_{bulk} \cdot q_B}{\varepsilon_{sub}} \quad (5.20)$$

where  $q_{inv}$  is

$$q_{inv} = -C_{oxeff} \cdot (V_{gsteffCV} - \Phi_\delta) \quad (5.21)$$

$C_{oxeff}$  is the unit-area gate oxide capacitance including the charge layer thickness effects as given by Eq. (5.9). The difference between  $\Phi_s$  and the strong inversion surface potential  $\varphi_s = 2\varphi_B$  defines a delta surface potential  $\Phi_\delta = -X_{bulk} \cdot q_B / \varepsilon_{sub} - 2\varphi_B$ . An analytical formulation for  $\Phi_\delta$  is obtained from the 1-D quantum mechanical solver [4, 5]. It is

$$\Phi_\delta = \frac{k_B T_{emp}}{q} \cdot \log \left[ \frac{V_{gsteffCV} \cdot (V_{gsteffCV} + 2 \cdot K1_{ox} \cdot \sqrt{\varphi_s})}{MOIN \cdot K1_{ox} \cdot \left( \frac{k_B T_{emp}}{q} \right)^2} + 1 \right] \quad (5.22)$$

Here, MOIN is a fitting parameter with a typical value of 15. Refer to Chapter 2 for the definition of  $K1_{ox}$ .

### 5.3.2 CAPMOD = 2 Charge Model Formulations

The CAPMOD = 2 charge and capacitance model uses the same formulations as CAPMOD = 1. The differences are that CAPMOD = 2 uses the effective gate dielectric capacitance  $C_{oxeff}$  (bias-dependent as in Eq. (5.9)) and the  $\Phi_\delta$  model to incorporate the finite charge thickness effects, whereas CAPMOD = 1 does not. Instead, CAPMOD = 1 uses the electrical gate oxide thickness TOXE or the equivalent gate dielectric thickness (EOT if not made of oxide) to compute a bias-independent gate dielectric capacitance.

The mathematical formulations of these two models are made compact as well as smooth and continuous for robust convergence. Their implementations into SPICE follow those guidelines in Section 1 of this chapter to ensure accuracy and further reduce runtime cost. The CAPMOD = 2 charge model formulations are discussed below.

In the accumulation region, the accumulation charges  $Q_{acc}$  are isolated from the reverse source-body and drain-body P-N junctions. The charges are simply proportional to the gate-body voltage minus the flat-band voltage of the device. With the finite charge thickness included,  $Q_{acc}$  is modeled by

$$Q_{acc} = -C_{oxeff} \cdot W_{effCV} \cdot L_{effCV} \cdot NF \cdot (V_{fbeff} - V_{fbzb}) \quad (5.23)$$

$V_{fbeff}$  is the smooth function of the gate-body voltage minus the flat voltage. It yields the gate-body voltage when in accumulation. It approaches  $V_{fbzb}$  when the device is moving out of the accumulation bias condition such that Eq. (5.23) becomes zero. Thus, the same continuous equation is evaluated in simulation without a sudden stop or switching to a different equation, which could otherwise lead to difficult convergence in circuit simulation.  $V_{fbeff}$  uses the same smoothing function

$$V_{fbeff} = V_{fbzb} - \frac{1}{2} \cdot \left[ V_0 + \sqrt{V_0^2 - 0.08V_{fbzb}} \right] \quad (5.24)$$

where  $V_0 = -V_{gs\_eff} + V_{bseffCV} + V_{fbzb} - 0.02$ . Note that  $V_{fbzb}$  is used here instead of the model parameter VFB for reasons given previously.

In the depletion region, the bulk charge  $Q_b$  of a MOSFET transistor results from the body depletion layer underneath the interface.  $Q_b$  is proportional to the width of that depletion layer. For a long-channel device or when the drain bias  $V_{ds}$  is close to zero, the depletion layer width (thickness) is approximately constant from the source to the drain,

hence a relatively constant body charge  $Q_{b0}$ . For a short-channel device and/or as the drain bias increases, the depletion layer becomes thicker near the drain end, leading to the presence of extra body charges,  $\delta Q_b$ . Hence, the total body charge becomes  $Q_b = Q_{b0} + \delta Q_b$ .  $\delta Q_b$  reduces the saturation voltage, the effective channel length, and the channel charges. This effect is referred to as the bulk-charge effect, which is characterized, in the BSIM4 charge and capacitance model, by a bulk-charge effect coefficient  $A_{bulkCV}$ .

$$A_{bulkCV} = A_{bulk0} \cdot \left[ 1 + \left( \frac{CLC}{L_{effCV}} \right) \right]^{CLE} \quad (5.25)$$

CLC and CLE are two fitting parameters for the short-channel charge and capacitance modeling. Refer to Chapter 2 for details on  $A_{bulk0}$ .

By taking into account the delta surface potential  $\Phi_\delta$  owing to the finite charge layer thickness, the new saturation voltage for the charge and capacitance model becomes

$$V_{dsatCV} = \frac{V_{gsteffCV} - \Phi_\delta}{A_{bulkCV}} \quad (5.26)$$

which links the linear and saturation drain-source voltage by an effective  $V_{dseffCV}$  for the CV model

$$V_{dseffCV} = V_{dsatCV} - \frac{1}{2} \cdot \left[ V_1 + \sqrt{V_1^2 + 0.08V_{dsatCV}} \right] \quad (5.27)$$

where  $V_1 = V_{dsatCV} + V_{ds} - 0.02$ .

By solving the Poisson equation, the continuous BSIM4  $Q_{b0}$  model is readily obtained

$$Q_{b0} = \frac{-C_{oxeff} \cdot W_{effCV} \cdot L_{effCV} \cdot NF \cdot K1_{ox} \cdot \left[ -\frac{K1_{ox}}{2} + \sqrt{\frac{K1_{ox}^2}{4} + (V_{gs\_eff} - V_{fbeff} - V_{bseffCV} - V_{gsteffCV})} \right]}{(5.28)}$$

This equation may seem too complex to gain any insights from it. In fact, the terms in the brackets on the right side reduce to the familiar term  $\sqrt{\varphi_s - V_{bs}}$  in the inversion region where  $V_{gsteffCV}$  is  $(V_{gs\_eff} - V_{th})$ ,  $V_{fbeff}$  is  $V_{fb}$ , and  $V_{bseffCV}$  can be replaced with  $V_{bs}$ .

In the following, the terminal and channel charges are to be obtained by performing integration along the channel. To do so, one needs to relate  $dy$  to  $dV_y$  from the DC channel current model

$$dy = \frac{L_{effCV} \cdot (V_{gsteffCV} - \Phi_\delta - A_{bulkCV} V_y)}{(V_{gsteffCV} - \Phi_\delta - \frac{1}{2} A_{bulkCV} V_{dseffCV}) \cdot V_{dseffCV}} \cdot dV_y \quad (5.29)$$

where the delta surface potential  $\Phi_\delta$  is included as well. The integration gives the inversion channel charge, gate charge and body charge as

$$Q_{inv} = -C_{oxeff} W_{effCV} NF \cdot \int_0^{L_{effCV}} (V_{gsteffCV} - A_{bulkCV} V_y) \cdot dy \quad (5.30a)$$

$$Q_g = C_{oxeff} W_{effCV} NF \cdot \int_0^{L_{effCV}} (V_{gs\_eff} - V_{fbzb} - \varphi_s - V_y) \cdot dy \quad (5.30b)$$

From  $Q_b = -(Q_g + Q_{inv})$ , one obtains

$$Q_b = -C_{oxeff} W_{effCV} NF \cdot \int_0^{L_{effCV}} [V_{th} - V_{fbzb} - \varphi_s - (A_{bulkCV} - 1) \cdot V_y] \cdot dy \quad (5.30c)$$

The integration process itself is trivial and omitted here. From Eq. (5.30c),  $\delta Q_b = Q_b - Q_{b0}$ .  $Q_{inv}$  is partitioned into the source and drain charges  $Q_{inv} = Q_s + Q_d$  by following the three partitioning schemes discussed in Section 1 of this chapter.

In the SPICE implementation of the BSIM4 charge and capacitance models, the following charges are computed:  $Q_{acc}$ ,  $Q_{b0}$ ,  $\delta Q_b$ ,  $Q_{inv}$  and  $Q_s$ . Their derivatives with respect to the drain, gate and body node voltages are evaluated to get nine independent trans-nodal capacitances (Refer to Section 1 of this chapter). The terminal intrinsic charges (excluding parasitic charges such as those of junction, overlap, and fringing capacitances) are then derived from those charge components directly:

$$\begin{cases} Q_b = Q_{acc} + Q_{b0} + \delta Q_b \\ Q_g = -(Q_b + Q_{inv}) \\ Q_d = -(Q_g + Q_b + Q_s) \end{cases} \quad (5.31)$$

NMOS and PMOS share the same models discussed above. In addition, these models are developed for the forward-mode operation ( $V_{ds} \geq 0$ ). This is the practice for an intrinsic MOSFET model. Swapping source and drain for the reverse-mode operation ( $V_{ds} < 0$ ) and PMOS-type conversion are performed in loading the model into SPICE circuit matrices, both the Jacobian and the RHS (right-hand side current vectors). Refer to Chapter 10 for the details of the BSIM4 SPICE implementation methodology.

Figure 5.6 shows the gate-to-channel capacitance  $C_{gc}$  comparison between measured CV data and the CAPMOD = 2 model. This model works well with the high- $k$  metal-gate stacks as well [5].

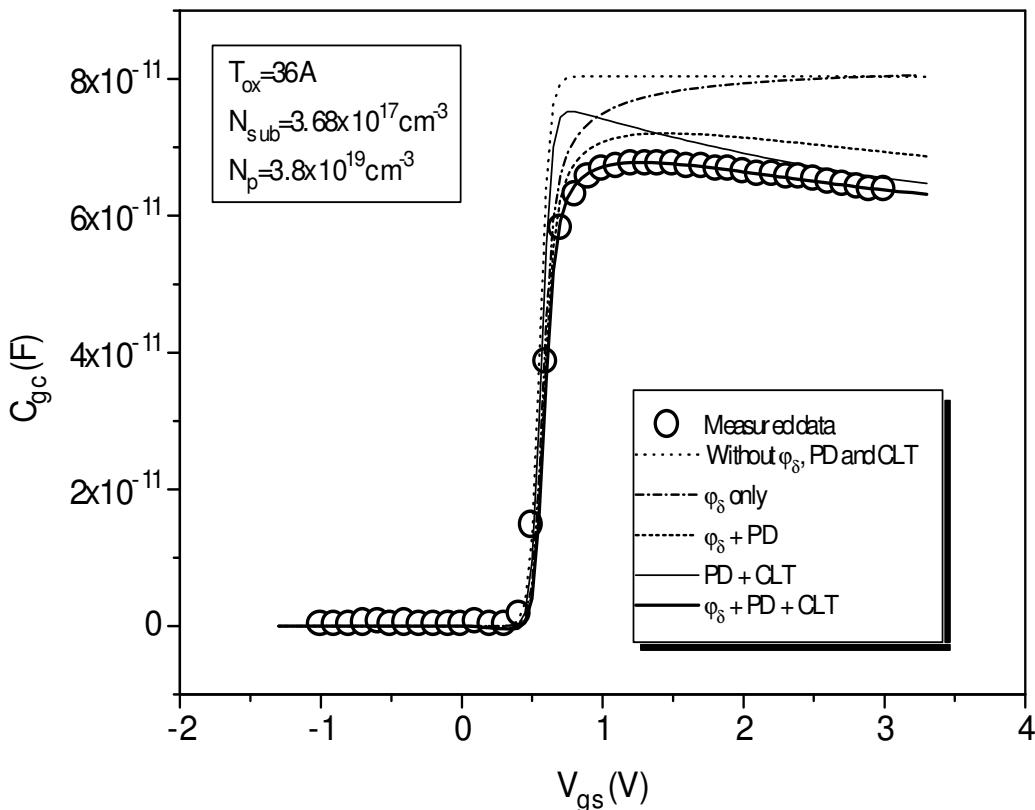


Fig. 5.6 Comparison of  $C_{gc}$  between measured data and BSIM4 CAPMOD = 2. The poly-silicon depletion (PD; poly doping:  $3.8 \times 10^{19} \text{ cm}^{-3}$ ), charge layer thickness model (CLT in the figure is another name for CTM in the text), and  $\Phi_\delta$  are turned on and off in the simulation to isolate the contribution of each effect to the capacitance [5].

## 5.4 Fringing and Overlap Capacitances

### 5.4.1 Fringing Capacitances

It is known that the gate electrode (either made of poly silicon or metal) and the source and drain diffusions are capacitively coupled with the source and drain diffusions through the gate sidewall spacers or the channel depletion region and the gate dielectric layer. This leads to, respectively, the outer and inner fringing capacitances. The inner fringing capacitance has weak dependencies on biases. It is relatively small in magnitude and difficult to separate from the intrinsic and overlap gate-to-source and gate-to-drain capacitances in measurement and parameter extractions. Hence, MOSFET compact models usually ignore the inner fringing capacitance. So does BSIM4.

BSIM4 models the outer fringing capacitance with a binnable model parameter **CF** in the default case (i.e., if it is specified in model cards). This parameter is given in the unit of Farad per unit device width. If not specified, the outer fringing capacitance will then be given in its default value computed with

$$CF = \frac{2 \cdot EPSROX \cdot \epsilon_0}{\pi} \cdot \log \left( 1 + \frac{4 \times 10^{-7}}{TOXE} \right) \quad (5.32)$$

**EPSROX** is the relative permittivity of the gate dielectrics. The constant  $4 \times 10^{-7}$  given in the unit of meters represents the typical gate electrode thickness. Note that there is no known theoretical derivation of Eq. (5.32). Instead, it was obtained by empirically fitting to TCAD numerical 2D simulation and measured data.

In the state-of-the-art process technologies, the outer fringing capacitance is quite significant relative to the intrinsic capacitances as the gate length is scaled down to sub-30 nanometers. In addition, the capacitances between the gate and the source and drain contacts (also through the sidewall spacers) becomes much more prominent. LPE (layout parasitics extraction) tools often treat this capacitance component within device models for better modeling accuracy and easy extraction. In BSIM4, this component can be merged into the fringing or overlap capacitance parameters specified in model card libraries. The BSIM4 overlap capacitance model is presented in the next sub-section.

### 5.4.2 Overlap Capacitances

The gate electrode and dielectric layer need to have a finite overlap with the source and drain diffusions to avoid unduly large source and drain series resistance. Intuitively, like the gate/source and gate/drain fringing capacitances, the overlap portions leads to an extra extrinsic/parasitic capacitance on top of the intrinsic capacitances of the transistor associated with the gate and channel. The overlap capacitances can be treated as regular, bias-independent capacitors in the case of heavily-doped source and drain diffusions, where the charge accumulation condition in the surface of the diffusions dominates in the operation of CMOS transistors. The same picture holds true for the gate-to-body overlap capacitance between the gate electrode/contact and the silicon substrate/body. The gate-to-body overlap capacitance is relatively small though.

Advanced MOSFETs are manufactured with shallow source and drain extensions, also known as LDD regions (Lightly-Doped (source and) Drain). Use of shallow junctions are required by CMOS scaling in order to suppress GIDL (gate-induced drain leakage) and short-channel effects (SCE) especially the subthreshold standby channel leakage current between the source and the drain. The doping concentrations of LDD are fairly low (typically of the order of  $10^{17} \text{ cm}^{-3}$ ) compared to those of the deep source/drain diffusion regions (in the mid  $10^{18} \text{ cm}^{-3}$  range), where the source and drain contacts are made. Because of the relatively low doping concentrations, the gate voltage can cause depletion of carriers in the surface of the LDD source or drain regions. Therefore, bias-dependent overlap capacitances result, a phenomenon very much like an MOS capacitor operating in its depletion region.

In BSIM4, if the simple, piece-wise intrinsic capacitance model (**CAPMOD = 0**) is selected, the overlap charges are computed from constant, bias-independent overlap capacitances as follows

$$Q_{gs,olap} = W_{effCV} \cdot NF \cdot CGSO \cdot v_{gs} \quad (5.33)$$

for the gate-to-source overlap capacitance,

$$Q_{gd,olap} = W_{effCV} \cdot NF \cdot CGDO \cdot v_{gd} \quad (5.34)$$

for the gate-to-drain overlap capacitance, and

$$Q_{gb,olap} = L_{effCV} \cdot NF \cdot CGBO \cdot v_{gb} \quad (5.35)$$

for the gate-to-body overlap capacitance. CGSO, CGDO and CGBO are global non-binnable model parameters, all given in the unit of Farad per meter.

Note that if CGSO is not specified in model cards, BSIM4 sets up the default value of CGSO as follows. If a positive DLC is given, then

$$CGSO = DLC \cdot C_{oxe} - CGSL \quad (5.36a)$$

else

$$CGSO = 0.6 \cdot XJ \cdot C_{oxe} \quad (5.36b)$$

The same holds true for the default setting for CGDO with CGSL to be replaced by CGDL. Eq. (5.36b) is an empirical treatment. In the case of CGBO, if it is not given, it defaults to  $2 \cdot DWC \cdot C_{oxe}$ . DLC and DWC are the gate-source or gate-drain overlap length and the gate-to-body overlap width, respectively (refer to the channel length and width computations given in Chapter 2]. CGSL and CGDL are discussed in the following.

When CAPMOD = 1 or 2 is specified, BSIM4 will use its bias-dependent overlap capacitance model instead. Taking the source charge as an example, the overlap charge is given by

$$Q_{gs,olap} = W_{effCV} \cdot NF \cdot \left\{ CGSO \cdot v_{gs} + CGSL \cdot \left[ v_{gs} - v_{gs,olap} - \frac{CKAPPAS}{2} \left( -1 + \sqrt{1 - \frac{4 \cdot v_{gs,olap}}{CKAPPAS}} \right) \right] \right\} \quad (5.37)$$

CGSL is a bias-dependent gate-to-source overlap capacitance parameter in the unit of Farad per meter in the width direction. CKAPPAS is the gate-to-source voltage coefficient in the unit of volts. It functions similarly as the body-bias coefficient  $K1_{ox}$  of the  $V_{th}$  model: It describes how the depletion width of the source LDD region changes with  $v_{gs}$ . [Refer to Eq. (5.28) for the long-channel zero- $V_{ds}$  uniform body charge  $Q_{b0}$  model]. In Eq. (5.37),  $v_{gs,olap}$  is a smooth function for the gate-to-source voltage  $v_{gs}$ . It is given by

$$v_{gs,olap} = \frac{1}{2} \left[ (v_{gs} + 0.02) - \sqrt{(v_{gs} + 0.02)^2 + 0.08} \right] \quad (5.38)$$

This equation states that in the case of NMOS, when  $v_{gs}$  is negative,  $v_{gs,olap} = v_{gs}$  and therefore the overlap charge is approximately proportional to the square root of the voltage drop between gate and source (refer to Eq. (5.37)). When  $v_{gs}$  is positive,  $v_{gs,olap}$  is always zero and thus no bias-dependent capacitance will result and be modeled. This is in agreement with the fact that the source LDD region is now in its accumulation mode (for NMOS). A similar situation holds for PMOS.

It is noteworthy that CKAPPAS and CKAPPAD can never be less than or equal to zero. Inappropriate, non-positive CKAPPAS and CKAPPAD parameter extraction are frequently seen in practice, which leads to a divide-by-zero error in circuit simulation.

In SPICE modeling and implementation, extrinsic charges and capacitances such as those associated with overlaps are not swappable and need not be interchanged between source and drain for the reverse-mode operations (i.e.,  $V_{ds} < 0$ ). This is because they are the charges and capacitances attached to the actual, physical source and drain as specified in SPICE net-lists/decks. In order to underscore this distinction, the equations above use lower-case  $v$ 's such as in  $v_{gs}$  to distinguish them from the capital  $V$ 's employed in the intrinsic models. Furthermore, the overlap capacitances, like the MOSFET junction capacitances, are reciprocal with respect to the two terminals of the device. This is in contrast to the intrinsic trans-nodal capacitances such as  $C_{gd}$  and  $C_{dg}$  (refer to the discussions in Section 5.1).

Finally, the complete charges found at the terminals of BSIM4 are the sum of the overlap and fringing charges and the intrinsic charges associated with each of those terminals.

## 5.5 Chapter Summary

This chapter presented the BSIM4 charge-based quasi-static capacitance models with a focus on the modeling of the finite charge-layer thickness effects that are important for sub-90nm process technologies. We also reviewed and analyzed the fundamentals of the MOSFET charge and capacitance theory and their implications to circuit operation including channel charge partitioning. It presented a holistic view of MOSFET charging and trans-nodal capacitive coupling and currents as depicted in Fig. 5.2. The charge model has the rigorous basis of the Electromagnetic Theory based on the Coulomb Law of electrical force on electron charge in an electric field and potential. The non-quasi-static and RF models of BSIM4 will be presented in Chapter 6.

## 5.6 Parameter Table

Name (type)	Description and default	Can be binned?	Note
CAPMOD (Global; integer)	Charge and capacitance model selector.  Default = 2; dimensionless. The other optional values are 0 and 1.	No	-
XPART (Global; double)	Channel charge partitioning selector.  Default = 0.0 (the 40/60 partition); dimensionless. The other optional values are 0.5 (50/50) and 1.0 (0/100).	No	If it is set less than 0 (a very rare or mistaken setting), no intrinsic charge and capacitance model will be evaluated.
CVCHARGEMOD (Global; integer)	$V_{gsteff}$ model selector for CAPMOD = 1 and 2.  Default = 0; dimensionless. The other optional value is a non-zero value, such as 1.	No	The default uses a simple $V_{gsteff}$ .
ACDE (Global; double)	A fitting parameter of the CAPMOD = 2 charge and capacitance model. It improves the charge and capacitance fitting accuracy from accumulation to depletion.  Default = 1.0; dimensionless.	Yes	Warning messages to be issued if its binned value is either less than 0.1 or greater than 1.6.
ADOS (Global; double)	Inversion charge thickness parameter to account for the density of states, used in both the DC channel current and the CAPMOD = 2 charge and capacitance models.  Default = 1.0; dimensionless.	No	-
BDOS (Global; double)	Inversion charge thickness power parameter to account for the density of states, used in both the DC channel current and the CAPMOD = 2 charge and capacitance models.  Default = 1.0; dimensionless.	No	-

VOFFCV (Global; double)	Offset voltage parameter to shift CV curves with respect to the gate voltage for CAPMOD = 1 and 2.  Default = 0.0; dimensionless.	Yes	-
VOFFCVL (Global; double)	Length-dependence parameter for the offset voltage VOFFCV for CAPMOD = 1 and 2.  Default = 0.0; dimensionless.	No	-
MINCVV (Global; double)	Parameter to adjust the CV curvature in the moderate inversion region of CAPMOD = 1 and 2. It is activated if CVCHARGEMOD is not zero.  Default = 0.0; dimensionless.	Yes	-
MOIN (Global; double)	A CAPMOD = 2 model parameter of the bias-dependent surface potential model. It is used for more accurate modeling of the charges and capacitance of the moderate inversion region of operation.  Default = 15.0; dimensionless.	Yes	If its binned value is less than 5 or greater than 25, a warning message will be issued.
CLC (Global; double)	Length-dependence coefficient parameter of the bulk charge effect parameter $A_{bulkCV}$ of the charge and capacitance models.  Default = $0.1 \times 10^{-6}$ in [m].	Yes	If the binned value is negative, a fatal message will be issued.
CLE (Global; double)	Length-dependence power parameter of the bulk charge effect parameter $A_{bulkCV}$ of the charge and capacitance models.  Default = 0.6; dimensionless.	Yes	-
CF (Global; double)	Outer fringing capacitance per unit width.  Default = Computed if not given in [Farad per meter].	Yes	-
CGDO (Global; double)	Bias-independent gate-to-drain overlap capacitance per unit width, given in the unit of [Farad/m].  Default: Refer to the text for various possible scenarios.	No	-

CGSO (Global; double)	Bias-independent gate-to-source overlap capacitance per unit width, given in the unit of [Farad/m].  Default: Refer to the text for various possible scenarios.	No	-
CGBO (Global; double)	Bias-independent gate-to-body overlap capacitance per unit length, given in the unit of [Farad/m].  Default: Refer to the text for various possible scenarios.	No	-
CGDL (Global; double)	Bias-dependent gate-to-drain overlap capacitance parameter per unit width.  Default = 0.0 in [Farad/m].	Yes	-
CGSL (Global; double)	Bias-dependent gate-to-source overlap capacitance parameter per unit width.  Default = 0.0 in [Farad/m].	Yes	-
CKAPPAS (Global; double)	Gate-to-source voltage-dependence coefficient of the bias-dependent gate-to-source overlap capacitance model.  Default = 0.6 in [V].	Yes	If its binned value is less than 0.02, a warning message will be issued and it is reset to 0.02.
CKAPPAD (Global; double)	Gate-to-drain voltage-dependence coefficient of the bias-dependent gate-to-drain overlap capacitance model.  Default = CKAPPAS in [V].	Yes	If its binned value is less than 0.02, a warning message will be issued and it is reset to 0.02.

## References

- [1] D. E. Ward, and R. W. Dutton, “A charge-orient model for MOS transistor capacitances,” *IEEE J. Solid-State Circuits*, vol. 13, pp. 703-708, 1978.

- [2] P. Yang, "Capacitance Modeling for MOSFET," in Advances in CAD for VLSI, vol. 3, Edited by A. E. Ruehli. Amsterdam, The Netherlands: North Holland, pp. 107-130, 1986.
- [3] Weidong Liu, Michael Orshansky, Xiaodong Jin, Kai Chen, and Chenming Hu, "MOSFET intrinsic-capacitance related inaccuracy in CMOS circuit speed simulation," IEEE 1997 International Semiconductor Device Research Symposium, pp. 337-340, Virginia, 1997.
- [4] Weidong Liu, and Chenming Hu, "BSIM3v3 MOSFET Model" — Silicon and Beyond: Advanced Device Models and Circuit Simulators, edited by Michael S. Shur and Tor A. Fjeldly, pp. 1-31, ISBN: 981-02-4280-8, World Scientific, 2000.
- [5] Weidong Liu, Xiaodong Jin, Yachin King, and Chenming Hu, "An efficient and accurate compact model for thin-oxide-MOSFET intrinsic capacitance considering the finite charge thickness for circuit simulation," IEEE Trans. on Electronic Devices, pp.1070-1072, May 1999.
- [6] Weidong Liu, Xiaodong Jin, James Chen, Min-Chie Jeng, Zhihong Liu, Yuhua Cheng, Kai Chen, Mansun Chan, Kelvin Hui, Jianhui Huang, Robert Tu, Ping K. Ko, and Chenming Hu, "BSIM3v3.2 MOSFET model — Users' manual", Memorandum No. UCB/ERL M98/51. Electronics Research Laboratory, College of Engineering, University of California, Berkeley, August 21, 1998. Give total number of pages.
- [7] Weidong Liu, Xiaodong Jin, Kanyu M. Cao, and Chenming Hu, "BSIM4.0.0 MOSFET Model — User's Manual", Memorandum No. UCB/ERL M00/38, Electronics Research Laboratory, College of Engineering, University of California, Berkeley, August 3, 2000. Give total number of pages
- [8] Mohan V. Dunga, Wenwei (Morgan) Yang, Xuemei (Jane) Xi, Jin He, Weidong Liu, Kanyu, M. Cao, Xiaodong Jin, Jeff J. Ou, Mansun Chan, Ali M. Niknejad, and Chenming Hu, "BSIM4.6.1 MOSFET Model — User's Manual," University of California, Berkeley, 2007.
- [9] Ya-Chin King, H. Fujioka, S. Kamohara, W.-C. Lee, and Chenming Hu, "AC charge centroid model for quantization of inversion layer in NMOSFET," Int. Symp. VLSI Technology, Systems and Applications, Proc. of Tech. Papers, Taipei, Taiwan, pp. 245-249, June 1997.

**This page intentionally left blank**

## **Chapter 6**

# **Non-Quasi-Static and Parasitic Gate and Body Resistances**

### **6.1 Introduction and Chapter Objectives**

In the preceding chapters, all the charge, current, conductance, and capacitance models were developed under DC bias conditions. However they are also routinely used when the terminal voltages are time varying. The implicit assumption here is that the channel charge carriers can respond to time-varying terminal biases instantaneously. This means that the charge movement, distributions and even channel-charge source and drain partition are able to follow exactly, with no dependence on the history of the terminal voltages. It is as if the biases at that particular moment had been static for a long time, an assumption known as the quasi-static approximation. In reality, these charges may not follow the changes in voltages instantaneously. Therefore, the accurate modeling of transistor operations at very high frequencies (in the giga-hertz range) requires more sophisticated non-quasi-static (NQS) treatment of the charges and currents.

TCAD device simulation tools can perform self-consistent numerical analysis of the time-varying Poisson and transport equations. The result is correct and useful for validating non-quasi-static compact models provided the transport parameters used in the TCAD tools have been calibrated with measured data already.

TCAD tools' ability to provide NQS analyses is based on two differences from quasi-static compact models. One is that the capacitances between the channel and the gate and the substrate are not treated in a lump but in a distributed fashion. The other is that the carrier

density in each segment of the channel is determined by detailed analysis of the drift, diffusion, and current continuity equations. The former leads to a frequency dependent capacitances and conductances determined by the distributed RC network, whereas the latter, at the basic level, leads to a finite carrier transit time for the carriers to traverse the entire channel region before collected by drain. The consequence is that the channel charge distribution is not just a function of instantaneous device terminal biases but a function of their history of the past pico to nano seconds depending on the channel length.

The gate and substrate resistances also form distributed RC network and, together with the NQS effect (with the root cause in the channel), determine the high-frequency or high-speed behaviors of MOSFETs. In this chapter, after a brief discussion of the gate electrode resistance model, an intrinsic-input gate resistance concept and its model development are presented to account for the distributed nature of the gate-channel RC network. This is followed by the derivations and analyses of two alternative NQS models. One is intended for the modeling of transistor operations in fast transients and the other is intended for modeling the small-signal AC operation at high frequencies. It is known that under radio-frequency (RF) operations, it is imperative to incorporate the parasitic, distributed body resistance in SPICE modeling. The BSIM4 body resistance network model is thus finally presented. Whenever possible during the course of the presentation, attentions are given to how the modes are derived and the meaning of key model parameters. This knowledge is useful for the understanding and development of parameter extraction.

SPICE implementation of an NQS model calls for special implementation techniques and skills. This will be presented in Chapter 10.

## 6.2 Gate Electrode Resistance

The (poly-silicon) gate electrode introduces a finite resistance. In SPICE compact modeling, this gate electrode resistance cannot be ignored for high-speed circuits, whether digital or analog or RF. To model the

resistance value, one needs to take into account the fact that the gate electrode resistance is distributive and the current (transient or AC) is not uniform in the gate electrode. Moreover, it forms a complex RC network with the channel resistance and the gate dielectric capacitance as illustrated in Fig. 6.1. The non-uniformity results directly from the presence of the distributed RC network.

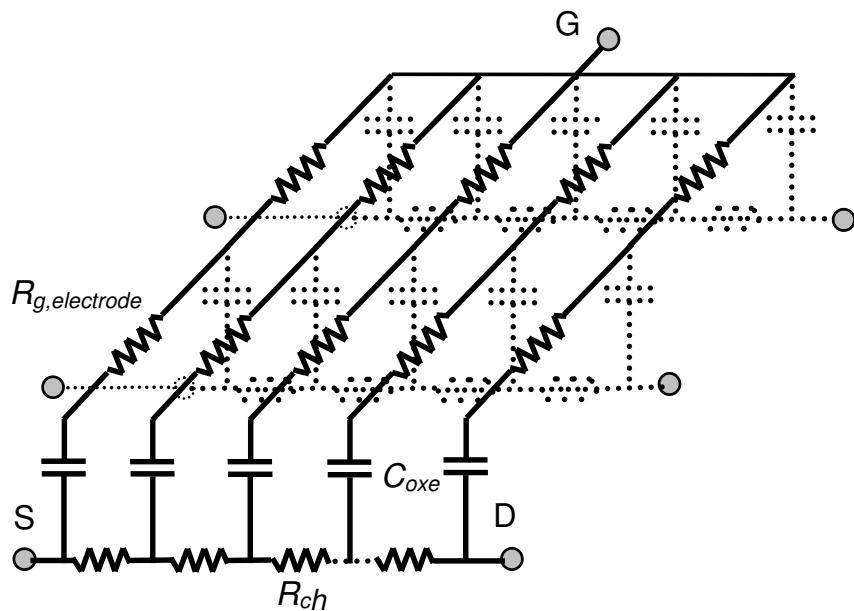


Fig. 6.1 Distributed-RC network associated with the poly-silicon gate, the gate oxide layer, and the (inverted) channel conduction layer. The transistor source is on the left side and drain on the right. The gate electrode is shown as many segments lined up along the device width direction. They are coupled through the gate capacitance to the channel segments that are lined up along the device length direction.

For simplicity BSIM4 models the gate resistance by assuming the channel is a (AC) ground plane. It can be shown that the lumped gate electrode resistance can be derived as [1]

$$R_{g,electrode} = \frac{R_{SHG} \cdot \left( X_{GW} + \frac{W_{effJCT}}{3 \cdot NGCON} \right)}{NGCON \cdot (L_{designed} - X_{GL})} \cdot \frac{1}{NF} \quad (6.1)$$

by applying the transmission line theory. The effect of the channel resistance is modeled in the next section. **RSHG** is the poly-silicon sheet resistance provided by process development teams or extracted from device  $Y$  parameters; it is a constant model parameter in BSIM4 although it could be formulated to account for the bias dependencies due to poly-silicon gate depletion. For a typical MOSFET with  $W_{drawn} = 20\mu\text{m}$  and  $L_{drawn} = 0.1\mu\text{m}$ ,  $R_{g,electrode}$  is typically around a few tens of ohms for  $\text{NF} = 1$  and  $\text{NGCON} = 1$ . [Note that  $R_{g,electrode}$  is automatically set to  $10^{-3}$  ohm if it is not positive in BSIM4.]

**XGW** is the distance between the center of the gate contact and the edge of the device in the width direction. It can be a global parameter but is often used as an instance parameter because it is different for different layouts and is supplied by layout parasitics extraction (LPE) tools. **XGL** is a (global) model parameter of a fab process that accounts for the offset between the actual gate length (not the channel length) and the designed gate length. Although **XGL** is typically small, it is necessary to include it in the  $L_{designed}$  term in the denominator as the two can be comparable in magnitude. **XGL** is similar to the model parameter **XL**, which is a separate parameter to calculate effective channel-lengths (refer to Chapter 2).

**NGCON** of Eq. (6.1) is both an instance and a model parameter. It represents the gate electrode contact scenarios with two optional numerical settings: **NGCON** = 1 designates the gate contact being made on only one side of the electrode, whereas **NGCON** = 2 means that there are two electrode contacts, one on each side. With this, one expects Eq. (6.1) to give a coefficient of  $1/3$  for **NGCON** = 1 and  $1/12$  in the case of **NGCON**=2 to account for the distributive nature of the gate electrode resistance. BSIM4 chooses to use the source and drain junction width  $W_{effJCT}$  in Eq. (6.1) instead of  $W_{eff}$  of the IV model or  $W_{effCV}$  of the CV model. This is justifiable following the fact that the RC time constant concerns more about the geometrical width of the device than the electrical width.  $W_{eff}$  or  $W_{effCV}$ , is often smaller than  $W_{effJCT}$  due to the

existence of the fringing electric field effects associated with the field oxide (bird's beaks) or shallow-trench isolations (STI).

To evaluate the effects of  $R_{g,electrode}$  in SPICE simulations, set the model flag **RGATEMOD** = 1 locally in transistor element cards or globally in model card libraries. An internal gate node  $gNodePrime$  is generated to connect the intrinsic gate node to the external gate node  $gNodeExt$  through the lumped resistance  $R_{g,electrode}$ . This is shown in Fig. 6.2.

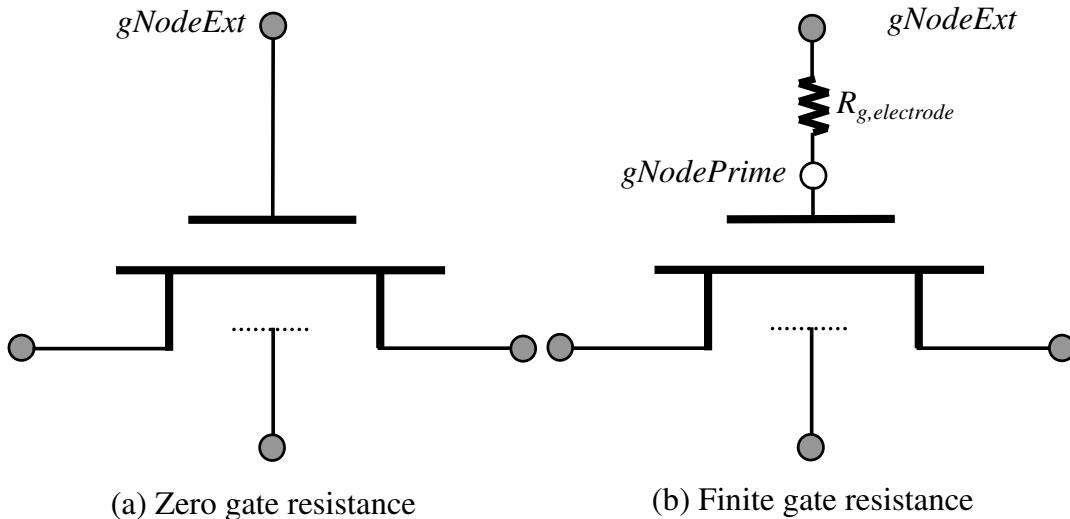


Fig. 6.2 BSIM4 model topologies with “Zero gate resistance” (a) and with “Finite gate resistance” (b). The zero-resistance mode is the default (**RGATEMOD** = 0) where the internal gate node  $gNodePrime$  is merged into the external,  $gNodeExt$ , in SPICE simulation. In (b), the resistance is composed of the bias-independent poly-silicon gate electrode resistance and an internal gate node  $gNodePrime$  that serves as the intrinsic gate.

### 6.3 Gate Intrinsic-Input Resistance for Non-Quasi-Static Modeling

Even if the gate electrode material is metal with negligible sheet resistance, an input signal source would still sense a resistive component in the input impedance. (There is of course a capacitive component.) BSIM4 introduced this resistive component and called it the *Intrinsic*

*Input Resistance.* Including the *Intrinsic Input Impedance* is a convenient computationally efficient way to model the basic non-quasi-static effect.

Intrinsic input resistance has its root in the channel resistance. The channel resistance has a similar distributive attribute as the (poly-silicon) gate electrode resistance discussed in the preceding section. It introduces another RC delay with the gate as shown in Fig. 6.1. However, unlike the gate electrode, the channel resistance is non-uniform between the source and the drain. In modeling the effect of the channel resistance on NQS in the following, it is assumed that the gate has zero resistance for ease of model derivation.

The traditional representation of the gate and channel regions of a MOSFET for SPICE modeling as sketched in Fig. 6.3 is overly ideal. In this drawing, the channel resistance is “invisible” to the input gate current, resulting in a zero time constant associated with the gate capacitance and channel resistance. While this assumption holds fairly well when the susceptance  $j\omega C_{ox}$  of the gate oxide capacitance is much smaller than the channel conductance of the device  $Y$  parameters, the “invisibility” of the channel becomes a serious deficiency in modeling high-speed circuits.

There are two possible remedies. Slicing the device into multiple shorter transistor segments that are connected in series by additional source and drain nodes (Fig. 6.4) is one way to model the distributed capacitive and resistive nature of the gate and channel region (Fig. 6.5). Although the accuracy improves as the slices become finer, there is a large SPICE simulation runtime overhead because of the additional nodes, which can make circuit matrices very large. Besides, the model that accurately represents the real short-channel transistors (see Chapters 2, 3 and 5) with their myriad of short-channel effects cannot accurately represent these imaginary “short” transistors. This approach presents large practical problems.

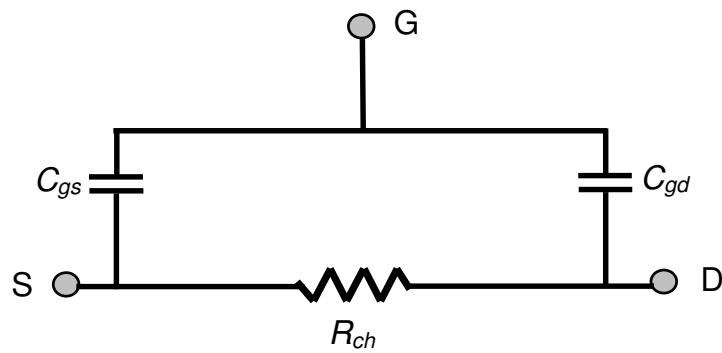


Fig. 6.3 A quasi-static MOSFET model topology, where the attribute of the distributed gate and channel RC is simply missing.

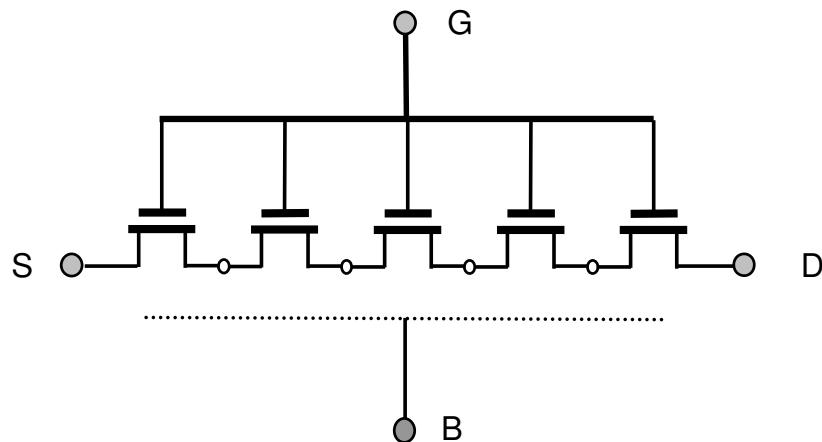


Fig. 6.4 Transistor-series representation of the gate-channel RC network. It can be accomplished either in SPICE netlisting or in model code implementation. The accuracy improves as the number of transistor segments increases. This approach is very computationally expensive.

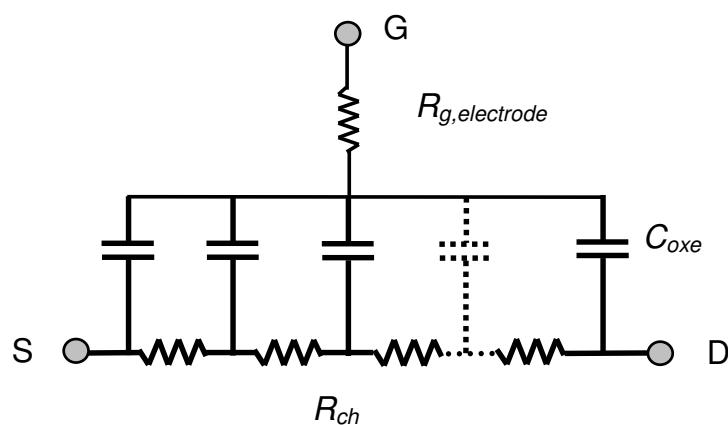


Fig. 6.5 BSIM4 compact representation of the distributed gate and channel RC network.

BSIM4 instead chooses to model NQS with an intrinsic-input gate resistance ( $R_{g,ii}$ ) [1]. In this approach, an appropriate fraction of the channel resistance that, if inserted between the gate topological nodes and the gate capacitances, can model the effective time constant for the non-quasi-static effects associated with the gate-channel RC network. The channel resistance  $R_{ch,Ohmic}$  is obtained by integrating over the channel the infinitesimal static channel resistance. It is expressed as

$$R_{ch,Ohmic} = \int_0^{V_{ds}} dR_{ch,Ohmic}(y) = \int_0^{V_{ds}} \frac{dV_y}{I_{ch}(y)} = \frac{1}{I_{ch}} \cdot \int_0^{V_{ds}} dV_y \quad (6.2)$$

with  $I_{ch}(y) = I_{ch}$  everywhere from source to drain to satisfy the channel current continuity requirement. Moreover, it is known that

$$R_{ch,Ohmic} = \begin{cases} \frac{V_{ds}}{I_{ch}} & \text{if } V_{ds} < V_{dsat} \\ \frac{V_{dsat}}{I_{ch}} & \text{if } V_{ds} \geq V_{dsat} \end{cases} \quad (6.3)$$

which is equivalent to

$$R_{ch,Ohmic} = \frac{V_{dseff}}{I_{ch}} \quad (6.4)$$

in terms of the BSIM4 continuous single-equation channel current model, where the effective drain-to-source voltage  $V_{dseff}$  term in the  $I_{ch}$  formulation (refer to Chapter 3 for details) conveniently cancels the numerator (that is,  $V_{dseff}$ ) of Eq. (6.4). Therefore, potential divide-by-zero errors in the  $R_{ch,Ohmic}$  code implementation at  $V_{ds} = 0$  is eliminated.

Equation (6.4) implies that the channel resistance is attributed only to the channel region from the source end to where the channel voltage reaches  $V_{dseff}$ , beyond which point the carrier transport to the drain end is not governed by resistive drift anymore.

It is known that a more accurate diffusion resistance model can be derived for the sub-threshold region by starting from the diffusion current equation

$$I_{ch} = -\frac{W_{eff} \mu_{eff} kT}{q} \cdot \frac{dQ_{ch}(y)}{dy} \approx \frac{W_{eff} \mu_{eff} kT}{q} \cdot \frac{Q_{ch}(y=0) - Q_{ch}(y=L_{eff})}{L_{eff}} \quad (6.5)$$

Recalling, from the section of the channel DC current modeling, the unified channel charge density  $Q_{ch}(y)$  that was developed with Taylor series expansion, one has

$$Q_{ch}(y) = C_{oxeff} \cdot V_{gsteff} \cdot \left[ 1 - \frac{\varphi_f(y)}{V_{bulk,q}} \right] \quad (6.6)$$

which can be simplified to

$$Q_{ch}(y=0) = C_{oxeff} \cdot V_{gsteff} \quad (6.6a)$$

with the quasi-Fermi potential  $\varphi_f = 0$  at the source side and

$$Q_{ch}(y=L_{eff}) = C_{oxeff} \cdot (V_{gsteff} - V_{dseff}) \quad (6.6b)$$

by approximating  $\varphi_f \approx V_{dseff}$  and the bulk-charge voltage  $V_{bulk,q} \approx V_{gsteff}$  at the drain side. Eq. (6.5) now becomes

$$I_{ch} \approx \frac{W_{eff} \mu_{eff} C_{oxeff} kT}{q L_{eff}} \cdot V_{dseff} \quad (6.7)$$

which yields the diffusion channel resistance

$$R_{ch,diffusion} \approx \frac{q L_{eff}}{kT \cdot W_{eff} \mu_{eff} C_{oxeff}} \quad (6.8)$$

The resistances given in Eq. (6.4) and Eq. (6.8) determine, in parallel, the channel resistance due to drift and diffusion.  $R_{g,ii}$  is a fraction of the channel resistance.

$$R_{g,ii} = \frac{1}{NF \cdot XRCRG1} \cdot \frac{1}{\frac{1}{R_{ch,Ohmic}} + XRCRG2 \cdot \frac{1}{R_{ch,diffusion}}} \quad (6.9a)$$

If expressed in its reciprocal (i.e., in conductance) for the convenience of SPICE implementation,

$$\begin{aligned} \frac{1}{R_{g,ii}} &= NF \cdot XRCRG1 \cdot \left( \frac{1}{R_{ch,Ohmic}} + XRCRG2 \cdot \frac{1}{R_{ch,diffusion}} \right) \\ &= NF \cdot XRCRG1 \cdot \left( \frac{I_{ch}}{V_{ds,eff}} + XRCRG2 \cdot \frac{kT \cdot W_{eff} \mu_{eff} C_{ox,eff}}{qL_{eff}} \right) \end{aligned} \quad (6.9b)$$

$NF$  is an instance parameter — the number of device fingers. As one expects,  $XRCRG2$  is found to be in the neighborhood of 1 and is introduced to allow the  $R_{g,ii}$  model to fit the  $Y$  parameter better. On the other hand, the model parameter  $XRCRG1$  (also dimensionless) requires more explanations. It is expected to be much greater than 1 and is needed to account for the distributiveness of the drift and diffusion channel resistance because not all the gate capacitance flows through the entire channel.  $XRCRG1$  is theoretically equal to 12 when the channel potential distribution is uniform or when  $V_{ds}$  is around zero, a situation analogous to the gate electrode resistance  $R_{g,electrode}$  formulation Eq. (6.1), where the contacts are made at both ends of the gate electrode.

A comparison of  $R_{g,ii}$  between 2-D TCAD simulations and Eq. (6.9b) for both linear and saturation regions is shown in Fig. 6.6. It demonstrates the good modeling accuracy and the need for the diffusion component to be incorporated in the  $R_{g,ii}$  modeling. This model is further validated with measurement as shown in Fig. 6.7, where the total gate resistance  $R_{g,total} = R_{g,electrode} + R_{g,ii}$  is considered from the linear region through the saturation region. Note that  $R_{g,electrode}$  of Fig. 6.7 is less than 1 ohm as the transistor used in the study is a wide device with ten fingers.

The small-signal input gate resistance  $R_{in}$  is defined by

$$R_{in} = \text{Real}\left(\frac{1}{Y_{11}}\right) \quad (6.9c)$$

where the input  $Y_{11}$  parameter is, in practice, converted from measured scattering  $S$ -parameters for various  $V_{ds}$ ,  $V_{gs}$ , and device geometries.  $R_{in}$  is composed of the gate electrode ( $R_{g,electrode}$ ), source/drain diffusion, and intrinsic-input gate resistances ( $R_{g,ii}$ ). The source and drain diffusion resistances are usually obtained from DC IV extractions. Thus,  $R_{g,ii}$  can now be extracted from  $R_{in}$  either analytically from  $Y_{11}$  or using global parameter extraction optimization over different  $V_{ds}$ . It is found that the  $R_{g,ii}$  formulation is independent of frequencies as expected.

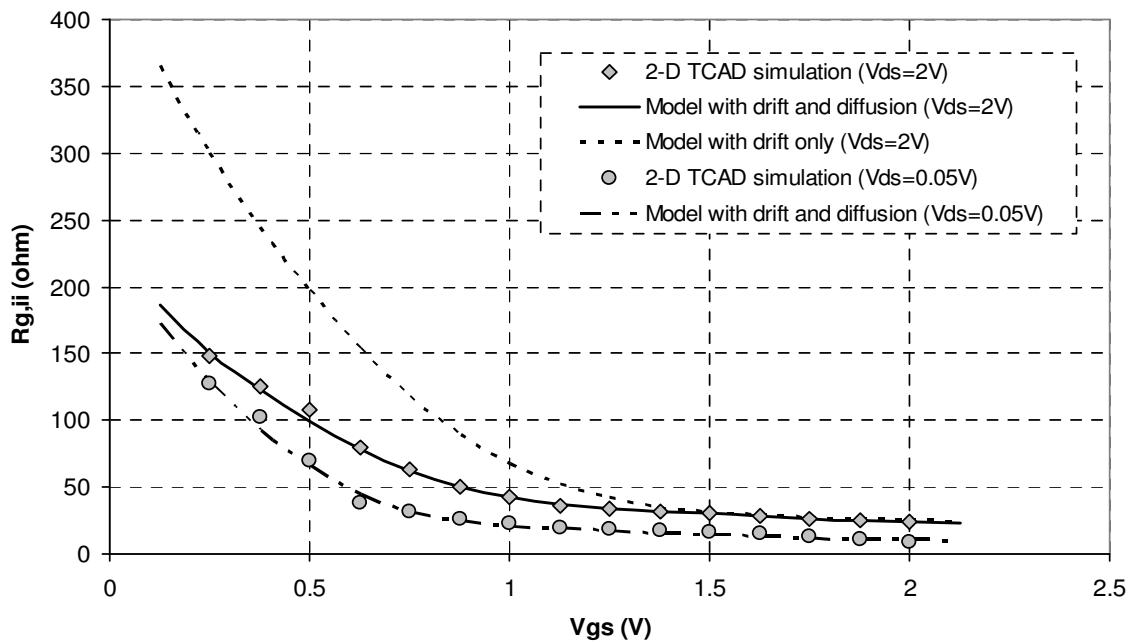


Fig. 6.6 Comparisons of the intrinsic-input gate resistance  $R_{g,ii}$  versus  $V_{gs}$  between the model (Eq. (6.9b)) and two-dimensional TCAD simulations in both linear and saturation regions. Ignoring the diffusion contribution leads to errors when the device operates in the moderate and sub-threshold regions. An NMOS device that is 10  $\mu\text{m}$  wide and 0.5  $\mu\text{m}$  long with a threshold voltage of 0.45 V is used here with  $XRCRG1 = 14$  and  $XRCRG2 = 1$ . The device has a single finger and one gate contact. No poly-silicon gate electrode resistance (i.e., an ideal conductor) was assumed in the TCAD simulations.

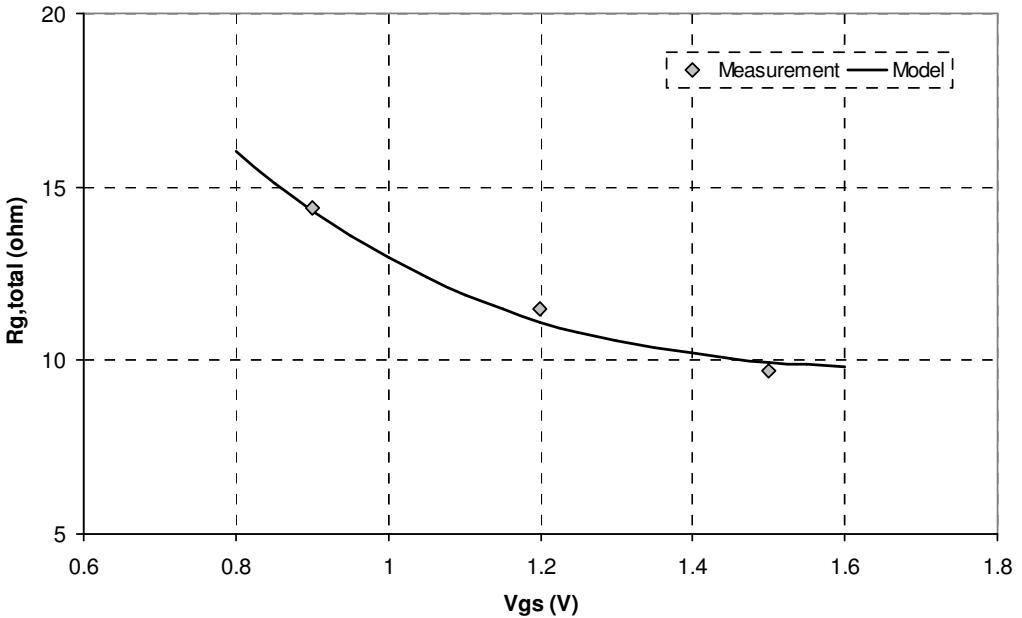


Fig. 6.7 Measured and modeled total gate resistances (the sum of  $R_{g,electrode}$  and  $R_{g,ii}$ ) versus gate biases for  $V_{ds} = 1$  V at a frequency of 2.3 GHz for an NMOSFET transistor of a drawn width of 160  $\mu\text{m}$ , a length of 0.35  $\mu\text{m}$ , NGCON = 1 (one-sided gate contact), and NF = 10. XRCRG1 = 11 and XRCRG2 = 1 were extracted.  $R_{g,electrode}$  is small, less than 1 ohm for this multi-finger device.

One can certainly argue that the channel resistance discussed above is also “visible” from the body node of the device through the depletion layer capacitance and that a portion of this resistance reflected into that body node can presumably be expressed in the form similar to Eq. (6.9a). This is, however, a much weaker effect as the body depletion capacitance is rather small in comparison with the gate oxide capacitance. Therefore, the NQS effects associated with body-channel coupling is usually neglected. However, just as the source and drain and the gate electrode resistances, the parasitic bias-independent body resistance needs to be modeled for RF circuits. This will be discussed later.

BSIM4 provides two connections of the intrinsic-input gate resistance  $R_{g,ii}$  as shown in Fig. 6.8. The *variable-resistance* option (by setting RGATEMOD = 2) considers the lump sum of  $R_{g,electrode}$  (if any) and  $R_{g,ii}$ . It is connected to the intrinsic gate of the transistor via the internal gate node *gNodePrime* (similar to Fig. 6.2 (b)). The *two-node* connection is

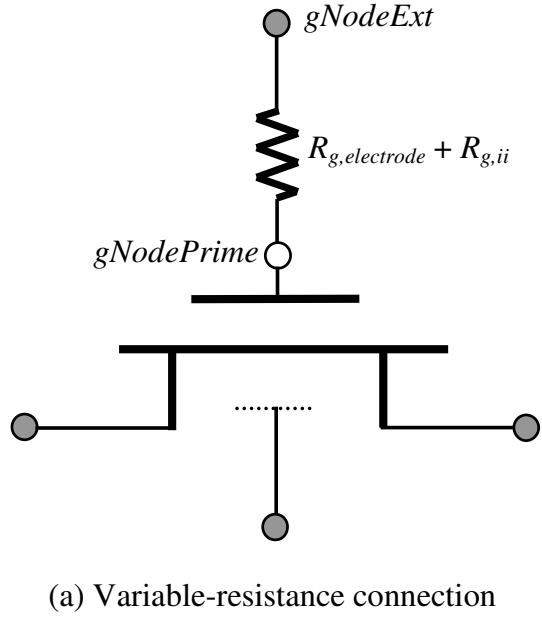
realized by specifying **RGATEMOD** = 3 to connect  $R_{g,electrode}$  and  $R_{g,ii}$  in series via one additional internal mid-gate node (*gNodeMid*). The latter (**RGATEMOD** = 3) of the two connection options considers the fact that the charging currents generated by the gate-source or gate-drain overlap and fringing capacitances flow through the overlap region, not the channel region. This is important for accurate modeling of the input gate impedance.

**RGATEMOD** = 3 is more accurate than **RGATEMOD** = 2 but it introduces one more internal gate node *gNodeMid*. Its accuracy advantage is more noticeable for smaller channel lengths where the extrinsic gate capacitances are comparable to the intrinsic ones. However, **RGATEMOD** = 3 comes with extra CPU runtime overhead in SPICE simulation because of that additional gate node. In either case of **RGATEMOD** = 2 and 3, connecting a highly nonlinear resistance such as  $R_{g,ii}$  in a SPICE model makes circuit matrices denser to solve (because of more matrix non-zero fillings associated with the *gNodePrime* node, which makes the matrix LU decomposition/factorization computationally more expensive). This is a similar situation in which the internal source and drain nodes are created for the accurate modeling of the source and drain diffusion and LDD resistances for advanced process technology nodes.

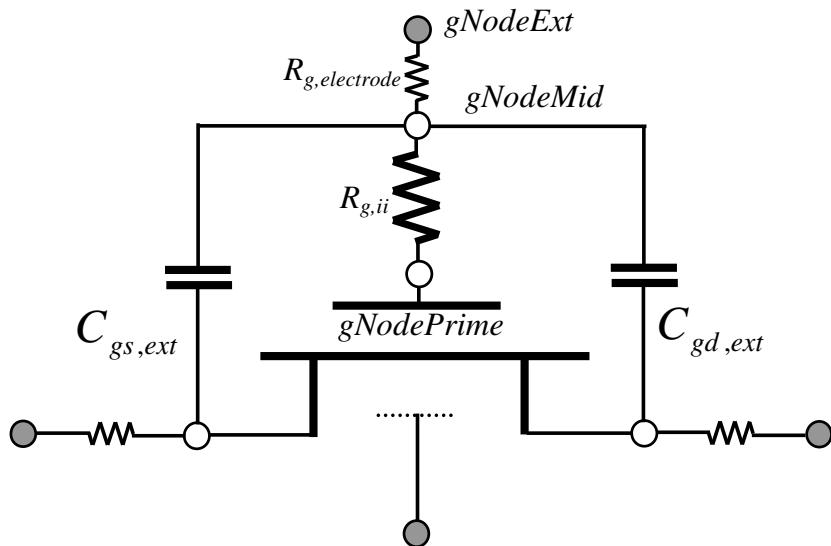
## 6.4 Charge-Deficit Transient and AC NQS Models

### 6.4.1 Charge-Deficit Transient NQS Model

Channel charges have finite mobilities and diffusion constants. When  $V_g$  is suddenly raised, it takes time for additional inversion charges to flow from the source into the channel to establish a new static or quasi-static channel charge distribution. The charge-deficit model provides a first-order model for modeling this transient delay for arbitrary terminal voltage waveforms.



(a) Variable-resistance connection



(b) Two-node connection

Fig. 6.8 Two connections of the intrinsic-input gate resistance  $R_{g,ii}$ : (a) the variable-resistance option which can be turned on by setting **RGATEMOD** = 2 globally in a model card library or for some transistors individually in their element lines. The name “*variable resistance*” is given in contrast to the option **RGATEMOD** = 1, where the same topology is employed but only the constant gate electrode resistance  $R_{g,electrode}$  is incorporated to give model users the maximum flexibility; (b) the two-node option if **RGATEMOD** is set to 3 to permit connecting the extrinsic gate-source and gate-drain overlapping and fringing capacitances to the internal gate node *gNodeMid*. Note that the connections of these extrinsic capacitances to *gNodePrime* are not shown in (a) for simplicity. The terminal charging current for all **RGATEMOD** options is computed by the quasi-static capacitance model from **CAPMOD** = 0, 1, or 2. Note also that the internal and external source nodes will collapse if no source resistance is present; the same holds true for the drain side.

channel charge distribution. The charge-deficit model provides a first-order model for modeling this transient delay for arbitrary terminal voltage waveforms.

While the previous paragraph present an accurate and detailed picture, it is easier and may be sufficient to think of the NQS as basically the charge transit time from source to drain. In both pictures, it is obvious that long-channel devices have more prominent NQS. High-speed digital, analog and RF CMOS ICs require accurate NQS models. BSIM4 models this finite time constant with the intrinsic-input gate resistance  $R_{g,ii}$  model as developed in the previous section.

BSIM4 provides an alternative NQS model, known as the *charge deficit* model. The channel charge deficit  $Q_{def}$  is the difference between the DC (quasi-static) channel charge  $Q_{ch_qs}$  and the actual NQS channel charge  $Q_{ch_nqs}$ .

$$Q_{def} = Q_{ch_qs} - Q_{ch_nqs} \quad (6.10)$$

Rearranging it leads to

$$Q_{ch_nqs} = Q_{ch_qs} - Q_{def} \quad (6.11)$$

where  $Q_{ch_qs}$  represents the DC channel charge model developed in the previous chapters; the subscript  $qs$  stands for quasi static. Taking time derivatives on both sides of Eq. (6.11)

$$\frac{dQ_{ch_nqs}}{dt} = \frac{dQ_{ch_qs}}{dt} - \frac{dQ_{def}}{dt} \quad (6.12)$$

The second term on the right-hand side is approximated by

$$\frac{dQ_{def}}{dt} \approx \frac{Q_{def}}{\tau_{nqs}} \quad (6.13)$$

using the relaxation-time approximation.  $\tau_{nqs}$  is a time constant to be defined shortly. Eq. (6.12) can be represented by an equivalent circuit shown in Fig. 6.9. The current source and the current through the resistor represent the first and second term on the right-hand side of Eq. (6.12), respectively. The difference between these two currents, i.e., the current through the capacitor, is the NQS capacitive current given by Eq. (6.13).

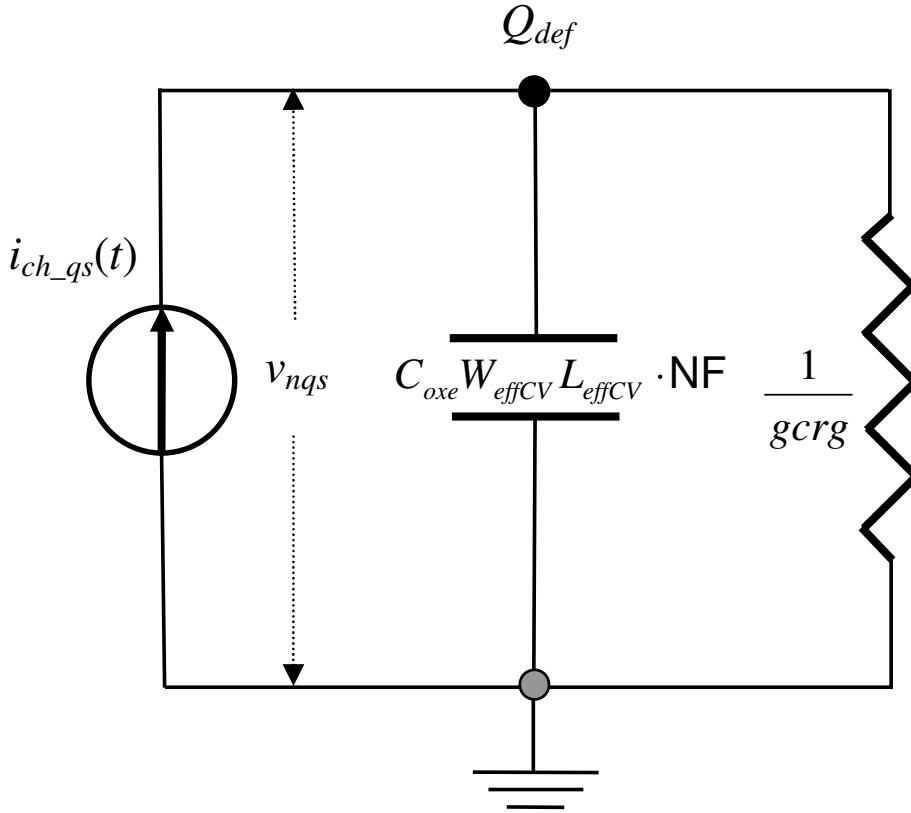


Fig. 6.9 Auxiliary charge-deficit NQS model.  $Q_{def}$ , the deficit of the quasi-static channel charge, is determined by the quasi-static charging current and the effective time constant given by Eq. (6.18).

With Eq. (6.13), it is possible to write the total currents for the source and drain terminals as

$$i_s(t) = I_s(DC) + \frac{dQ_{ch\_nqs,s}}{dt} \approx I_s(DC) + S_{xpart} \cdot \frac{Q_{def}}{\tau_{nqs}} \quad (6.14)$$

and

$$i_d(t) = I_d(DC) + \frac{dQ_{ch\_nqs,d}}{dt} \approx I_d(DC) + D_{xpart} \cdot \frac{Q_{def}}{\tau_{nqs}} \quad (6.15)$$

in which the channel charge  $Q_{ch\_nqs}$  is partitioned into  $Q_{ch\_nqs,s}$  for the source terminal and  $Q_{ch\_nqs,d}$  for the drain terminal. In doing so,  $Q_{ch\_nqs} = Q_{ch\_nqs,s} + Q_{ch\_nqs,d}$  and  $S_{xpart} + D_{xpart} = 1$ .  $S_{xpart}$  and  $D_{xpart}$  are

called the NQS channel charge partitioning coefficients for the source and drain terminals, respectively. They are formulated to be consistent with those of the quasi-static charge-capacitance models presented in Chapter 5. In the forward-mode operation,

$$D_{xpart} = -\frac{Q_d}{Q_g + Q_b} \quad (6.16a)$$

In the reverse mode, the same partition ratio is assigned to

$$S_{xpart} = -\frac{Q_d}{Q_g + Q_b} \quad (6.16b)$$

to observe the source and drain swapping practice. Please note that  $-(Q_g + Q_b)$  represents the channel charge because of the charge neutrality requirements, in which all four terminal (intrinsic) charges must satisfy  $(Q_s + Q_d) = -(Q_g + Q_b)$ . Note also that these intrinsic terminal charges are quasi static and are functions of the four internal terminal nodal voltages. [Please refer to Chapter 5 for details.]

The total gate terminal current, when the charge-deficit transient NQS consideration is applied, can be similarly obtained

$$i_g(t) = I_g(DT) + \frac{dQ_{ch\_nqs,g}}{dt} \approx I_g(DT) + G_{xpart} \cdot \frac{Q_{def}}{\tau_{nqs}} \quad (6.17)$$

in which  $I_g(DT)$  is a symbol for the gate direct-tunneling current contribution as discussed in Chapter 4, and  $G_{xpart}$  signifies the partitioning of the charge-deficit NQS current  $\frac{Q_{def}}{\tau_{nqs}}$  into the gate terminal. The rest,  $(1 - G_{xpart}) \cdot \frac{Q_{def}}{\tau_{nqs}}$ , flows into or from the body terminal. In BSIM4, however,  $G_{xpart} = -(S_{xpart} + D_{xpart}) = -1$  is always chosen, which means that the charge-deficit transient NQS effects on the body terminal are neglected.

The effective time constant determined by the distributed channel resistance and gate oxide capacitance can be written

$$\tau_{nqs} = \frac{C_{oxe} W_{effCV} L_{effCV} \cdot NF}{gcrg} \quad (6.18)$$

where the numerator uses the electrical gate oxide capacitance computed with the electrical gate oxide thickness  $TOXE$  and the effective channel width and length derived for the intrinsic charge-capacitance model. The quantity  $gcrg$  is a conductance, the inverse of the gate resistance, defined by

$$\frac{1}{gcrg} = R_{g,electrode} + R_{g,ii} \quad (6.18a)$$

when  $RGATEMOD = 2$  as shown in Fig. 6.8 (a) and

$$\frac{1}{gcrg} = R_{g,ii} \quad (6.18b)$$

if  $RGATEMOD = 3$  as illustrated in Fig. 6.8 (b).

The charge-deficit state variable  $Q_{def}$  in Eqs. (6.14), (6.15) and (6.17) is obtained through a Newton-Raphson iterative solution process. This is the same as solving for other device and circuit nodal voltages. The iterative approach requires an auxiliary circuit that has an internal charge node that provides the numeric values for  $Q_{def}$ . This circuit was shown in Fig. 6.9. Its RC time constant is given by Eq. (6.18).  $Q_{def}$  is

$$Q_{def} = (C_{oxe} W_{effCV} L_{effCV} \cdot NF) \cdot v_{nqs} \quad (6.19)$$

where  $v_{nqs}$  is the NQS model internal node voltage. The channel charging current in Fig. 6.9 is expressed by

$$i_{ch\_qs}(t) = \frac{d(Q_s + Q_d)}{dt} = -\frac{d(Q_g + Q_b)}{dt} \quad (6.20)$$

which is known from the intrinsic quasi-static charge-capacitance model ( $CAPMOD = 0, 1$ , or  $2$ ) and serves as the stimulus to the auxiliary NQS sub-circuit.

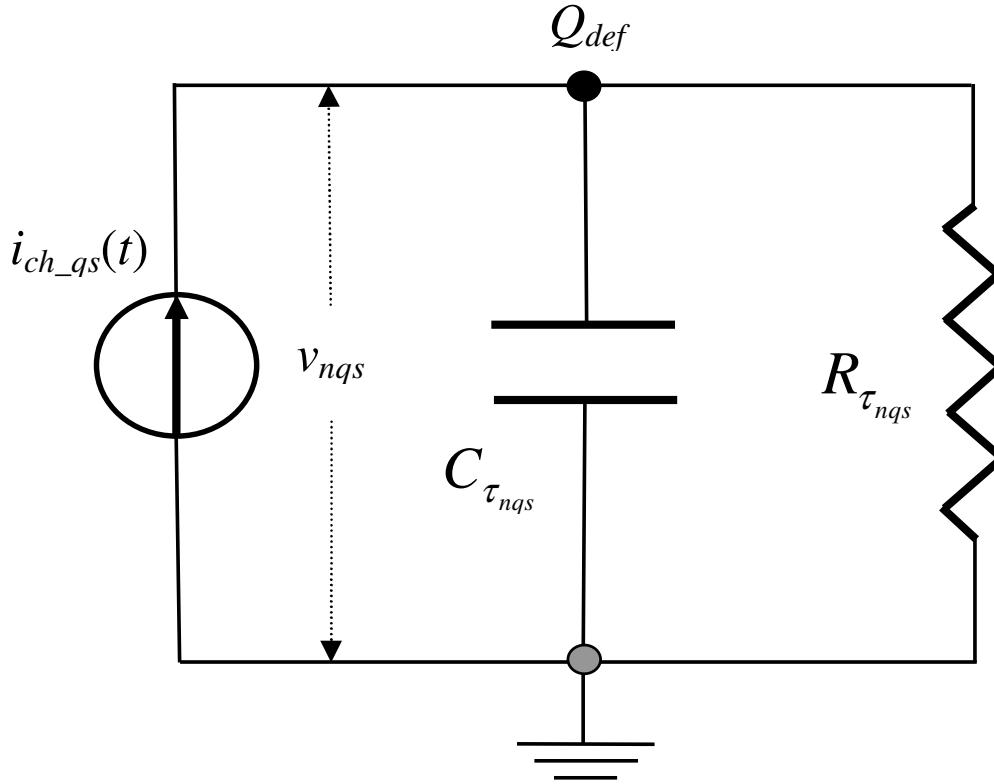


Fig. 6.10 A more generic implementation of the charge-deficit NQS auxiliary sub-circuit of Fig. 6.9.

A more generic form of the auxiliary sub-circuit of Fig. 6.9 is shown in Fig. 6.10, where  $C_{\tau_{nqs}}$  is a constant capacitance of arbitrary values and  $R_{\tau_{nqs}}$  is a non-linear resistance such that their product, an RC time constant, is equal to  $\tau_{nqs}$  of Eq. (6.18); i.e.,

$$R_{\tau_{nqs}} \cdot C_{\tau_{nqs}} \equiv \tau_{nqs} \quad (6.20a)$$

which then gives

$$R_{\tau_{nqs}} = \frac{\tau_{nqs}}{C_{\tau_{nqs}}} = \frac{1}{C_{\tau_{nqs}}} \cdot \frac{C_{oxe} W_{effCV} L_{effCV} \cdot NF}{gcrg} \quad (6.20b)$$

It is obvious that by choosing  $C_{\tau_{nqs}} = C_{oxe} W_{effCV} L_{effCV} \cdot NF$ , Fig. 6.10 reduces to Fig. 6.9. In the following, it will be demonstrated that the NQS terminal charging currents that are proportional to  $\frac{Q_{def}}{\tau_{nqs}}$  are largely independent of the choice of  $C_{\tau_{nqs}}$ . Referring to Fig. 6.10, one can show

$$\begin{aligned}
\frac{Q_{def}}{\tau_{nqs}} &= \frac{v_{nqs} \cdot C_{\tau_{nqs}}}{\tau_{nqs}} = \frac{C_{\tau_{nqs}}}{\tau_{nqs}} \cdot i_{ch\_qs}(t) \cdot \left( \frac{1}{\frac{C_{\tau_{nqs}}}{\Delta t} + \frac{1}{R_{\tau_{nqs}}}} \right) \\
&= \frac{i_{ch\_qs}(t)}{\tau_{nqs}} \cdot \left( \frac{1}{\frac{1}{\Delta t} + \frac{gcrg}{C_{oxe} W_{effCV} L_{effCV} \cdot NF}} \right)
\end{aligned} \tag{6.20c}$$

$C_{\tau_{nqs}}$  cancels out indeed.  $\Delta t$  is the time step of Newton iteration. In fact, it has been found that a choice of  $C_{\tau_{nqs}} \equiv 1 \times 10^{-9}$  Farads can speed up simulation and retain good modeling accuracy. It has also been observed that too large a  $C_{\tau_{nqs}}$  ( $C_{\tau_{nqs}} > 1 \times 10^{-6}$  Farads) can lead to poor accuracy and long CPU runtime. When  $C_{\tau_{nqs}}$  is smaller than  $1 \times 10^{-10}$  Farads, convergence difficulties can result. For this reason, a hard coded  $C_{\tau_{nqs}} \equiv 1 \times 10^{-9}$  is chosen in BSIM4. Table 6.1 is given to demonstrate the impacts of  $C_{\tau_{nqs}}$  on SPICE simulation overhead and modeling accuracy. These observations are useful for SPICE modeling and implementation of similar RC network such as the auxiliary network of the SOI-MOSFET thermal resistance and capacitance for modeling self-heating effects.

The charging current components of the source, drain, and gate terminals that result from the charge-deficit transient NQS model can be developed from Fig. 6.10. They are

$$i_s(t) = S_{xpart} \cdot \frac{v_{nqs} \cdot C_{\tau_{nqs}}}{C_{oxe} W_{effCV} L_{effCV} \cdot NF} \cdot gcrg \tag{6.20d}$$

for the source terminal. Similarly

$$i_d(t) = D_{xpart} \cdot \frac{v_{nqs} \cdot C_{\tau_{nqs}}}{C_{oxe} W_{effCV} L_{effCV} \cdot NF} \cdot gcrg \quad (6.20e)$$

for the drain terminal and

$$\begin{aligned} i_g(t) &= G_{xpart} \cdot \frac{v_{nqs} \cdot C_{\tau_{nqs}}}{\tau_{nqs}} \\ &= G_{xpart} \cdot \frac{v_{nqs} \cdot C_{\tau_{nqs}}}{C_{oxe} W_{effCV} L_{effCV} \cdot NF} \cdot gcrg \end{aligned} \quad (6.20f)$$

for the gate terminal.

Table 6.1 The impact of  $C_{\tau_{nqs}}$  on CPU and iteration performance and model accuracy: A 17-stage RO (ring oscillator) with TRNQSMOD = 1. When  $C_{\tau_{nqs}}$  is greater than  $10^{-7}$  or less than  $10^{-9}$  Farads, the simulations will either abort or have poor accuracy.

$C_{\tau_{nqs}}$ (Farads)	CPU runtime (s)	Total iteration count	RO period (ns)
1e-7	1158.46	2671751	0.98
1e-8	60.78	151252	1.07
1e-9	23.14	57099	1.07

A comparison of the transient drain current between the TRNQSMOD = 1 model and measurement is plotted in Fig. 6.11 by applying a rapid ramping gate input voltage to a large MOSFET. The typical non-quasi-static behavior is clearly seen: No drain terminal current exists at the very beginning of the  $V_g$  ramping until after a finite lapse of time for both  $V_{dd} = 1.5$  and 3V. Furthermore, in the case of  $V_{dd} = 1.5$ V, the drain current needs to take additional time to be able to settle down to its steady state value even after the 1 nano-second time point.

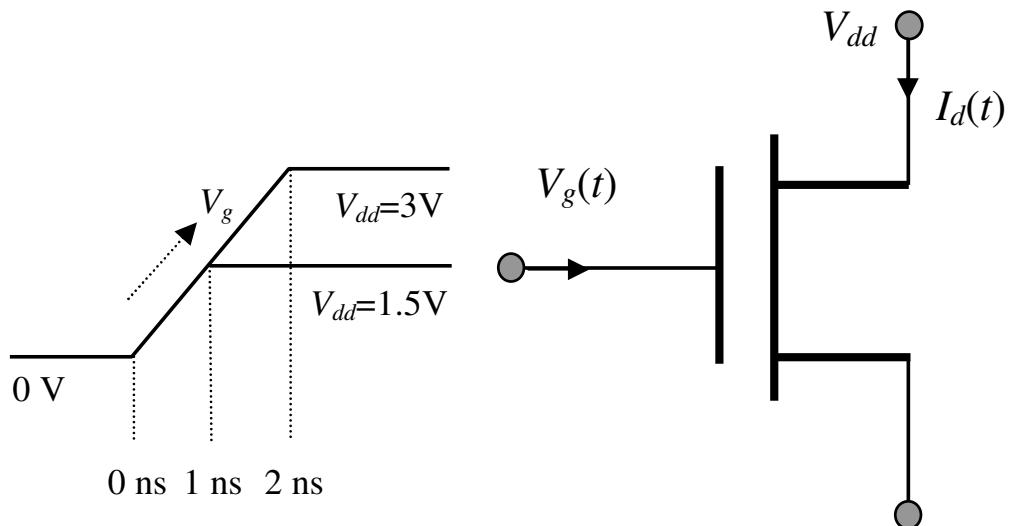
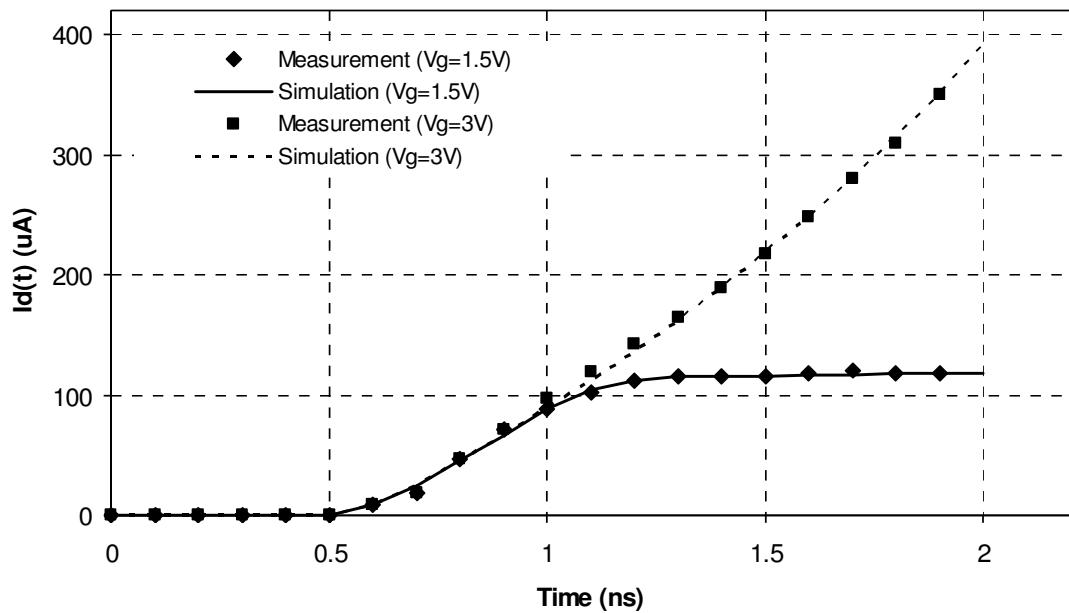


Fig. 6.11 Drain current as a function of time when TRNQSMOD is turned on; a rapid ramp voltage is applied to the gate. A large transistor with a channel length of  $3\mu\text{m}$  is employed to produce non-quasi-static behavior for easy observation. A 0/100 drain and source charge partition ( $\text{XPART} = 1$ ) is used in conjunction with TRNQSMOD.

It is worthwhile to note that approximating the NQS source, drain, and gate terminal charging currents with  $\frac{Q_{def}}{\tau_{nqs}}$  as made in Eqs. (6.14), (6.15) and (6.17) is mathematically unnecessary. For example, one can rewrite Eq. (6.14) in a more exact form

$$i_s(t) = I_s(DC) + \frac{dQ_{ch\_nqs,s}}{dt} \equiv I_s(DC) + S_{xpart} \cdot \left( \frac{dQ_{ch\_qs}}{dt} - \frac{dQ_{def}}{dt} \right) \quad (6.21)$$

where  $Q_{ch\_qs}$  and its time derivative is known from the quasi-static charge capacitance model.  $Q_{def}$  can still be solved for from the method demonstrated in Fig. 6.10 and its time derivative  $\frac{dQ_{def}}{dt}$  can be computed

numerically as  $\frac{\Delta Q_{def}}{\Delta t} \equiv \frac{\Delta(v_{nqs} \cdot C_{\tau_{nqs}})}{\Delta t}$  where  $C_{\tau_{nqs}}$  is constant and  $\Delta v_{nqs}$  is the  $v_{nqs}$  difference between the current and past time points. Here, the computation process is analogous to that for  $\frac{dQ_{ch\_qs}}{dt}$ . Eq. (6.21) has

better accuracy than the  $\frac{Q_{def}}{\tau_{nqs}}$  approximation in many cases.

### 6.4.2 Charge-Deficit AC NQS Model

It is known that the small-signal AC analysis with SPICE requires the circuit operating-point (OP) information such as the conductances and capacitances of the circuit elements that are obtained under the OP bias conditions. For this reason, the charge-deficit transient NQS auxiliary circuit model that involves time-domain iteration of  $Q_{def}$  is not applicable in the frequency domain. However, it is possible to transform the transient model into its AC version by performing a few mathematical manipulations. Repeat Eq. (6.13) for convenience as

$$\frac{dQ_{ch\_nqs}}{dt} \approx \frac{Q_{def}}{\tau_{nqs}} \quad (6.22)$$

Substituting  $Q_{def} = Q_{ch\_qs} - Q_{ch\_nqs}$  of Eq. (6.10) into the right side of Eq. (6.22), one obtains [2]

$$\frac{dQ_{ch\_nqs}}{dt} \approx \frac{Q_{ch\_qs} - Q_{ch\_nqs}}{\tau_{nqs}} \quad (6.23)$$

which, in the frequency domain, becomes

$$Q_{ch\_nqs} \approx \frac{Q_{ch\_qs}}{1 + j\omega\tau_{nqs}} \quad (6.24)$$

with the time-derivative operator  $\frac{d}{dt}$  replaced by the complex angular frequency  $j\omega$ . Applying Eq. (6.24) to the channel current model yields

$$I_{ch\_nqs} \approx \frac{I_{ch}}{1 + j\omega\tau_{nqs}} \quad (6.25)$$

which further leads to the following channel trans-conductances

$$G_{ch\_nqs\_j} = \left[ \frac{G_{ch\_j}}{1 + (\omega\tau_{nqs})^2} - \frac{2\omega^2\tau_{nqs} \cdot I_{ch}}{(1 + \omega^2\tau_{nqs}^2)^2} \cdot \frac{d\tau_{nqs}}{dV_j} \right] \\ - j \cdot \left[ \frac{G_{ch\_j}\omega\tau_{nqs}}{1 + (\omega\tau_{nqs})^2} + \frac{\omega \cdot (1 - \omega^2\tau_{nqs}^2) \cdot I_{ch}}{(1 + \omega^2\tau_{nqs}^2)^2} \cdot \frac{d\tau_{nqs}}{dV_j} \right] \quad (6.26)$$

where the subscript  $j$  of  $V_j$  denotes the device terminal indices; that is,  $j = d, g, s$ , or  $b$ . For example, Eq. (6.26) gives the NQS gate transconductance  $G_{m\_nqs}$  when the derivatives are computed with respect to  $V_g$ . Analogously, the trans-capacitances of the intrinsic terminal charges are

$$C_{nqs\_i,j} = \left[ \frac{C_{i,j}}{1 + (\omega\tau_{nqs})^2} - \frac{2\omega^2\tau_{nqs} \cdot Q_i}{(1 + \omega^2\tau_{nqs}^2)^2} \cdot \frac{d\tau_{nqs}}{dV_j} \right] \\ - j \cdot \left[ \frac{C_{i,j}\omega\tau_{nqs}}{1 + (\omega\tau_{nqs})^2} + \frac{\omega \cdot (1 - \omega^2\tau_{nqs}^2) \cdot Q_i}{(1 + \omega^2\tau_{nqs}^2)^2} \cdot \frac{d\tau_{nqs}}{dV_j} \right] \quad (6.27)$$

Eq. (6.27) computes the total gate capacitance  $C_{gg\_nqs}$  under NQS if the subscripts  $i$  and  $j$  both refer to the gate node.

Equations (6.26) and (6.27) provide useful insights. They state that under the NQS conditions, the channel conductances and terminal capacitances become complex quantities that can be split and stamped into the real and imaginary parts of an AC circuit matrix. It is found that the second terms in the brackets on the right side of Eqs. (6.26) and (6.27) become significant only when the operating frequency  $\omega$  is much greater than the cut-off frequency  $f_T$ ; hence, they can be ignored for simplicity. The terms of Eqs. (6.26) and (6.27) that have been implemented in the BSIM4 charge-deficit AC NQS model are

$$G_{ch\_nqs\_j} = \frac{G_{ch\_j}}{1 + (\omega\tau_{nqs})^2} - j \cdot \frac{G_{ch\_j}\omega\tau_{nqs}}{1 + (\omega\tau_{nqs})^2} \quad (6.28)$$

and

$$C_{nqs\_i,j} = \frac{C_{i,j}}{1 + (\omega\tau_{nqs})^2} - j \cdot \frac{C_{i,j}\omega\tau_{nqs}}{1 + (\omega\tau_{nqs})^2} \quad (6.29)$$

The charge-deficit transient and AC NQS model can be turned on by setting parameters **TRNQSMOD** = 1 and **ACNQSMOD** = 1, respectively. These parameters are both instance and global model parameters with their default values being zero. Turning on either or both of them while setting **RGATEMOD** = 2 or 3 would be a mistake — it would take the NQS effects into account twice and hence result in significant modeling errors.

Applications of the gate resistance and charge-deficit NQS models require extractions of the model parameters RSHG, XGW, XGL, XRCRG1 and XRCRG2 from either transistor network  $Y$  parameters or from fast-transient inverter voltage waveforms or ring oscillator frequencies. The rest of the model parameters can be obtained regularly from typical DC and capacitance extractions.

## 6.5 Body Resistance Network

The reason that a MOSFET body resistance network is needed for SPICE modeling and simulation is obvious. For high-speed digital/analog or RF IC operations, the MOSFET parasitic body resistance becomes comparable to the reactance of the source/drain-body junction capacitances  $\frac{1}{j\omega C_j}$  (where  $C_j$  denotes the junction capacitance without

distinguishing the bottom-area, and isolation sidewall and gate-edge peripheral components). For RF IC designs that use 180nm and 130nm process technologies, the general practice with BSIM3v3 has been a sub-circuit macro model wrapper built by disabling and replacing the BSIM3v3 built-in junction diodes with external ones in series with a customized body resistance model [3], [4]. This is illustrated in Fig. 6.12. This approach works but it is very inconvenient as it requires additional efforts in the development of models and parameter extraction strategies or even modifications to foundry model libraries.

This section is devoted to the BSIM4 distributed body-resistance network model, its applications, and the parameter extraction methodology. This model distinguishes itself from those published previously in the following aspects. It is integrated inside BSIM4 to eliminate the need for the sub-circuit approach that was just discussed. It considers the geometry and layout dependencies. The impact of a multi-finger device configuration, such as the narrow-width, parasitic terminal resistance, and junction leakage and CV effects, are all taken into account in the formulation. This model is the result of a successful collective effort by the BSIM team and the industry. It is portable to future technologies and non-bulk MOSFET device structures as well.

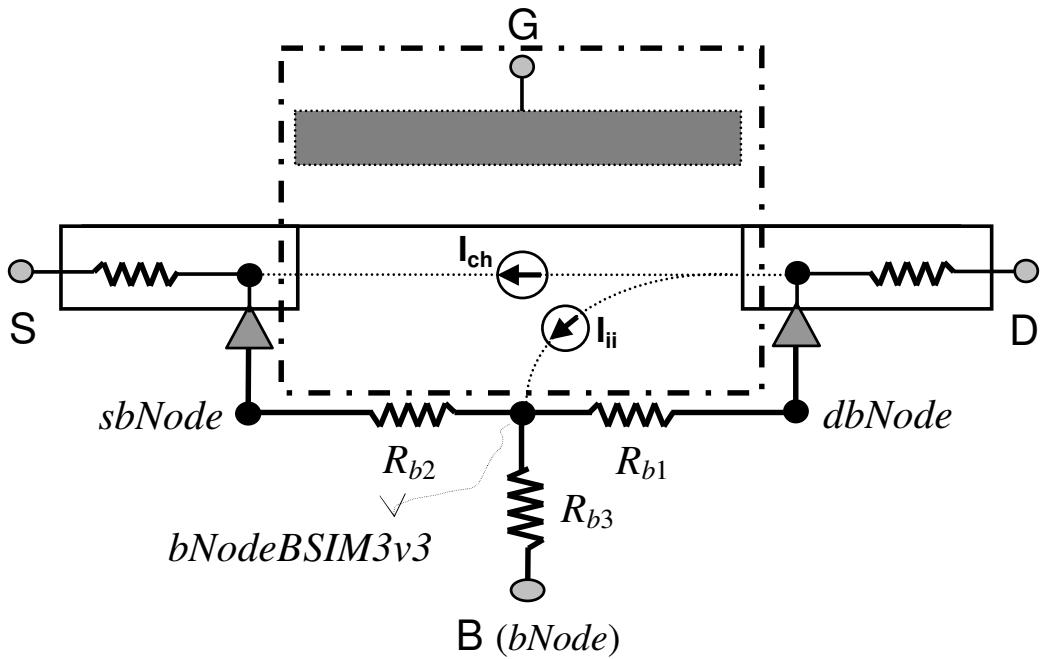


Fig. 6.12 A typical customized BSIM3v3 body resistance network topology. The BSIM3v3 model, as enclosed inside the dash-dotted line, has its built-in junction diode models disabled and replaced with external ones. The *bNodeBSIM3v3* node is the body node of the BSIM3v3 model whereas the *sbNode*, *dbNode*, and *bNode* are the new external body nodes for this customized sub-circuit RF model. User-defined geometrical-dependence expressions for  $R_{b1}$ ,  $R_{b2}$ , and  $R_{b3}$  can be specified. This is often implemented in the form of a SPICE sub-circuit macro model wrapper.

The BSIM4 body-resistance network model provides several options. The model selector **RBODYMOD** is both a local instance and a global model parameter. It has three numeric optional settings. When **RBODYMOD** is set to zero, no body-resistance network will be generated; when set to 1, a non-scalable resistance network is constructed; and when set to 2, the resistances of the network are set up with the geometrical dependencies on the channel length and width and the number of fingers **NF**. When the network is turned on (namely, **RBODYMOD** = 1 or 2), three additional internal body nodes *bNodePrime*, *sbNode* and *dbNode* will be generated in addition to the external body node *bNode*; this is illustrated in Fig. 6.13. A more optimized code implementation can be made to merge some of these nodes, when permissible, to improve simulation efficiency. For instance, *bNodePrime* and *dbNode* can be merged into one if their connecting conductance  $G_{rbpd}$  is large enough, say, great than  $10^3$  mho. In the

following, the BSIM4 body-resistance model and code implementation for RBODYMOD = 1 or 2 are presented.

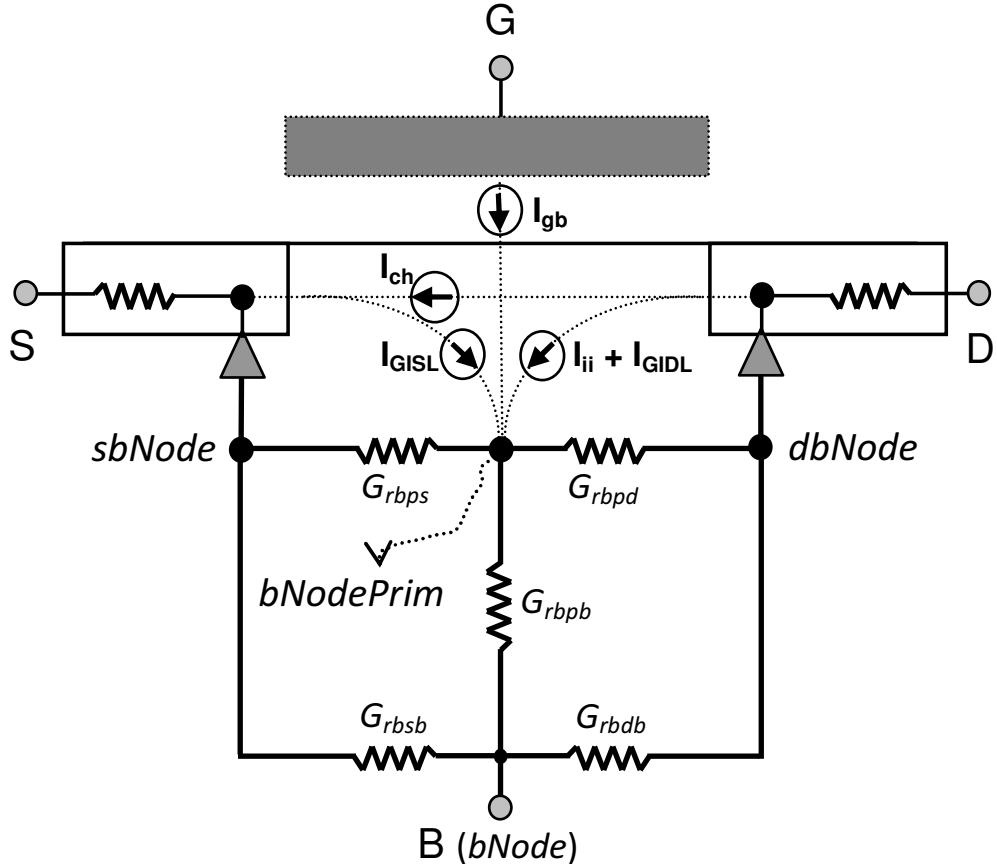


Fig. 6.13 The BSIM4 body-resistance network topology when RBODYMOD = 1 or 2, where the internal body node *bNodePrim* serves as the body reference for the BSIM4 intrinsic model. The gate-body direct tunneling current, and the impact ionization and gate-induced drain/source leakage (GIDL or GISL on the drain and source side, respectively) currents all originate from or are collected by this node. When RBODYMOD = 2, this five-resistance network may be simplified to a 3-R or 1-R network as discussed in the text. Note that the body-resistance network is connected via the junction diodes to the internal source and drain nodes when these nodes are present.

### 6.5.1 RBODYMOD = 1: A Local Network

When RBODYMOD is equal to 1, the five conductances are computed directly from the resistance parameters with no geometrical dependencies considered:

$$G_{rpd} = \text{GBMIN} + \frac{1}{\text{RBPD}} \quad (6.30)$$

where **GBMIN** is a global model parameter with a default of  $10^{-12}$  mho and **RBPD** is both an instance and model parameter having a default of 50 ohm. To prevent potential convergence difficulties from too small an **RBPD**,  $G_{rbpd}$  will be set to  $10^3$  mho if **RBPD** is less than  $10^{-3}$  ohm, irrespective of the value of **GBMIN**. Identical settings and treatments are applied for the other network resistances in the case of **RBODYMOD** = 1.  $G_{rbps}$  is thus written

$$G_{rbps} = \text{GBMIN} + \frac{1}{\text{RBPS}} \quad (6.31)$$

The same is done for  $G_{rbpb}$

$$G_{rbpb} = \text{GBMIN} + \frac{1}{\text{RBPB}} \quad (6.32)$$

Similarly, the conductances  $G_{rbdb}$  and  $G_{rbsb}$  of Fig. 6.13 are computed by

$$G_{rbdb} = \text{GBMIN} + \frac{1}{\text{RBDB}} \quad (6.33)$$

and

$$G_{rbsb} = \text{GBMIN} + \frac{1}{\text{RBSB}} \quad (6.34)$$

respectively. These five conductances are stamped into a circuit matrix to establish the body-resistance network **RBODYMOD** = 1.

One shortcoming of this **RBODYMOD** = 1 model is the efforts needed to generate a model set for every transistors that have different layouts including different channel length, width, the number of fingers, or even transistor orientations. It is thus often inevitable for one to revert to the sub-circuit methodology in which the core intrinsic BSIM4 model is wrapped with possibly many and lengthy mathematical expressions for the body-resistance dependencies on geometries and layouts. This still looms as a cumbersome strategy as was the case with BSIM3v3, although the direct passing of the parameterized expressions into the BSIM4 core via such instance parameters as **RBPD**, **RBPS**, **RBPB**, **RBSB**, and **RBDB** makes the sub-circuit modeling task easier as exemplified in Fig. 6.14.

```

.Subckt BSIM4_Rbody_macro D G S B W = 10U L = 0.05U
+ NF = 20 RBPD = 55 RBPS = 45 RBPB = 120 RBSB = 30
+ RBDB = 35
+ ... $ Other sub-circuit instance parameters specified here.

.Parameter ... $ Parameterized expressions are defined from here
.Parameter RBPD_eval = function_1(W, L, NF)
.Parameter RBPS_eval = function_2(W, L, NF)
.Parameter RBPB_eval = function_3(W, L, NF)
.Parameter RBSB_eval = function_4(W, L, NF)
.Parameter RBDB_eval = function_5(W, L, NF)
. ....
Minstance D G S B Model_BSIM4 W = 10U L = 0.05U
+ NF = 20RBPD = RBPD_eval RBPS = RBPS_eval
+ RBPB = RBPB_eval RBSB = RBSB_eval
+ RBDB = RBDB_eval ...
.Model Model_BSIM4 NMOS LEVEL = 14 VERSION = 4.4
+ TOXE = 1.5e-9 VTH0 = 0.21 RBODYMOD = 1
+ ... $ The BSIM4 model parameter cards

.Ends $ The end of the sub-circuit BSIM4_Rbody_macro

```

Fig. 6.14 A sub-circuit macro model definition block that computes the BSIM4 RBODYMOD = 1 body-resistance parameters using user-defined expressions with geometry and layout dependencies. The computed body-resistance parameter values are then passed along to the BSIM4 body-resistance instance parameters RBPD, RBPS, RBPB, RBSB, and RBDB that are specified in the BSIM4 transistor element lines. These values are to be taken into the BSIM4 model equations and circuit matrix via Eq. (6.30) through Eq. (6.34). No BSIM4 model topology changes to the junction diodes and internal body nodes are needed with this approach.

### 6.5.2 RBODYMOD = 2: A Scalable Network

The body resistance is three dimensional in nature. A scalable body resistance network model deals not only with the geometrical dependencies on the channel length and width and the number of fingers,

but also with body-contact layouts, in orientation and proximity with respect to channel regions. Having a simple and efficient parameter extraction methodology is another requirement for such a model. The BSIM4 RBODYMOD = 2 model that was first introduced into BSIM4.5.0 is designed to take these factors into account. It provides three options to permit a 5-R, 3-R, or 1-R body resistance connection with 3-R being the most practical in terms of both accuracy and ease of use.

### 6.5.2.1 The 5-R Model

In this case, the model topology shown in Fig. 6.13 is employed. The resistances in the denominators of Eq. (6.30) through Eq. (6.34) take on their numerical values from geometry and layout-dependence expressions. To illustrate this point, change Eq. (6.30) to

$$G_{rbpd} = \text{GBMIN} + \frac{1}{rbpd} \quad (6.35)$$

where the quantity  $rbpd$  (not a parameter of the model) is

$$rbpd = \text{RBPDO} \cdot \left( \frac{L_{eff}}{10^{-6}} \right)^{\text{RBPDL}} \cdot \left( \frac{W_{eff}}{10^{-6}} \right)^{\text{RBDW}} \cdot \text{NF}^{\text{RBDNF}} \quad (6.35a)$$

The factor  $10^{-6}$  is introduced to scale the dependencies on  $L_{eff}$  and  $W_{eff}$  that are given in meters. The global model parameter RBPDO, with a default of 50 ohm, denotes the horizontal y-direction resistance from the internal body reference node (*bNodePrime*) to the drain-body junction edge when  $L_{eff} = W_{eff} = 10^{-6}$  meter and NF = 1 (namely, for a single-finger transistor). Heuristically, it is understood that  $rbpd$  is approximately proportional to  $L_{eff}$  and NF and to the inverse of  $W_{eff}$ ; hence, three global model parameters as exponents, RBPDL, RBDW, and

and RBPDNF, are introduced to fit the geometrical dependencies that could deviate from the above approximate relationships. To prevent potential convergence difficulties due to too small an  $r_{bpd}$ ,  $G_{r_{bpd}}$  will be set to  $10^3$  mho if  $r_{bpd}$  is less than  $10^{-3}$  ohm, which is again irrespective of the value of GDMIN.

The same implementations and analyses above are made equally applicable to  $G_{rbps}$  except that RBPS0 measures in the  $y$  direction to the source-body junction edge for a single-finger transistor.

$$G_{rbps} = \text{GDMIN} + \frac{1}{rbps} \quad (6.36)$$

with

$$rbps = \text{RBPS0} \cdot \left( \frac{L_{eff}}{10^{-6}} \right)^{\text{RBPSL}} \cdot \left( \frac{W_{eff}}{10^{-6}} \right)^{\text{RBPSW}} \cdot \text{NF}^{\text{RBPSNF}} \quad (6.36a)$$

Modeling  $G_{rbpb}$  of Fig. 6.13 demands more efforts. The complexity arises from the possibility of the co-existence of both the horizontal (in the  $y$  or channel length direction) and vertical (in the  $x$  or channel width direction) body contacts of Fig. 6.15. A compact expression in the form of

$$G_{rbpb} = \text{GDMIN} + \frac{1}{rbpb} \quad (6.37)$$

still works with  $r_{rbpb}$  consisting of two resistance components in parallel;  $r_{rbpby}$ , a symbol for the resistance component connected to the horizontal contacts, is found to be

$$r_{rbpby} = \text{RBPBY0} \cdot \left( \frac{L_{eff}}{10^{-6}} \right)^{\text{RBPBYL}} \cdot \left( \frac{W_{eff}}{10^{-6}} \right)^{\text{RBPBYW}} \cdot \text{NF}^{\text{RBPBYNF}} \quad (6.37a)$$

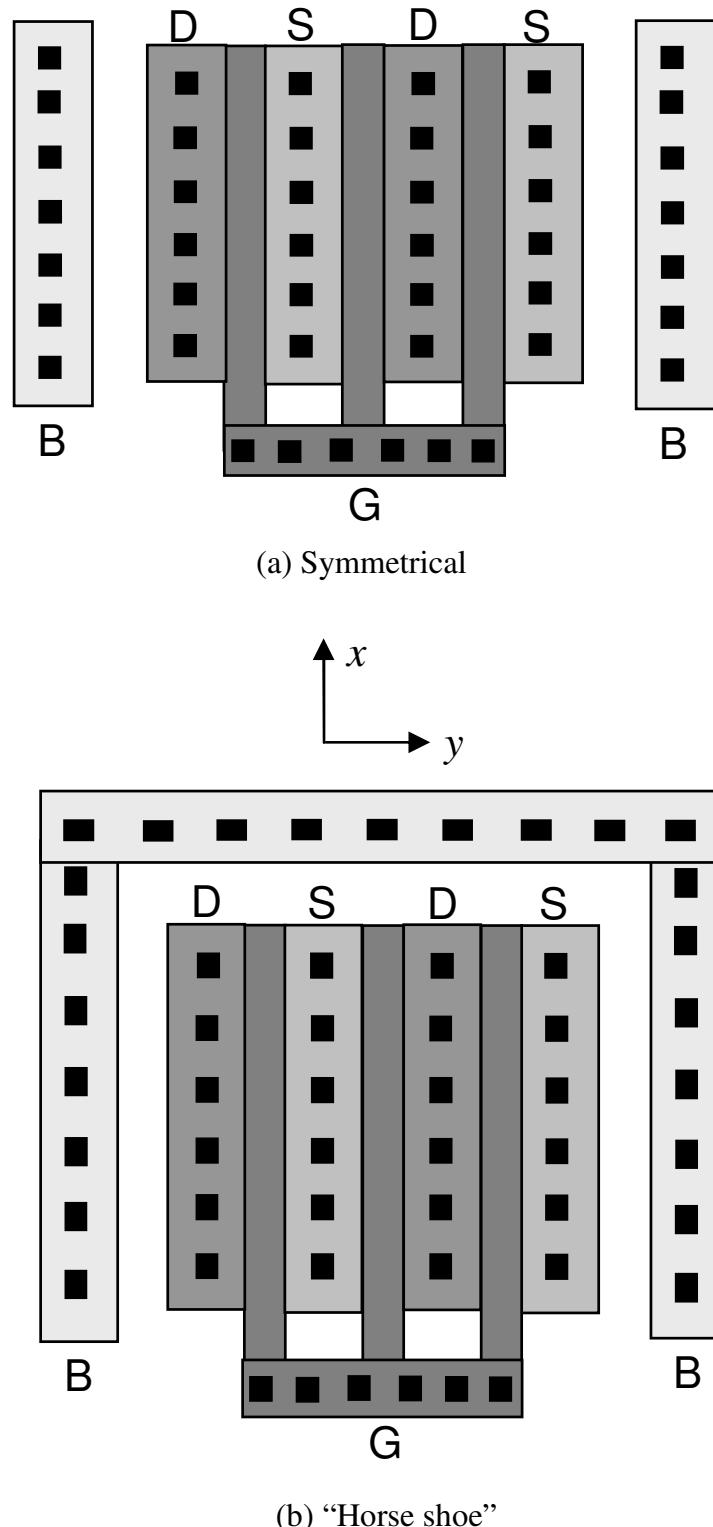


Fig. 6.15 Schematic top-view illustration of the symmetrical (a) and “horse-shoe” (b) substrate contact layouts of a multi-finger MOSFET transistor typically found in an RF CMOS IC design. Alternating source and drain between gate fingers is a common practice. One-sided gate contact NGCON = 1 is assumed.

which, at the first order, is proportional to  $W_{eff}$  and inversely to  $L_{eff}$  with a weak dependence on NF. New parameters in Eq. (6.37a) retain similar interpretations given above. More details will be presented in a parameter table shortly. The other component  $rpbpx$  that accounts for the channel length and width dependencies of the resistance to the vertical running contacts on the source/drain side is expressed in an analogous form

$$rpbpx = RBPBX0 \cdot \left( \frac{L_{eff}}{10^{-6}} \right)^{RBPBXL} \cdot \left( \frac{W_{eff}}{10^{-6}} \right)^{RBPBXW} \cdot NF^{RBPBXNF} \quad (6.37b)$$

The resistances of Eqs. (6.37a) and (6.37b) are connected in parallel to produce a combined equivalent,  $rpbp$  of Eq. (6.37):

$$rpbp = \frac{rpbpx \cdot rpbpy}{rpbpx + rpbpy} \quad (6.37c)$$

Similarly,  $G_{rbpb}$  of Eq. (6.37) will be set to  $10^3$  mho if  $rpbp$  is less than  $10^{-3}$  ohm, regardless of GBMIN.

The same implementations and considerations as those for  $G_{rbpb}$  apply to  $G_{rbdb}$  and  $G_{rbsb}$  of Fig. 6.13. Consider first the case of  $G_{rbdb}$ ; it is given by

$$G_{rbdb} = GBMIN + \frac{1}{rbdb} \quad (6.38)$$

with

$$rbdb = \frac{rbdbx \cdot rbdby}{rbdbx + rbdby} \quad (6.38a)$$

Moreover, the  $x$  and  $y$  components are formulated in a familiar form

$$rdbbx = RBDBX0 \cdot \left( \frac{L_{eff}}{10^{-6}} \right)^{RBSDBXL} \cdot \left( \frac{W_{eff}}{10^{-6}} \right)^{RBSDBXW} \cdot NF^{RBSDBXNF} \quad (6.38b)$$

and

$$rdbby = RBDBY0 \cdot \left( \frac{L_{eff}}{10^{-6}} \right)^{RBSDBYL} \cdot \left( \frac{W_{eff}}{10^{-6}} \right)^{RBSDBYW} \cdot NF^{RBSDBYNF} \quad (6.38c)$$

Assume that  $G_{rbsb}$  has similar parametric dependencies as  $G_{rbdb}$ ,

$$G_{rbsb} = GBMIN + \frac{1}{rbsb} \quad (6.39)$$

and

$$rbsb = \frac{rbsbx \cdot rbsby}{rbsbx + rbsby} \quad (6.39a)$$

one can write

$$rbsbx = RBSBX0 \cdot \left( \frac{L_{eff}}{10^{-6}} \right)^{RBSDBXL} \cdot \left( \frac{W_{eff}}{10^{-6}} \right)^{RBSDBXW} \cdot NF^{RBSDBXNF} \quad (6.39b)$$

and

$$rbsby = RBSBY0 \cdot \left( \frac{L_{eff}}{10^{-6}} \right)^{RBSDBYL} \cdot \left( \frac{W_{eff}}{10^{-6}} \right)^{RBSDBYW} \cdot NF^{RBSDBYNF} \quad (6.39c)$$

They are identical to Eqs. 6.38b and 6.38c, respectively, except for the coefficients,  $RBSBX0$  and  $RBSBY0$ , versus their counterparts ( $RBDBX0$  and  $RBDBY0$ ).

### 6.5.2.2 The 3-R Model

It was found that ignoring  $G_{rbsb}$  and  $G_{rbdb}$  from the 5-R network led to no significant accuracy loss while simplifying parameter extraction. This can be accomplished by not specifying RBSBX0 and RBSBY0 or by not specifying RBDBX0 and RBDBY0 in the model cards. Under such circumstances, the conductances  $G_{rbsb}$  and  $G_{rbdb}$  are made to take on GBMIN, an extremely small conductance in default ( $10^{-12}$  mho), which effectively makes  $G_{rbsb}$  and  $G_{rbdb}$  an open branch. This is illustrated in Fig. 6.16. It is certainly true that in this case, the circuit matrix conductance entries that are related to the *sbNode* and *bNode* nodes and the *dbNode* and *bNode* nodes can be filled with zeros for a sparser circuit matrix.

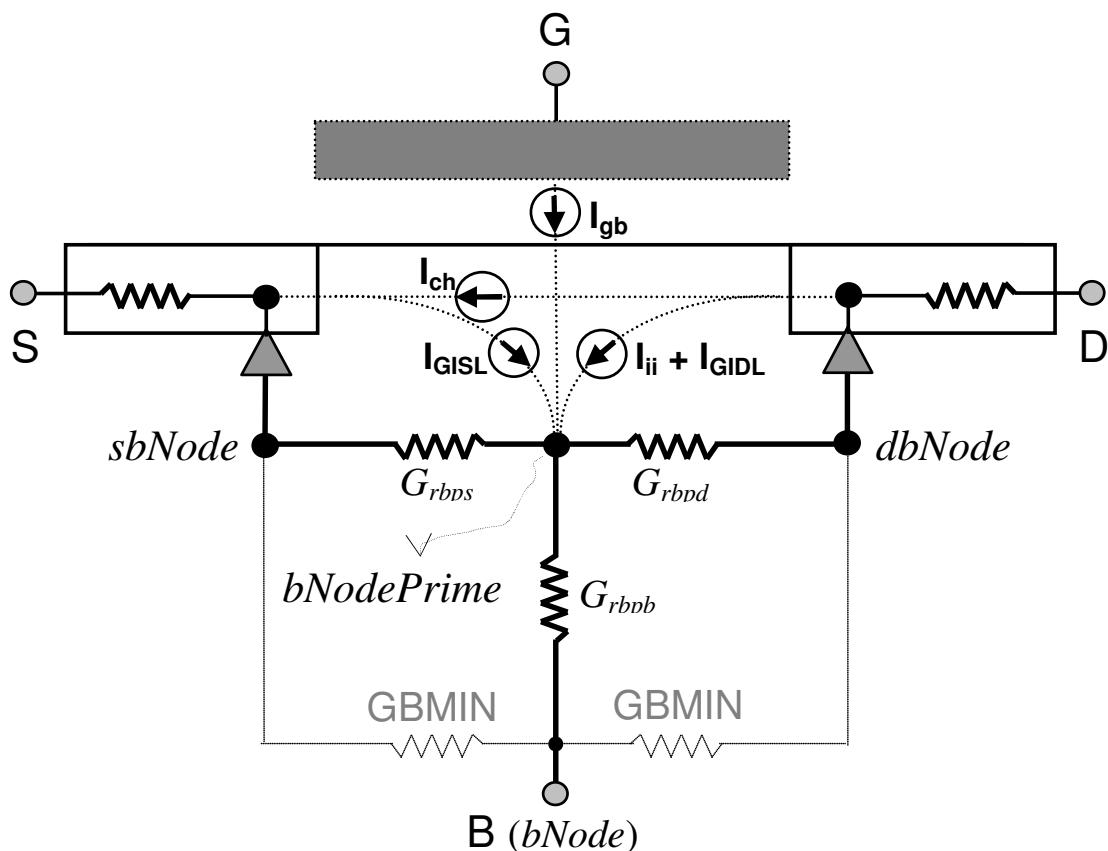


Fig. 6.16 The 3-R model of RBODYMOD = 2 where the two resistance *rbsb* and *rbdb* terms are removed from  $G_{rbsb}$  and  $G_{rbdb}$ , respectively, to simplify the resistance-network parameter extractions. GBMIN has a default value of  $10^{-12}$  mho.

### 6.5.2.3 The 1-R Model

An even more aggressive option of RBODYMOD = 2 can be made to let the 5-R network reduce further to a one-resistor configuration that has only  $G_{rbpb}$  given by Eq. (6.37) and Eq. (6.37a).  $G_{rbsb}$  and  $G_{rbdb}$  are made to take on GBMIN to effectively make their respective branch an open circuit.  $G_{rbps}$  and  $G_{rbpd}$  are both hard coded to be  $10^3$  mho to produce an effective short between *bNodePrime* and *sbNode*, and between *bNodePrime* and *dbNode*. This configuration is shown in Fig. 6.17 and is realized by not specifying RBPS0 and RBD0 in model cards. Admittedly, an optimal code implementation should drop the two internal body nodes as well, namely *sbNode* and *dbNode*, and the four conductances  $G_{rbsb}$ ,  $G_{rbdb}$ ,  $G_{rbps}$  and  $G_{rbpd}$  completely.

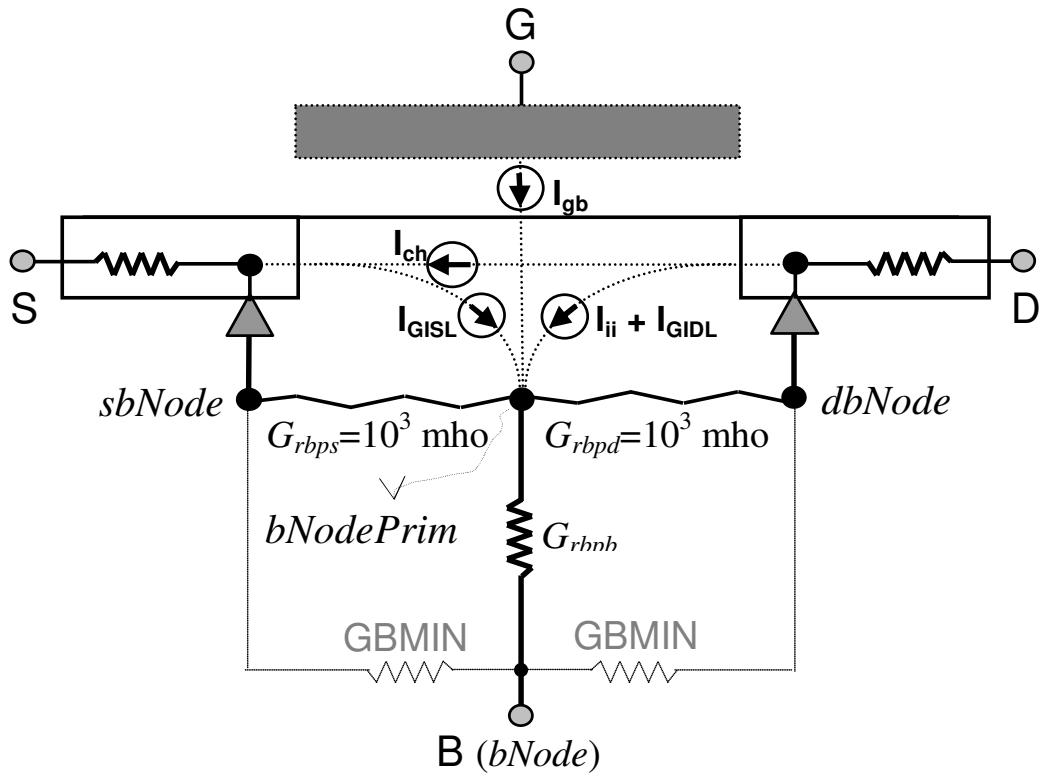


Fig. 6.17 The 1-R connection topology of RBODYMOD = 2 where the resistance *rbsb* and *rbdb* terms are dropped from  $G_{rbsb}$  and  $G_{rbdb}$ , respectively, and where  $G_{rbpd}$  and  $G_{rbps}$  are both set to a constant  $10^3$  mho conductance, to further simplify parameter extractions. It is anticipated that this connection scenario is less accurate and flexible to use.

As with the extractions of the intrinsic-input gate resistance  $R_{g,ii}$ , the body-resistance network model parameters should be extracted from device  $Y$  parameters as well, which in turn are computed from measured  $S$ -parameters. A two-step de-embedding procedure is required to obtain the “intrinsic”  $S$ -parameters that are associated only with the transistor in question, by excluding the unwanted “contributions” from, for instance, measurement equipment, device pin pads, and other peripheries around the transistor. To accomplish this goal, one often needs to design opens and shorts in on-wafer test structures for the purpose of subsequent de-embedding process. The interested readers in this topic are referred to the approach published in [5].

$Y_{22}$  is an important  $Y$  parameter (with the drain terminal being the output port of the transistor network) from which the body-resistance model parameters need to be extracted. It is often a difficult task to derive a close-form analytical expression for  $Y_{22}$  for an arbitrary  $V_{ds}$  value. Thus, a global optimization technique by solving a frequency domain circuit matrix system is the one to utilize in a parameter extraction tool. For this purpose, measured data with wide ranges of biases and transistor geometries as well as different layout combinations needs to be supplied for parameter extraction optimization.

## 6.6 Chapter Summary

This chapter presented the BSIM4 models for high-speed CMOS IC applications. It provides comprehensive analyses of the modeling concept, methodology, formulations, and parameter extraction for the SPICE simulation of the MOSFET gate electrode resistance, gate intrinsic-input resistance  $R_{g,ii}$  and charge-deficit non-quasi-static effects, and a scalable MOSFET body resistance network. The authors would like to emphasize that compact models for advanced, high-speed CMOS technologies must take into account the distributed RC nature of the gate and channel regions. BSIM4 initiates and enables this in SPICE modeling in a truly compact fashion. Also essential to the high-speed CMOS IC design are accurate MOSFET noise models. The next chapter is devoted to the BSIM4 noise modeling.

## 6.7 Parameter Table

Name (type)	Description and default	Can be binned?	Note
<b>RGATEMOD</b> (Local and global; integer)	Gate resistance model selector.  Default = 0; dimensionless: No gate resistance <i>is</i> to be generated. Other optional values are 1, 2, or 3.	No	-
<b>RBODYMOD</b> (Local and global; integer)	Body-resistance network model selector.  Default = 0; dimensionless: No network is to be generated. Other optional values are 1 or 2.	No	-
<b>TRNQSMOD</b> (Local and global; integer)	Charge-deficit transient NQS model selector.  Default = 0; dimensionless: Charge-deficit transient NQS model is turned off. Other optional value is 1 (turned on).	No	When <b>TRNQSMOD</b> is set to 1, do not make <b>RGATEMOD</b> equal to 2 or 3; vice versa.
<b>ACNQSMOD</b> (Local and global; integer)	Charge-deficit AC NQS model selector.  Default = 0; dimensionless: Charge-deficit AC NQS model is turned off. Other optional value is 1 (turned on).	No	When <b>ACNQSMOD</b> is set to 1, do not make <b>RGATEMOD</b> equal to 2 or 3; vice versa.
<b>NF</b> (Local; integer)	The number of fingers that a multi-finger device structure has.  Default = 1; dimensionless.	No	Reset to 1 if <b>NF</b> $\leq$ 1 with a fatal error.
<b>RSHG</b> (Global; double)	Poly-silicon gate sheet resistance.  Default = 0.1 [ohm].	No	If <b>RSHG</b> $\leq$ 0.0, a warning will be issued.
<b>NGCON</b> (Local and global; double)	Poly-silicon gate contact type selector.  Default = 1; dimensionless: Gate contact made only on one side of the gate. Other optional value is 2 (two-sided contact).	No	If <b>NGCON</b> is less than 1, a fatal error will be issued.
<b>XGW</b> (Local and global; double)	Distance from the gate contact to the device edge.  Default = 0.0 [m].	No	-

XGL (Global; double)	The offset between the actual poly-silicon gate (not the channel) length and the designed length of the poly-silicon gate.  Default = 0.0 [m].	No	If $(L_{drawn} + XL) < XGL$ , a fatal error will be issued.
XRCRG1 (Global; double)	The intrinsic-input gate resistance $R_{g,ii}$ fitting parameter for both drift and diffusion components.  Default = 12.0; dimensionless.	Yes	If $XRCRG1 < 0.0$ , a warning will be given.
XRCRG2 (Global; double)	The intrinsic-input gate resistance $R_{g,ii}$ fitting parameter for the diffusion component only when $V_{gs}$ is close to or less than $V_{th}$ .  Default = 1.0; dimensionless.	Yes	-
GBMIN (Global; double)	The minimum conductance set for each body-resistance network branch to prevent too high resistances that would otherwise make convergence difficult.  Default = $10^{-12}$ [mho].	No	If it is less than $10^{-20}$ mho, a warning will be given.
RBPD (Local and global; double)	The body resistance component connecting <i>sbNodePrime</i> and <i>dbNode</i> .  Default = 50.0 [ohm].	No	Used for RBODYMOD = 1 only.
RBPS (Local and global; double)	The body resistance component connecting <i>sbNodePrime</i> and <i>sbNode</i> .  Default = 50.0 [ohm].	No	Used for RBODYMOD = 1 only.
RBPB (Local and global; double)	The body resistance component connecting <i>sbNodePrime</i> and <i>bNode</i> .  Default = 50.0 [ohm].	No	Used for RBODYMOD = 1 only.
RBDB (Local and global; double)	The body resistance component connecting <i>dbNode</i> and <i>bNode</i> .  Default = 50.0 [ohm].	No	Used for RBODYMOD = 1 only.
RBSB (Local and global; double)	The body resistance component connecting <i>sbNode</i> and <i>bNode</i> .  Default = 50.0 [ohm].	No	Used for RBODYMOD = 1 only.

RBPD0 (Global; double)	The body resistance measured in the horizontal channel length direction from <i>bNodePrime</i> to <i>dbNode</i> when $L_{eff} = W_{eff} = 10^{-6}$ meter and $NF = 1$ .  Default = 50.0 [ohm].	No	Used for 3-R and 5-R options of RBODYMOD = 2 only.
RBPDL (Global; double)	The exponent fitting parameter for the channel-length dependence of the body resistance RBPD0 measured in the horizontal channel length direction from <i>bNodePrime</i> to <i>dbNode</i> when $L_{eff} = W_{eff} = 10^{-6}$ meter and $NF = 1$ .  Default = 0.0 [ohm].	No	Used for 3-R and 5-R options of RBODYMOD = 2 only.
RBPDW (Global; double)	The exponent fitting parameter for the channel-width dependence of the body resistance RBPD0 measured in the horizontal channel length direction from <i>bNodePrime</i> to <i>dbNode</i> when $L_{eff} = W_{eff} = 10^{-6}$ meter and $NF = 1$ .  Default = 0.0 [ohm].	No	Used for 3-R and 5-R options of RBODYMOD = 2 only.
RBPDNF (Global; double)	The exponent fitting parameter for the number-of-finger NF dependence of the body resistance RBPD0 measured in the horizontal channel length direction from <i>bNodePrime</i> to <i>dbNode</i> when $L_{eff} = W_{eff} = 10^{-6}$ meter and $NF = 1$ .  Default = 0.0 [ohm].	No	Used for 3-R and 5-R options of RBODYMOD = 2 only.
RBPS0 (Global; double)	The body resistance measured in the horizontal channel length direction from <i>bNodePrime</i> to <i>sbNode</i> when $L_{eff} = W_{eff} = 10^{-6}$ meter and $NF = 1$ .  Default = 50.0 [ohm].	No	Used for 3-R and 5-R options of RBODYMOD = 2 only.
RBPSL (Global; double)	The exponent fitting parameter for the channel-length dependence of the body resistance RBPS0 measured in the horizontal channel length direction from <i>bNodePrime</i> to <i>sbNode</i> when $L_{eff} = W_{eff} = 10^{-6}$ meter and $NF = 1$ .  Default = 0.0 [ohm].	No	Used for 3-R and 5-R options of RBODYMOD = 2 only.

RBPSW (Global; double)	The exponent fitting parameter for the channel-width dependence of the body resistance RBPS0 measured in the horizontal channel length direction from <i>bNodePrime</i> to <i>sbNode</i> when $L_{eff} = W_{eff} = 10^{-6}$ meter and NF = 1.  Default = 0.0 [ohm].	No	Used for 3-R and 5-R options of RBODYMOD = 2 only.
RBPSNF (Global; double)	The exponent fitting parameter for the number-of-finger NF dependence of the body resistance RBPS0 measured in the horizontal channel length direction from <i>bNodePrime</i> to <i>sbNode</i> when $L_{eff} = W_{eff} = 10^{-6}$ meter and NF = 1.  Default = 0.0 [ohm].	No	Used for 3-R and 5-R options of RBODYMOD = 2 only.
RBPBX0 (Global; double)	The body resistance measured in the <i>x</i> or channel width direction from <i>bNodePrime</i> to <i>bNode</i> when $L_{eff} = W_{eff} = 10^{-6}$ meter and NF = 1.  Default = 100.0 [ohm].	No	Used for 1-R, 3-R and 5-R options of RBODYMOD = 2 only.
RBPBXL (Global; double)	The exponent fitting parameter for the channel-length dependence of the body resistance RBPBX0 measured in the <i>x</i> or channel width direction from <i>bNodePrime</i> to <i>bNode</i> when $L_{eff} = W_{eff} = 10^{-6}$ meter and NF = 1.  Default = 0.0 [ohm].	No	Used for 1-R, 3-R and 5-R options of RBODYMOD = 2 only.
RBPBXW (Global; double)	The exponent fitting parameter for the channel-width dependence of the body resistance RBPBX0 measured in the <i>x</i> or channel width direction from <i>bNodePrime</i> to <i>bNode</i> when $L_{eff} = W_{eff} = 10^{-6}$ meter and NF = 1.  Default = 0.0 [ohm].	No	Used for 1-R, 3-R and 5-R options of RBODYMOD = 2 only.
RBPBXNF (Global; double)	The exponent fitting parameter for the number-of-finger NF dependence of the body resistance RBPBX0 measured in the <i>x</i> or channel width direction from <i>bNodePrime</i> to <i>bNode</i> when $L_{eff} = W_{eff} = 10^{-6}$ meter and NF = 1.  Default = 0.0 [ohm].	No	Used for 1-R, 3-R and 5-R options of RBODYMOD = 2 only.

<b>RBPBY0</b> (Global; double)	The body resistance measured in the y or channel length direction from <i>bNodePrime</i> to <i>bNode</i> when $L_{eff} = W_{eff} = 10^{-6}$ meter and $NF = 1$ .  Default = 100.0 [ohm].	No	Used for 1-R, 3-R and 5-R options of <b>RBODYMOD</b> = 2 only.
<b>RBPBYL</b> (Global; double)	The exponent fitting parameter for the channel-length dependence of the body resistance <b>RBPBY0</b> measured in the y or channel length direction from <i>bNodePrime</i> to <i>bNode</i> when $L_{eff} = W_{eff} = 10^{-6}$ meter and $NF = 1$ .  Default = 0.0 [ohm].	No	Used for 1-R, 3-R and 5-R options of <b>RBODYMOD</b> = 2 only.
<b>RBPBYW</b> (Global; double)	The exponent fitting parameter for the channel-width dependence of the body resistance <b>RBPBY0</b> measured in the y or channel length direction from <i>bNodePrime</i> to <i>bNode</i> when $L_{eff} = W_{eff} = 10^{-6}$ meter and $NF = 1$ .  Default = 0.0 [ohm].	No	Used for 1-R, 3-R and 5-R options of <b>RBODYMOD</b> = 2 only.
<b>RBPBYNF</b> (Global; double)	The exponent fitting parameter for the number-of-finger $NF$ dependence of the body resistance <b>RBPBY0</b> measured in the y or channel length direction from <i>bNodePrime</i> to <i>bNode</i> when $L_{eff} = W_{eff} = 10^{-6}$ meter and $NF = 1$ .  Default = 0.0 [ohm].	No	Used for 1-R, 3-R and 5-R options of <b>RBODYMOD</b> = 2 only.
<b>RBDBX0</b> (Global; double)	The body resistance measured in the <i>x</i> or channel width direction from <i>dbNode</i> (drain-body junction) to <i>bNode</i> (horizontal body contact) when $L_{eff} = W_{eff} = 10^{-6}$ meter and $NF = 1$ .  Default = 100.0 [ohm].	No	Used for the 5-R option of <b>RBODYMOD</b> = 2 only.
<b>RBDBY0</b> (Global; double)	The body resistance measured in the y or channel length direction from <i>dbNode</i> (drain-body junction) to <i>bNode</i> (vertical body contact) when $L_{eff} = W_{eff} = 10^{-6}$ meter and $NF = 1$ .  Default = 100.0 [ohm].	No	Used for the 5-R option of <b>RBODYMOD</b> = 2 only.

RBSBX0 (Global; double)	The body resistance measured in the $x$ or channel width direction from $sbNode$ (source-body junction) to $bNode$ (horizontal body contact) when $L_{eff} = W_{eff} = 10^{-6}$ meter and $NF = 1$ .  Default = 100.0 [ohm].	No	Used for the 5-R option of RBODYMOD = 2 only.
RBSBY0 (Global; double)	The body resistance measured in the $y$ or channel length direction from $sbNode$ (source-body junction) to $bNode$ (horizontal body contact) when $L_{eff} = W_{eff} = 10^{-6}$ meter and $NF = 1$ .  Default = 100.0 [ohm].	No	Used for the 5-R option of RBODYMOD = 2 only.
RBSDXL (Global; double)	The exponent fitting parameter for the channel-length dependence of the body resistances RBSBX0 and RBDBX0.  Default = 0.0 [ohm].	No	Used for the 5-R option of RBODYMOD = 2 only.
RBSDXW (Global; double)	The exponent fitting parameter for the channel-width dependence of the body resistances RBSBX0 and RBDBX0.  Default = 0.0 [ohm].	No	Used for the 5-R option of RBODYMOD = 2 only.
RBSDXNF (Global; double)	The exponent fitting parameter for the number-of-finger NF dependence of the body resistances RBSBX0 and RBDBX0.  Default = 0.0 [ohm].	No	Used for the 5-R option of RBODYMOD = 2 only.
RBSDYL (Global; double)	The exponent fitting parameter for the channel-length dependence of the body resistances RBSBY0 and RBDBY0.  Default = 0.0 [ohm].	No	Used for the 5-R option of RBODYMOD = 2 only.
RBSDYW (Global; double)	The exponent fitting parameter for the channel-width dependence of the body resistances RBSBY0 and RBDBY0.  Default = 0.0 [ohm].	No	Used for the 5-R option of RBODYMOD = 2 only.
RBSDYNF (Global; double)	The exponent fitting parameter for the number-of-finger NF dependence of the body resistances RBSBY0 and RBDBY0.  Default = 0.0 [ohm].	No	Used for the 5-R option of RBODYMOD = 2 only.

## References

- [1] Xiaodong Jin, Jia-Jiunn Ou, Chih-Hung Chen, Weidong Liu, M. Jamal Deen, Paul R. Gray, and Chenming Hu, “An effective gate resistance model for CMOS RF and noise modeling,” *Tech. Dig. of IEDM*, pp. 961-964, San Francisco, December 1998.
- [2] Xiaodong Jin, Kanyu Cao, Jia-Jiunn Ou, Weidong Liu, Yuhua Cheng, Mishel Matloubian, and Chenming Hu, “An accurate non-quasi-static MOSFET model for simulation of RF and high speed circuits,” *VLSI Technology, Dig. of Technical Papers*, pp. 196-197, 2000.
- [3] Jia-Jiunn Ou, Xiaodong Jin, Ingrid Ma, Chenming Hu, and Paul R. Gray, “CMOS RF modeling for GHz communication IC’s,” *VLSI Technology, Dig. of Technical Papers*, pp. 94-95, 1998.
- [4] W. Liu, R. Gharpurey, M. C. Chang, U. Erdogan, R. Aggarwal, and J.P. Mattia, “R.F. MOSFET modeling accounting for distributed substrate and channel resistances with emphasis on the BSIM3v3 SPICE model,” *Tech. Dig. of IEDM*, pp. 309-312, Washington D. C., December 1997.
- [5] M. C. A. M. Koolen, J. A. M. Geelen, and M. P. J. G. Versleijen, “An improved de-embedding technique for on-wafer high-frequency characterization,” *IEEE Proc. Bipolar Circuit and Technology Meeting*, pp. 188-191, 1991.

**This page intentionally left blank**

## Chapter 7

# Noise Models

### 7.1 Introduction and Chapter Objectives

MOS transistors, like other semiconductor devices, generate electrical noise. Noise comes from multiple physical mechanisms. They are flicker (also known as  $1/f$ ) noise, thermal (often referred to as white) noise, and shot noise. In the giga-hertz frequency range, induced gate noise become important. It originates from the channel thermal noise and it is partially correlated with the channel thermal noise.

Electrical noise is generated in electronic devices by the random fluctuations in the number and velocities of the charge carriers. These fluctuations give rise to small fluctuations in voltages and currents (in the range of nano to micro volts and amperes). They set a lower limit in discerning small signals and acceptable power supplies. Minimizing noise is particularly crucial for such circuits as low-noise amplifiers, mixers, oscillators, and A/D converters.

The subject of noise, including its SPICE models, has received intensive research over the decades. This chapter presents the BSIM4 noise models. These models have been successfully used for many generations and types of CMOS technologies worldwide. The noise representation and SPICE noise analyses will be briefly reviewed in Section 7.2. Section 7.3 is devoted to the BSIM4 flicker noise models. Section 7.4 describes the BSIM4 channel thermal noise models, including the charge-based channel thermal noise model. In particular, the BSIM4 holistic channel thermal noise model will be discussed. This holistic model, taking into account the distributed RC nature of the channel and using a noise partitioning approach, combines the modeling

of channel thermal noise, induced gate noise and their correlation into one single compact noise model. The noise attributed to parasitic resistors and the shot noise produced by the junction diodes are described in Section 7.5. Section 7.6 summarizes this chapter. A complete BSIM4 noise model parameter table is given in Section 7.7.

## 7.2 Noise Representations and Parameters

### 7.2.1 Noise and Power Spectral Intensity

One attribute of a random variable  $X(t)$ , for instance, sampled over time  $t$ , is its average value  $\bar{X}$  over a period of time. However,  $\bar{X}$  is often found to be zero when averaged over a long period of time, as in the case of averaged random noise voltages and currents. Another attribute, the mean square average  $\overline{X^2}$ , is the average of the square of  $X(t)$ , which does not vanish over time, and hence becomes more useful. In addition, the square links to the power or the energy per unit time. It is the practice to introduce a power spectral intensity function  $S_x(f)$ , which is defined by the following equation (7.1), from averaging  $X^2(t)$  over a sufficiently long time, measured through a narrow-bandwidth signal filter with a bandwidth  $\Delta f$  centered around the signal frequency  $f$ :

$$\overline{X^2} \triangleq \overline{(X - \bar{X})^2} = S_x(f) \cdot \Delta f \quad (7.1)$$

MOSFET noise sources produce fluctuating noise voltages  $v_n$  and noise currents  $i_n$ . Both quantities also have zero average values

$$\overline{v_n} = 0 \quad (7.2)$$

and

$$\overline{i_n} = 0 \quad (7.3)$$

The power spectral intensities of these noise voltages and currents are then defined using Eq. (7.1) and given by the following equations:

$$\overline{v_n^2(f)} = S_v(f) \cdot \Delta f \quad (7.4)$$

and

$$\overline{i_n^2(f)} = S_i(f) \cdot \Delta f \quad (7.5)$$

$S_v(f)$  and  $S_i(f)$  have the units of volt squared per hertz and ampere squared per hertz, respectively. Over a one-hertz frequency interval, the noise voltage (in volts) and current (in amperes) at frequency  $f$ , are understood to be the square-root values of Eqs. (7.4) and (7.5), namely, the root-mean-square noise voltage and noise current per unit bandwidth or hertz:

$$\sqrt{\overline{v_n^2(f)}} = \sqrt{S_v(f)} \quad (7.4a)$$

and

$$\sqrt{\overline{i_n^2(f)}} = \sqrt{S_i(f)} \quad (7.5b)$$

Over a finite frequency range from  $f_1$  to  $f_2$ , the total mean square noise voltage and current can be found by integrating Eqs. (7.4) and (7.5)

$$\overline{v_n^2} = \int_{f_1}^{f_2} S_v(f) \cdot df \quad (7.6)$$

and

$$\overline{i_n^2} = \int_{f_1}^{f_2} S_i(f) \cdot df \quad (7.7)$$

If  $S_v(f)$  and  $S_i(f)$  are constant over the frequency range  $(f_2 - f_1)$ , as in the case of MOSFET thermal noise (also known as white noise, which is independent of the frequency in the frequency range, hence “white”), Eqs. (7.6) and (7.7) are simply proportional to  $(f_2 - f_1)$ , referred to as the noise bandwidth  $B_{eff}$ .

Note that the power spectral intensities  $S_v(f)$  and  $S_i(f)$  throughout this chapter are taken to be stationary, i.e., time invariant or constant when measured at different times, each over a sufficiently long time interval. They are intended for frequency-domain noise analyses. When the noise fluctuation rate or fluctuation frequency is comparable to the inverse of the length of the measurement time interval,  $S_v(f)$  and  $S_i(f)$  are no longer stationary. Transient noise analysis in the time-domain requires

the development of time-dependent spectral intensity models,  $S_v(f,t)$  and  $S_i(f,t)$ . BSIM4 does not support transient noise modeling.

$S_v(f)$  and  $S_i(f)$  are usually bias dependent. This will be discussed in the following sections. In addition, the bias dependencies are generally different for different noise types. For instance, the flicker and thermal noises have different bias dependencies because they are generated by different physical mechanisms: Flicker noise results from random carrier trapping and thermal noise from random carrier scattering.

### 7.2.2 SPICE Noise Representations

To perform SPICE noise analyses in the frequency domain, the noise voltage and current given by Eqs. (7.4a) and (7.5b) are incorporated into a linear, small-signal equivalent circuit as shown in Fig. 7.1. They are treated as regular small-signal voltage and current sources, i.e., electric excitations to that circuit. This is accurate because noise usually has a very small amplitude that falls into the small-signal range in which the small-signal analysis of semiconductor devices is valid (noise voltage  $\ll$  thermal voltage,  $k_B T/q \sim 25$  mV). To do SPICE noise analyses, any externally-applied signal sources are removed from the equivalent circuit. An open-circuit branch will be left from removing a small-signal voltage source and a short-circuit branch from the removal of a small-signal current source.

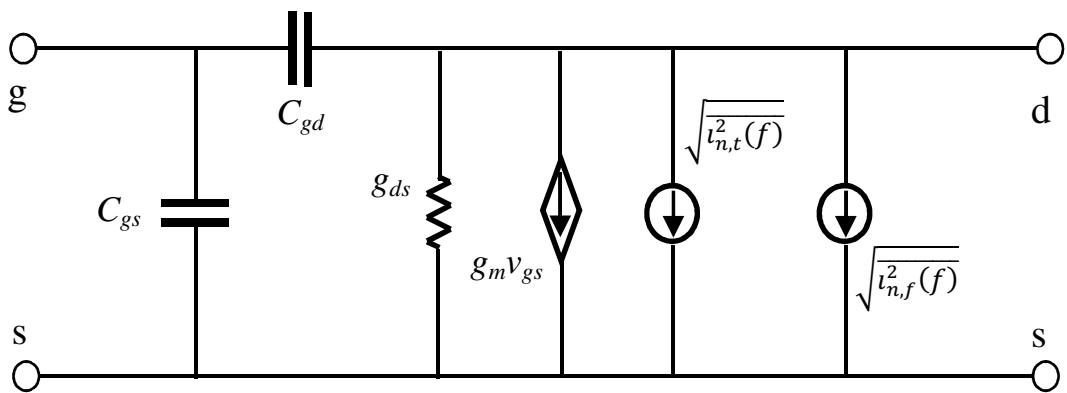


Fig. 7.1 A MOSFET small-signal equivalent circuit with the channel thermal and flicker noise current sources. The  $g_m v_{gs}$  current source here represents another noise that results from those two sources; i.e., noise  $v_{gs}$  is produced by the noise current flowing through  $C_{gd}$  and results in a noise in the drain current,  $g_m v_{gs}$ . The body terminal and some other trans-capacitances and conductances of the transistor are omitted for ease of illustration.

Different types of noise may or may not be correlated. Consider an aggregate noise current  $i_n$ , which is composed of two components  $i_{n,t}$  and  $i_{n,f}$  that is,  $i_n = i_{n,t} + i_{n,f}$ . The mean-square value of  $i_n$  is then

$$\overline{i_n^2} = \overline{i_{n,t}^2} + \overline{i_{n,f}^2} + 2 \cdot \overline{i_{n,t} \cdot i_{n,f}} \quad (7.8)$$

The noise correlation coefficient  $c$  is defined as [1]

$$c = \frac{\overline{i_{n,t} \cdot i_{n,f}}}{\sqrt{\overline{i_{n,t}^2} \cdot \overline{i_{n,f}^2}}} \quad (7.9)$$

The correlation coefficient,  $c$ , can take on a value within the range of  $-1 \leq c \leq 1$ . If the noise current product  $\overline{i_{n,t} \cdot i_{n,f}} = 0$  (hence  $c = 0$ ),  $i_{n,t}$  and  $i_{n,f}$  are said to be uncorrelated, which is the case between the thermal noise and flicker noise of MOSFET because they arise from totally unrelated mechanisms. The effects of these two noise sources are simply additive. In other words, their joint contributions to circuit nodal voltages or currents can be obtained by superposing the contribution of one on that of the other, as shown in Fig. 7.1. This is because of the fact that these two noise sources are generated from independent physical mechanisms. Noise sources of this type have random phases. Hence, the cross product averages to zero,  $\overline{i_{n,t} \cdot i_{n,f}} = 0$ .

When  $i_{n,t}$  and  $i_{n,f}$  are fully correlated positively, then  $c=+1$ , or negatively, then  $c=-1$ . In all other cases ( $c \neq \pm 1, 0$ )  $i_{n,t}$  and  $i_{n,f}$  are said to be partially correlated. Moreover, in the event  $i_{n,t}$  and  $i_{n,f}$  are complex numbers due to for example, phase differences,  $c$  may be a complex number.

### 7.2.3 Noise Representation and Parameters of a Two-Port Network

In a more general representation, a MOSFET transistor can be described in a two-port active network that generates noise. A common-source configuration is illustrated in Fig. 7.2. In this figure, the input and output admittances  $Y_i$  and  $Y_o$  of the network are complex numbers. For instance,  $Y_i = G_i + jB_i$  with  $G_i$  the conductance and  $B_i$  the susceptance. The source of the noise current  $i_{n,i}$  may be the thermal noise of the conductance  $G_i$ , the shot noise in the gate tunneling currents, and/or the channel-induced gate noise.  $i_{n,o}$  is the sum of the channel flicker and thermal noise. Note that the noise voltage  $v_{n,i}$  at the input port,

approximately equaling  $\sqrt{i_{n,l}^2(f)}$  divided by  $Y_i$  (assuming  $C_{gd}$  is small in the saturation region), is amplified by the network trans-admittance  $Y_m$  into the output noise voltage  $v_{n,o}$ . Of course,  $Y_m$  becomes the transconductance  $g_m$  at low frequencies.

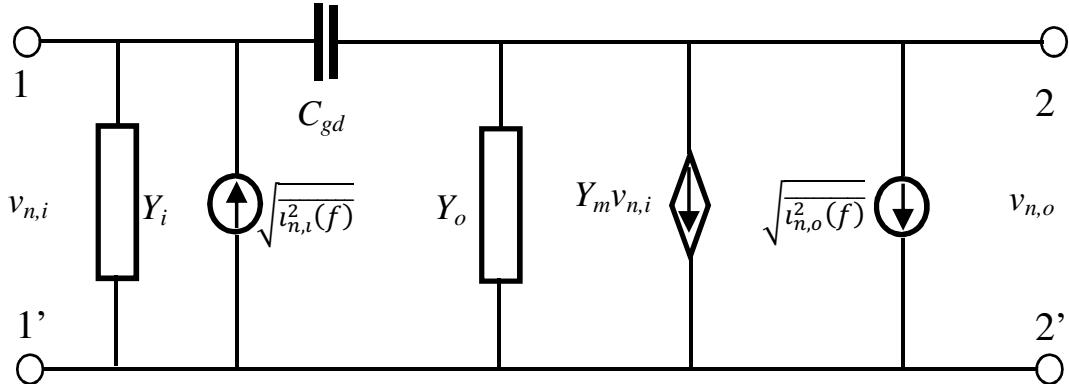


Fig. 7.2 A two-port network representation of a noisy MOS transistor.

Circuit designers often use the schematic as shown in Fig. 7.3 in lieu of Fig. 7.2. In this equivalent representation, the transistor enclosed in the dash-line box is treated as “noiseless”. The noise sources of the device are now all moved out and represented by the input-referred noise current  $i_{n,i}$  and noise voltage  $e_{n,i}$  [1]. At low frequencies, the noise voltage  $\sqrt{e_{n,l}^2(f)}$  is again roughly  $\sqrt{i_{n,o}^2(f)}$  divided by the magnitude of  $Y_m$ . As device operation frequencies increase to gigahertz range,  $i_{n,i}$  and voltage  $e_{n,i}$  can be partially correlated.

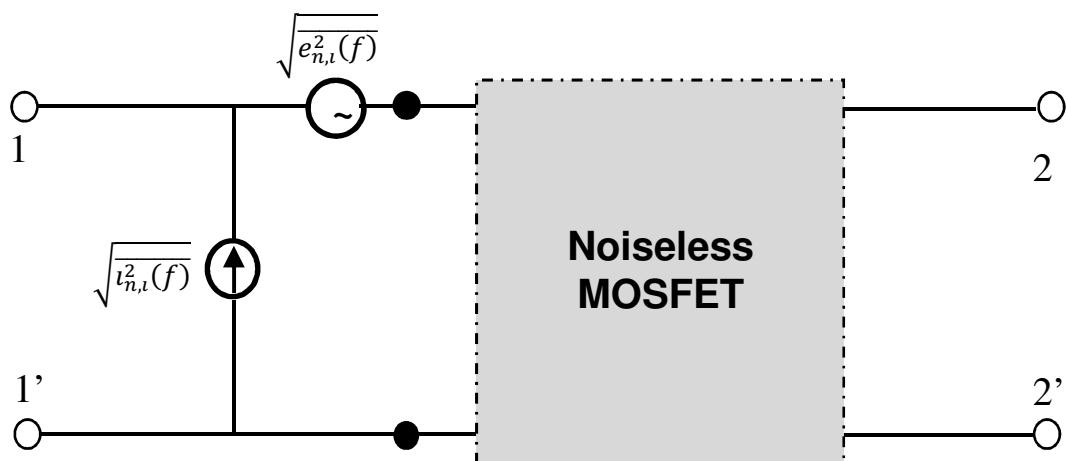


Fig. 7.3 Input-referred noise sources to present a noiseless MOS transistor.

In laboratories, noise is measured by connecting a diode with very low series resistance to the input port of Figs. 7.2 and 7.3. A dc current  $I_{eq}$  is forced through this diode such that the output noise power doubles that from the MOS device under test (DUT). The noise generated by this diode is mainly shot noise with a noise current power given by  $2q \cdot I_{eq} \cdot \Delta f$ . The forced diode current  $I_{eq}$  is called the equivalent diode noise current of that DUT. The input-referred noise current power of Fig. 7.3 can now be measured and represented by  $\overline{i_{n,i}^2(f)} = 2q \cdot I_{eq,i} \cdot \Delta f$ .

As mentioned earlier, the input-referred noise voltage is related to the channel noise current  $i_{n,o}$  via the network trans-admittance  $Y_m$

$$\overline{e_{n,i}^2} \approx \frac{\overline{i_{n,o}^2(f)}}{|Y_m|^2} \quad (7.10)$$

Therefore,  $e_{n,i}$  can also be measured and represented in terms of  $I_{eq,o}$ , the equivalent diode current of  $i_{n,o}$ , where  $\overline{i_{n,o}^2(f)} = 2q \cdot I_{eq,o} \cdot \Delta f$ . One can then obtain the measured value of  $I_{eq,o}$  by short-circuiting the input port while connecting a similar diode at the output port of Fig. 7.2, a similar approach to obtaining  $I_{eq,i}$ .

According to Nyquist's theorem, the thermal or white (independent of the measurement frequencies of band-pass filters) noise voltage power of an ohmic resistor  $R$  at the absolute temperature  $T$  is  $4kT\Delta f$ . The thermal noise current power is  $4kTG\Delta f$ , where the conductance is given by  $G=1/R$ . Assuming the device operating temperature is known, such as  $T_0=290K$ , then an equivalent noise resistance  $R_n$  and conductance  $G_n$  of  $\overline{e_{n,i}^2(f)}$  and  $\overline{i_{n,i}^2(f)}$  of Fig. 7.3 can be defined

$$\overline{e_{n,i}^2(f)} = 4kT_0 \cdot R_n \cdot \Delta f \quad (7.11a)$$

and

$$\overline{i_{n,i}^2(f)} = 4kT_0 \cdot G_n \cdot \Delta f \quad (7.11b)$$

In addition to  $R_n$  and  $G_n$ , there is another noise parameter, the equivalent noise temperature  $T_n$ . It is advantageous to use this parameter when the device under test is expected to produce thermal noise and the exact device operating temperature is unknown or difficult to determine. The equivalent noise temperature is calculated by substituting the actual resistance  $R$  and conductance  $G$  values for  $R_n$  and  $G_n$  and  $T_n$  for  $T_0$  of Eqs. (7.11a) and (7.11b), which yields

$$T_n = \frac{R_n}{R} \cdot T_0 = \frac{G_n}{G} \cdot T_0 \quad (7.12)$$

Like  $R_n$  and  $G_n$ ,  $T_n$  can be measured via the equivalent saturated diode noise current  $I_{eq}$  by letting  $4kT_n \cdot G$  equal  $2q \cdot I_{eq}$ . Commercial SPICE simulators provide a way to output  $T_n$  as well as  $R_n$  and  $G_n$ .

These two-port noise parameters differ in one transistor in different dc bias configurations, such as the common source and common drain dc bias configurations. The inconvenience imposed on device characterization and circuit design is obvious. Noise figure is used to circumvent this inconvenience.

The noise figure  $NF$  of a single-stage two-port network is defined as the signal-to-noise power ratio at the input port divided by the signal-to-noise power ratio at the output port which is dimensionless.

$$NF \equiv \frac{\text{Signal-to-noise power ratio at the input}}{\text{Signal-to-noise power ratio at the output}} = \frac{S_i/N_i}{S_o/N_o} \quad (7.13)$$

It signifies how the desired signal strength is weakened relative to the noise after network amplification (Note  $NF$  is different from the number of finger parameter  $NF$ .). According to network theory, Eq. (7.13) can be transformed into a more analytical equivalent

$$NF \equiv \frac{S_i}{kT_0 \cdot B_{eff}} \cdot \frac{k(T_0 + \Delta T) \cdot B_{eff} \cdot G_{network}}{S_i \cdot G_{network}} = 1 + \frac{\Delta T}{T_0} \quad (7.14)$$

where  $T_0 = 290$  K is the noise reference temperature,  $\Delta T$  is the shift of the device noise temperature from  $T_0$ ,  $G_{network}$  is the power gain of that one-stage network, and  $B_{eff}$  is again the noise bandwidth of the network.  $\Delta T$  is positive and  $NF$  is equal or greater than 1.

The noise figure is expressed in decibel (dB). Eq. (7.14) can be converted into dB by  $NF(\text{in dB}) = 10 \cdot \log_{10}(NF)$ . For example,  $NF = 100$  from Eq. (7.14) means a noise figure of 20 db.

For an  $n$ -stage cascaded network, one can make use of the method of deriving Eq. (7.14) to obtain the total noise figure

$$NF_{\text{total}} = NF_{\text{stage}_1} + \frac{NF_{\text{stage}_2} - 1}{G_{\text{stage}_1}} + \dots + \frac{NF_{\text{stage}_n} - 1}{G_{\text{stage}_1} \cdot G_{\text{stage}_2} \dots \cdot G_{\text{stage}_n}} \quad (7.15)$$

where the gain of each stage in the denominator denotes the power gain. The first two terms on the right-hand side of Eq. (7.15) are particularly important for minimizing  $NF$ . This is to say  $NF_{\text{stage}_1}$  should be made small whereas  $G_{\text{stage}_1}$  should be made as large as possible.

For characterizing the noise figure of a DUT (Device Under Test) and for designing a low noise circuit, a noise matching network is usually constructed and connected to the input terminals of the transistor. A

## Section 7.2 Noise Representations and Parameters

noise source such as a diode is then connected to the other side of the matching network. This is as shown in Fig. 7.4. The output noise of the DUT will then undergo low-noise amplification (LNA) before it is transmitted to a noise figure meter.

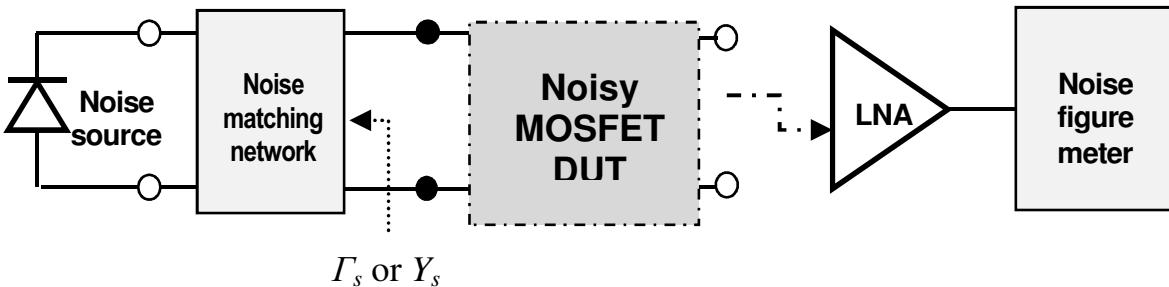


Fig. 7.4 Schematic diagram of noise figure measurement.

The measured noise figure  $NF$  varies with the noise source reflection or the impedance mismatch between the noise source and the transistor input. It is known that  $NF$  is related to three noise parameters by

$$NF = NF_{min} + \frac{4 \cdot R_n}{Z_0} \cdot \frac{|\Gamma_s - \Gamma_{opt}|^2}{|1 + \Gamma_{opt}|^2 \cdot (1 - |\Gamma_s|^2)} \quad (7.16)$$

The three parameters are  $NF_{min}$ ,  $R_n$ , the minimum noise figure and the equivalent noise resistance of the transistor, and  $\Gamma_{opt}$ , the optimum complex reflection coefficient of the network.  $Z_0$  is the impedance of the measurement system, typically 50 ohms.  $r_n = R_n/Z_0$  is referred to as the normalized equivalent noise resistance.  $\Gamma_s$  in Fig. 7.4 and Eq. (7.16) are the complex reflection coefficient of the noise source. Setting  $\Gamma_s$  to  $\Gamma_{opt}$  yields the minimum noise figure  $NF_{min}$ .  $\Gamma_s$  is

$$\Gamma_s = \frac{R_s + j \cdot X_s - Z_0}{R_s + j \cdot X_s + Z_0} \quad (7.17)$$

with  $R_s$  and  $X_s$  being the real part (resistance) and imaginary part (reactance) of the matching network. Alternatively, the matching network can be represented in the form of admittance  $Y_s$ , which is a function of  $\Gamma_s$  as

$$Y_s = \frac{1}{Z_0} \cdot \frac{1 - \Gamma_s}{1 + \Gamma_s} \quad (7.18)$$

In an analogous fashion, the optimum admittance of the matching network is given by

$$Y_{opt} = \frac{1}{Z_0} \cdot \frac{1 - \Gamma_{opt}}{1 + \Gamma_{opt}} \quad (7.19)$$

With Eqs. (7.18) and (7.19), one can prove that

$$NF = NF_{min} + \frac{R_n}{G_s} \cdot |Y_s - Y_{opt}|^2 \quad (7.20)$$

is equivalent to Eq. (7.16) if  $G_s$  is the real part of  $Y_s$ , i.e.,  $\text{Re}\{Y_s\}$ .

In order to find  $NF_{min}$ ,  $R_n$  and  $\Gamma_{opt}$  from measurement or SPICE simulation,  $Y_s$  or  $\Gamma_s$  needs to be swept to find where the minimum noise figure of the transistor  $NF_{min}$  is attained. In principle, since four scalar variable noise parameters (i.e.,  $NF_{min}$ ,  $R_n$ ,  $\Gamma_{opt}$  and either  $Y_s$  or  $\Gamma_s$ ) are to be found, only four measurements at different  $Y_s$  need to be performed. In real life however, more measurements for other favorable  $Y_s$  or  $\Gamma_s$  values are pursued. This process may utilize, for instance, the least-mean-square optimization technique for parameter extraction and optimization of those four parameters.

There exist many possible values of  $\Gamma_s$  that yield a given noise figure  $NF$ . They can be found by tuning the admittance of the noise matching network. In fact, when plotted in a complex  $\Gamma_s$  plane, these  $\Gamma_s$  values form a circle, called a constant noise figure circle. A family of such circles (known as noise figure contours) can be obtained for various  $NF$  as shown in Fig. 7.5. This information can be generated with SPICE simulation. Circuit designers can plot both the noise figure circles and the circuit voltage or power gains and then determine graphically the range of the matching impedance that can potentially give the lowest noise and highest gain.

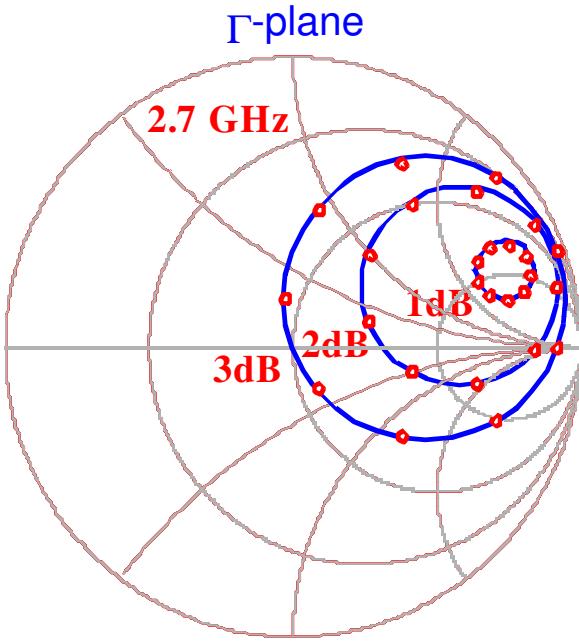


Fig. 7.5 Noise figure contours of a MOSFET in the  $\Gamma_s$  plane for  $NF = 1, 2$  and  $3$  dB at 2.7 GHz. The dark gray dots are measured values, with which the SPICE BSIM4 simulations (the gray lines) are in excellent agreement.

The equation of the noise figure circles can be developed as follows. Reproduced below for convenience is Eq. (7.16)

$$NF = NF_{min} + \frac{4 \cdot R_n}{Z_0} \cdot \frac{|\Gamma_s - \Gamma_{opt}|^2}{|1 + \Gamma_{opt}|^2 \cdot (1 - |\Gamma_s|^2)} \quad (7.16)$$

Let

$$N_c = \frac{Z_0 \cdot (NF - NF_{min})}{4R_n} \cdot |1 + \Gamma_{opt}|^2 \quad (7.21)$$

Then, Eq. (7.16) becomes

$$\frac{|\Gamma_s - \Gamma_{opt}|^2}{(1 - |\Gamma_s|^2)} = N_c \quad (7.22a)$$

which gives

$$(\Gamma_s - \Gamma_{opt}) \cdot (\Gamma_s^* - \Gamma_{opt}^*) = N_c - N_c \cdot |\Gamma_s|^2 \quad (7.22b)$$

By regrouping the terms of Eq. (7.22b), we have

$$(1 + N_c)|\Gamma_s|^2 + |\Gamma_{opt}|^2 - 2 \cdot \operatorname{Re}\{\Gamma_s \Gamma_{opt}^*\} = N_c \quad (7.23)$$

Multiplying both sides of Eq. (7.23) by  $(1 + N_c)$ , we have

$$\left| \Gamma_s - \frac{\Gamma_{opt}}{1+N_c} \right|^2 = \left[ \frac{\sqrt{N_c^2 + N_c(1 - |\Gamma_{opt}|^2)}}{1+N_c} \right]^2 \quad (7.24)$$

In the  $\Gamma_s$  plane, these circles are centered at  $\Gamma_{opt}/(1+N_c)$  with a radius of  $[N_c^2 + N_c(1 - |\Gamma_{opt}|^2)]^{1/2}/(1+N_c)$ . In the remaining sections of this chapter, the BSIM4 flicker noise and thermal noise models will be discussed. Particular attention will be paid to the modeling of the noise current power spectral intensity  $S_i(f)$  of Eq. (7.5b) (i.e., noise current power in a unit frequency interval). One can then employ SPICE simulators to perform frequency-domain noise analyses (refer to Fig. 7.1) and derive the two-port network noise parameters presented above.

### 7.3 BSIM4 Flicker Noise Models

The MOSFET flicker (or 1/f) noise results from random charge trapping and de-trapping in the gate dielectrics and at the gate dielectric/silicon interface. The random fluctuation in the trap charge induces a corresponding fluctuation in the channel surface potential and therefore the density (number) of channel charges carriers. In addition, the random fluctuation of trap charge causes mobility fluctuations through Coulombic scattering [S10]. The trap charges fluctuations are discrete in nature.

For MOS transistors with small channel areas, there may be only one trap (in the vicinity of surface Fermi level) that significantly participates in the random trapping and de-trapping. This one-trap or few-trap condition was reported in [S1]. Experiments of small-size MOSFETs, capturing and emission of carriers by these traps, result in channel current pulses/spikes which resemble a telegraph signal [2]. For so few traps, modeling the noise becomes a very difficult task. So far, compact models only consider the average effects of a large number of traps (ensemble average).

In manufacturing process development, flicker noise can be reduced by improving the gate dielectric and interface qualities and thus reducing the oxide and interface trap densities. These observations give strong support for attributing the flicker noise to oxide and interface charge trapping and de-trapping.

Measurements on recent production MOSFETs have also shown that the flicker noise spectral intensity is approximately inversely proportional to frequency up to about 100 kHz, above which the channel thermal noise power starts to dominate. For this reason, the flicker noise is also known as the  $1/f$  noise. This characteristic can be modeled in compact modeling by a uniform trap density through the thickness of the oxide (at least near the oxide/semiconductor interface) [2] which was first suggested by McWhorter's MIT PhD thesis and extensively proven by Fu [S11]. Hence, in a log-log scale, this noise power intensity is largely a straight line as a function of frequencies. The slope may differ somewhat from unity (i.e.,  $1/f^{\text{EF}}$ ) depending on device manufacturing processes (modeled by the exponent parameter  $\text{EF}$ ), and the material and integrity of the gate dielectrics. Furthermore, the slope may vary somewhat for different frequency ranges, i.e., deviating from a simple power-law relationship. This is because the time constants of the oxide traps can vary with the trap location and energy level [2], just like those modeled by Fu 40 years ago [S11]. However,  $1/f$  is a good approximate description of the flicker noise power spectral intensity in compact modeling of the MOSFET.

Flicker noise has impact on circuit noise performance directly at low frequencies. However, it is also a serious source of noise in RF circuits such as mixers, oscillators, and PLL (phase-locked loops). In these circuits, the flicker noise is up-converted to the vicinity of the carrier frequency. This degrades the signal-to-noise ratio or generates phase noise in these high-frequency circuits [3].

BSIM4 provides two channel flicker noise models. They can be selected by specifying the BSIM4 noise model flag parameter **FNOIMOD**. When **FNOIMOD** is set to 0, a simple, empirical flicker noise model from SPICE2g6 [4], SPICE3 [5], [6], [7], and BSIM3v3 [8], [9] is invoked. This model is convenient for hand calculations. When **FNOIMOD** is set to 1 (the default option), a physics-based, unified flicker noise model [10] will be triggered. The BSIM4 **FNOIMOD = 1** model finds its production origin in BSIM3v3 [6] and received bug fixes and enhancements including updates of default parameter values, superior smoothness across all operating regions and improved bulk charge effects.

### 7.3.1 The **FNOIMOD = 0** Simple Flicker Noise Model

This model computes the channel current flicker noise spectral intensity measured in square amperes per hertz as

$$S_{i,ds}(f) = KF \cdot \frac{I_{ds}^{AF}}{C_{oxe} \cdot L_{eff}^2 \cdot f^{EF}} \quad (7.25)$$

where **KF** is the flicker noise coefficient with default value of  $0.0\text{ A}^{2-AF} \cdot \text{s}^{1-EF} \cdot \text{Farad}$ , **AF** is the exponent of the channel DC current, and **EF** is the frequency exponent. The default values of **AF** and **EF**, dimensionless, are both set to be 1.0. All these parameters are extracted by way of fitting the measured flicker noise data.  $C_{oxe}$  is the electrical gate dielectric unit-area capacitance. The number of device fingers **NF** is implicitly included the channel current  $I_{ds}$ . Eq. (7.25) can be rough as multiple slopes of  $S_{i,ds}(f)$  possibly exist because of the co-existence of traps with various trapping and de-trapping time constants. Hence, constant **AF** and **EF** do not suffice. For a better curve fitting, their values can be made piece-wise constants for various frequency ranges. This leads to the BSIM4 physics-based, unified flicker noise model in the next sub-section.

### 7.3.2 The **FNOIMOD = 1** Physics-Based, Unified Flicker Noise Model

This is the default flicker noise model of BSIM4. It originated from the work published in [7]. It is implemented into BSIM4 with several enhancements for accurate and robust SPICE simulations. The development of this model will be briefly presented to demonstrate how the device physics and flicker noise modeling methodology can be engineered into a production-worthy SPICE MOSFET model for accuracy, ease of use, and simulation robustness.

There are two origins of flicker noise, the charge number fluctuation and the mobility fluctuation. The unified flicker noise theory combines the two into a more general theory. From the preceding description, we know that oxide and interface traps constantly trap and de-trap electrons (and holes) of the silicon surface channel and substrate. The fluctuation in the number of trapped charges causes fluctuations in the channel carrier number and in the carrier mobility through Coulombic scattering

## Section 7.3 BSIM4 Flicker Noise Models

Consider a small channel length segment  $\Delta y$ , with the y-axis from source to drain. From these two fluctuations, the fractional change of the channel DC current is then

$$\frac{\delta I_{ds}}{I_{ds}} = \left( \frac{1}{\Delta N_{ch}} \cdot \frac{\delta \Delta N_{ch}}{\delta \Delta N_t} \pm \frac{1}{\mu_{eff}} \cdot \frac{\delta \mu_{eff}}{\delta \Delta N_t} \right) \cdot \delta \Delta N_t \quad (7.26)$$

where  $\Delta N_{ch} = N_{ch} \cdot W_{eff} \cdot \Delta y$ , and  $\Delta N_t = N_t \cdot W_{eff} \cdot \Delta y$ .  $N_{ch}$  and  $N_t$  are the number of channel carriers and charged oxide and interface traps, or their equivalent, per unit area. The first term on the right side of Eq. (7.26) represents the inversion channel current carrier charge fluctuation and the second term the mobility fluctuation due to the random scattering of the carrier velocity. The plus and minus sign in front of the mobility term of Eq. (7.26) was a tentative indicator employed by the 1990 theory [2] which was replaced by  $\alpha$  which takes into account of the correlation of the carrier number and carrier mobility fluctuations at the same physical location of the trap, namely, the fluctuating trapping and detrapping of the carriers at the oxide and interface traps. Thus, a ratio  $R(y)$  is defined, which is less than unity.

$$R(y) = \frac{\delta \Delta N_{ch}}{\delta \Delta N_t} = \frac{N_{ch}(y)}{N_{ch}(y) + N^*} \quad (7.27)$$

Here  $N^* = \frac{kT \cdot (C_{oxe} + C_{dep} + CIT)}{q^2}$ .  $C_{dep}$  and  $CIT$  are the unit-area depletion-layer and interface trap capacitances. They were discussed in Chapter 2.

The inverse of the mobility is the rate of carrier scattering. The rate of scattering contains scattering by the ionized interface and oxide traps, and also all other scattering mechanisms, such as phonon scattering and rough gate-metal/gate-oxide interface. Considering all of these, then one can write

$$\frac{\delta \mu_{eff}}{\delta \Delta N_t} = - \frac{\alpha \cdot \mu_{eff}^2}{W_{eff} \cdot \Delta y} \quad (7.28)$$

where  $\alpha$  is the trap scattering coefficient. Substituting Eqs. (7.27) and (7.28) into Eq. (7.26) gives

$$\frac{\delta I_{ds}}{I_{ds}} = \left( \frac{R(y)}{N_{ch}(y)} \pm \alpha \mu_{eff} \right) \cdot \frac{\delta \Delta N_t}{W_{eff} \cdot \Delta y} \quad (7.29)$$

From this equation, one can derive the local channel current flicker noise spectral intensity, which is

$$S_{\Delta I_{ds}}(y, f) = \left(\frac{I_{ds}}{W_{eff} \cdot \Delta y}\right)^2 \cdot \left(\frac{R(y)}{N_{ch}(y)} \pm \alpha \mu_{eff}\right)^2 \cdot S_{\Delta N_t}(y, f) \quad (7.30)$$

where  $S_{\Delta N_t}(y, f)$  is the power spectral intensity of the mean-square fluctuations in the number of occupied traps over the channel segment area  $W_{eff} \cdot \Delta y$ .

As indicated earlier, the major contributions to the noise by these traps are those whose effective trap energy levels are near or at the electron quasi-Fermi energy level  $E_{fn}$  for nMOSFET's electron channel. Those traps with energy levels far below  $E_{fn}$  are always filled by electrons, hence do not contribute to fluctuations. Similarly, those located far above  $E_{fn}$  are always empty and do not contribute charge fluctuations either. Therefore,  $S_{\Delta N_t}(y, f)$  can be approximated by

$$S_{\Delta N_t}(y, f) = N_t(E_{fn}) \cdot \frac{kT W_{eff} \Delta y}{\gamma f^{\text{EF}}} \quad (7.31)$$

where the parameter **EF** with a default value of 1.0 in BSIM4, provides the flexibility to model various process technologies including high- $k$  gate stacks.

With these established, the total channel-current flicker noise current spectral intensity is obtained by integrating over the channel

$$S_{I_{ds}}(f) = \frac{1}{L_{eff}^2} \int_0^{L_{eff}} S_{\Delta I_{ds}}(y, f) \cdot \Delta y \cdot dy \quad (7.32)$$

Substituting those that are already developed into the above equation gives

$$S_{I_{ds}}(f) = \frac{kT I_{ds}^2}{\gamma f^{\text{EF}} W_{eff} L_{eff}^2} \cdot \int_0^{L_{eff}} N_t(E_{fn}) \cdot \left[1 \pm \alpha \mu_{eff} \cdot \frac{N_{ch}(y)}{R(y)}\right]^2 \cdot \frac{R_{ch}^2(y)}{N_{ch}^2(y)} \cdot dy \quad (7.33)$$

The  $N_t(E_{fn}) \cdot \left[1 \pm \alpha \mu_{eff} \cdot \frac{N_{ch}(y)}{R(y)}\right]^2$  may be approximated by a polynomial of up to the second order,  $\text{NOIA} + \text{NOIB} \cdot N_{ch}(y) + \text{NOIC} \cdot N_{ch}(y)^2$  which was found to be scalable. The three parameters **NOIA**, **NOIB** and **NOIC** are called the equivalent trap density parameters. They allow the BSIM4 unified flicker noise model to scale down to the latest 20nm process node. The default values and units are given in the

## Section 7.3 BSIM4 Flicker Noise Models

parameter table at the end of this chapter. Upon completing the integration of Eq. (7.33), the BSIM4 FNOIMOD = 1 flicker noise spectral intensity in the strong inversion region becomes

$$S_{I_{ds,inv}}(f) = \frac{kTq^2\mu_{eff}I_{ds}}{\gamma f^{EF}A_{bulk}C_{oxe}L_{eff}^2} \cdot \left[ NOIA \cdot \log \left( \frac{N_{ch}(0)+N^*}{N_{ch}(L_{eff})+N^*} \right) + NOIB \cdot \left( N_{ch}(0) - N_{ch}(L_{eff}) \right) + NOIC \cdot \frac{N_{ch}(0)^2 - N_{ch}(L_{eff})^2}{2} \right] + \frac{kTI_{ds}^2}{\gamma f^{EF}L_{eff}^2 \cdot NF \cdot W_{eff}} \cdot \Delta L \cdot \frac{NOIA + NOIB \cdot N_{ch}(L_{eff}) + NOIC \cdot N_{ch}(L_{eff})^2}{[N_{ch}(L_{eff})+N^*]^2} \quad (7.34)$$

The first term on the right-hand side is derived for the linear operation region. The parameters therein are extracted from long channel devices and under low  $V_{ds}$ . The second term models the saturation operation region. Accordingly, the related parameters need to be extracted from both large and small devices and under low and high drain biases.

The carrier densities at the source and drain ends of the channel are

$$N_{ch}(0) = \frac{C_{oxe} \cdot V_{gsteff}}{q} \quad (7.35a)$$

and

$$N_{ch}(L_{eff}) = \frac{C_{oxe} \cdot V_{gsteff}}{q} \cdot \left( 1 - \frac{A_{bulk}}{V_{gsteff} + 2kT/q} \cdot V_{dseff} \right) \quad (7.35b)$$

which takes into account the bulk-charge effect. These two density terms were employed as the lower and upper limits of integration that leads to Eq. (7.34). The  $\Delta L$  term of Eq. (7.34) represents the reduction of the channel length caused by the channel length modulation. In that channel segment, the equivalent trap density and the amount of the channel charges are both different from those of the linear region [refer to Eq. (7.33)].  $\Delta L$  can be obtained by solving a quasi-2D Gaussian equation in the velocity saturation region near the drain. It is

$$\Delta L = l_c \cdot \log \left( \frac{EM + (V_{ds} - V_{dseff})/l_c}{E_{sat}} \right) \quad (7.36)$$

where  $\text{EM}$  and  $E_{sat}$  are the maximum and saturation electric field strengths of the channel.  $E_{sat} = 2 \cdot \text{VSAT}/\mu_{eff}$  [Refer to Chapters 2 and 3 for details about velocity saturation].  $l_c$  is the characteristic length of a CMOS technology (the smaller  $l_c$ , the better controlled the short-channel effects). It is proportional to the square root of the product of the junction depth and the gate dielectric thickness

$$l_c = \sqrt{3 \cdot XJ \cdot TOXE} \quad (7.37)$$

Throughout the development in the preceding sections of this chapter, the effective channel length  $L_{eff}$  for the flicker noise modeling should be understood to be the effective channel length used in the DC current models minus  $2 \cdot \text{LINTNOI}$ . The model parameter  $\text{LINTNOI}$  is defined as the gate-to-source/drain overlap length for the noise models. Introducing this parameter lends additional flexibility in accomplishing noise modeling accuracy.

In the sub-threshold region, the diffusion current component dominates. The BSIM4 model uses a single and continuous channel current  $I_{ds}$  formulation for all regions of operation. Hence, by employing the same approach as used for Eq. (7.32), the  $\text{FNOIMOD} = 1$  channel flicker noise current spectral intensity of the sub-threshold region is

$$S_{I_{ds,sub-vth}}(f) = \frac{\text{NOIA} \cdot kT \cdot I_{ds}^2}{\gamma f^{\text{EF}} \cdot L_{eff} \cdot \text{NF} \cdot W_{eff} \cdot N^{*2}} \quad (7.38)$$

Finally, the BSIM4  $\text{FNOIMOD} = 1$  flicker noise current spectral intensity model is

$$S_{I_{ds}}(f) = \frac{S_{I_{ds,inv}}(f) \times S_{I_{ds,sub-vth}}(f)}{S_{I_{ds,inv}}(f) + S_{I_{ds,sub-vth}}(f)} \quad (7.39)$$

Note that this model is smooth and continuous for all regions of operation and accounts for detailed device physics of MOSFET flicker noise. It has been used successfully in production for many generations of CMOS process technologies of the industry.

## 7.4 BSIM4 Channel Thermal Noise Models

There are multiple thermal noise sources in a MOSFET transistor such as the parasitic source and drain resistance. However, the channel thermal noise is the most significant and interesting and needs to be accounted for accurately in high-frequency analog or RF IC designs. (At low frequencies, the flicker noise dominates.) In the past two decades, many research efforts have been devoted to finding an effective way to model the channel thermal noise for SPICE simulations.

One model is the “ $\gamma$ -factor” based model [1], [11]. Note this  $\gamma$  is a different parameter from the attenuation coefficient of the electron wave introduced in the preceding section. This approach follows Nyquist’s theorem, which states that the noise current spectral intensity of an ohmic resistance is equal to  $4kT \cdot G_R = 4kT \cdot 1/R$ . Thus, the channel current thermal noise spectral intensity could be expressed in terms of a channel conductance. For instance, it can be given in terms of the channel transfer conductance as  $4kT \cdot \gamma \cdot g_m$ . The shortcoming of using  $g_m$  is apparent: When  $V_{ds} = 0$ ,  $g_m = 0$ , but the channel noise actually will be finite. Another similar model gives the noise spectral intensity as  $4kT \cdot \gamma \cdot g_{ds0}$  with  $g_{ds0}$  taken at  $V_{ds} = 0$ . In the latter case,  $\gamma$  was reported to be close to 1 in the linear region and approximately 2/3 in saturation. These were well-known theoretically and experimentally for nearly 50 years [S3, S8]. As channel length decreases in the recent and future technologies, these models are found to require improvements. In particular, the value of  $\gamma$  is found to become bias and gate length dependent, even if the conductance term such as  $g_m$  and  $g_{ds0}$  is taken to be the measured conductance. Another similar example is the BSIM4 TNOIMOD = 0 thermal noise model, which will be presented shortly.

All these models have one thing in common: The MOSFET channel is treated as a standalone resistor, albeit it is not ohmic. The BSIM4 TNOIMOD = 1 thermal noise model attempts to provide a more holistic model.

#### 7.4.1 The TNOIMOD=0 Charge-Based Model

This is the default channel thermal noise model of BSIM4. Its first version did not consider the bias-dependent LDD resistance effects which were first available in BSIM3v3. The thinking behind this charge-based model is simple: Integrating the channel resistance from the source end to the drain end of the channel by utilizing the BSIM4 inversion charge formulation.

Taking the  $y$ -axis in the source to drain direction with its origin located at the source end of the channel and using the charge carrier drift transport theory and constant current law by disregarding the very small generation-recombination-trapping DC currents at the interface traps, the channel current is given by

$$I_{ds} = W_{eff}\mu_{eff}C_{oxe} \cdot (V_{ggesteff} - A_{bulk} \cdot V_y) \cdot \frac{dV_y}{dy} \quad (7.40)$$

Since everywhere in the channel the differential channel resistance is  $dR_{ch} = dV_y/I_{ds}$ , multiplying both sides of Eq. (7.40) with  $dy/I_{ds}$  and then integrating from the source to the drain end of the channel will yield the total channel resistance  $R_{ch}$ , when assuming long channel and linear region of operation,

$$R_{ch} = \frac{L_{eff}^2}{\mu_{eff} \cdot |Q_{inv}|} \quad (7.41)$$

where  $Q_{inv}$  is the inversion channel charge in the entire channel area. After including the LDD bias-dependence resistance  $R_{ds}(V)$  that is in series with  $R_{ch}$ , the channel current thermal noise spectral intensity is

$$S_{I_{ds}}(f) = 4kT_{emp} \cdot \frac{1}{R_{ds}(V) + \frac{L_{eff}^2}{\mu_{eff} \cdot |Q_{inv}|}} \cdot \text{NTNOI} \quad (7.42)$$

where the parameter  $\text{NTNOI}$  is introduced to improve the model flexibility and accuracy. It defaults to 1. Note that  $Q_{inv}$  of Eq. (7.42) is similar to the inversion charge formulation of the BSIM4  $\text{CAPMOD} = 1$  model. It is

$$|Q_{inv}| = \text{NF} \cdot W_{eff,CV} \cdot L_{eff,CV} \cdot \left[ V_{gsteff} - \frac{A_{bulk,CV} \cdot V_{dseff}}{2} + \frac{A_{bulk,CV}^2 \cdot V_{dseff}^2}{12 \cdot (V_{gsteff} - 0.5 \cdot A_{bulk,CV} \cdot V_{dseff})} \right] \quad (7.43)$$

The LDD resistance  $R_{ds}(V)$  is taken into account in Eq. (7.42). In the BSIM4 model implementation in SPICE, that term is considered only when the BSIM4 model selector  $\text{RDSMOD}$  is set to 0 (the resistor is included in the channel current model). When  $\text{RDSMOD}$  is set to 1 in the model cards, that resistor is split into two components,  $R_s(V)$  and  $R_d(V)$ , each to be placed between the external and internal source nodes and the external and internal drain nodes, respectively. In this case ( $\text{RDSMOD}=1$ ), the contribution of  $R_s(V)$  and  $R_d(V)$  to noise is computed separately, as the noise voltage of any regular resistor,  $4kT_{emp} \cdot R$ . Therefore, the  $R_{ds}(V)$  term is dropped from Eq. (7.42) to avoid double counting when  $\text{RDSMOD} = 1$ .

This  $\text{TNOIMOD} = 0$  model still suffers from the same accuracy limitation of the other  $\gamma$ -based models. From the model derivation above, one can see that it is intended for long channel devices operating in the linear region. Secondly, the channel resistance is now treated as a regular resistor. However, the MOSFET channel is essentially a non-uniform, and RC-distributed network. This network introduces induced gate noise that is (partially) correlated via the gate capacitance with the channel noise above GHz. This will be discussed and modeled in the BSIM4  $\text{TNOIMOD} = 1$  thermal noise model in the next section.

#### 7.4.2 The $\text{TNOIMOD} = 1$ Holistic Thermal Noise Model

In the GHz range, the MOSFET channel must be considered as a series of distributed RC segments along the channel parallel to the y-axis shown in Fig. 7.6. The distributed gate capacitance represents capacitive coupling. The infinitesimal resistors are small chunks of the channel

resistance. Each resistance segment  $R_{ch,y}$  produces a thermal noise voltage  $\Delta v_{dy}$ , which can be modulated by the drain conductance and the gate and body trans-conductances specific to that segment ( $g_{dy}$ ,  $g_{my}$  and  $g_{mby}$ ) to produce a noise current at the drain,  $\Delta i_{dy}$ . Significant portion of this channel noise current can flow into the gate and the external circuit at GHz frequencies, producing a gate noise current  $i_g$  and therefore a gate noise voltage  $v_g$ . This gate noise is referred to as the induced gate noise. It can be partially correlated with the channel noise in phase and amplitude, because they are generated from the same source  $R_{ch,y}$ .

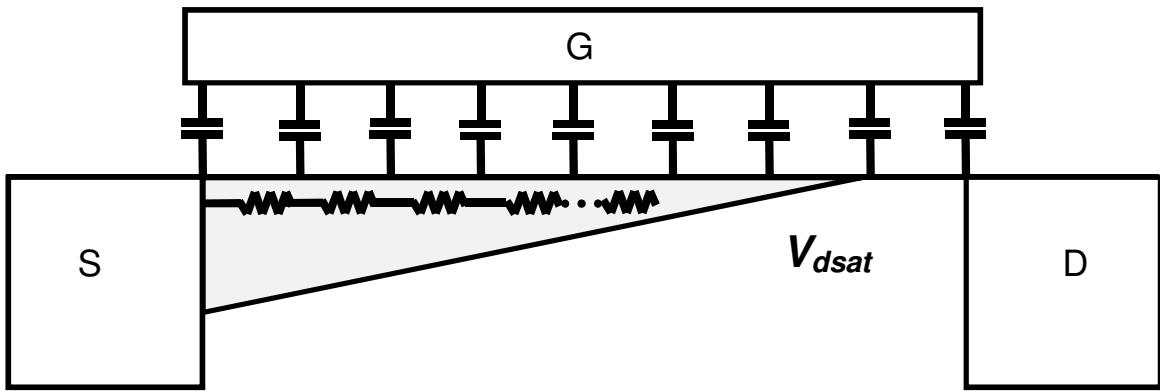


Fig. 7.6 The distributed gate and channel RC model of BSIM4 holistic thermal noise model.

Consider now the trans-conductance amplification effect. The noise current produced at drain by the resistor segment  $R_{ch,y}$  is

$$\Delta i_{dy} = \Delta v_{dy} \cdot (g_{dy} + g_{my} + g_{mby}) \quad (7.44)$$

The total output noise power  $\bar{i_d^2}$  at the drain is obtained by summing the noise power of each individual resistor segment. This requires the integration of Eq. (7.44) over the channel which is not an easy task.  $g_{dy}$  is independent of  $y$  because each delta  $v_{dy}$  simply appears at the drain. In order to carry out the integration, assume that the conductance term is independent of the channel location  $y$ . Eq. (7.44) thus gives the drain noise current power which is

$$\bar{i_d^2} \approx 4kT \cdot \Delta f \cdot (g_d + \beta \cdot g_m + \beta \cdot g_{mb})^2 \cdot \frac{V_{dseff}}{I_{ds}} \quad (7.45)$$

## Section 7.4 BSIM4 Channel Thermal Noise Models

where  $\beta$  is called the effective trans-conductance amplification coefficient.  $\beta g_m$  and  $\beta g_{mb}$  are the “average” front and back gate transconductances over all  $y$ .  $\beta$  is found to be geometry and bias dependent. This will be discussed shortly.

Note that both  $V_{dseff}$  and  $I_{ds}$  are readily known from the BSIM4 DC channel current model. This is advantageous since all the physical effects that have been included in the DC model are now automatically incorporated into the holistic noise model formulation.  $V_{dseff}/I_{ds}$  is simply the resistance of the channel, which must be distinguished from the small-signal channel resistance,  $g_{ds}$ . In the linear region, this resistance is found to be approximately equal to  $V_{ds}/I_{ds}$  whereas in saturation, it is simply clamped at  $V_{dsat}/I_{ds}$ . Note that the channel “pinch-off” region  $\Delta L$  in saturation does not produce thermal noise. Refer to [S1] and [S2].

Traditionally, the channel thermal noise model implementation in SPICE has been practiced by lumping the noise current between the source and drain nodes (terminals). The disadvantage of this topology treatment is obvious. Even at high frequencies, none of the noise current could flow into the gate through the gate capacitances, such as  $C_{gs}$ . Instead, all of it flows into and out of the source and drain terminals.

One approach has been suggested to deal with this problem: Adding a noisy bias-dependent resistor in series with  $C_{gs}$  between the gate and source nodes [1]. The approach was cited by many researchers but it has not found a good production use [12], [13]. The shortcomings are the following. From the device physics point of view, that resistor is empirically created to produce an induced gate noise, but the input circuit may be slowed down as a victim. In model parameter extractions, curve fitting determines the resistance without guidance on bias and device size effects. In terms of SPICE element topologies, one additional internal node would need to be created to connect that resistor to  $C_{gs}$ .

BSIM4 takes a different approach. The channel thermal noise gives a total noise voltage power of  $4kT_{emp} \cdot V_{dseff}/I_{ds} \cdot \Delta f$ . In the consideration of the distributed-RC nature of the channel region, it is rational to split that total amount into two components: One is placed on

the source side for the forward-mode operation or on the drain side for the reverse mode operation (i.e., when  $V_{ds} < 0$ ). That is, this component is placed outside the channel region. The other component still remains inside. Let this process be termed noise partitioning. The partitioning is intended to model the induced gate noise current through the distributed RC network between the channel and the gate. To make the discussion of this approach simple to understand, refer to Fig. 7.7.

Suppose that there exists a noise partitioning coefficient  $\theta$  and that the noise voltage power component to be placed outside the channel is  $4kT_{emp} \cdot \theta^2 \cdot V_{dseff}/I_{ds} \cdot \Delta f$  [The model for  $\theta$  will be discussed soon]. In the saturation region for instance, the gate noise current power induced by this noise voltage source is

$$\overline{i_g^2} \approx 4kT_{emp} \cdot \theta^2 \cdot \frac{V_{dseff}}{I_{ds}} \cdot \omega^2 C_{gs}^2 \cdot \Delta f \quad (7.46)$$

Note that the contribution from  $C_{gd}$  is ignored because it is very small compared to  $C_{gs}$  in the saturation operation region (note that  $C_{gs}$  includes both the intrinsic and extrinsic capacitances such as the overlap capacitance). The drain noise current due to the noise source partitioned to the source side, after the amplification by the transistor small-signal conductances, is

$$\overline{i_{d,part,sd}^2} = 4kT \cdot \theta^2 \cdot \frac{V_{dseff}}{I_{ds}} \cdot (g_d + g_m + g_{mb})^2 \cdot \Delta f \quad (7.47)$$

The amplification here is modeled in a fashion analogous to a simple source follower. The total drain thermal noise current caused by the channel thermal noise has been derived earlier and is given in Eq. (7.45). Thus, the amount of the noise current power that needs to be placed in the intrinsic channel region is obtained by subtracting Eq. (7.47) from Eq. (7.45)

$$\overline{i_{d,part,ch}^2} = \overline{i_d^2} - \overline{i_{d,part,sd}^2} \quad (7.48)$$

The source or drain noise partition implementation into SPICE cannot be carried out in the form of a voltage source as  $4kT \cdot \theta^2 \cdot V_{dseff}/I_{ds}$ . It is known that implementing a voltage source in SPICE is usually more expensive than implementing a current source in terms of circuit matrix solving. In addition, it would require an additional internal node in order to connect to the source or drain diffusion or LDD resistors, in the case

## Section 7.4 BSIM4 Channel Thermal Noise Models

of the TNOIMOD = 1 model. In the BSIM4 SPICE implementation, the resistance term  $\theta^2 \cdot V_{dseff}/I_{ds}$  of that voltage source is converted to a voltage-controlled current source, namely a non-linear current source in parallel with the source (or drain depending on the sign of  $V_{ds}$ ) conductance. This is illustrated in Fig. 7.7(b). This avoids creating the internal node that would otherwise be required. The non-linear current source is activated only when a SPICE noise analysis is to be performed. Unlike Eq. (7.48), the non-linear current source is implemented without the trans-conductance amplification term ( $g_d + g_m + g_{mb}$ ). This is because the amplification action takes place inherently through the intrinsic MOSFET equivalent circuit inside SPICE.

Our discussion will now turn to the modeling of the effective transconductance amplification coefficient  $\beta$  of Eq. (7.45) and the noise partitioning coefficient  $\theta$  of Eqs. (7.46) and (7.47). For a long channel device, it is found that they are about 0.577 and 0.5164 (dimensionless), respectively. These two numbers can be obtained by equating  $i_d^2$  to that computed from the BSIM4 TNOIMOD = 0 charge-based thermal noise model and by equating  $i_g^2$  to the induced gate noise current of  $16kT\omega^2C_{gs}^2/15g_d$  [1], respectively. For short channel devices, both coefficients are dependent on biases, increasing as  $V_{gs}$  increases as shown in Fig. 7.8. This is because the velocity saturation effect starts to kick in and the channel-length modulation becomes stronger, which makes the transconductance amplification as well as the impedance composed of the series  $C_{gs}$  and channel resistance smaller.  $\beta$  and  $\theta$  are formulated as

$$\beta = \text{RNOIA} \cdot \left( 1 + \text{TNOIA} \cdot L_{eff} \cdot \frac{V_{gsteff}^2}{(E_{sat} \cdot L_{eff})^2} \right) \quad (7.49)$$

and

$$\theta = \text{RNOIB} \cdot \left( 1 + \text{TNOIB} \cdot L_{eff} \cdot \frac{V_{gsteff}^2}{(E_{sat} \cdot L_{eff})^2} \right) \quad (7.50)$$

where  $(E_{sat} \cdot L_{eff})$  is the saturation voltage of long-channel devices, which is equal to  $2L_{eff} \cdot \text{VSAT}/\mu_{eff}$ .

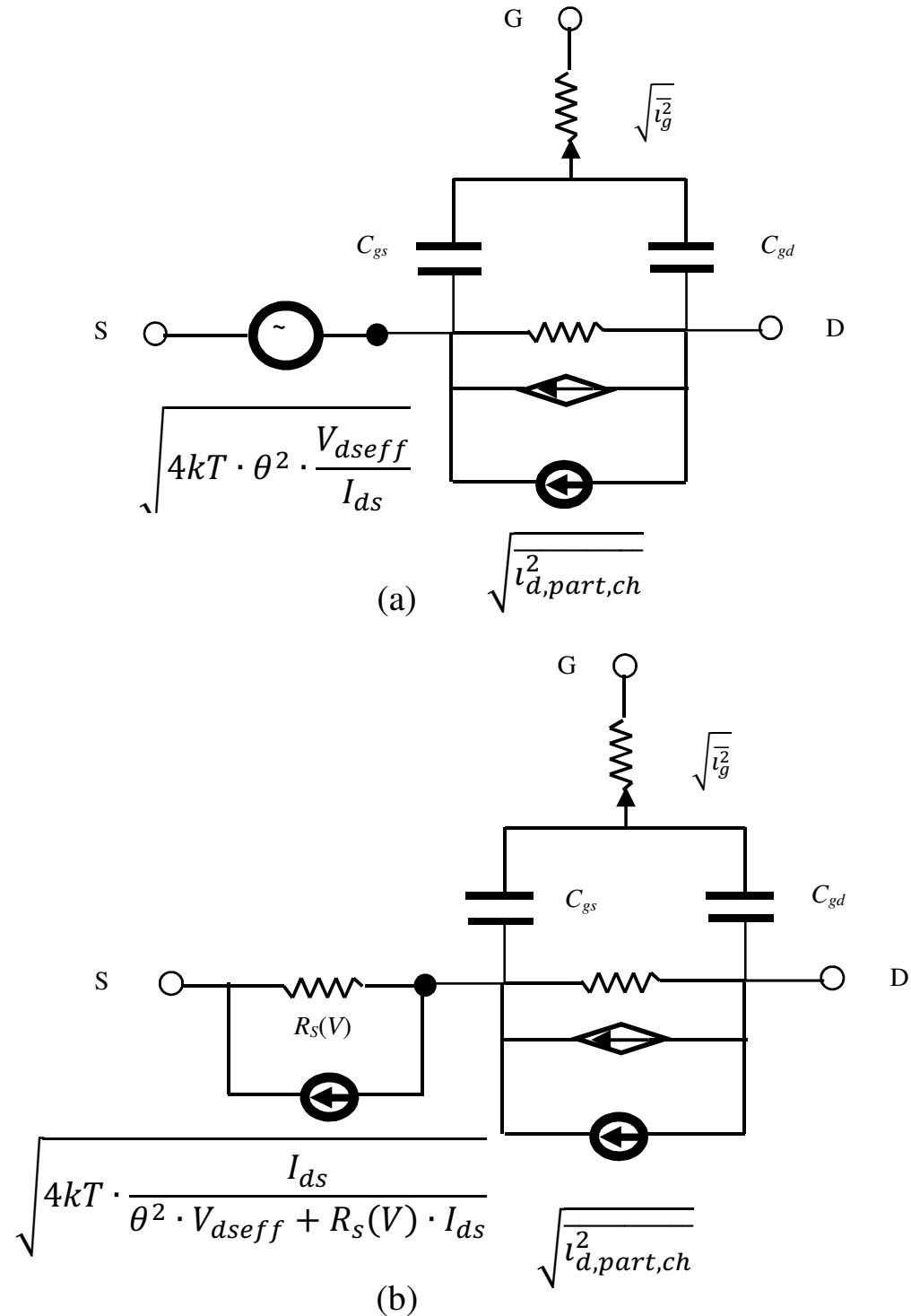


Fig. 7.7 The concept and topology of the BSIM4 TNOIMOD = 1 holistic thermal noise model. (a) The concept of noise voltage partitioning into the source and channel noises. (b) SPICE implementation topology in the form of noise current rather than noise voltage (The forward-bias mode is shown here).

## Section 7.4 BSIM4 Channel Thermal Noise Models

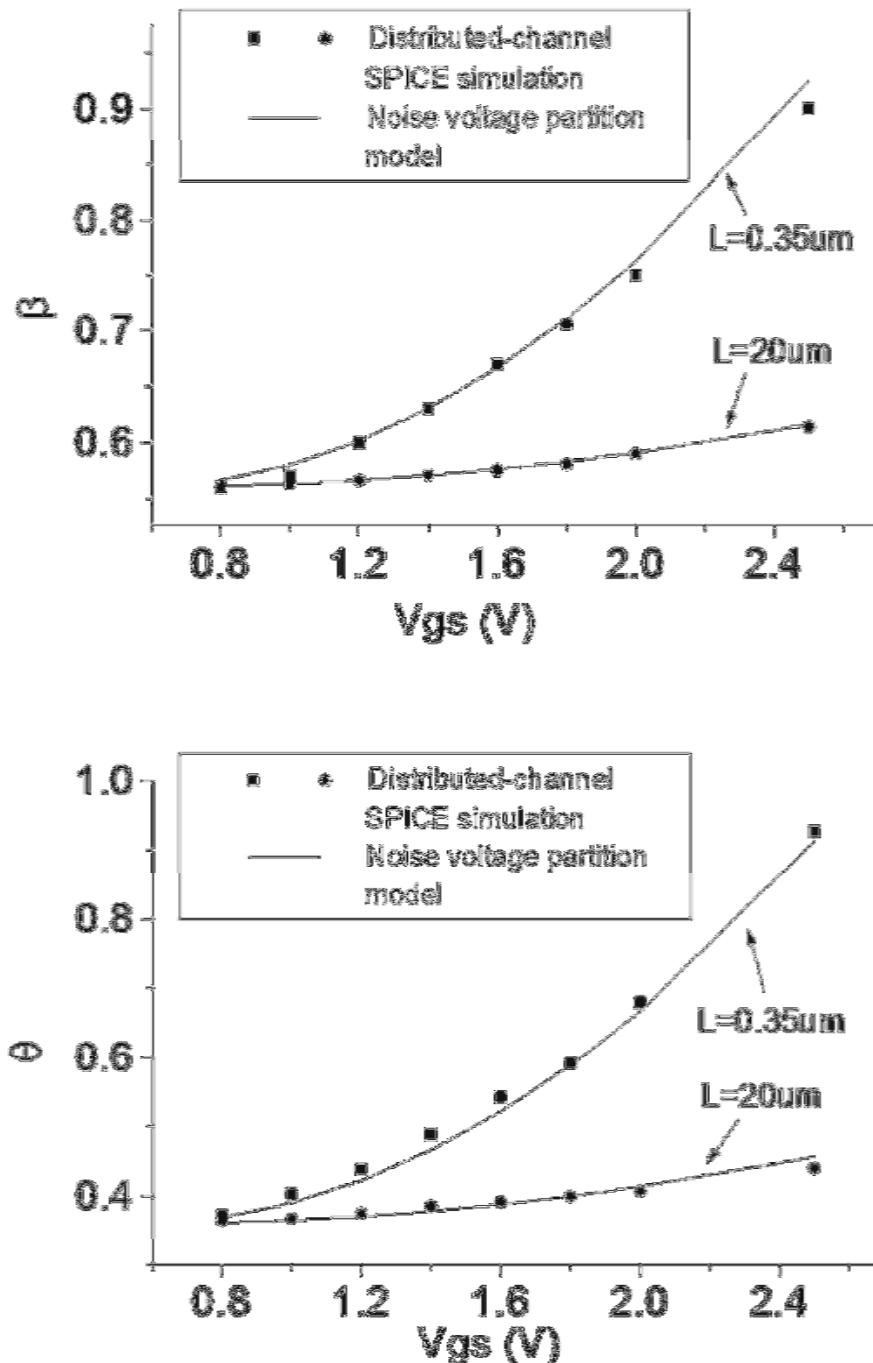


Fig. 7.8 The bias and geometry dependencies of  $\beta$  (at the top) and  $\theta$  (at the bottom). For short channel devices, these two coefficients increase with  $V_{gs}$  due to velocity saturation and channel length modulation.

Equations (7.49) and (7.50) can be verified with a SPICE simulation experiment. To simulate the RC-distributed nature of the MOSFET channel (Fig. 7.6), one can divide the channel into, for instance, ten

segments and represent them with a 10-MOSFET cascaded chain (Fig. 7.9) each having 1/10 the channel length of the original channel length. The cascade chain is simulated using SPICE with all short-channel effects ignored. This setup provides a way to check how the induced gate noise, the noise partitioning and the trans-conductance amplification work under the hood by sweeping the gate and drain voltage for various possible operation conditions of a MOSFET, in particular, the saturation region. After that, the transistor is simulated (in one segment) with the TNOIMOD = 1 holistic thermal noise model and the result should match the segmented (distributed) channel simulation result as shown in Fig. 7.8. The holistic thermal noise model has also been verified with measurement data of MOSFETs having different geometries and under a wide range of bias conditions.

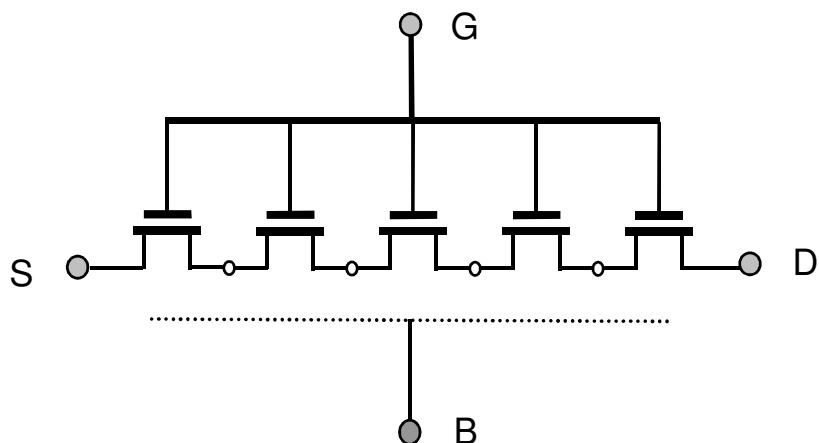


Fig. 7.9 Cascaded segments of a MOSFET transistor employed for the development and verification of the BSIM4 TNOIMOD = 1 holistic channel thermal noise model. The number of segments can be ten or more for better accuracy.

As mentioned earlier,  $\overline{i_g^2}$  and  $\overline{i_d^2}$  are correlated as given by Eq. (7.9). SPICE3 and some commercial SPICE simulators do not support the modeling of a general noise correlation problem but rather on an ad-hoc basis only when the model formulations of  $\overline{i_g^2}$  and  $\overline{i_d^2}$  are known a priori. Furthermore, the accurate formulation of the correlation coefficient becomes more difficult to develop for advanced MOSFETs owing to their complex model topologies and physical effects.

In the case of the BSIM4 TNOIMOD = 1 model, however, one can think of extracting the noise partitioning coefficient  $\theta$  from correlation data measured at high frequencies. This is in contrast to the extraction methodology proposed earlier, namely an extraction from the induced gate noise directly.

One alternative approach to extracting  $\theta$  from the noise correlation is to make use of the SPICE simulation setup of Fig. 7.9. The difference is that each MOSFET needs to be made noiseless and a very small noise voltage source with a random phase is inserted between any two consecutive MOSFETs. Random phases can be designated to each voltage source by using Monte Carlo and random phase number generation. After this, all the voltage sources are then made in phase and SPICE simulations are repeated to output the  $i_g^2$  and  $i_d^2$  simulation data. This is a good methodology in the research and development of MOSFET transistor noise models.

## 7.5 Other Noise Sources

In addition to the modeling of the channel flicker and thermal noise, induced gate noise, and noise correlation, BSIM4 also models the thermal noise contributed by the parasitic resistors of the drain, gate (in the case of RGATEMOD = 1), source, and substrate (for the RBODYMOD = 1 and 2 selections) in the form of  $4kT \cdot R \cdot \Delta f$ . Shot noise in the form of  $2q \cdot I \cdot \Delta f$  due to the gate direct tunneling currents and the junction diode DC currents is modeled as well.

## 7.6 Chapter Summary

This chapter first discussed the physics fundamentals of MOSFET noise, and then the parametric representation and characterization by a two-port circuit network and SPICE implementations. It focused on the methodology of the BSIM4 unified channel flicker noise, channel thermal noise, and induced gate noise modeling. Particular efforts were made in presenting and analyzing the BSIM4 holistic channel thermal noise model, where a novel and simulation efficient noise partitioning technique was developed for the first time in compact modeling to accurately model the correlation between the channel thermal noise and

the induced gate noise. This holistic thermal noise model considers the same fundamental property of MOSFET as does the intrinsic-input gate resistance  $R_{g,ii}$  model presented in Chapter 6: The MOSFET gate and channel region is a non-linear, distributed RC network. In addition, this chapter also presented useful noise measurement, characterization and model parameter extraction methodologies.

## 7.7 Parameter Table

Name (type)	Description and default	Can be binned?	Note
FNOIMOD (Global; integer)	Flicker noise model selector.  Default = 1; dimensionless. The other optional value is 1.	No	-
KF (Global; double)	The FNOIMOD = 0 flicker noise coefficient.  Default = 1.0 in $[A^{2-AF} \cdot s^{1-EF} \cdot \text{Farad}]$ .	No	-
AF (Global; double)	The FNOIMOD = 0 flicker noise channel current exponent.  Default = 1.0; dimensionless.	No	-
EF (Global; double)	Flicker noise frequency exponent.  Default = 1.0; dimensionless.	No	-
NOIA (Global; double)	The FNOIMOD = 1 flicker noise equivalent trap density parameter.  Default = NMOS: $6.25 \times 10^{41} (\text{eV})^{-1} \cdot \text{s}^{1-EF} \cdot \text{m}^{-3}$ ; PMOS: $6.188 \times 10^{40} (\text{eV})^{-1} \cdot \text{s}^{1-EF} \cdot \text{m}^{-3}$ .	No	-
NOIB (Global; double)	The FNOIMOD = 1 flicker noise equivalent trap density parameter.  Default = NMOS: $3.125 \times 10^{26} (\text{eV})^{-1} \cdot \text{s}^{1-EF} \cdot \text{m}^{-1}$ ; PMOS: $1.5 \times 10^{25} (\text{eV})^{-1} \cdot \text{s}^{1-EF} \cdot \text{m}^{-1}$ .	No	-

<b>NOIC</b> (Global; double)	The <b>FNOIMOD</b> = 1 flicker noise equivalent trap density parameter.  Default = $8.75 \times 10^9$ in [ $(\text{eV})^{-1} \cdot \text{s}^{1-\text{EF}} \cdot \text{m}$ ].	No	-
<b>EM</b> (Global; double)	Maximum channel electric field strength of the <b>FNOIMOD</b> = 1 flicker noise model.  Default = $4.1 \times 10^7$ in [ $\text{V} \cdot \text{m}^{-1}$ ].	No	-
<b>LINTNOI</b> (Global; double)	The gate-to-source/drain overlap length for the <b>FNOIMOD</b> = 1 flicker noise model.  Default = 0.0 in [m].	No	Fatal error if it makes the effective channel length negative for noise modeling and simulation.
<b>NTNOI</b> (Global; double)	A <b>TNOIMOD</b> = 0 channel thermal noise spectral intensity coefficient.  Default = 1.0; dimensionless.	No	If negative, a warning will be issued and its value will be reset to 0.
<b>RNOIA</b> (Global; double)	Noise partitioning coefficient for the <b>TNOIMOD</b> = 1 channel thermal noise model.  Default = 0.5164; dimensionless.	No	If negative, a warning will be issued and its value will be reset to 0.
<b>RNOIB</b> (Global; double)	Transconductance amplification coefficient for the <b>TNOIMOD</b> = 1 channel thermal noise model.  Default = 0.577; dimensionless.	No	If negative, a warning will be issued and its value will be reset to 0.
<b>TNOIA</b> (Global; double)	Bias and length dependence parameter for the <b>TNOIMOD</b> = 1 thermal noise model partitioning.  Default = 1.5; dimensionless.	No	If negative, a warning will be issued and its value will be reset to 0.
<b>TNOIB</b> (Global; double)	Bias and length dependence parameter for the noise transconductance amplification of the <b>TNOIMOD</b> = 1 thermal noise model.  Default = 3.5; dimensionless.	No	If negative, a warning will be issued and its value will be reset to 0.

## References

- [1] A. van der Ziel, "Noise in Solid State Devices and Circuits," New York: John Wiley & Sons, 1986.
- [2] K. K. Hung, P. K. Ko, C. Hu, and Y. C. Cheng, "Random telegraph noise of deep-submicrometer MOSFET's," *IEEE Electron Device Letters*, vol. 11, no. 2. February 1990.
- [3] D. K. Shaeffer, and T. H. Lee, "A 1.5V 1.5GHz CMOS low noise amplifier," *IEEE J. Solid-State Circuits*, vol. 32, pp. 745-759, 1997.
- [4] L. W. Nagel, "SPICE2: A Computer Program to Simulate Semiconductor Circuits," Memorandum No. UCB/ERL-M520, Electronics Research Laboratory, College of Engineering, University of California, Berkeley, May 1975.
- [5] Thomas L. Quarles, "The SPICE3 Implementation Guide," Memorandum No. UCB/ERL-M89/44, Electronics Research Laboratory, College of Engineering, University of California, Berkeley, April 24, 1989.
- [6] Thomas L. Quarles, "Analysis of Performance and Convergence Issues for Circuit Simulation," Memorandum No. UCB/ERL-M89/42, Electronics Research Laboratory, College of Engineering, University of California, Berkeley, April 1989.
- [7] Thomas L. Quarles, "Adding Devices to SPICE3," Memorandum No. UCB/ERL-M89/45, Electronics Research Laboratory, College of Engineering, University of California, Berkeley, April 24, 1989.
- [8] Weidong Liu, and Chenming Hu, "BSIM3v3 MOSFET Model" – Silicon and Beyond: Advanced Device Models and Circuit Simulators, edited by Michael S. Shur and Tor A. Fjeldly, pp. 1-31, ISBN: 981-02-4280-8, World Scientific, 2000.
- [9] Weidong Liu, Xiaodong Jin, James Chen, Min-Chie Jeng, Zhihong Liu, Yuhua Cheng, Kai Chen, Mansun Chan, Kelvin Hui, Jianhui Huang, Robert Tu, Ping K. Ko, and Chenming Hu, "BSIM3v3.2 MOSFET model — Users' manual," Memorandum No. UCB/ERL M98/51. Electronics Research Laboratory, College of Engineering, University of California, Berkeley, August 21, 1998.
- [10] K. K. Hung, P. K. Ko, C. Hu, and Y. C. Cheng, "A physics-based MOSFET noise model for circuit simulators," *IEEE Trans. Electron Devices*, vol. 37, no. 5. May 1990.
- [11] Y. P. Tsividis, "Operation and Modeling of the MOS Transistor," 2nd ed., McGraw-Hill, Inc., 1999.
- [12] C. H. Chen, and M. J. Deen, "A general noise and S-parameter de-embedding procedure for on-wafer high-frequency noise measurements of MOSFETs," *IEEE Trans. on Microwave Theory and Techniques*, vol. 49, no.5, pp. 1004-1005, May 2001.
- [13] C. Enz, and Y. Cheng, "MOS transistor modeling for RF IC design," *IEEE J. Solid-State Circuits*, vol. 35, pp. 186-201, Feb. 2000.

- [S1] Chih-Tang Sah, "Theory of low-frequency noise in junction-gate field-effect transistors," IEEE Transaction on Electron Devices, 11(7), 324-345, July 1964.
- [S2] C. T. Sah, S. W. Wu and F. H. Hielscher, "The effects of fixed bulk charge on the thermal noise in MOS transistors," IEEE Trans. Electron Devices, 13(4), 416-420, April 1966.
- [S3] Chih-Tang Sah and Frank H. Hielscher, "Evidence of the surface origin of the 1/f noise," Physical Review Letters, 17(18), 956-958, 31 October 1966.
- [S4] Leopaldo D. Yau and Chih-Tang Sah, "Theory and experiments of low-frequency generation-recombination noise in MOS transistors," IEEE Trans. Electron Devices, 16(2), 170-177, February 1969.
- [S5] H. S. Fu and C. T. Sah, "Lumped model analysis of the low-frequency generation noise in gold-doped silicon junction-gate field-effect transistors," Solid-State Electronics, 12(4), 605-618, April 1969.
- [S6] L. D. Yau and C. T. Sah, "Observation of the ideal generation-recombination noise spectrum and spectra with voltage variable relaxation time in gold-doped silicon," Applied Physics Letters, 14, 267-269, 1 May 1969.
- [S7] L. D. Yau and C. T. Sah, "Geometrical dependences of the generation-recombination noise in gold-doped silicon MOS transistors," Solid-State Electronics, 12(11), 903-905, November 1969.
- [S8] L. D. Yau and C. T. Sah, "On the excess white noise in MOS transistors," Solid-State Electronics, 12(12), 927-936, December 1969.
- [S9] L. D. Yau and C. T. Sah, "Temperature and field dependences of the generation-recombination noise and thermal emission rates at the gold acceptor center in silicon," Solid-State Electronics 13(9), 1213-1218, September 1970.
- [S10] Chih-Tang Sah, "Equivalent circuit models in semiconductor transport for thermal, optical, Auger-impact and tunneling recombination-generation-trapping processes," Physica Status Solidi, (a)7, 541-559, 16 October 1971.
- [S11] H. S. Fu and C. T. Sah, "Theory and experiments on surface 1/f noise," IEEE Transaction on Electron Devices, 19(2), 273-285, February 1972.

**This page intentionally left blank**

## Chapter 8

# Source and Drain Parasitics: Layout-Dependence Model

### 8.1 Introduction and Chapter Objectives

The layout-dependent effects (LDE) due to mechanical stress on the intrinsic characteristics of a MOS transistor have been discussed in Chapter 2. Similarly important are the unavoidable parasitic leakage currents, resistance and junction capacitances associated with the MOSFET source and drain. These parasitic effects have significant impacts on the device performance of the advanced CMOS technologies. This is a fun chapter that uses simple arithmetics to describe the geometries of source and drain. Multiple transistors may be connected in parallel and/or in series in many possible combinations with sources and drains.

A production-worthy MOSFET SPICE model must take into account the source, drain, and contact geometries and connectivity that are employed in today's IC designs. The model must ensure accurate, convenient and consistent modeling of the parasitic effects in both pre- and post-layout simulations.

The traditional approach of merely specifying the source and drain perimeters and areas, namely **PS**, **PD**, **AS**, and **AD**, is overly simplistic for the modeling of advanced CMOS technologies. This chapter discusses the source/drain/contact implementations that are used in practice and the comprehensive layout-dependence modeling methodology that BSIM4 brings forth. This methodology resulted from the joint efforts of the Compact Model Council and the BSIM research

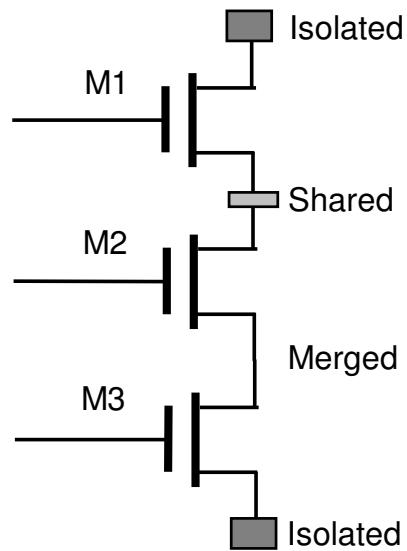
team at UC Berkeley. Together they produced an industry standard for modeling MOSFET parasitic effects for the first time.

In this chapter, three types of source and drain connections (that is, *isolated*, *shared*, and *merged*) are presented. The discussion first assumes the transistors are non-fingered (or interchangeably, single-fingered) MOSFETs that each have only a single gate finger. The single-fingered transistors are routinely used in silicon implementations for small width transistors. That will be followed by a discussion of multi-fingered devices (the number of device gate fingers is greater than one) that are used to implement large width transistors with the advantage of creating less parasitic effect. Various end-source and end-drain connection scenarios permitted by BSIM4 are then presented and analyzed. The source and drain perimeter and area geometry computations corresponding to each scenario are presented next. Applications of this BSIM4 geometry calculation method to junction saturation current and zero-bias capacitance are discussed subsequently. Finally, three contact types of the end-source and end-drain regions (i.e., *point*, *wide* and *no contacts*) are presented.

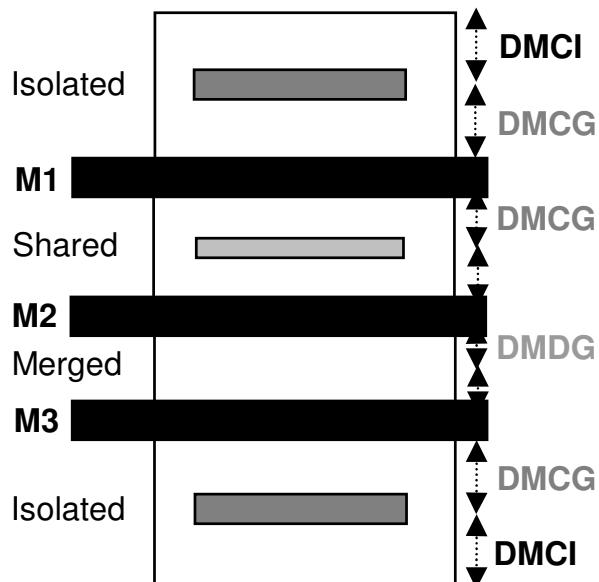
## 8.2 Connections of a Multi-Transistor Stack

In analog and RF circuits, several MOSFET transistors are often stacked in series to achieve the desired gains and output resistances. The source and drain contacts may be of one of three types: *Isolated*, *shared*, and *merged*. This is schematically illustrated in Fig. 8.1 (a) with the layout shown in Fig. 8.1 (b).

The gate and body (as opposed to source and drain) contacts and their connections, in the case of non-fingered devices, are relatively uninteresting and straightforward. In the case of multi-finger devices, the gate and body contacts have been discussed in Chapter 6. Therefore, they are not repeated in this chapter. In Fig. 8.1, the *isolated* source and drain connections at the top and bottom ends of the series are often required to connect to adjacent circuit blocks,  $V_{dd}$  and the ground. The *shared* source



(a)



(b)

Fig. 8.1 Possible MOSFET series connections considered in BSIM4. The black bars are gates, spaces between the gates are source or drain regions, and the gray bars represent the contact holes. (a) The schematics and (b) the layout. The BSIM4 parameters that specify the source and drain geometries for these connections are illustrated. By definition, the merged source and drain region has no contact. For ease of illustration, the circular contact (array) commonly used in CMOS processes and the stripe or slot contact found in the advanced CMOS processes are both represented with the gray bars.

and drain diffusion region between transistors M1 and M2 is employed to bring the output signal from this node to the next logic gate. The *merged* source and drain requires no contact and, hence, can use the minimum size source and drain diffusion region to reduce the parasitic capacitance and resistance. The region between transistors M2 and M3 is an example of a merged source and drain diffusion region.

Several BSIM4 geometrical parameters are noted in Fig. 8.1. **DMCI** measures the distance from the mid-contact (the contact center) of an isolated source or drain diffusion region (at the top or bottom of the transistor stack) to the isolation edge. **DMDG** designates the distance measured from the mid-diffusion region or the center of that region to the gate edge in the case of a *merged* source and drain connection. **DMCG** measures the distance from the mid-contact or the contact center of an isolated or shared source and/or drain diffusion region to the gate edge. In practice, **DMCG** is usually but not always made as small as allowed by design rules. Note that all these distances are measured in the channel-length direction.

One parameter of BSIM4 left out in Fig. 8.1 is **DMCGT**. It has the same definition as **DMCG**, but it is intended only for test device structures (rather than an actual transistor in a circuit) that are employed for model parameter extractions. In test device structures, the distance from the mid-contact or the contact center of an isolated or shared source and/or drain diffusion region to its gate edge, namely **DMCGT**, can be made the smallest possible to minimize any other parasitic effects associated with the source and drain diffusions on the characterization and model parameter extraction of the intrinsic portion of a device. Nevertheless, one may and should subtract **DMCGT** from the actual values of **DMCG** to avoid double counting in calculating source and drain geometry and resistance when using BSIM4 since the **DMCGT** portion of source and drain diffusions is already taken into account in the intrinsic BSIM4 model extractions.

One should use these BSIM4 geometrical parameters in lieu of the traditional source and drain perimeter and area parameters such as **PS**,

PD, AS and AD for better accuracy. The latter set of parameters does not provide sufficient information to satisfy the accuracy requirement of analog and RF CMOS circuit simulations today. A description of how to use them to calculate the effective perimeter, area, and resistance calculations will follow shortly.

The structure shown in Fig. 8.1 contains three series connected transistors, each of which has only one gate finger (the black bars.) Each has one source region and one drain region, although its source/ drain region may be shared with neighbor transistors. In Section 8.3, a single transistor will employ multiple gate fingers and multiple source and/or drain regions.

### 8.3 Source and Drain of a Transistor With Multiple Gate Fingers

A multi-finger transistor is simply a wide width, say 10 micron  $W$ , transistor that is comprised of ten 1 micron wide transistors connected in parallel. Doing so can reduce the source/drain area and therefore reduce the parasitic capacitance. It also reduces the gate electrode resistance (Chapter 6) significantly. Both improve the speed of the transistor. Multi-finger design is almost always used for wide  $W$  transistors and is indispensable in high-speed logic and analog circuits including, of course, RF circuits.

Fig. 8.2, for example, may be two 5-micron MOFSET connected in parallel to function as a 10-micron  $W$  MOSFET. In the case of multi-finger transistors, all inside (not end) diffusion regions have contact holes and therefore are *shared*. Each diffusion region is either a source or a drain, but not both simultaneously. Also, all the drain regions, all the gate fingers, and all the source regions are electrically tied together, respectively. A multi-finger device is a single device and has one gate node, one drain node, and one source node. SPICE simulators also treat it as one transistor for efficient SPICE simulation with four common nodes (that is, drain, gate, source and body). Of course, a multi-finger device will reduce to a non-finger device when it has a single finger (the number of its fingers  $NF = 1$ ).

Assume that  $NF$  is two (or any even number). The number of gate stripes is two as sketched in Fig. 8.2. It is easy to see that the end diffusion regions of this device can be both source regions (Fig. 8.2 (a)), or both drain regions (Fig. 8.2 (b)). One thus needs to distinguish them by having a local instance parameter (**MIN**) in the element card to indicate if the end diffusions of this particular device are considered to be source (by setting  $\text{MIN} = 0$ , which is the default case) or drain ( $\text{MIN} = 1$ ).

The parameter **MIN** is not required for a device that has an odd number of fingers because the parasitics of the two nodes are identical. The two cases in Fig. 8.3 are equivalent electrically. Therefore there is no need to distinguish them.

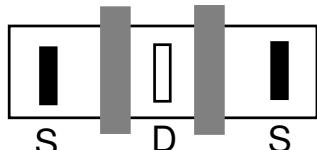
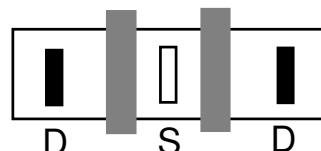
(a)  $NF = 2$  and  $\text{MIN} = 0$ (b)  $NF = 2$  and  $\text{MIN} = 1$ 

Fig. 8.2 A multi-finger MOSFET device with an even number of fingers ( $NF = 2$  in this case) has “symmetrical” end contacts (black bars): Either both are sources or both are drains at the same time.

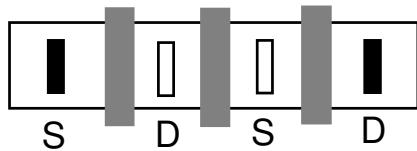
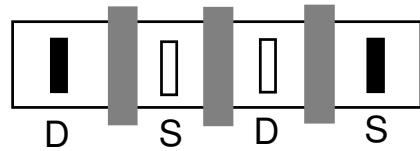
(a)  $NF = 3$ (b)  $NF = 3$ 

Fig. 8.3 A multi-finger MOSFET device with an odd number of fingers ( $NF = 3$  in this case) is “asymmetrical”.

## 8.4 GEOMOD: The End-Source and End-Drain of a Multi-Finger Transistor

Recall from the preceding section that each diffusion region of a multi-finger transistor is either a source or a drain and has a contact (*shared*). A multi-finger device is usually part (one transistor) of a multi-transistor stack. For example it may take the position of M1, M2, and/or M3 in Fig. 8.1(b). For this reason, one needs a way to specify the contact type of the end diffusions of a multi-finger device. BSIM4 provides eleven possible combinations as listed in Table 8.1. Each of them can be selected by means of specifying a connection type parameter called **GEOMOD** globally in a model card or locally in an element statement of a SPICE net-list. Some of them (**GEOMOD** = 0, 3 and 8) apply to any **NF** values, even or odd, with the rest to either even **NF** only or odd **NF** only. These end-source and end-drain connection types in conjunction with **NF** and **MIN** determine the number of end and inner diffusions. They also help to determine the effective values of the source and drain junction perimeters and areas, as well as the effective source and drain diffusion resistances. The next section will show how the perimeters and areas are computed based upon the **GEOMOD** selections. Examples of **GEOMOD** applications to cascoding non-finger and multi-finger devices will then follow.

Table 8.1 GEOMOD and the end source and drain connection scenarios of BSIM4.

GEOMOD	Connection type of end sources	Connection type of end drains	Applicable NF
0	Isolated (if any)	Isolated (if any)	Even or odd
1	Isolated	Shared	Odd
2	Shared	Isolated	Odd
3	Shared (if any)	Shared (if any)	Even or odd
4	Isolated	Merged	Odd
5	Shared	Merged	Odd
6	Merged	Isolated	Odd
7	Merged	Shared	Odd
8	Merged (if any)	Merged (if any)	Even or odd
9	Shared/Isolated	No end drains	Even only
10	No end sources	Shared/Isolated	Even only

## 8.5 Source and Drain Area and Perimeter Calculation

Suppose that a GEOMOD value is known (either by default or specified). The numbers of the end-source ( $nEndS$ ) and end-drain ( $nEndD$ ) diffusions as well as the numbers of any inner ones ( $nInS$  and  $nInD$ ) can be obtained as follows.

If the device has an odd number of fingers (including  $NF = 1$ ), it is certain that this device has both end-source and end-drain diffusions and that  $nEndS = nEndD = 1$ . One can also prove that the numbers of the inner source and the inner drain diffusions in this case are equal and can be mathematically expressed by

$$nInS = nInD = 2 \cdot \text{MAX}\left(\frac{NF-1}{2}, 0\right) \quad (8.1)$$

In the case of an even  $NF$ , one would then need to specify if both end-diffusions are the source or the drain. If they are the source, specifying  $\text{MIN} = 0$  in the instance line obtains these results:  $nEndS = 2$ ,  $nEndD = 0$ ,  $nInD = NF$ , and

$$nInS = 2 \cdot \text{MAX}\left(\left(\frac{NF}{2}-1\right), 0\right) \quad (8.2)$$

In a similar way, if both end-diffusions are drain, specifying  $\text{MIN} = 1$  leads to  $nEndS = 0$ ,  $nEndD = 2$ ,  $nInS = NF$ , and

$$nInD = 2 \cdot \text{MAX}\left(\left(\frac{NF}{2}-1\right), 0\right) \quad (8.3)$$

Note that the above discussion applies only to  $\text{GEOMOD} = 0$  through 8, because  $\text{GEOMOD} = 9$  and 10 are concerned with an even  $NF$  exclusively.

Before proceeding to computing the total effective junction perimeter and area of all source and drain diffusion regions of a device, it is useful to formulate these quantities for the diffusion regions of *isolated*, *shared*, and *merged* connection scenarios. For this purpose, Fig. 8.1 (b) is redrawn into Fig. 8.4 for ease of reference.

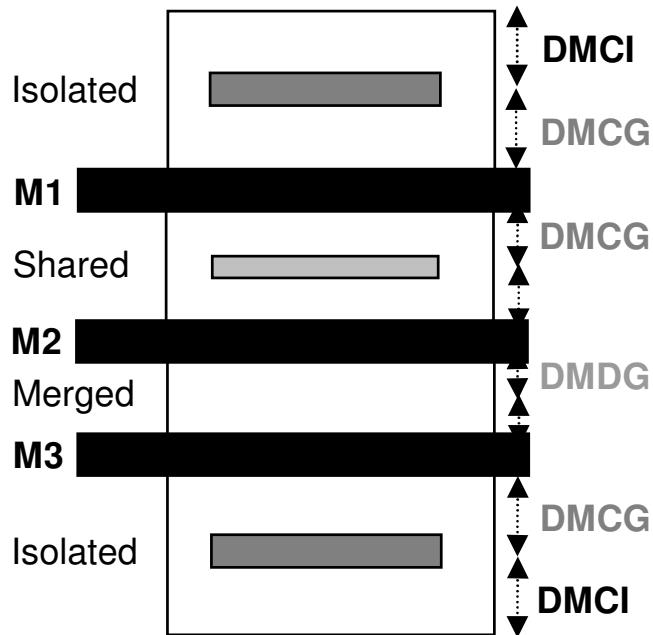


Fig. 8.4 Illustrations of MOSFET end source/drain diffusion geometries in the case of *isolated*, *shared*, and *merged* connection scenarios.

The perimeter of an *isolated* source or drain diffusion is

$$P_{isolated} = 2 \cdot (DMCI + DMCG) + W_{effJCT} \quad (8.4)$$

where  $W_{effJCT}$  is the effective width of source and drain junctions. Similarly, the perimeters of *shared* and *merged* diffusions are given by

$$P_{shared} = 2 \cdot DMCG \quad (8.5)$$

and

$$P_{merged} = 2 \cdot DMDG \quad (8.6)$$

respectively. Note that, unlike the *isolated* source or drain, a *shared* or a *merged* source or drain does not have the isolation-sidewall perimeter component parallel to the device width. The gate-edge perimeter component is not included in Eqs. (8.4) through (8.6) for the sake of convenience in the modeling of the isolation- and gate-edge junction leakage currents and capacitances. More on this topic will be presented shortly.

As can be seen in Fig. 8.4, the junction bottom area of an *isolated* source or drain junction is

$$A_{\text{isolated}} = (\text{DMCI} + \text{DMCG}) \cdot W_{\text{effJCT}} \quad (8.7)$$

Similarly, the bottom areas of a *shared* and a *merged* source or drain junction are

$$A_{\text{shared}} = \text{DMCG} \cdot W_{\text{effJCT}} \quad (8.8)$$

and

$$A_{\text{merged}} = \text{DMDG} \cdot W_{\text{effJCT}} \quad (8.9)$$

respectively.

With these source and drain diffusion perimeter and area formulations for a given NF and MIN, the total perimeter and area values of a device, whether fingered or not, are given next for a GEOMOD selection. This is discussed with simple layout illustrations. Note that for a multi-finger device, all inner source or drain diffusions are assumed to be *shared*.

**GEOMOD = 0** (end-source *isolated* and/or end-drain *isolated*): In this case, the end-diffusions can either be an *isolated* source or an *isolated* drain, or both; the device can also have either an odd or even number of fingers. The total effective source and drain perimeters are then found to be the sum in the form

$$PS_{\text{eff}} = nEndS \cdot P_{\text{isolated}} + nInS \cdot P_{\text{shared}} \quad (8.10a)$$

and

$$PD_{\text{eff}} = nEndD \cdot P_{\text{isolated}} + nInD \cdot P_{\text{shared}} \quad (8.10b)$$

respectively. The total source and drain bottom areas are

$$AS_{\text{eff}} = nEndS \cdot A_{\text{isolated}} + nInS \cdot A_{\text{shared}} \quad (8.10c)$$

and

$$AD_{eff} = nEndD \cdot A_{isolated} + nInD \cdot A_{shared} \quad (8.10d)$$

Possible layouts for  $GEOMOD = 0$  are illustrated in Fig. 8.5 for  $NF = 2$  and 3. As in Figs. 8.2 and 8.3, as well as in Fig. 8.5 and other figures to be given in the rest of this section, a black rectangle represents the contact of an isolated source and drain diffusion, whereas an open unfilled rectangle denotes the contact of a shared source and drain diffusion.

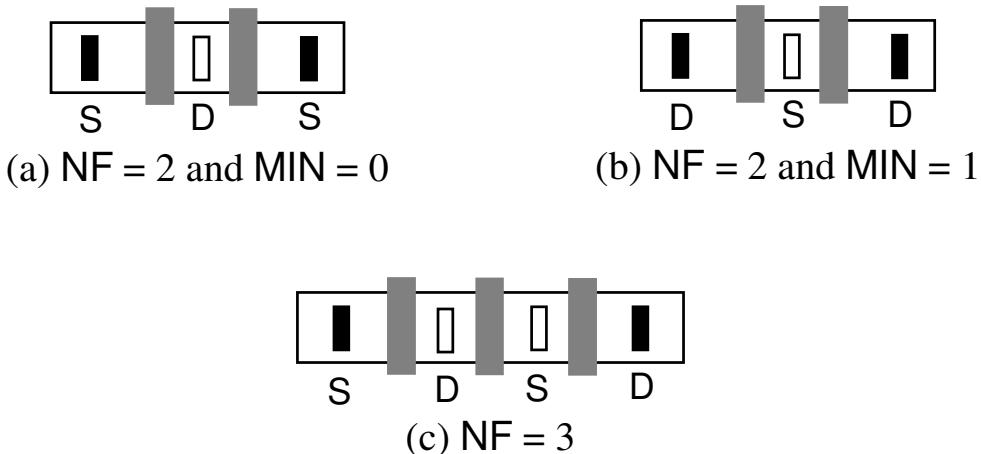


Fig. 8.5 Example layouts for  $GEOMOD = 0$  for  $NF = 2$  and 3.

$GEOMOD = 1$  (end-source *isolated* and end-drain *shared*): In this case, the end-diffusions must be an *isolated* source and a *shared* drain. Hence, the device should have an odd number of fingers. The total effective source and drain perimeters are

$$PS_{eff} = P_{isolated} + nInS \cdot P_{shared} \quad (8.11a)$$

and

$$PD_{eff} = (1 + nInD) \cdot P_{shared} \quad (8.11b)$$

The total source and drain junction bottom areas are

$$AS_{eff} = A_{isolated} + nInS \cdot A_{shared} \quad (8.11c)$$

and

$$AD_{eff} = (1 + nInD) \cdot A_{shared} \quad (8.11d)$$

An example layout for **GEOMOD** = 1 is provided in Fig. 8.6 for **NF** = 3, where the black solid rectangle is drawn as the contact of the isolated source diffusion and the open rectangles are the contacts of the shared source and drain diffusions. As the number of fingers increases, the layouts can be drawn similarly.

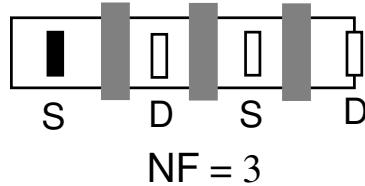


Fig. 8.6 An example layout for **GEOMOD** = 1 for **NF** = 3.

**GEOMOD** = 2 (end-source *shared* and end-drain *isolated*): In this case, the end-diffusions must be a *shared* source and an *isolated* drain, an opposite of **GEOMOD** = 1. Therefore, the device should have an odd number of fingers as well. The total effective source and drain perimeters are then found to be the sum

$$PS_{eff} = (1 + nInS) \cdot P_{shared} \quad (8.12a)$$

and

$$PD_{eff} = P_{isolated} + nInD \cdot P_{shared} \quad (8.12b)$$

The total source and drain areas are

$$AS_{eff} = (1 + nInS) \cdot A_{shared} \quad (8.12c)$$

and

$$AD_{eff} = A_{isolated} + nInD \cdot A_{shared} \quad (8.12d)$$

An example layout for **GEOMOD** = 2 are illustrated in Fig. 8.7 for **NF** = 3.

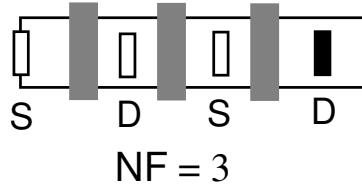


Fig. 8.7 An example layout for GEOMOD = 2 for  $\mathbf{NF} = 3$ .

**GEOMOD = 3 (end-source *shared* and/or end-drain *shared*):** In this case, the end-diffusions can either be *shared* source or *shared* drain, or one of each. The device can have either an odd or even number of fingers. The total source and drain perimeters are

$$PS_{\text{eff}} = (nEndS + nInS) \cdot P_{\text{shared}} \quad (8.13\text{a})$$

and

$$PD_{\text{eff}} = (nEndD + nInD) \cdot P_{\text{shared}} \quad (8.13\text{b})$$

The total source and drain junction bottom areas are

$$AS_{\text{eff}} = (nEndS + nInS) \cdot A_{\text{shared}} \quad (8.13\text{c})$$

and

$$AD_{\text{eff}} = (nEndD + nInD) \cdot A_{\text{shared}} \quad (8.13\text{d})$$

Possible layouts for GEOMOD = 3 are shown in Fig. 8.8 for  $\mathbf{NF} = 2$  and 3.

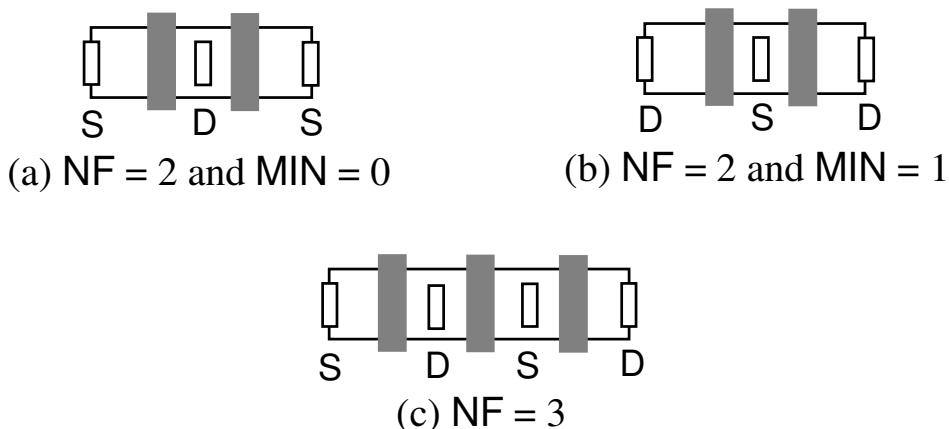


Fig. 8.8 Example layouts for GEOMOD = 3 for  $\mathbf{NF} = 2$  and 3. Whether the drain diffusions are minimized for the case of  $\mathbf{NF} = 2$  (or any even  $\mathbf{NF}$ ) is also shown.

**GEOMOD = 4** (end-source *isolated* and end-drain *merged*): In this case, the end-diffusions must be an *isolated* source and a *merged* drain; hence, the device should have an odd number of fingers. The total source and drain perimeters are found to be the sum

$$PS_{eff} = P_{isolated} + nInS \cdot P_{shared} \quad (8.14a)$$

and

$$PD_{eff} = P_{merged} + nInD \cdot P_{shared} \quad (8.14b)$$

The total effective source and drain bottom areas are

$$AS_{eff} = A_{isolated} + nInS \cdot A_{shared} \quad (8.14c)$$

and

$$AD_{eff} = A_{merged} + nInD \cdot A_{shared} \quad (8.14d)$$

An example layout for **GEOMOD = 4** is illustrated in Fig. 8.9 for **NF = 3**. As the number of fingers increases, the layouts can be made analogously.

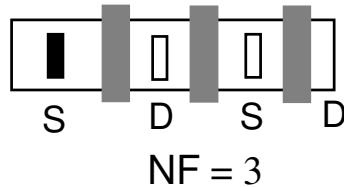


Fig. 8.9 An example layout for GEOMOD = 4 for NF = 3.

**GEOMOD = 5** (end-source *shared* and end-drain *merged*): In this case, the end-diffusions are a *shared* source and a *merged* drain. Therefore, the device has an odd number of fingers. The total source and drain perimeters are

$$PS_{eff} = (1 + nInS) \cdot P_{shared} \quad (8.15a)$$

and

$$PD_{eff} = P_{merged} + nInD \cdot P_{shared} \quad (8.15b)$$

The total effective source and drain bottom areas are

$$AS_{eff} = (1 + nInS) \cdot A_{shared} \quad (8.15c)$$

and

$$AD_{eff} = A_{merged} + nInD \cdot A_{shared} \quad (8.15d)$$

An example layout for  $GEOMOD = 5$  is provided in Fig. 8.10 for  $NF = 3$ . As the number of fingers increases, the layouts can be edited similarly.

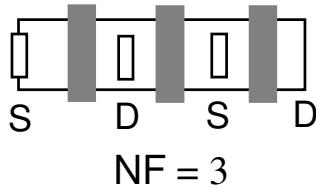


Fig. 8.10 An example layout for  $GEOMOD = 5$  for  $NF = 3$ .

$GEOMOD = 6$  (end-source *merged* and end-drain *isolated*): This is the opposite to  $GEOMOD = 4$ . In this case, the end-diffusions must be a *merged* source and an *isolated* drain. The device has an odd number of fingers. The total source and drain perimeters are

$$PS_{eff} = P_{merged} + nInS \cdot P_{shared} \quad (8.16a)$$

and

$$PD_{eff} = P_{isolated} + nInD \cdot P_{shared} \quad (8.16b)$$

The total source and drain junction areas are

$$AS_{eff} = A_{merged} + nInS \cdot A_{shared} \quad (8.16c)$$

and

$$AD_{eff} = A_{isolated} + nInD \cdot A_{shared} \quad (8.16d)$$

The layout for  $GEOMOD = 6$  is typified in Fig. 8.11 for  $NF = 3$ .

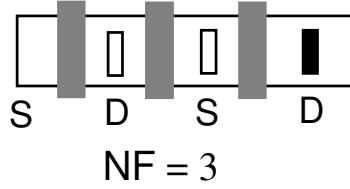


Fig. 8.11 An example layout for GEOMOD = 6 for NF = 3.

**GEOMOD = 7** (end-source *merged* and end-drain *shared*). In this case, the end-diffusions must be a *merged* source and a *shared* drain. The device must have an odd number of fingers. The total source and drain perimeters are then

$$PS_{eff} = P_{merged} + nInS \cdot P_{shared} \quad (8.17a)$$

and

$$PD_{eff} = (1 + nInD) \cdot P_{shared} \quad (8.17b)$$

The total effective source and drain junction areas are respectively given by

$$AS_{eff} = A_{merged} + nInS \cdot A_{shared} \quad (8.17c)$$

and

$$AD_{eff} = (1 + nInD) \cdot A_{shared} \quad (8.17d)$$

An example layout for GEOMOD = 7 is illustrated in Fig. 8.12 for NF = 3. As the number of fingers increases, the layouts can be drawn similarly.

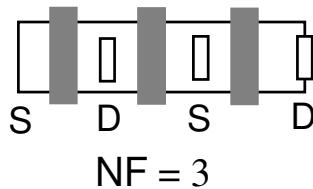


Fig. 8.12 An example layout for GEOMOD = 7 for NF = 3.

**GEOMOD = 8 (end-source *merged* and end-drain *merged*):** The end-source and the end-drain, if present, must be merged. The device can have either an odd or an even number of fingers. The total source and drain perimeters are

$$PS_{eff} = nEndS \cdot P_{merged} + nInS \cdot P_{shared} \quad (8.18a)$$

and

$$PD_{eff} = nEndD \cdot P_{merged} + nInD \cdot P_{shared} \quad (8.18b)$$

The total source and drain bottom areas are

$$AS_{eff} = nEndS \cdot A_{merged} + nInS \cdot A_{shared} \quad (8.18c)$$

and

$$AD_{eff} = nEndD \cdot A_{merged} + nInD \cdot A_{shared} \quad (8.18d)$$

Possible layouts for GEOMOD = 8 are shown in Fig. 8.13 for NF = 2 and 3. As the number of fingers increases, the layouts can be extended similarly.

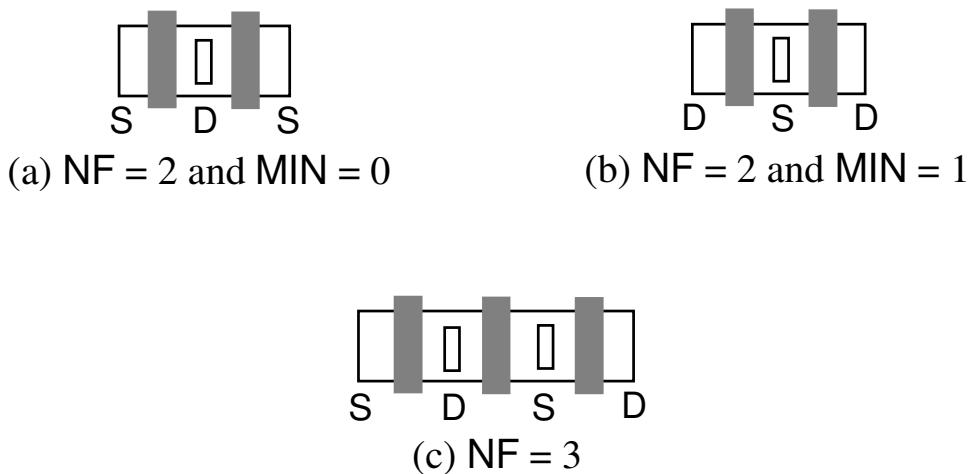


Fig. 8.13 Possible layouts for GEOMOD = 8 for NF = 2 and 3.

**GEOMOD = 9 (end-sources only, one *shared* and the other *isolated*; there are no end-drains):** In this case, the end-diffusions are the source diffusions only, with one source *shared* and the other *isolate*. For this reason, the device must have an even number of fingers. The total source and drain perimeters are

$$PS_{eff} = P_{isolated} + (\text{NF} - 1) \cdot P_{shared} \quad (8.19a)$$

and

$$PD_{eff} = \text{NF} \cdot P_{shared} \quad (8.19b)$$

The total effective source and drain bottom areas are

$$AS_{eff} = A_{isolated} + (\text{NF} - 1) \cdot A_{shared} \quad (8.19c)$$

and

$$AD_{eff} = \text{NF} \cdot A_{shared} \quad (8.19d)$$

An example layout for  $\text{GEOMOD} = 9$  is illustrated in Fig. 8.14 for  $\text{NF} = 2$ .

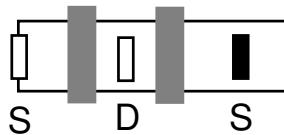


Fig. 8.14 An example layout for  $\text{GEOMOD} = 9$  for  $\text{NF} = 2$ .

$\text{GEOMOD} = 10$  (end-drains only, one *shared* and the other *isolated*; there are no end-sources): In this case, the end-diffusions must be drain and cannot be source, with one drain *shared* and the other *isolated*. The device must have an even number of fingers. The total source and drain perimeters are then

$$PS_{eff} = \text{NF} \cdot P_{shared} \quad (8.20a)$$

and

$$PD_{eff} = P_{isolated} + (\text{NF} - 1) \cdot P_{shared} \quad (8.20b)$$

The total effective source and drain bottom areas are

$$AS_{eff} = \text{NF} \cdot A_{shared} \quad (8.20c)$$

and

$$AD_{eff} = A_{isolated} + (\text{NF} - 1) \cdot A_{shared} \quad (8.20d)$$

An example layout for  $\text{GEOMOD} = 10$  is illustrated in Fig. 8.15 for  $\text{NF} = 2$ . As the number of fingers increases, the layouts can be extended and drawn similarly.

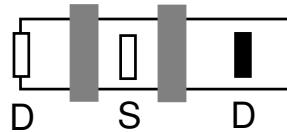
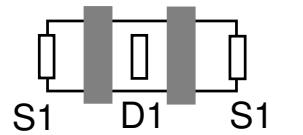


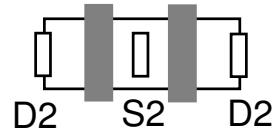
Fig. 8.15 An example layout for  $\text{GEOMOD} = 10$  for  $\text{NF} = 2$ .

With multiple combinations of end-source and end-drain connections permitted by the BSIM4 layout-dependence  $\text{GEOMOD}$  model, it is possible to describe accurately the layouts of cascaded single-finger or multi-finger MOSFET devices for various design needs. One good example would be the “mingling” of two double-finger ( $\text{NF} = 2$ ) transistors in series. When two multi-finger devices are mingled, the fingers of one transistor are dispersed among the fingers of the second transistor. Doing so can improve the high-frequency performance by minimizing the parasitic resistances and capacitances associated with the source and drain diffusion regions, junctions, and overlap regions. One example is shown in Fig. 8.16. To take the full advantage of this capability, one may compare and evaluate the three different approaches to connecting the diffusions in SPICE net-listing: *isolated*, *shared*, and *merged*.

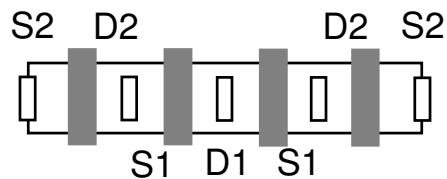
Suppose there are two transistors M1 and M2 that need to be mingled by cascading them in layout implementations. Each transistor has two fingers. Cascading these two transistors with their joined end-diffusions *shared*, for example, can be implemented as shown in Fig. 8.16 (a), whose equivalent schematics is given in Fig. 8.16 (b). M1 and M2 need to be specified in a SPICE net-list with  $\text{GEOMOD} = 3$  and  $\text{NF} = 2$  for both, with  $\text{MIN} = 0$  for M1 and 1 for M2. As another example, in order to cascode them with their joined end-diffusions *merged* as illustrated in Fig. 8.17 (a) and (b), one needs to specify  $\text{GEOMOD} = 8$  and  $\text{NF} = 2$  for both M1 and M2, with  $\text{MIN} = 0$  and 1, respectively. Cascading multi-finger devices with other  $\text{GEOMOD}$  options can be accomplished similarly.



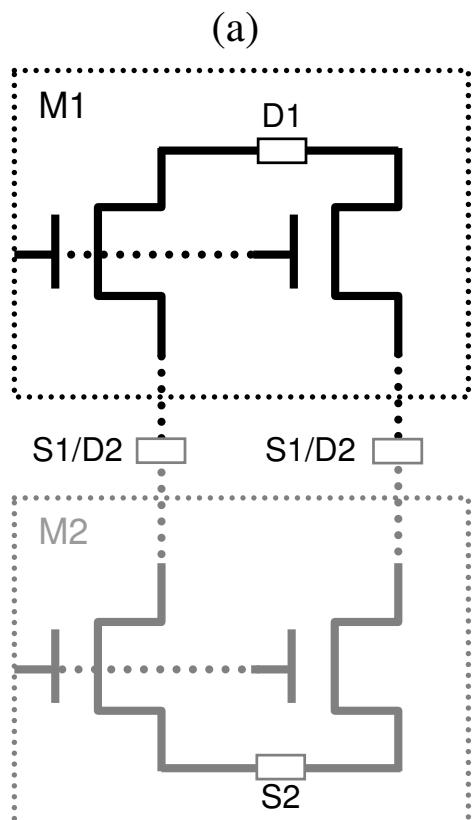
M1: GEOMOD = 3,  
NF = 2 and MIN = 0



M2: GEOMOD = 3,  
NF = 2 and MIN = 1

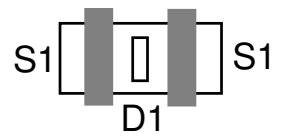


M1 and M2 cascoded

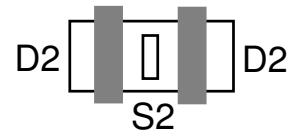


(b)

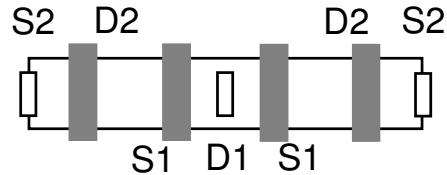
Fig. 8.16 Layout (a) and schematic (b) representations of cascoding two double-finger devices (M1 and M2) by *sharing* their joint end diffusions.



M1: GEOMOD = 8,  
 NF = 2 and MIN = 0

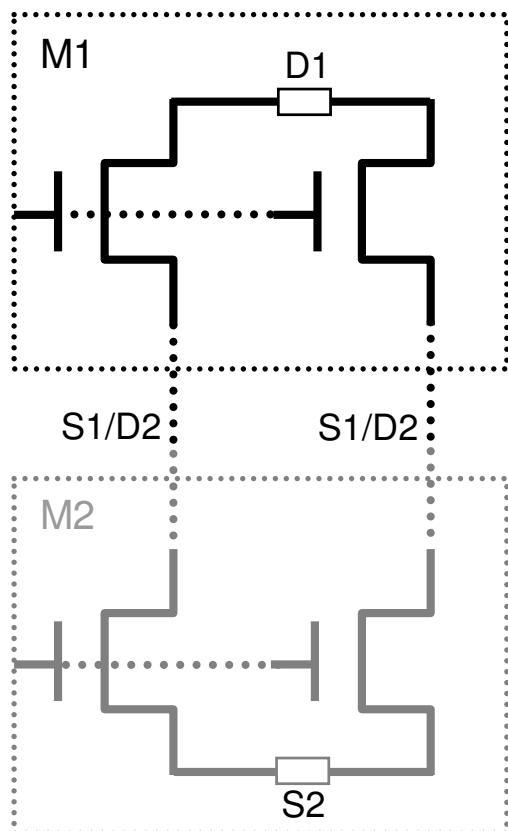


M2: GEOMOD = 8,  
 NF = 2 and MIN = 1



M1 and M2 cascoded

(a)



(b)

Fig. 8.17 Layout (a) and schematic (b) representations of cascoding two double-finger devices (M1 and M2) by *merging* their joint end diffusions.

## 8.6 Saturation Junction Leakage Current and Zero-Bias Capacitance Models

In addition to the GEOMOD source and drain layout model, BSIM4 provides the traditional simple geometry-dependence models, where the source and drain perimeters (**PS** and **PD**, respectively) and areas (**AS** and **AD**, respectively) are explicitly specified in BSIM4 element statements.

If **PS** is given (in element cards), the effective source junction perimeter is

$$PS_{eff} = PS \quad (8.21a)$$

for **PERMOD** = 0, or

$$PS_{eff} = PS - W_{effJCT} \cdot NF \quad (8.21b)$$

when **PERMOD** = 1. **PERMOD** is a global model parameter and flag, which means that **PS** will not include the gate-edge perimeter component when **PERMOD** = 0 (the default setting). Note that from BSIM4.6.5 onwards, two more constraints are added for  $PS_{eff}$  regardless of the setting of **PERMOD**: If **PS** is given to be zero,  $PS_{eff}$  is made equal to zero; and if a negative **PS** is given (not expected in reality),  $PS_{eff}$  will also be set to zero together with a warning message issued that states so. These treatments are equally applied to the computation of the effective drain junction perimeter.

If **PS** is not specified, the GEOMOD layout-dependence model will be used to compute  $PS_{eff}$ . Recall the discussion in the preceding section that  $PS_{eff}$  excludes the gate-edge perimeter component — whether GEOMOD is employed or not. The same holds true for  $PD_{eff}$ .

Likewise, if **AS** is given (in element statement cards), the total effective source junction area is simply

$$AS_{eff} = AS \quad (8.22)$$

Otherwise (**AS** not specified), the GEOMOD layout-dependence model is used to calculate  $AS_{eff}$ . This treatment applies to the  $AD_{eff}$  computation as well.

With  $PS_{eff}$  and  $AS_{eff}$ , BSIM4 computes the saturation leakage current of the source junction diode by summing up the three leakage components attributed to the three junctions at the bottom, the isolation/STI sidewall, and the gate edge:

$$I_{js,sat} = AS_{eff} \cdot J_{s,bottom}(T) + PS_{eff} \cdot J_{s,sidewall}(T) + W_{effJCT} \cdot NF \cdot J_{s,gate-edge}(T) \quad (8.23)$$

$W_{effJCT}$  is the junction width of each finger stripe. The saturation current densities on the right side include their temperature dependencies. They will be presented in Chapter 9. The total saturation current for the drain junction can be written analogously

$$I_{jd,sat} = AD_{eff} \cdot J_{d,bottom}(T) + PD_{eff} \cdot J_{d,sidewall}(T) + W_{effJCT} \cdot NF \cdot J_{d,gate-edge}(T) \quad (8.24)$$

The bottom, sidewall, and gate-edge source junction capacitance components are

$$C_{js,bottom}(V_{bs} = 0) = AS_{eff} \cdot C_{js,bottom,unit-area}(T) \quad (8.25a)$$

for the bottom area,

$$C_{js,sidewall}(V_{bs} = 0) = PS_{eff} \cdot C_{js,sidewall,unit-length}(T) \quad (8.25b)$$

for the sidewall/STI along isolations, and

$$C_{js,gate-edge}(V_{bs} = 0) = W_{effJCT} \cdot NF \cdot C_{js,gate-edge,unit-length}(T) \quad (8.25c)$$

for the gate-edge side. Note that Eq. (8.25b) considers the contribution from the isolation sidewall only, none from the gate edge.

The drain-body junction capacitances are related to the perimeter and area as follows

$$C_{jd,bottom}(V_{bd} = 0) = AD_{eff} \cdot C_{jd,bottom,unit-area}(T) \quad (8.26a)$$

for the bottom area,

$$C_{jd,sidewall}(V_{bd} = 0) = PD_{eff} \cdot C_{jd,sidewall,unit-length}(T) \quad (8.26b)$$

for the sidewall along isolations, and

$$C_{jd, \text{gate-edge}}(V_{bd} = 0) = W_{\text{effJCT}} \cdot \text{NF} \cdot C_{jd, \text{gate-edge, unit-length}}(T) \quad (8.26c)$$

for the gate edge. Again, the effective source and drain perimeter excludes the gate-edge components. Hence, the junction saturation current and zero-bias junction capacitance need to include the ( $W_{\text{effJCT}} \cdot \text{NF}$ ) term. The bias and temperature dependence models of the junction capacitance will be discussed in Chapter 9.

## 8.7 Source and Drain Contact Scenarios and Diffusion Resistances

As discussed in the previous sections, the inner source and drain diffusions of a multi-finger MOSFET device are *shared* between the two neighboring finger stripes. It is further noted here that the contacts of these *shared* diffusions are also made shared in practice. As a result, the inner source and drain diffusion resistances are modeled as

$$R_{inS} = \frac{\text{RSH} \cdot \text{DMCG}}{W_{\text{effJCT}} \cdot nInS} \quad (8.27a)$$

and

$$R_{inD} = \frac{\text{RSH} \cdot \text{DMCG}}{W_{\text{effJCT}} \cdot nInD} \quad (8.27b)$$

$\text{RSH}$ , in the unit of ohm per square, is a global model parameter that designates the source and drain diffusion sheet resistance. It is assumed to be independent of voltage (true for higher doping concentrations). The terms  $nInS$  and  $nInD$  (the number of the inner source and drain diffusions of a multi-finger device, respectively) in the denominators result from the GEOMOD model discussed in the previous sections.  $R_{inS}$  and  $R_{inD}$  of Eqs. (8.27a) and (8.27b) are hard-coded to be zero in BSIM4 if  $nInS$  or  $nInD$  is found to be zero. This avoids potential divide-by-zero errors in SPICE simulation.

In addition to this wide-contact scenario, the end-source and end-drain diffusions of a MOSFET device, fingered or not, may have a point contact or no contact at all in practice. This is as shown in Fig. 8.18.

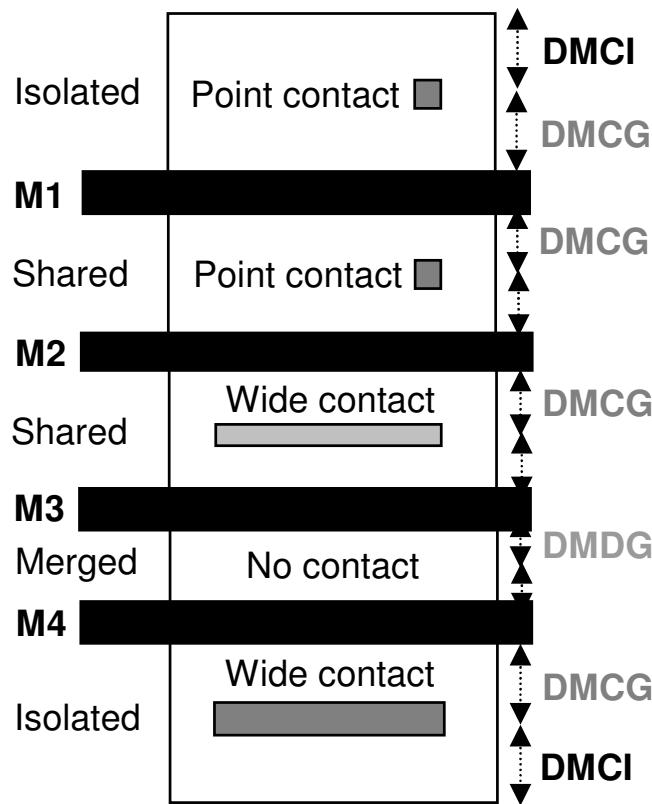


Fig. 8.18 Various end-source and end-drain connections and contact scenarios modeled by BSIM4.

A point contact in an *isolated* end source or drain diffusion gives rise to an end diffusion resistance, the product of the sheet resistance and the number of source/drain squares

$$R_{endS} = \frac{RSH \cdot W_{effJCT}}{3 \cdot nEndS \cdot (DMCI + DMCG)} \quad (8.28a)$$

and

$$R_{endD} = \frac{RSH \cdot W_{effJCT}}{3 \cdot nEndD \cdot (DMCI + DMCG)} \quad (8.28b)$$

Here, the resistance does not include the contact resistance itself.  $nEndS$  and  $nEndD$ , the numbers of the end-source and end-drain diffusions, respectively, are determined from the GEOMOD model discussed above. The constant, 3, in the denominators results from the assumption that the transistor current is uniform across the width of the diffusion regions. The mathematical analysis that leads to the constant 3 is similar to the gate electrode resistance analysis in Chapter 6. Divide-by-zero errors in the above expressions and below can all be avoided by setting  $R_{endS}$  and  $R_{endD}$  to zero if a zero denominator is detected.

Likewise, a point contact made in a *shared* end source or drain diffusion (not shown in Fig. 8.18) gives rise to a diffusion resistance given by

$$R_{endS} = \frac{RSH \cdot W_{effJCT}}{6 \cdot nEndS \cdot DMCG} \quad (8.29a)$$

upon setting DMCI equal to DMCG in Eq. (8.28a). Similarly,

$$R_{endD} = \frac{RSH \cdot W_{effJCT}}{6 \cdot nEndD \cdot DMCG} \quad (8.29b)$$

The impact of a wide contact on an *isolated* end-source or end-drain diffusion resistance is similar to that on an inner *shared* source or *shared* drain diffusion (given by Eqs. (8.27a) and (8.27b)). Thus

$$R_{endS} = \frac{RSH \cdot DMCG}{W_{effJCT} \cdot nEndS} \quad (8.30a)$$

and

$$R_{endD} = \frac{RSH \cdot DMCG}{W_{effJCT} \cdot nEndD} \quad (8.30b)$$

irrespective of how far, in the channel length direction, the contact is from the isolation. Eqs. (8.30a) and (8.30b) are also valid for the case of a *shared* end-source or end-drain diffusion with a wide contact.

A *merged* end-source or end-drain diffusion connection, as designated by GEOMOD = 4, 5, 6, 7, and 8, requires no contacts. However, it does introduce resistances for the two transistors that are connected in series

through the diffusion. This end-diffusion resistance, for an end-source or an end-drain alike, is simply

$$R_{endS} = R_{endD} = \frac{RSH \cdot DMDG}{W_{effJCT}} \quad (8.31a)$$

This model is applicable for the GEOMOD = 4, 5, 6, and 7. In the case of GEOMOD = 8, Eq. (8.31a) needs to be extended. In this case, both end-diffusions can be the source or the drain for an even number of fingers, depending upon the parameter MIN. Therefore

$$R_{endS} = \frac{RSH \cdot DMDG}{W_{effJCT} \cdot nEndS} \quad (8.31b)$$

for any end-source, and

$$R_{endD} = \frac{RSH \cdot DMDG}{W_{effJCT} \cdot nEndD} \quad (8.31c)$$

for any end-drain.

Special attention need be paid when GEOMOD = 9 (end-sources *shared* and *isolated*; no end-drains) or 10 (end-drains *shared* and *isolated*; no end-sources). For simplicity, wide contacts are assumed for these end-diffusions. For GEOMOD = 9, the total resistance of both *shared* and *isolated* end-source diffusions is half of the resistance found for an inner diffusion

$$R_{endS} = \frac{RSH \cdot DMCG}{2 \cdot W_{effJCT}} \quad (8.32a)$$

A wide contact is assumed. Similarly for GEOMOD = 10, the resistance of both *shared* and *isolated* end-drain diffusions is

$$R_{endD} = \frac{RSH \cdot DMCG}{2 \cdot W_{effJCT}} \quad (8.32b)$$

## 8.8 RGEOMOD: Selecting A Source and Drain Contact Scenario for GEOMOD

In the previous section, the models of the end-source and end-drain diffusion resistance with various contact types were presented. In circuit design, it is often desired to designate a particular combination of end-source and end-drain contact types as well as a particular end-source or end-drain connection type (through GEOMOD, NF, and MIN settings). Table 8.2 gives possible combinations available in BSIM4. Each of them can be selected with a local instance parameter RGEOMOD.

Table 8.2 RGEOMOD and corresponding end source and drain contact types of BSIM4.

RGEOMOD	Contact type of end-sources (if any)	Contact type of end-drains (if any)
0 (default)	No contact type to be selected	No contact type to be selected
1	Wide	Wide
2	Wide	Point
3	Point	Wide
4	Point	Point
5	Wide	No contact/Merged
6	Point	No contact/Merged
7	No contact/Merged	Wide
8	No contact/Merged	Point

Note that GEOMOD = 9 and 10 do not require an RGEOMOD designation because both already assume wide contacts for their end-diffusions, as discussed in the preceding section. Moreover, GEOMOD = 8 (*merged* end-source and/or *merged* end-drain) requires no contacts for both end-diffusions. Therefore, RGEOMOD does not have to and cannot be specified in the case of GEOMOD = 8.

Only some RGEOMOD values are permitted for a given GEOMOD. As an example, consider GEOMOD = 1 (where the end-diffusions must be an *isolated* source and a *shared* drain and the device must have an odd number of fingers). Possible RGEOMOD values are 1 through 4 of Table 8.2. To illustrate this point, RGEOMOD = 3 (a point contact for the end-source and a wide one for the end-drain) is shown in Fig. 8.19.

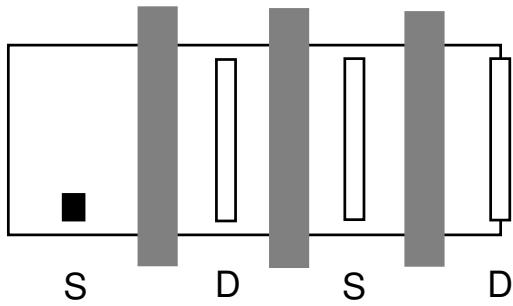


Fig. 8.19 An example layout for GEOMOD = 1 and RGEOMOD = 3 for NF = 3.

It is worth noting that RGEOMOD is set to zero by default. Therefore, the end-source and end-drain diffusion resistances will be computed as  $(RSH \cdot NRS)$  and  $(RSH \cdot NRD)$  if the numbers of the source and drain squares NRS and NRD are specified in element statement cards. Thus, the default model is the traditional one, which is a less accurate resistance model as noted earlier.

The total source diffusion resistance of a multi-finger device is the parallel combination of the inner source and end-source diffusion resistances

$$R_{S,total} = \frac{R_{inS} \cdot R_{endS}}{R_{inS} + R_{endS}} \quad (8.33a)$$

with  $R_{S,total} = R_{endS}$  if  $R_{inS}$  is not positive (by mistake somehow),  $R_{S,total} = R_{inS}$  if  $R_{endS}$  is not positive, or no source diffusion resistance to be modeled if both are not positive. The total drain diffusion resistance is modeled in a similar form

$$R_{D,total} = \frac{R_{inD} \cdot R_{endD}}{R_{inD} + R_{endD}} \quad (8.33b)$$

Again,  $R_{D,total} = R_{endD}$  if  $R_{inD}$  is not positive,  $R_{D,total} = R_{inD}$  if  $R_{endD}$  is not positive, and no drain diffusion resistance will be modeled if neither is positive.

Finally, the authors wish to caution that all the source and drain diffusion resistance computations performed above do not include contact resistances. These resistances need to be modeled and extracted separately, in both pre- and post-layout simulations. Most of commercial SPICE simulators and LPE (layout parasitics extraction) tools provide the contact resistance models of their own.

## 8.9 Chapter Summary

This chapter presented and discussed the BSIM4 layout-dependent source and drain parasitics models. These models provide a comprehensive way of defining (via the GEOMOD, NF and MIN parameters) all possible source and drain connections (*isolated*, *shared* and *merged*) for transistor stacks and for multi-finger MOSFET devices. Various source and drain contact scenarios such as *point*, *wide* and no contacts can be accurately described (via the RGEOMOD parameter) for those source and drain connections. With these, the effective source and drain junction perimeters and areas, parasitic (diffusion) resistances, and geometry-dependent junction saturation currents and zero-bias junction capacitances are accurately modeled for both schematic net-listing and layout implementation.

## 8.10 Parameter Table

Name (type)	Description and default	Can be binned?	Note
GEOMOD (Local and global; integer)	<p>End source/drain connection type selector for non-finger or multi-finger devices.</p> <p>Default = 0; dimensionless: The end diffusions can either be an isolated source or an isolated drain, or both; NF can be an odd or even number. Other optional values are 1 through 10.</p>	No	-

<b>PERMOD</b> (global; integer)	Flag to indicate whether the gate-edge source and drain perimeter components are to be included in the source and drain perimeters PS and PD. This flag applies to the layout-dependence model where PS and PD are explicitly specified in BSIM4 element statements, not to the GEOMOD model.  Default = 1; dimensionless: The BSIM4 source and drain perimeters PS and PD that are explicitly specified in element statements include the gate-edge perimeter components.	No	-
<b>RGEOMOD</b> (Local; integer)	End source/drain diffusion contact type selector for non-fingered or multi-fingered devices.  Default = 0; dimensionless: No source and drain diffusion contact type is considered. Other optional values are 1 through 8.	No	-
<b>NF</b> (Local; integer)	The number of fingers that a multi-finger device structure has.  Default = 1; dimensionless.	No	Reset to 1 if <b>NF</b> ≤ 1 with a fatal error to be issued.
<b>DMCG</b> (Global; double)	The distance measured in the channel-length direction from the mid-contact or the contact center of a source or drain diffusion region to its gate edge.  Default = 0.0 [m].	No	-
<b>DMCGT</b> (Global; double)	DMCG of test device structures used in device characterization and model parameter extractions.  Default = 0.0 [m].	No	-
<b>DMCI</b> (Global; double)	The distance measured in the channel-length direction from the mid-contact or the contact center of an isolated source or drain diffusion region to an (STI) isolation edge.  Default = DMCG [m].	No	-

## 300 BSIM4 AND MOSFET MODELING FOR IC SIMULATION

By Weidong Liu and Chenming Hu

DMDG (Global; double)	The distance measured in the channel-length direction from the source or drain mid-diffusion point to its gate edge.  Default = 0.0 [m].	No	-
MIN (Local; integer)	An integer flag to indicate whether the end diffusions are source or drain when the number of finger ( <b>NF</b> ) is even.  Default = 0; the end diffusions serve as the source terminal.	No	-
PS (Local; double)	The source junction perimeter. It is meant to include the gate-edge perimeter component when <b>PERMOD</b> = 1. The <b>GEOMOD</b> model will be used to compute the total effective source perimeter value when <b>PS</b> is not explicitly specified in element statements.  Default = 0.0 [m].	No	From BSIM4.5.0 onwards, a warning message will be issued if a negative <b>PS</b> is given. The effective source perimeter $PS_{eff}$ will then be set to zero regardless of the setting of <b>PERMOD</b> .
PD (Local; double)	The drain junction perimeter. It is meant to include the gate-edge perimeter component when <b>PERMOD</b> = 1. The <b>GEOMOD</b> model will be used to compute the total effective drain perimeter value when <b>PD</b> is not explicitly specified in element statements.  Default = 0.0 [m].	No	From BSIM4.5.0 onwards, a warning message will be issued if a negative <b>PD</b> is given. The effective drain perimeter $PD_{eff}$ will then be set to zero regardless of the setting of <b>PERMOD</b> .

AS (Local; double)	The source junction bottom area. The GEOMOD model will be used to compute the effective source bottom area value when AS is not explicitly specified in element statements.  Default = 0.0 [m <sup>2</sup> ].	No	-
AD (Local; double)	The drain junction bottom area. The GEOMOD model will be used to compute the effective drain bottom area value when AD is not explicitly specified in element statements.  Default = 0.0 [m <sup>2</sup> ].	No	-
RSH (Global; double)	Sheet resistance of source and drain diffusions.  Default = 0.0 [ohm-square].	No	-
NRS (Local; double)	The number of squares of the source diffusion region.  Default = 1.0 [square].	No	-
NRD (Local; double)	The number of squares of the drain diffusion region.  Default = 1.0 [square].	No	-

**This page intentionally left blank**

## Chapter 9

# Junction Diode IV and CV Models

### 9.1 Introduction and Chapter Objectives

The MOSFET source and drain junction diodes are an integral part of the MOSFET transistor. They support the normal function of the transistor but, in addition, introduce unwanted parasitic effects, affecting or, even under certain circumstances, potentially disrupting the normal operation of a circuit. The basic parasitic effects are the leakage current, including trap-assisted tunneling and reverse bias breakdown, and capacitances associated with the junctions.

This chapter will first present and discuss the various physical mechanisms that are responsible for the junction diode currents and their relationships with the diode voltage. This will be followed by the presentation and analysis of several BSIM4 junction diode IV models and options. The discussion of the BSIM4 junction CV model will be given subsequently. The temperature-dependence models of the junction IV and CV must be accurate in a production-worthy compact MOSFET model. They will be presented afterwards. Finally, the junction IV and CV model parameter table will conclude this chapter.

### 9.2 Physical Mechanisms of Diode DC Currents

It is known that the ideal junction diode current is related to the junction voltage  $V_j$  by

$$I_j = I_{js0} \cdot \left[ \exp\left(\frac{qV_j}{k_B T_{emp}}\right) - 1 \right] \quad (9.1)$$

where  $T_{emp}$  is the device junction operating temperature and  $I_{js0}$  is the reverse-bias junction saturation current. This saturation current is made up of three components as illustrated in Fig. 9.1.

$$I_{js0} = A_{eff} \cdot J_{bottom} + W_{effJCT} \cdot J_{gate-edge} + P_{eff} \cdot J_{STI} \quad (9.2)$$

The three terms on the right-hand side are the currents through the junction bottom, the junction/gate edge, and the junction/STI (shallow-trench isolation) edge, respectively. The current densities,  $J_{bottom}$ ,  $J_{gate\_edge}$ , and  $J_{STI}$ , are different because of their different junction doping profiles.

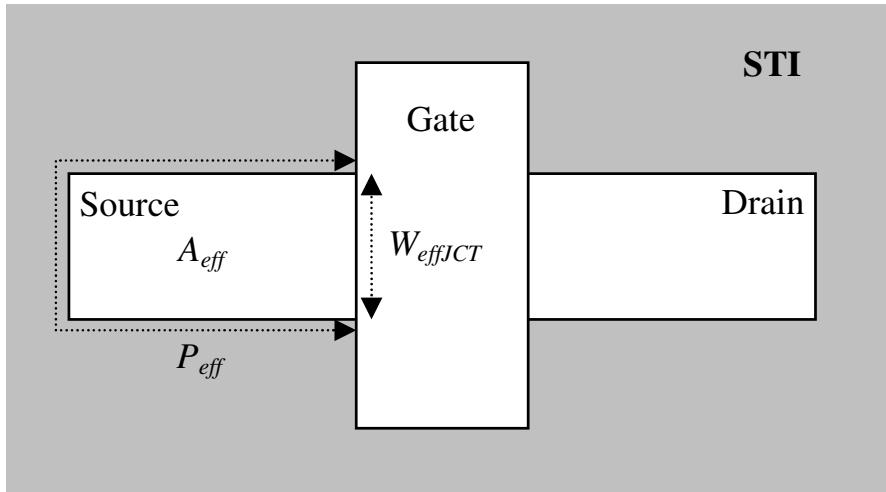


Fig. 9.1 Three components of a MOSFET junction diode contribute to the diode IV and CV: The gate edge denoted by  $W_{effJCT}$ , the STI (shallow-trench isolation) edge with the length  $P_{eff}$ , and the junction bottom area  $A_{eff}$ . Note that older CMOS process technologies use LOCOS (LOCal Oxidation of Silicon) instead of STI for electrical isolation of integrated MOS devices. The junction leakage current at the LOCOS edge is much larger than at the STI edge because of LOCOS induced defects.

Equation (9.1) is simplistic and has the following shortcomings. (1) It ignores the parasitic junction diode series resistance associated with the real junction diode. Therefore, it produces a pure exponential dependence on  $V_j$  in the forward-bias regime, which can cause numerical overflow and convergence difficulties for SPICE simulations. (2) When the junction voltage  $V_j$  has a large negative value, the ideal junction diode current is nearly a constant and the small-signal conductance, i.e., the derivative of  $I_j$  with respect to  $V_j$ , is nearly zero. This can also potentially lead again to SPICE convergence difficulties. In practice, a  $G_{min} \cdot V_j$  term is added to prevent this from being a problem. (3)

Breakdown is not modeled. (4) Depletion-region Shockley-Read-Hall (SRH) generation and recombination, trap-assisted tunneling (TAT), and band-to-band tunneling (BTBT) leakage mechanisms are not taken into account.

In the remainder of this section, the SRH, TAT, and BTBT theory is briefly reviewed. The BSIM4 junction diode IV model will be presented in the next section.

### 9.2.1 Shockley-Read-Hall (SRH) Generation and Recombination

It is known that electron and hole generation-recombination-trapping (GRT) centers exist in silicon. These GRT centers introduce trap energy states in the silicon energy-band gap. According to the Shockley-Read-Hall (SRH) theory, generation and recombination of silicon electrons and holes take place through trap centers. Energy is exchanged with silicon lattice through the absorption or emission of phonons or photons. These GRT centers arise from the physical crystal defects such as silicon vacancies and from chemical impurities, such as gold, iron, and other noble and transition metal elements.

The density of the GRT centers is usually very small thanks to highly developed purification and annealing techniques. However, they are numerous enough of concern to low-power, portable and memory devices. In the forward mode of junction diode operation, some of these trap centers can facilitate electron and hole recombination and result in a net recombination current. In the reverse-bias mode of operation, electrons and holes can be generated through these trap sites, resulting in leakage currents.

Figure 9.2 illustrates an SRH generation process. An electron trap at the energy level  $E_{trap}$  in the forbidden gap of the energy bands first captures an electron from the valence band (the valence electron is indicated by the gray dot in the silicon valence band), thereby leaving a hole in the valence band. The hole in the valence band is represented, in Fig. 9.2, by the open circle. This captured or trapped electron at the SRH center is then released (or emitted) into the conduction band, thereby producing an electron. The electric field from the voltage drop  $V_j$  through the junction separates the electron-hole pair and creates a reverse current component. The opposite of this two-step SRH generation process is the recombination of a conduction-band electron with a hole at such a trap

center or in the valence band. This results in a forward current passing through the p/n junction [1].

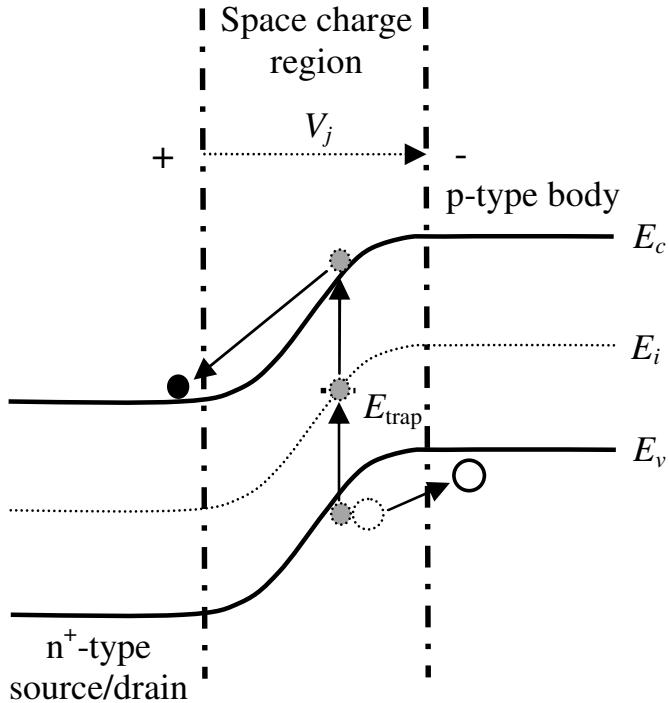


Fig. 9.2 SRH generation process under a reverse junction bias condition. An electron is captured from the Si valence band ( $E_v$ ) into the trap site ( $E_{trap}$ ) and then it is released into the conduction band ( $E_c$ ). This leaves a hole in the valence band. Both the electron and hole then drift to the neutral region.

The SRH generation-recombination process causes the junction diode current to deviate from the simple, ideal model of Eq. (9.1). Mathematically, these processes can be modeled with a non-ideality factor  $NJ$ . Eq. (9.1) now becomes

$$I_j = I_{js0} \cdot \left[ \exp\left( \frac{qV_j}{NJ \cdot k_B T_{emp}} \right) - 1 \right] \quad (9.3)$$

where  $NJ$  defaults to 1 and is determined by junction doping profiles and fab process conditions.

### 9.2.2 Trap-Assisted Tunneling (TAT)

Trap centers in the energy gap can also serve as stepping stones that facilitate electron and hole tunneling from one side of the p/n junction to the other. This trap-assisted tunneling current is insignificant in the forward bias regime in comparison with the large forward-bias junction diode current. In the reverse bias regime, it can produce a significant additional junction leakage current. This phenomenon is commonly known as the trap-assisted tunneling or TAT [1].

TAT often takes place in highly-doped p/n junctions where the junction depletion layer is so thin that electrons or holes can easily tunnel from one side to the other. In this tunneling process, an electron in the valence band of the p-type silicon (Fig. 9.3), for instance, tunnels into the trap at the energy level  $E_{trap}$  leaving a hole behind. This trapped electron then tunnels into the conduction band of n<sup>+</sup>-type silicon. This two-step tunneling process is much more efficient than the one-step tunneling directly from the valence band to the conduction band (without the trap assistance) because of the smaller tunneling distance via the trap. A decrease in the distance increases the tunneling rate exponentially. In other words, the trap at  $E_{trap}$  serves as a stepping stone that makes tunneling easier.

It thus creates an additional junction leakage current component. For advanced CMOS process technologies, where the p/n junctions are heavily doped, TAT is significant because more lattice defects are unavoidably created by the doping process and the electric field is stronger. The quantum-mechanical analysis of the tunneling probability as a function of  $E_{trap}$ , the junction voltage, and dopant impurity concentrations is quite complex and is not presented here. BSIM4 employs an empirical model, which will be discussed in the next section. Note that as the electric field strength increases, the tunneling rate increases. Hence, TAT is a strong function of the voltage applied across the junction.

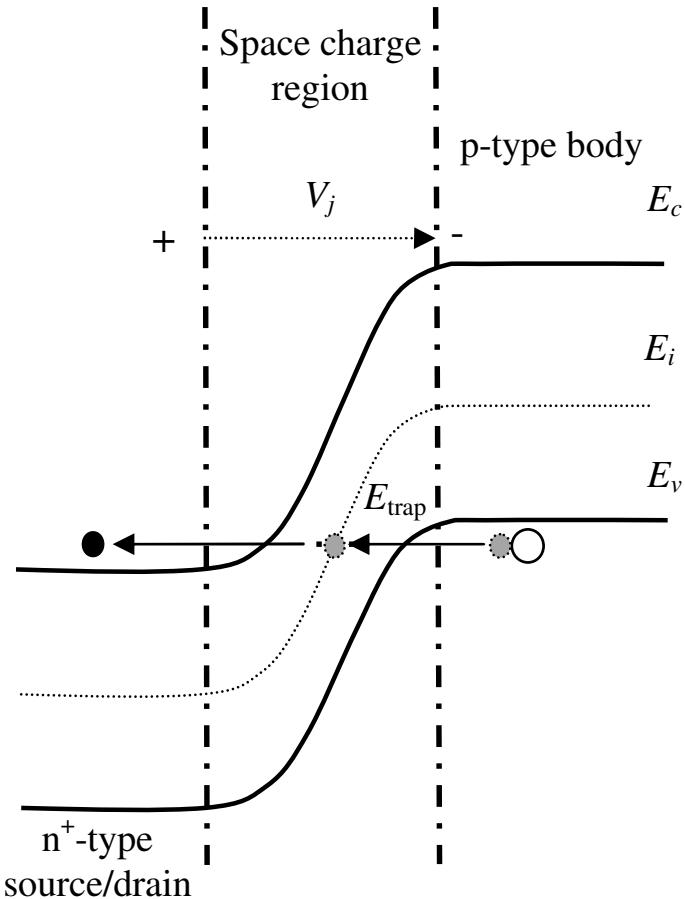


Fig. 9.3 Illustration of the trap-assisted tunneling (TAT) process in the reverse-bias regime. In this case, an electron in the valence band  $E_v$  of the p-type silicon first tunnels into the trap site  $E_{trap}$  and, from there, tunnels into the conduction band of the n<sup>+</sup>-type silicon under the influence of the junction electric field produced by a reverse bias  $V_j$ .

### 9.2.3 Band-To-Band Tunneling (BTBT)

When the junction is heavily doped, additional current can result from band-to-band tunneling (BTBT). This process is illustrated in Fig. 9.4. Under a reverse bias, the electrons in the valence band of the p<sup>+</sup>-type silicon can tunnel, without any stepping-stone trap assistance, directly into the conduction band of the n<sup>+</sup>-type silicon.

BTBT becomes significant when the electric field of the junction approaches  $10^6$  V/cm. This tunneling can take place in both the forward and reverse modes of diode operation. In the case of MOSFET junction diodes, the band-to-band tunneling is not significant in comparison with

the other junction diode leakage current components. For this reason, the band-to-band tunneling current in the source/body and drain/body junctions is not considered in BSIM4. However, BSIM4 model another BTBT current, namely the gate-induced drain/source leakage current (GIDL/GISL) in the surface regions of the drain and the source. This leakage current is induced by the gate-to-drain or the gate-to-source voltage rather than the bulk p/n junction voltage. This current component was already discussed in Chapter 4.

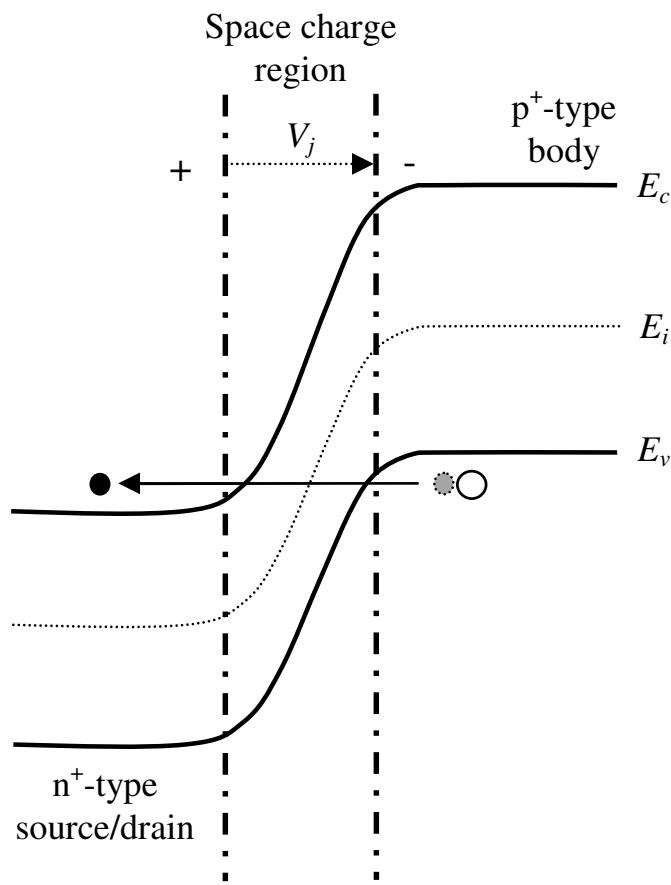


Fig. 9.4 Illustration of the band-to-band tunneling (BTBT) process in a heavily-doped junction diode in the reverse-bias regime.

#### 9.2.4 Diode Breakdown

When the reverse-bias voltage  $V_j$  across a p-n junction diode is large enough, the reverse current will suddenly shoot up. This is known as “junction breakdown”. The junction is no longer resistant to the flow of

the reverse current. The breakdown mechanism can be either band-to-band tunneling, or avalanche multiplication due to impact ionization in the junction depletion region, or both in some occasions. In the case of a breakdown voltage that is less than approximately  $4E_g/q$  where  $E_g$  is the energy-band gap ( $E_g = 1.12$  eV for silicon at  $T_{emp} = 300$  K), band-to-band tunneling is the dominant mechanism. If the breakdown voltage is higher than  $6E_g/q$ , avalanche multiplication is responsible. A mixture of both mechanisms is involved when the breakdown voltage falls in between.

It is known that the energy-band gap  $E_g$  of silicon decreases with increasing temperatures. The junction breakdown voltage due to tunneling (refer to Fig. 9.4) therefore is slightly lower at higher temperature. Impact ionization rate, on the other hand, decreases at higher temperatures due to more frequent phonon scattering which prevents the carriers from gaining sufficient energy to produce impact ionization. Therefore the breakdown voltage due to avalanche is higher at higher temperatures.

MOSFET junction diode breakdown is usually triggered by avalanche multiplication [1]. A thorough theoretical analysis of the breakdown voltage is complicated. A simple derivation is given in [2]. For compact modeling, one could use an empirical expression for the breakdown voltage  $V_B$ .  $V_B$  is basically determined by the doping concentration of the more lightly doped side of the p/n junction. It is given by [3]

$$V_B \approx 60 \cdot \left[ \frac{E_g(300K)}{1.1} \right]^{3/2} \cdot \left( \frac{N_B}{10^{16}} \right)^{-3/4} \quad (9.4a)$$

where  $N_B$  is the lighter doping concentration in the unit of  $\text{cm}^{-3}$ . For linearly graded junctions,  $V_B$  is found to be

$$V_B \approx 60 \cdot \left[ \frac{E_g(300K)}{1.1} \right]^{6/5} \cdot \left( \frac{a}{3 \times 10^{20}} \right)^{-2/5} \quad (9.4b)$$

with  $a$  denoting the gradient of the doping profile in  $\text{cm}^{-4}$ .

For the SPICE modeling purpose, it is convenient to take  $V_B$  as a global model parameter, which can be measured using device test structures. BSIM4 takes this approach.

### 9.3 BSIM4 Diode DC IV Model [4]

BSIM4 provides three options for modeling the source-bulk and drain-bulk p/n junction diode current-voltage (IV) characteristics. Each option can be invoked by setting a global model parameter DIOMOD. When DIOMOD = 0, a “resistance-free” (namely, no current limiting in the forward region) IV model is selected. In this case, whether to model the junction breakdown behavior is optional as detailed below. If DIOMOD = 1 (the default option), the junction diode current in the forward-bias region is limited at high forward bias to a linear IV relationship for better SPICE simulation convergence. No breakdown modeling is available in the reverse-bias region. The DIOMOD = 2 option includes breakdown in the reverse-bias region and current limiting in both the forward and reverse operations.

In the following subsections, these three options are presented. The current limiting and breakdown modeling techniques are explained in details. In the course of the discussion, only the source-bulk p/n junction diode is considered since the drain-bulk junction uses an analogous set of formula and parameters. The modeling of the additional leakage current due to the trap-assisted tunneling is given in Section 9.4.

#### 9.3.1 ***DIOMOD = 0***

By taking into account the impact of the depletion region recombination on the ideal diode IV model, the source-bulk junction diode current becomes

$$I_{js} = I_{js,sat} \cdot \left[ \exp\left(\frac{qv_{bs}}{\mathbf{NJS} \cdot k_B T_{emp}}\right) - 1 \right] \quad (9.5)$$

$\text{NJS}$  is the non-ideality factor whose default value is 1.  $v_{bs}$  is the bulk-to-source voltage (a lower-case  $v$  is employed to indicate that no source and drain swapping is needed in SPICE implementation when the drain-to-source voltage changes sign).  $I_{js,sat}$  is the reverse junction saturation current, which was introduced in Chapter 8:

$$\begin{aligned} I_{js,sat} = & AS_{eff} \cdot J_{s,bottom}(T) + PS_{eff} \cdot J_{s,sidewall}(T) \\ & + W_{effJCT} \cdot \text{NF} \cdot J_{s,gate-edge}(T) \end{aligned} \quad (9.5a)$$

The three saturation current density components, for the junction bottom, STI sidewall ( $J_{s,sidewall}$ ), and gate-edge, on the right side of the equation are temperature dependent. Eq. (9.5a) is implemented with safeguards against two rare but possible error conditions. If both  $AS_{eff}$  and  $PS_{eff}$  have non-positive values,  $I_{js,sat}$  will be set to  $1 \times 10^{-14}$  A (Note under this condition,  $I_{js,sat}$  will be set to 0 for more accurate junction diode leakage current modeling from the BSIM4 releases VERSION  $\geq 4.6.5$ ). If  $I_{js,sat}$  is not positive,  $I_{js}$  is set equal to  $G_{\min} \cdot v_{bs}$  for better simulation convergence [refer to Eq. (9.6) below]. The same treatment is applied to the drain-bulk junction diode IV modeling.

Equation (9.5) is graphically sketched in Fig. 9.5 (a). It is conceivable that as  $\text{NJS}$  increases, the slope of the forward-bias IV curve becomes smaller. Furthermore, to permit the modeling of junction breakdown in the reverse-bias region (i.e.,  $v_{bs} < 0$  for NMOS and  $v_{bs} > 0$  for PMOS), BSIM4 chooses to change Eq. (9.5) to

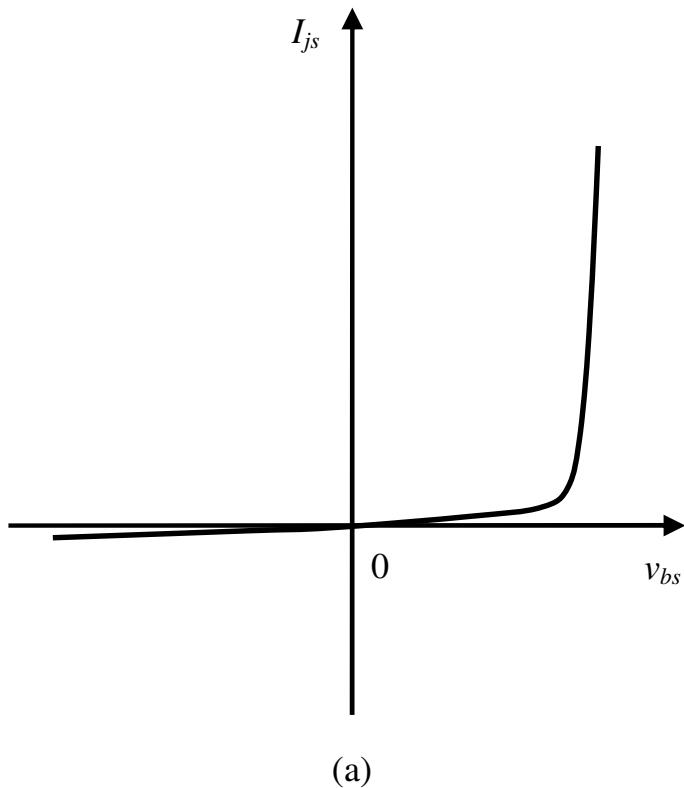
$$I_{js} = I_{js,sat} \cdot \left[ \exp\left(\frac{qv_{bs}}{\text{NJS} \cdot k_B T_{emp}}\right) - 1 \right] \cdot F_{breakdown} + G_{\min} \cdot v_{bs} \quad (9.6)$$

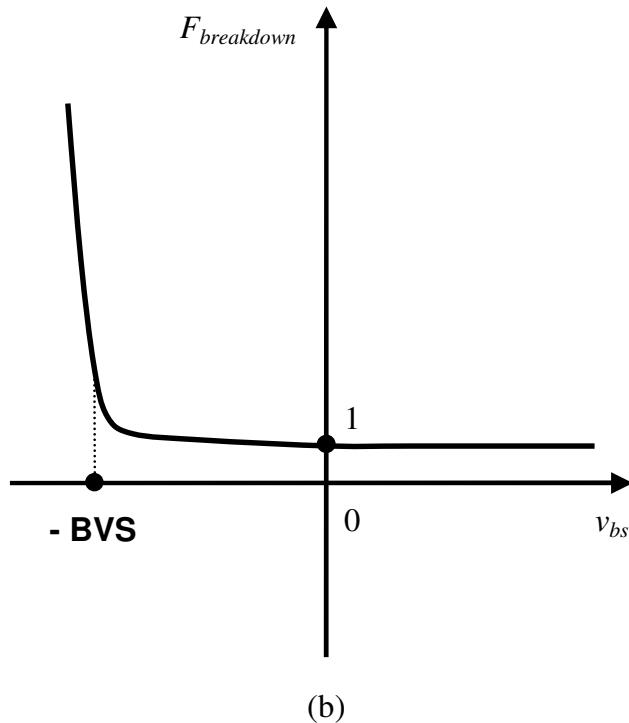
$G_{\min}$  is a small, artificial conductance added in parallel with the junction to aid SPICE convergence in reverse-bias operation. The conductance of the junction itself in this region is usually so small (around  $10^{-12}$  siemens or even less) that convergence difficulty can arise otherwise.  $F_{breakdown}$  is called the breakdown factor and is empirically taken to be

$$F_{breakdown} = 1 + \text{XJBVS} \cdot \exp\left(-\frac{q \cdot (\text{BVS} + v_{bs})}{\text{NJS} \cdot k_B T_{emp}}\right) \quad (9.7)$$

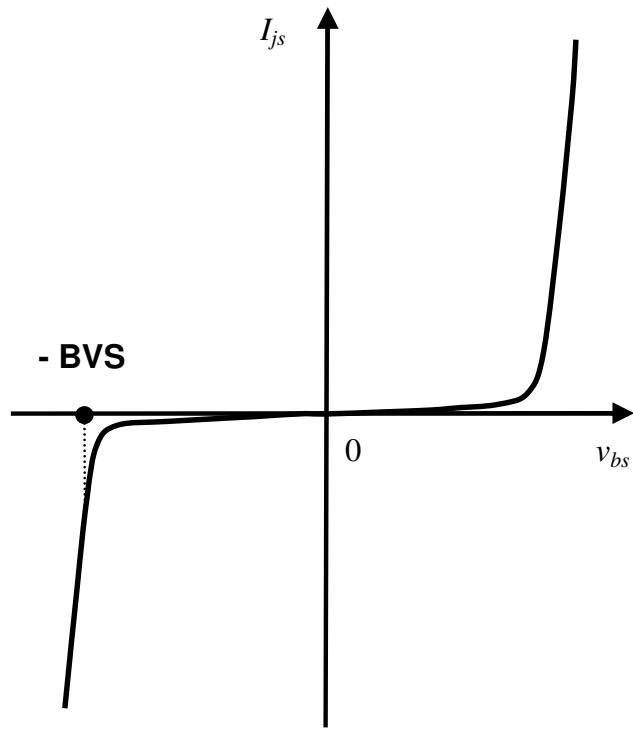
**XJBVS** is called the breakdown coefficient. It has the default value of 1. **BVS** is the source-bulk junction breakdown voltage with a default value of 10 volts. These two parameters cannot be negative. In the case of **DIOMOD** = 0, **XJBVS** will be reset to 1 if it is less than zero. When **XJBVS** = 0, there will be no breakdown to model.

$F_{breakdown}$  of Eq. (9.7) introduces a very useful approach to modeling device breakdown continuously and robustly for SPICE simulation, which had not been possible prior to BSIM4.  $F_{breakdown}$  and Eq. (9.6) are plotted in Fig. 9.5 (b) and (c), respectively.





(b)



(c)

Fig. 9.5 (a)  $I_{js}$  versus  $v_{bs}$  in the case of NMOS without the breakdown factor  $F_{breakdown}$ . (b) The breakdown factor  $F_{breakdown}$  increases exponentially when  $v_{bs} < -BVS$  and approaches 1 when  $v_{bs} > -BVS$ . (c)  $I_{js}$  versus  $v_{bs}$  when a breakdown is present and modeled.

### 9.3.2 DIOMOD = 1

A real junction diode unavoidably has a finite parasitic series resistance, which makes the forward-bias IV characteristics deviate from the ideal exponential dependence on the junction voltage. The BSIM4 DIOMOD = 0 model, as discussed above, ignores this resistance. This usually does not give rise to accuracy concerns as in practice the junction diodes of bulk CMOS are rarely biased into the forward operation mode.

For robust SPICE simulation, however, such large forward currents must be curbed. Otherwise it could give rise to a singular circuit matrix exception in the numerical iterative process (such as Newton-Raphson) during SPICE simulations. Inserting a series resistance, on the other hand, requires an additional node with extra cost to simulation time. For these reasons, the BSIM4 DIOMOD = 1 model (the default option) simply linearize the forward exponential IV dependence beyond a given current level. This treatment provides an effective and efficient current limiting capability for better SPICE simulation convergence when the forward junction voltage is very large.

For the sake of efficient SPICE simulation, the BSIM4 DIOMOD = 1 model does not contain a breakdown model. It is formulated as

$$I_{js} = I_{js,sat} \cdot \left[ \exp\left(\frac{qv_{bs}}{\text{NJS} \cdot k_B T_{emp}}\right) - 1 \right] + G_{\min} \cdot v_{bs} \quad (9.8)$$

The exponential term is linearized such that  $I_{js}$  becomes

$$I_{js} = \text{IJTHSFWD} + k_{slope} \cdot (v_{bs} - VjsmFwd) + G_{\min} \cdot v_{bs} \quad (9.9)$$

when  $I_{js}$  is greater than **lJTHSFWD**, a model parameter that represents the forward current beyond which the source-bulk junction forward-bias current is linearized to improve convergence.  $VjsmFwd$  of Eq. (9.9) is the  $v_{bs}$  that makes the exponential term in Eq. (9.8) equal to **lJTHSFWD**

$$VjsmFwd = \frac{\text{NJS} \cdot k_B T_{emp}}{q} \cdot \log\left(1 + \frac{\text{IJTHSFWD}}{I_{js,sat}}\right) \quad (9.10)$$

$k_{slope}$  of Eq. (9.9) is the derivative of  $I_{js}$  (Eq. (9.8)) with respect to  $v_{bs}$  at  $VjsmFwd$ . It is

$$k_{slope} = \frac{q \cdot I_{js,sat}}{\mathbf{NJS} \cdot k_B T_{emp}} \cdot \exp\left(\frac{q \cdot VjsmFwd}{\mathbf{NJS} \cdot k_B T_{emp}}\right) \quad (9.11)$$

These parameters ensure that Eqs. (9.8) and (9.9) are continuous at  $VjsmFwd$  up to the first-order derivative of  $I_{js}$  with respect to  $v_{bs}$ . Fig. 9.6 illustrates the concept of the linearization performed for  $\mathbf{DIOMOD} = 1$ .

One would think that Eq. (9.9) is mathematically sound as a robust linear extension of Eq. (9.8) when  $v_{bs}$  is greater than  $VjsmFwd$ . In fact, implementing Eq. (9.9) as is into a SPICE program is problematic. This is because a typical numerical overflow protection that needs to be applied to the exponential term of Eq. (9.8) can make Eq. (9.9) no longer continuous with Eq. (9.8) at  $v_{bs} = VjsmFwd$ . As a remedy, a new version of Eq. (9.9) that is mathematically equivalent but numerically robust for SPICE implementation is employed in BSIM4. It is

$$\begin{aligned} I_{js} &= I_{js,sat} \cdot \left[ \exp\left(\frac{q \cdot VjsmFwd}{\mathbf{NJS} \cdot k_B T_{emp}}\right) - 1 \right] \\ &+ k_{slope} \cdot (v_{bs} - VjsmFwd) + G_{min} \cdot v_{bs} \end{aligned} \quad (9.12)$$

when  $v_{bs}$  is greater than  $VjsmFwd$ .

### 9.3.3 DIOMOD = 2

The BSIM4  $\mathbf{DIOMOD} = 2$  option is a comprehensive junction diode model. Junction diode breakdown is always modeled. Moreover, current limiting by linearization is always applied to both the forward and reverse IV equations. Fig. 9.7 illustrates these features of  $\mathbf{DIOMOD} = 2$  graphically. In this section, the  $\mathbf{DIOMOD} = 2$  IV formulations with breakdown and current limiting are presented. Only the source-bulk junction diode model is presented. The drain-bulk diode has an analogous set of formulations and they are omitted in the following.

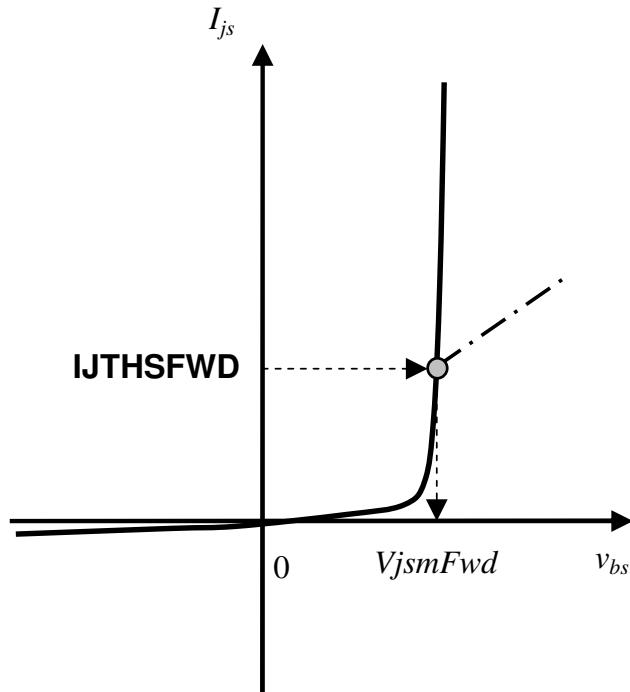


Fig. 9.6 Linearization of  $I_{js}$  of DIOMOD = 1 at large forward current. When the forward current is greater than IJTHSFWD or, equivalently,  $v_{bs}$  is larger than  $VjsmFwd$ , the IV curve becomes linearly dependent on  $v_{bs}$  (the dash-dotted line).

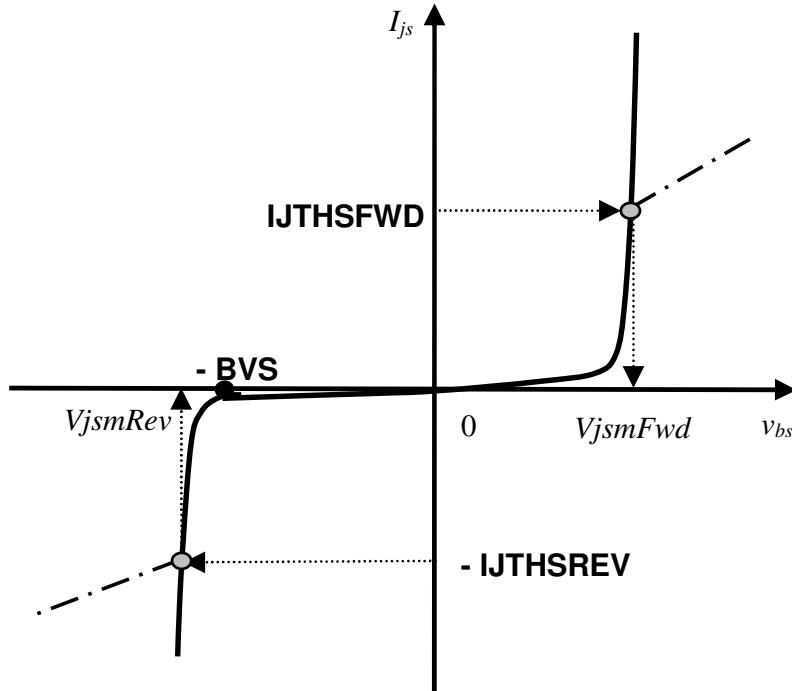


Fig. 9.7 Junction diode current breakdown modeling and linearization with DIOMOD = 2. When the forward junction current is greater than a user-specified value IJTHSFWD, it is linearized, at which point, the exponential and the linear curves have the same slope. The same linearization technique is applied to the reverse bias regime with the parameter IJTHSREV.

As can be seen from Fig. 9.7, the IV formulations can be divided into the following three distinctive regions. When the junction current is between **IJTHSREV** and **IJTHSFWD** or, equivalently,  $v_{bs}$  is between  $V_{jsmRev}$  and  $V_{jsmFwd}$ , the junction current is still described by Eqs. (9.6) and (9.7) and they are repeated here with the  $G_{min}$  term omitted for convenience of analysis:

$$I_{js} = I_{js,sat} \cdot \left[ \exp\left(\frac{qv_{bs}}{\text{NJS} \cdot k_B T_{emp}}\right) - 1 \right] \cdot F_{breakdown} \quad (9.13)$$

and

$$F_{breakdown} = 1 + \text{XJBVS} \cdot \exp\left(-\frac{q \cdot (\text{BVS} + v_{bs})}{\text{NJS} \cdot k_B T_{emp}}\right) \quad (9.14)$$

When the junction current is greater than **IJTHSREV** or **IJTHSFWD** (or equivalently  $v_{bs}$  is less than  $V_{jsmRev}$  or larger than  $V_{jsmFwd}$ , respectively), the junction current is linearized.

Consider first the forward bias region where  $I_{js}$  is larger than **IJTHSFWD**. To solve for  $V_{jsmFwd}$ , first substitute Eq. (9.14) into Eq. (9.13) and equate Eq. (9.13) and **IJTHSFWD**. After regrouping similar terms, one obtains

$$\left[ \exp\left(\frac{q \cdot V_{jsmFwd}}{\text{NJS} \cdot k_B T_{emp}}\right) \right]^2 - B \cdot \exp\left(\frac{q \cdot V_{jsmFwd}}{\text{NJS} \cdot k_B T_{emp}}\right) - C = 0 \quad (9.15)$$

where

$$B = 1 + \frac{\text{IJTHSFWD}}{I_{js,sat}} - \text{XJBVS} \cdot \exp\left(-\frac{q \cdot \text{BVS}}{\text{NJS} \cdot k_B T_{emp}}\right) \quad (9.15a)$$

and

$$C = \text{XJBVS} \cdot \exp\left(-\frac{q \cdot \text{BVS}}{\text{NJS} \cdot k_B T_{emp}}\right) \quad (9.15c)$$

Note that  $C$  is positive since  $XJBVS$  has to be positive. The solution to Eq. (9.15) is

$$\exp\left(\frac{q \cdot VjsmFwd}{NJS \cdot k_B T_{emp}}\right) = \frac{B \pm \sqrt{B^2 + 4C}}{2} \quad (9.16)$$

The minus sign preceding the square root must be dropped since the left-hand side must be positive. Therefore,

$$VjsmFwd = \frac{NJS \cdot k_B T_{emp}}{q} \cdot \log\left(\frac{B + \sqrt{B^2 + 4C}}{2}\right) \quad (9.17)$$

Note that Eq. (9.17) reduces to Eq. (10) of  $DIOMOD = 0$  if  $XJBVS$  is zero. The source-bulk junction current  $IVjsmFwd$  at  $v_{bs} = VjsmFwd$  and its first-order derivative  $k_{slopeFwd}$  with respect to  $v_{bs}$  can be obtained without much effort by substituting  $v_{bs} = VjsmFwd$  into Eq. (9.13). Therefore, the linearized  $I_{js}$  in the forward-bias region when  $I_{js}$  is greater than  $IJTFSFWD$  becomes

$$\begin{aligned} I_{js} &= IVjsmFwd \\ &+ k_{slopeFwd} \cdot (v_{bs} - VjsmFwd) + G_{min} \cdot v_{bs} \end{aligned} \quad (9.18)$$

This approach is similar to that employed in deriving Eq. (9.12). Note that Eq. (9.18) has the  $G_{min}$  term added back, which was temporarily dropped earlier. One can verify that Eq. (9.13) with the  $G_{min}$  term and Eq. (9.18) are smooth and continuous at  $v_{bs} = VjsmFwd$  up to their first-order derivative.

The linearization of the source-bulk junction diode current in the reverse body bias region for  $DIOMOD = 2$  is needed because of the presence of the exponential breakdown factor  $F_{breakdown}$  in Eqs. (9.13) and (9.14). Both are repeated here with the  $G_{min}$  term temporarily omitted for convenience

$$I_{js} = I_{js,sat} \cdot \left[ \exp\left(\frac{qv_{bs}}{NJS \cdot k_B T_{emp}}\right) - 1 \right] \cdot F_{breakdown} \quad (9.13)$$

and

$$F_{breakdown} = 1 + \text{XJBVS} \cdot \exp\left(-\frac{q \cdot (\text{BVS} + v_{bs})}{\text{NJS} \cdot k_B T_{emp}}\right) \quad (9.14)$$

A quick examination of Eq. (9.13) leads to the discovery that the exponential term becomes extremely small in magnitude for a large negative  $v_{bs}$  whereas  $F_{breakdown}$  increases exponentially at large negative  $v_{bs}$ . This can be seen graphically in Fig. 9.5 (b). Therefore, one desires to perform the linearization on the product of  $I_{js,sat}$  and  $F_{breakdown}$  once the junction diode current becomes larger in magnitude than the user-specified value of the parameter **IJTHSREV**. In order to find the voltage  $VjsmRev$  that corresponds to **IJTHSREV**, start with the equation

$$I_{js,sat} \cdot F_{breakdown} = \text{IJTHSREV} \quad (9.19)$$

This yields

$$VjsmRev = -\text{BVS} - \frac{\text{NJS} \cdot k_B T_{emp}}{q} \cdot \log\left(\frac{\text{IJTHSREV} - I_{js,sat}}{I_{js,sat} \cdot \text{XJBVS}}\right) \quad (9.20)$$

upon substituting Eq. (9.14) into Eq. (9.19). Note from Eq. (9.20) that a numerically robust linearization requires **IJTHSREV** to be greater than  $I_{js,sat}$ ; otherwise a logarithmic domain error will occur.

Substituting  $VjsmRev$  into the left-hand side of Eq. (9.19) yields the current  $IVjsmRev$  that corresponds to  $VjsmRev$

$$IVjsmRev = I_{js,sat} \cdot \left\{ 1 + \text{XJBVS} \cdot \exp\left(-\frac{q \cdot (\text{BVS} + VjsmRev)}{\text{NJS} \cdot k_B T_{emp}}\right) \right\} \quad (9.21)$$

The first-order derivative of the left-hand side of Eq. (9.19) with respect to  $v_{bs}$  when  $v_{bs} = VjsmRev$  is

$$k_{slopeRev} = -I_{js,sat} \cdot \frac{q \cdot \text{XJBVS}}{\text{NJS} \cdot k_B T_{emp}} \cdot \exp\left[-\frac{q(\text{BVS} + VjsmRev)}{\text{NJS} \cdot k_B T_{emp}}\right] \quad (9.22)$$

With this, the linearized junction breakdown current (when the magnitude of the junction current is greater than **IJTHSREV**) is now found to be

$$I_{js} = \left[ \exp\left( \frac{qv_{bs}}{\text{NJS} \cdot k_B T_{emp}} \right) - 1 \right] \cdot [IVjsmRev + k_{slopeRev} \cdot (v_{bs} - VjsmRev)] + G_{\min} \cdot v_{bs} \quad (9.23)$$

Again, one can verify that Eq. (9.13) with the  $G_{\min}$  term and Eq. (9.23) are smooth and continuous at  $v_{bs} = VjsmRev$  up to their first-order derivative.

The same breakdown model and linearization technique are applied to the drain-bulk junction diode current model, for which BSIM4 uses a set of formulas and parameters that are analogous to those of the source-bulk junction diode IV model. For the sake of brevity, these formulas and parameters are not presented here. Refer to the parameter table at the end of this chapter for details.

## 9.4 BSIM4 Junction Leakage Due to Trap-Assisted Tunneling [4]

Junction leakage due to trap-assisted tunneling is modeled in BSIM4, in addition to what has already been presented in the preceding section. By including TAT, the total source-bulk junction current modeled by BSIM4 becomes

$$\begin{aligned} I_{js,total} &= I_{js} \\ &- AS_{eff} \cdot J_{s,bottom,TAT}(T) \\ &\cdot \left\{ \exp\left[ -\frac{q \cdot v_{bs}}{NJS_{TAT}(T) \cdot k_B \cdot TNOM} \cdot \frac{VTSS}{VTSS - v_{bs}} \right] - 1 \right\} \\ &- PS_{eff} \cdot J_{s,sidewall,TAT}(T) \\ &\cdot \left\{ \exp\left[ -\frac{q \cdot v_{bs}}{NJSSW_{TAT}(T) \cdot k_B \cdot TNOM} \cdot \frac{VTSSWS}{VTSSWS - v_{bs}} \right] - 1 \right\} \\ &- W_{effJCT} \cdot NF \cdot J_{s,gate-edge,TAT}(T) \\ &\cdot \left\{ \exp\left[ -\frac{q \cdot v_{bs}}{NJSSWG_{TAT}(T) \cdot k_B \cdot TNOM} \cdot \frac{VTSSWGS}{VTSSWGS - v_{bs}} \right] - 1 \right\} \end{aligned} \quad (9.24)$$

Here  $I_{js}$  is the current computed from the equations given in the previous sections including the  $(G_{\min} \cdot v_{bs})$  term. VTSS, VTSSWS, and VTSSWGS are the junction bias dependence parameters of the bottom, STI sidewall and gate-edge TAT current components.  $NJS_{TAT}(T)$ ,  $NJSSW_{TAT}(T)$ , and  $NJSSWG_{TAT}(T)$  are the temperature-dependent non-ideality factors of the bottom, STI sidewall and gate-edge TAT currents.  $J_{s,bottom,TAT}(T)$ ,  $J_{s,sidewall,TAT}(T)$ , and  $J_{s,gate-edge,TAT}(T)$  are the temperature-dependent TAT saturation current densities of the junction bottom, the STI edge, and the gate-edge, respectively. The temperature-dependence will be discussed further in the temperature-dependence section later in this chapter.

As discussed previously, trap-assisted tunneling is only significant in the reverse body bias region. Thus, the three terms in the curly brackets on the right-hand side of Eq. (9.24) are positive in the reverse bias regime and diminish to zero in the forward bias regime. Hence, trap-assisted tunneling currents will add to and raise the leakage current  $I_{js}$ . In the BSIM4 implementation, the TAT current component does not change the breakdown and current limiting modeling.

## 9.5 BSIM4 Diode Charge and Capacitance [4]

The source/drain-bulk junction diodes of floating-body MOS transistors such as SOI-MOSFET may be forward biased and the junction *diffusion* capacitance may not be ignored. However, in bulk CMOS, the junction diodes are rarely forward biased in circuit operation. Therefore, it is sufficient to only consider the junction *depletion* capacitance component for the junction capacitances in SPICE modeling. According to device physics, this capacitance is determined by the bias-dependent junction depletion-layer thickness or width.

From Poisson equation, one can derive the junction depletion width

$$W_{dep} = W_{dep}(V_j = 0) \cdot \left(1 - \frac{V_j}{V_{bi}}\right)^{MJ} \quad (9.25)$$

$W_{dep}(V_j = 0)$  is the zero-bias junction depletion width.  $V_j$  is the externally applied junction voltage.  $V_{bi}$  is the built-in voltage.  $MJ$  is the junction grading coefficient, approximately equaling 0.5 for a step junction and 1/3 for a linearly graded junction; and it is usually extracted from measured CV data. The junction depletion capacitance per unit length or area (depending upon whether a periphery or an area component of the junction capacitance is under consideration) is

$$c_j = \frac{\epsilon_{Si}}{W_{dep}} = \frac{c_{j0}}{\left(1 - \frac{V_j}{V_{bi}}\right)^{MJ}} \quad (9.26)$$

where  $\epsilon_{Si}$  is the permittivity of silicon and  $c_{j0}$  is the zero-bias junction depletion capacitance per unit length or area

$$c_{j0} = \frac{\epsilon_{Si}}{W_{dep}(V_j = 0)} \quad (9.26a)$$

The junction depletion charge density (per unit length or area) is obtained by integrating Eq. (9.26) over  $V_j$  from zero to any junction diode voltage applied

$$q_j = \int_0^{V_j} \frac{c_{j0}}{\left(1 - \frac{V_j}{V_{bi}}\right)^{MJ}} dV_j = \frac{c_{j0} \cdot V_{bi}}{1 - MJ} \left[ 1 - \left(1 - \frac{V_j}{V_{bi}}\right)^{1-MJ} \right] \quad (9.27)$$

From Eqs. (9.25) through (9.27), one realizes that  $MJ$  must not be greater than or equal to 1 and  $V_j$  must not be positive.

### 9.5.1 BSIM4 Diode CV Model [4]

In the following, the BSIM4 model of source-bulk diode charge and capacitance as a function of the diode voltage is detailed for the case of n-channel MOSFET transistors. A similar model and analysis exists for the drain-bulk junction diode with a separate set of model parameters to permit the modeling of asymmetrical source and drain structures. (Refer

to the parameter table section at the end of this chapter.) Therefore, the drain-bulk junction diode CV model will not be presented. Furthermore, the same model is used for p-channel transistors. In SPICE implementation, the opposite voltage polarities are used for n-type and p-type transistor junction diodes, details of which are presented in Chapter 10. Hence, the junction CV model of a p-channel MOSFET will also be skipped in the following.

The source-bulk diode charge and capacitance are also divided into three components: Those attributed to the junction bottom, the STI sidewall, and the gate edge. The junction depletion capacitance and charge components due to the bottom junction are

$$C_{js,bottom} = C_{js,bottom}(v_{bs} = 0) \cdot \left[ 1 - \frac{v_{bs}}{PBS(T)} \right]^{-MJS} \quad (9.28a)$$

and

$$Q_{js,bottom} = \int_0^{v_{bs}} C_{js,bottom} \cdot dv_{bs} = \frac{C_{js,bottom}(v_{bs} = 0) \cdot PBS(T)}{1 - MJS} \cdot \left[ 1 - \left( 1 - \frac{v_{bs}}{PBS(T)} \right)^{1-MJS} \right] \quad (9.28b)$$

$MJS$  and  $PBS(T)$  are the source-bulk junction grading coefficient and built-in potential. The zero-bias bottom junction capacitance was given in Chapter 8. It is repeated here

$$C_{js,bottom}(v_{bs} = 0) = AS_{eff} \cdot C_{js,bottom,unit-area}(T) \quad (8.25a)$$

in which  $AS_{eff}$  is the effective bottom junction area. In the SPICE BSIM4 code implementation,  $C_{js,bottom}$  and  $Q_{js,bottom}$  are both set to zero if  $C_{js,bottom}(v_{bs} = 0)$  is not positive. The temperature dependencies in Eqs. (9.28a) and (8.25a) will be discussed later in this chapter.

In the case of the STI junction sidewall edge, the junction depletion capacitance and charge components are computed in a similar way

$$C_{js, sidewall} = C_{js, sidewall}(v_{bs} = 0) \cdot \left[ 1 - \frac{v_{bs}}{PBSWS(T)} \right]^{-MJSWS} \quad (9.29a)$$

and

$$Q_{js, sidewall} = \int_0^{v_{bs}} C_{js, sidewall} \cdot dv_{bs} = \frac{C_{js, sidewall}(v_{bs} = 0) \cdot PBSWS(T)}{1 - MJSWS} \cdot \left[ 1 - \left( 1 - \frac{v_{bs}}{PBSWS(T)} \right)^{1-MJSWS} \right] \quad (9.29b)$$

$MJSWS$  and  $PBSWS(T)$  are the grading coefficient and temperature-dependent built-in potential of the junction along the STI edge. The zero-bias temperature-dependent STI source-bulk junction capacitance, first given in Chapter 8, is

$$C_{js, sidewall}(v_{bs} = 0) = PS_{eff} \cdot C_{js, sidewall, unit-length}(T) \quad (8.25b)$$

$PS_{eff}$  is the effective STI-edge junction periphery length. In the SPICE implementation of BSIM4,  $C_{js, sidewall}$  and  $Q_{js, sidewall}$  both will be set to zero if  $C_{js, sidewall}(v_{bs} = 0)$  is not positive.

Similarly, in the case of the gate edge of the junction, the junction depletion capacitance and charge components are modeled as

$$C_{js, gate-edge} = C_{js, gate-edge}(v_{bs} = 0) \cdot \left[ 1 - \frac{v_{bs}}{PBSWGS(T)} \right]^{-MJSWGS} \quad (9.30a)$$

and

$$Q_{js, gate-edge} = \int_0^{v_{bs}} C_{js, gate-edge} \cdot dv_{bs} = \frac{C_{js, gate-edge}(v_{bs} = 0) \cdot PBSWGS(T)}{1 - MJSWS} \cdot \left[ 1 - \left( 1 - \frac{v_{bs}}{PBSWGS(T)} \right)^{1-MJSWGS} \right] \quad (9.30b)$$

$MJSWGS$  and  $PBSWGS(T)$  are the grading coefficient and temperature-dependent built-in potential of the gate-edge junction. The zero-bias gate-edge source-bulk junction capacitance from Eq. (8.25c) is repeated

$$C_{js,gate-edge}(V_{bs} = 0) = W_{effJCT} \cdot NF \cdot C_{js,gate-edge,unit-length}(T) \quad (8.25c)$$

In the SPICE implementation,  $C_{js,gate-edge}$  and  $Q_{js,gate-edge}$  are both set to zero if  $C_{js,gate-edge}(v_{bs} = 0)$  is not positive. The temperature dependencies given in Eqs. (9.30a) through (8.25c) above will be analyzed in the next section.

The depletion junction capacitances and charges derived above are correct only for the reverse body bias regime. Assume the source-bulk junction voltage  $v_{bs}$  is greater than zero (i.e., the junction is in the forward biased mode). The junction depletion capacitance and charge components contributed by the junction bottom are

$$C_{js,bottom} = C_{js,bottom}(v_{bs} = 0) \cdot \left[ 1 + MJS \cdot \frac{v_{bs}}{PBS(T)} \right] \quad (9.31a)$$

and

$$Q_{js,bottom} = \int_0^{v_{bs}} C_{js,bottom} \cdot dv_{bs} = C_{js,bottom}(v_{bs} = 0) \cdot v_{bs} \left[ 1 + \frac{v_{bs} \cdot MJS}{2 \cdot PBS(T)} \right] \quad (9.31b)$$

Note that the junction capacitances of Eqs. (9.31a) and (9.28a) and the junction charge of Eqs. (9.31b) and (9.28b) are continuous at  $v_{bs} = 0$ . This is illustrated in Fig. 9.8. Similar models exist for the junction capacitances and charges at the STI sidewall edge and the gate edge.

The junction depletion capacitance and charge components contributed by the STI sidewall edge are modeled with

$$C_{js,sidewall} = C_{js,sidewall}(v_{bs} = 0) \cdot \left[ 1 + MJSWS \cdot \frac{v_{bs}}{PBSWS(T)} \right] \quad (9.32a)$$

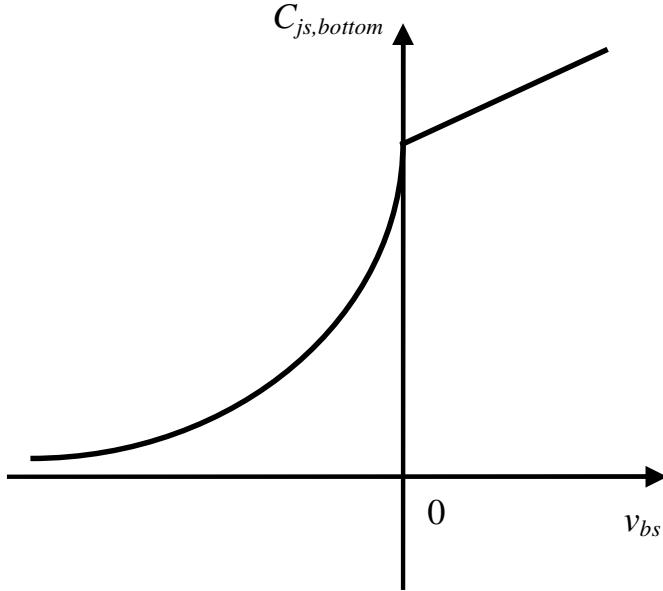


Fig. 9.8 The bottom-junction component of the source-body junction depletion capacitance versus  $v_{bs}$ . The piece-wise depletion capacitance formulations in Eqs. (9.28a) and (9.31a) are continuous at  $v_{bs} = 0$ , so are the other capacitance components and the junction depletion charges.

and

$$Q_{js,\text{sidewall}} = \int_0^{v_{bs}} C_{js,\text{sidewall}} \cdot dv_{bs} = C_{js,\text{sidewall}}(v_{bs} = 0) \cdot v_{bs} \left[ 1 + \frac{v_{bs} \cdot \text{MJSWS}}{2 \cdot \text{PBSWS}(T)} \right] \quad (9.32\text{b})$$

when the junction is forward biased.

In the same fashion, the junction depletion capacitance and charge components contributed by the gate edge are modeled as

$$C_{js,\text{gate-edge}} = C_{js,\text{gate-edge}}(v_{bs} = 0) \cdot \left[ 1 + \text{MJSWGS} \cdot \frac{v_{bs}}{\text{PBSWGS}(T)} \right] \quad (9.33\text{a})$$

and

$$Q_{js,\text{gate-edge}} = \int_0^{v_{bs}} C_{js,\text{gate-edge}} \cdot dv_{bs} = C_{js,\text{gate-edge}}(v_{bs} = 0) \cdot v_{bs} \left[ 1 + \frac{v_{bs} \cdot \text{MJSWGS}}{2 \cdot \text{PBSWGS}(T)} \right] \quad (9.33\text{b})$$

for a forward-biased source-bulk junction.

The total depletion capacitance and charge of the source-bulk junction diode are the sum of the components contributed by the bottom, the STI sidewall, and the gate edge:

$$C_{js} = C_{js,bottom} + C_{js,sidewall} + C_{js,gate-edge} \quad (9.34a)$$

and

$$Q_{js} = Q_{js,bottom} + Q_{js,sidewall} + Q_{js,gate-edge} \quad (9.34b)$$

for all values of  $v_{bs}$ , whether in the forward or reverse bias regime.

Note that the depletion capacitance and charge of the drain junction diode are modeled similarly but with a different set of parameters because the source and drain junctions may have different doping profiles and geometries because of process variations or certain circuit design needs.

## 9.6 Diode Temperature-Dependence Model [4]

As mentioned before, the MOSFET junction diode IV and CV models need to include the temperature dependencies of the parameters and variables such as the saturation leakage current densities, built-in potentials, and zero-bias junction depletion capacitances.

The temperature dependencies of these parameters and variables can be traced to the temperature dependences of the intrinsic carrier density, the silicon energy band gap  $E_g(T)$ , the Fermi level, and the densities of energy states. A complete description of all these temperature dependencies involves complex formulations. In practice, the temperature dependences are modeled semi-empirically for ease of model parameter extraction and fast SPICE simulation.

In the following two sub-sections, the BSIM4 temperature-dependence models of the junction diodes are presented for IV and CV. The effort here will be focused on the source diode. The same approach can be applied to derive the drain diode IV and CV models.

### 9.6.1 Temperature-Dependence Model for Diode IV

It is known from Eq. (9.5a) that the total source junction saturation current

$$I_{js,sat} = AS_{eff} \cdot J_{s,bottom}(T) + PS_{eff} \cdot J_{s,sidewall}(T) + W_{effJCT} \cdot NF \cdot J_{s,gate-edge}(T) \quad (9.5a)$$

consists of the components attributed to the junction bottom, the STI sidewall, and the gate edge. They are temperature dependent, which is modeled as

$$J_{s,bottom}(T) = JS \cdot \exp \left[ \frac{\left( \frac{q \cdot E_g(T_{NOM})}{k_B \cdot T_{NOM}} - \frac{q \cdot E_g(T_{emp})}{k_B \cdot T_{emp}} + XTIS \cdot \log \left( \frac{T_{emp}}{T_{NOM}} \right) \right)}{NJS} \right] \quad (9.35)$$

for the bottom junction.

$$J_{s,sidewall}(T) = JSWS \cdot \exp \left[ \frac{\left( \frac{q \cdot E_g(T_{NOM})}{k_B \cdot T_{NOM}} - \frac{q \cdot E_g(T_{emp})}{k_B \cdot T_{emp}} + XTIS \cdot \log \left( \frac{T_{emp}}{T_{NOM}} \right) \right)}{NJS} \right] \quad (9.36)$$

for the STI sidewall junction, and

$$J_{s,gate-edge}(T) = JSWGS \cdot \exp \left[ \frac{\left( \frac{q \cdot E_g(T_{NOM})}{k_B \cdot T_{NOM}} - \frac{q \cdot E_g(T_{emp})}{k_B \cdot T_{emp}} + XTIS \cdot \log \left( \frac{T_{emp}}{T_{NOM}} \right) \right)}{NJS} \right] \quad (9.37)$$

for the gate-edge junction. Note that in order to achieve numerical stability,  $J_{s,bottom}(T)$ ,  $J_{s,sidewall}(T)$  and  $J_{s,gate-edge}(T)$  are implemented in BSIM4 in a way such that each of these components is set to zero if that component becomes negative. In Eqs. (9.35) through (9.37), the silicon energy band gap  $E_g(T)$  is given in the unit of volts instead of electron-volts (eV) as typically given in text books for the sake of efficient model formulation. It is formulated as

$$E_g(T_{NOM}) = 1.16 - \frac{7.02 \times 10^{-4} \cdot T_{NOM}^2}{T_{NOM} + 1108} \quad (9.38a)$$

at the nominal temperature,  $T_{NOM}$ , at which the BSIM4 model parameters are extracted, and

$$E_g(T_{emp}) = 1.16 - \frac{7.02 \times 10^{-4} T_{emp}^2}{T_{emp} + 1108} \quad (9.38b)$$

at the device/circuit operating temperature  $T_{emp}$ .  $T_{NOM}$  and  $T_{emp}$  have the unit of degree Kelvin.

The temperature dependencies of the saturation leakage current and the non-ideality factor are also modeled in BSIM4. Recall from Eq. (9.24) that the total leakage current of the source-bulk junction diode that includes the trap-assisted tunneling current component is

$$\begin{aligned} I_{js,total} &= I_{js} \\ &- AS_{eff} \cdot J_{s,bottom,TAT}(T) \\ &\cdot \left\{ \exp \left[ - \frac{q \cdot v_{bs}}{NJS_{TAT}(T) \cdot k_B \cdot T_{NOM}} \cdot \frac{VTSS}{VTSS - v_{bs}} \right] - 1 \right\} \\ &- PS_{eff} \cdot J_{s,sidewall,TAT}(T) \\ &\cdot \left\{ \exp \left[ - \frac{q \cdot v_{bs}}{NJSSW_{TAT}(T) \cdot k_B \cdot T_{NOM}} \cdot \frac{VTSSWS}{VTSSWS - v_{bs}} \right] - 1 \right\} \\ &- W_{effJCT} \cdot NF \cdot J_{s,gate-edge,TAT}(T) \\ &\cdot \left\{ \exp \left[ - \frac{q \cdot v_{bs}}{NJSSWG_{TAT}(T) \cdot k_B \cdot T_{NOM}} \cdot \frac{VTSSWGS}{VTSSWGS - v_{bs}} \right] - 1 \right\} \end{aligned} \quad (9.24)$$

$NJS_{TAT}(T)$ ,  $NJSSW_{TAT}(T)$ , and  $NJSSWG_{TAT}(T)$  are the temperature-dependent non-ideality factors of the bottom, STI sidewall and gate-edge components of the TAT current.  $J_{s,bottom,TAT}(T)$ ,  $J_{s,sidewall,TAT}(T)$ ,  $J_{s,gate-edge,TAT}(T)$  are the temperature-dependent TAT saturation current densities of the junction bottom, the STI sidewall and the gate edge. Their temperature dependencies are

$$NJS_{TAT}(T) = NJTS \cdot \left[ 1 + TNJTS \cdot \left( \frac{T_{emp}}{TNOM} - 1 \right) \right] \quad (9.39)$$

for the non-ideality factor of the bottom junction;

$$NJSSW_{TAT}(T) = NJTSSW \cdot \left[ 1 + TNJTSSW \cdot \left( \frac{T_{emp}}{TNOM} - 1 \right) \right] \quad (9.40)$$

for the non-ideality factor of the STI sidewall junction;

$$NJSSWG_{TAT}(T) = NJTSSSWG \cdot \left[ 1 + TNJTSSSWG \cdot \left( \frac{T_{emp}}{TNOM} - 1 \right) \right] \quad (9.41)$$

for the non-ideality factor of the gate-edge junction;

$$J_{s,bottom,TAT}(T) = JTSS \cdot \exp \left[ XTSS \cdot \frac{q \cdot E_g(TNOM)}{k_B T_{emp}} \cdot \left( \frac{T_{emp}}{TNOM} - 1 \right) \right] \quad (9.42)$$

for the saturation current of the bottom junction due to TAT;

$$J_{s,sidewall,TAT}(T) = JTSSWS \cdot \exp \left[ XTSSWS \cdot \frac{q \cdot E_g(TNOM)}{k_B T_{emp}} \cdot \left( \frac{T_{emp}}{TNOM} - 1 \right) \right] \quad (9.43)$$

for the saturation current of the STI junction due to TAT; and

$$J_{s,gate-edge,TAT}(T) = JTSSWGS \cdot \exp \left[ XTSSWGS \cdot \frac{q \cdot E_g(TNOM)}{k_B T_{emp}} \cdot \left( \frac{T_{emp}}{TNOM} - 1 \right) \right] \quad (9.44)$$

for the saturation current density of the gate-edge junction due to TAT.

### 9.6.2 Diode CV Temperature-Dependence Model

The zero-bias unit-length/area junction capacitance and the junction built-in potential are dependent on temperatures. This section describes their temperature dependence models. Again, only the source junction diode is treated here. The drain junction model can be readily derived by following the same procedure.

Take the source bottom junction as an example and repeat Eqs. (9.28a) and (9.28b) for ease of reference

$$C_{js,bottom} = C_{js,bottom}(v_{bs} = 0) \cdot \left[ 1 - \frac{v_{bs}}{PBS(T)} \right]^{-MJS} \quad (9.28a)$$

$$Q_{js,bottom} = \frac{C_{js,bottom}(v_{bs} = 0) \cdot PBS(T)}{1 - MJS} \cdot \left[ 1 - \left( 1 - \frac{v_{bs}}{PBS(T)} \right)^{1-MJS} \right] \quad (9.28b)$$

The temperature dependencies of the zero-bias unit-area junction capacitance  $C_{js,bottom,unit-area}(T)$  and the built-in potential  $PBS(T)$  of are modeled as

$$C_{js,bottom,unit-area}(T) = CJS \cdot [1 + TCJ \cdot (T_{emp} - TNOM)] \quad (9.45)$$

which will be set to zero if it is negative, and

$$PBS(T) = PBS - TPB \cdot (T_{emp} - TNOM) \quad (9.46)$$

which will be set to 0.01 if it is less than this value.

The temperature dependencies of the zero-bias unit-length STI sidewall capacitance  $C_{js,sidewall,unit-length}(T)$  and the built-in potential  $PBSWS(T)$  are modeled similarly

$$C_{js,sidewall,unit-length}(T) = CJSWS \cdot [1 + TCJSW \cdot (T_{emp} - TNOM)] \quad (9.47)$$

which will also be set to zero if it is negative.

$$PBSWS(T) = PBSWS - TPBSW \cdot (T_{emp} - TNOM) \quad (9.48)$$

which will also be set to 0.01 if it is less than this value.

Likewise, the temperature dependencies of the zero-bias unit-length gate-edge junction capacitance  $C_{js, \text{gate-edge}, \text{unit-length}}(T)$  and the built-in potential  $PBSWGS(T)$  are

$$C_{js, \text{gate-edge}, \text{unit-length}}(T) = CJSWGS \cdot [1 + TCJSWG \cdot (T_{\text{emp}} - TNOM)] \quad (9.49)$$

which will be set to zero if it is negative, and

$$PBSWGS(T) = PBSWGS - TPBSWG \cdot (T_{\text{emp}} - TNOM) \quad (9.50)$$

which will be set to 0.01 if it is less than this value.

Note that the junction diode IV and CV temperature-dependence models are independent of the setting of the temperature-dependence model selector **TEMPMOD**, which provides optional temperature-dependence models for such parameters as the carrier mobility.

## 9.7 Chapter Summary

This chapter presented and discussed the theory and BSIM4 models of modern MOSFET source and drain diode IV and CV. Leakage current mechanisms include the Shockley-Read-Hall generation and recombination, band-to-band tunneling and trap-assisted tunneling and they all need to be modeled. Junction breakdown is modeled for the avalanche mechanism. The derivations of the BSIM4 diode IV and CV models and their temperature dependencies were described in details.

## 9.8 Parameter Table

Name (type)	Description and default	Can be binned?	Note
DIOMOD (Global; integer)	Model selector for the source and drain junction diode IV model.  Default = 1; dimensionless. Other optional values are 0 and 2.	No	-

NF (Local; integer)	The number of fingers of a multi-finger device.  Default = 1; dimensionless.	No	Reset to 1 if smaller than 1 with a warning to be issued.
NJS (Global; double)	The non-ideality factor for the source-bulk junction diode IV model.  Default = 1.0; dimensionless.	No	A warning to be issued if it is negative.
NJD (Global; double)	The non-ideality factor for the drain-bulk junction diode IV model.  Default = NJS; dimensionless.	No	A warning to be issued if it is negative.
XJBVS (Global; double)	The source-bulk junction diode breakdown coefficient. In the case of DIOMOD = 0, XJBVS will be reset to 1 if it is negative. XJBVS = 0 means no breakdown will be modeled. In the case of DIOMOD = 2, XJBVS will be reset to 1 if it is not positive.  Default = 1.0; dimensionless.	No	-
XJBVD (Global; double)	The drain-bulk junction diode breakdown coefficient. In the case of DIOMOD = 0, XJBVD will be reset to 1 if it is negative. XJBVD = 0 means no breakdown will be modeled. In the case of DIOMOD = 2, XJBVD will be reset to 1 if it is not positive.  Default = XJBVS; dimensionless.	No	-
BVS (Global; double)	The source-bulk junction diode breakdown voltage. If BVS is not positive, it is reset to 10 V.  Default = 10.0 in [V].	No	-
BVD (Global; double)	The drain-bulk junction diode breakdown voltage. If BVD is not positive, it is reset to 10 V.  Default = BVS in [V].	No	-
IJTHSFWD (Global; double)	User-specified source junction current value at which the source junction current equation is linearized for the forward body bias regime. If IJTHSFWD is not positive, it will be reset to 0.1 A.  Default = 0.1 in [A].	No	-

<b>IJTHDFWD</b> (Global; double)	User-specified drain junction current value at which the drain junction current equation is linearized for the forward body bias regime. If IJTHDFWD is not positive, it will be reset to 0.1 A.  Default = IJTHSFWD in [A].	No	-
<b>IJTHSREV</b> (Global; double)	User-specified source junction current value at which the source junction current equation is linearized for the reverse body bias regime. If IJTHSREV is not positive, it will be reset to 0.1 A.  Default = 0.1 in [A].	No	-
<b>IJTHDREV</b> (Global; double)	User-specified drain junction current value at which the drain junction current equation is linearized for the reverse body bias regime. If IJTHDREV is not positive, it will be reset to 0.1 A.  Default = IJTHSREV in [A].	No	-
<b>VTSS</b> (Global; double)	Bias-dependence parameter for the trap-assisted tunneling current of the source junction bottom.  Default = 10.0 in [V].	No	Fatal error to be issued if it is negative.
<b>VTSD</b> (Global; double)	Bias-dependence parameter for the trap-assisted tunneling current of the drain junction bottom.  Default = VTSS in [V].	No	Fatal error to be issued if it is negative.
<b>VTSSWS</b> (Global; double)	Bias-dependence parameter for the trap-assisted tunneling current of the source (STI) sidewall junction.  Default = 10.0 in [V].	No	Fatal error to be issued if it is negative.
<b>VTSSWD</b> (Global; double)	Bias-dependence parameter for the trap-assisted tunneling current of the drain (STI) sidewall junction.  Default = VTSSWS in [V].	No	Fatal error to be issued if it is negative.
<b>VTSSWGS</b> (Global; double)	Bias-dependence parameter for the trap-assisted tunneling current of the gate edge of the source junction.  Default = 10.0 in [V].	No	Fatal error to be issued if it is negative.

336 BSIM4 AND MOSFET MODELING FOR IC SIMULATION  
By Weidong Liu and Chenming Hu

VTSSWGD (Global; double)	Bias-dependence parameter for the trap-assisted tunneling current of the gate edge of the drain junction.  Default = VTSSWGS in [V].	No	Fatal error to be issued if it is negative.
JSS (Global; double)	Saturation current density of the source junction bottom at TNOM.  Default = 1.0e-4 in [A/m <sup>2</sup> ].	No	-
JSD (Global; double)	Saturation current density of the drain junction bottom at TNOM.  Default = JSS in [A/m <sup>2</sup> ].	No	-
JSWS (Global; double)	Saturation current density of the source (STI) sidewall junction at TNOM.  Default = 0.0 in [A/m].	No	-
JSWD (Global; double)	Saturation current density of the drain (STI) sidewall junction at TNOM.  Default = JSWS in [A/m].	No	-
JSWGS (Global; double)	Saturation current density of the source gate-edge junction at TNOM.  Default = 0.0 in [A/m].	No	-
JSWGD (Global; double)	Saturation current density of the drain gate-edge junction at TNOM.  Default = JSWGS in [A/m].	No	-
XTIS (Global; double)	Exponent of the temperature dependence of the source-bulk junction saturation current density.  Default = 3.0; dimensionless.	No	-
XTID (Global; double)	Exponent of the temperature dependence of the drain-bulk junction saturation current density.  Default = XTIS; dimensionless.	No	-

MJS (Global; double)	Source-bulk bottom junction grading coefficient for depletion capacitance model.  Default = 0.5; dimensionless.	No	A warning to be issued if it is greater than or equal to 0.99; reset to 0.99 if it is greater than 0.99 for numerical stability.
MJD (Global; double)	Drain-bulk bottom junction grading coefficient for depletion capacitance model.  Default = MJS; dimensionless.	No	A warning to be issued if it is greater than or equal to 0.99; reset to 0.99 if it is greater than 0.99 for numerical stability.
MJSWS (Global; double)	Source-bulk STI sidewall junction grading coefficient for depletion capacitance model.  Default = 0.33; dimensionless.	No	A warning to be issued if it is greater than or equal to 0.99; reset to 0.99 if it is greater than 0.99 for numerical stability.
MJSWD (Global; double)	Drain-bulk STI sidewall junction grading coefficient for depletion capacitance model.  Default = MJSWS; dimensionless.	No	A warning to be issued if it is greater than or equal to 0.99; reset to 0.99 if it is greater than 0.99 for numerical stability.

338 BSIM4 AND MOSFET MODELING FOR IC SIMULATION  
By Weidong Liu and Chenming Hu

MJSWGS (Global; double)	Source-bulk gate-edge junction grading coefficient for depletion capacitance model.  Default = MJSWS; dimensionless.	No	A warning to be issued if it is greater than or equal to 0.99; reset to 0.99 if it is greater than 0.99 for numerical stability.
MJSWGD (Global; double)	Drain-bulk gate-edge junction grading coefficient for depletion capacitance model.  Default = MJSWGS; dimensionless.	No	A warning to be issued if it is greater than or equal to 0.99; reset to 0.99 if it is greater than 0.99 for numerical stability.
JTSS (Global; double)	Trap-assisted tunneling saturation current density of the source-bulk junction bottom at TNOM.  Default = 0.0 in [A/m <sup>2</sup> ].	No	-
XTSS (Global; double)	Exponent of the temperature dependence of the trap-assisted tunneling saturation current density of the source-bulk junction bottom.  Default = 0.02; dimensionless.	No	-
JTSD (Global; double)	Trap-assisted tunneling saturation current density of the drain-bulk junction bottom at TNOM.  Default = JTSS in [A/m <sup>2</sup> ].	No	-
XTSD (Global; double)	Exponent of the temperature dependence of the trap-assisted tunneling saturation current density of the drain-bulk junction bottom.  Default = XTSS; dimensionless.	No	-
JTSSWS (Global; double)	Trap-assisted tunneling saturation current density of the source-bulk STI sidewall junction at TNOM.  Default = 0.0 in [A/m].	No	-

<b>XTSSWS</b> (Global; double)	Exponent of the temperature dependence of the trap-assisted tunneling saturation current density of the source-bulk STI sidewall junction.  Default = 0.02; dimensionless.	No	-
<b>JTSSWD</b> (Global; double)	Trap-assisted tunneling saturation current density of the drain-bulk STI sidewall junction at TNOM.  Default = JTSSWS in [A/m].	No	-
<b>XTSSWD</b> (Global; double)	Exponent of the temperature dependence of the trap-assisted tunneling saturation current density of the drain-bulk STI sidewall junction.  Default = XTSSWS; dimensionless.	No	-
<b>JTSSWGS</b> (Global; double)	Trap-assisted tunneling saturation current density of the source-bulk gate-edge junction at TNOM.  Default = 0.0 in [A/m].	No	-
<b>XTSSWGS</b> (Global; double)	Exponent of the temperature dependence of the trap-assisted tunneling saturation current density of the source-bulk gate-edge junction.  Default = 0.02; dimensionless.	No	-
<b>JTSSWGD</b> (Global; double)	Trap-assisted tunneling saturation current density of the drain-bulk gate-edge junction at TNOM.  Default = JTSSWGS in [A/m].	No	-
<b>XTSSWGD</b> (Global; double)	Exponent of the temperature dependence of the trap-assisted tunneling saturation current density of the drain-bulk gate-edge junction.  Default = XTSSWGS; dimensionless.	No	-
<b>NJTS</b> (Global; double)	The trap-assisted tunneling current non-ideality factor for the source-bulk and drain-bulk bottom junctions.  Default = 20.0; dimensionless.	No	A warning to be issued if it makes the temperature-dependent $NJS_{TAT}(T)$ negative.

NJTSSW (Global; double)	The trap-assisted tunneling current non-ideality factor for the source-bulk and drain-bulk STI sidewall junctions.  Default = 20.0; dimensionless.	No	A warning to be issued if it makes the temperature-dependent $NJSSW_{TAT}(T)$ negative.
NJTSSWG (Global; double)	The trap-assisted tunneling current non-ideality factor for the source-bulk and drain-bulk gate-edge junctions.  Default = 20.0; dimensionless.	No	A warning to be issued if it makes the temperature-dependent $NJSSWG_{TAT}(T)$ negative.
TNJTS (Global; double)	Temperature-dependence coefficient of NJTS.  Default = 0.0; dimensionless.	No	A warning to be issued if it makes the temperature-dependent $NJS_{TAT}(T)$ negative.
TNJTSSW (Global; double)	Temperature-dependence coefficient of NJTSSW.  Default = 0.0; dimensionless.	No	A warning to be issued if it makes the temperature-dependent $NJSSW_{TAT}(T)$ negative.
TNJTSSWG (Global; double)	Temperature-dependence coefficient of NJTSSWG.  Default = 0.0; dimensionless.	No	A warning to be issued if it makes the temperature-dependent $NJSSWG_{TAT}(T)$ negative.
CJS (Global; double)	Zero-bias unit-area depletion capacitance of the source-bulk bottom junction at TNOM.  Default = $5.0 \times 10^{-4}$ in [Farad/m <sup>2</sup> ].	No	-
CJD (Global; double)	Zero-bias unit-area depletion capacitance of the drain-bulk bottom junction at TNOM.  Default = CJS in [Farad/m <sup>2</sup> ].	No	-

CJSWS (Global; double)	Zero-bias unit-length depletion capacitance of the source-bulk STI sidewall junction at TNOM.  Default = $5.0 \times 10^{-10}$ in [Farad/m].	No	-
CJSWD (Global; double)	Zero-bias unit-length depletion capacitance of the drain-bulk STI sidewall junction at TNOM.  Default = CJSWS in [Farad/m].	No	-
CJSWGS (Global; double)	Zero-bias unit-length depletion capacitance of the source-bulk gate-edge junction at TNOM.  Default = CJSWS in [Farad/m].	No	-
CJSWGD (Global; double)	Zero-bias unit-length depletion capacitance of the drain-bulk gate-edge junction at TNOM.  Default = CJSWGS in [Farad/m].	No	-
PBS (Global; double)	Built-in potential of the source-bulk bottom junction at TNOM.  Default = 1.0 in [V].	No	$PBS(T)$ reset to 0.01 if PBS makes it less than 0.01 at a given $T_{emp}$ .
PBD (Global; double)	Built-in potential of the drain-bulk bottom junction at TNOM.  Default = PBS in [V].	No	$PBD(T)$ reset to 0.01 if PBD makes it less than 0.01 at a given $T_{emp}$ .
PBSWS (Global; double)	Built-in potential of the source-bulk STI sidewall junction at TNOM.  Default = 1.0 in [V].	No	$PBSWS(T)$ reset to 0.01 if PBSWS makes it less than 0.01 at a given $T_{emp}$ .
PBSWD (Global; double)	Built-in potential of the drain-bulk STI sidewall junction at TNOM.  Default = PBSWS in [V].	No	$PBSWD(T)$ reset to 0.01 if PBSWD makes it less than 0.01 at a given $T_{emp}$ .

PBSWGS (Global; double)	Built-in potential of the source-bulk gate-edge junction at TNOM.  Default = PBSWS in [V].	No	$PBSWGS(T)$ reset to 0.01 if PBSWGS makes it less than 0.01 at a given $T_{emp}$ .
PBSWGD (Global; double)	Built-in potential of the drain-bulk gate-edge junction at TNOM.  Default = PBSWGS in [V].	No	$PBSWGD(T)$ reset to 0.01 if PBSWGD makes it less than 0.01 at a given $T_{emp}$ .
TPB (Global; double)	Temperature-dependence coefficient for PBS and PBD.  Default = 0.0 in [V/K].	No	$PBS(T)$ reset to 0.01 if TPB makes it less than 0.01 at a given $T_{emp}$ .
TPBSW (Global; double)	Temperature-dependence coefficient for PBSWS and PBSWD.  Default = 0.0 in [V/K].	No	$PBSWS(T)$ reset to 0.01 if TPBSW makes it less than 0.01 at a given $T_{emp}$ .
TPBSWG (Global; double)	Temperature-dependence coefficient for PBSWGS and PBSWGD.  Default = 0.0 in [V/K].	No	$PBSWGS(T)$ reset to 0.01 if TPBSWG makes it less than 0.01 at a given $T_{emp}$ .
TCJ (Global; double)	Temperature-dependence coefficient for CJS and CJD.  Default = 0.0 in [V/K].	No	-
TCJSW (Global; double)	Temperature-dependence coefficient for CJSWS and CJSWD.  Default = 0.0 in [V/K].	No	-
TCJSWG (Global; double)	Temperature-dependence coefficient for CJSWGS and CJSWGD.  Default = 0.0 in [V/K].	No	-

## References

- [1] Chenming Calvin Hu, “Modern Semiconductor Devices for Integrated Circuits,” Chapter 4, pp. 89-146, Pearson Prentice Hall, 2010.
- [2] Chih-Tang Sah, “Fundamentals of Solid-State Electronics,” Section 536, pp. 441-450, World Scientific Publishing Co., Singapore, 1991.
- [3] S. M. Sze, “Physics of Semiconductor Devices,” John Wiley, New York, 1981.
- [4] Weidong Liu, Xiaodong Jin, Kanyu M. Cao, and Chenming Hu, “BSIM4.0.0 MOSFET model – User’s manual,” Memorandum No. UCB/ERL M00/38, Electronics Research Laboratory, College of Engineering, University of California, Berkeley. August 3, 2000.

**This page intentionally left blank**

## **Chapter 10**

# **SPICE Implementation Example: The Methodology with BSIM4 Transient NQS**

### **10.1 Introduction and Chapter Objectives**

SPICE modeling has employed increasingly more sophisticated mathematical formulations and implementation strategies to model complex physical effects of advanced CMOS transistors. In the previous chapters, many examples of such cases have been presented.

A robust and efficient SPICE implementation is critical to the success of a sophisticated, complex semiconductor device model for circuit simulation. Making such an implementation requires good understanding of device physics, operation and modeling as well as the framework and numerical algorithms of circuit simulation. This chapter details the implementation methodologies and techniques that are useful to compact modeling in general.

The authors will discuss SPICE model implementation through the example of the BSIM4 charge-deficit transient non-quasi-static (NQS) model in SPICE3. The physics and formulation of the model was presented in Chapter 6. This NQS model is a good example to use as it involves rather complex a model circuit topology with an auxiliary non-linear capacitive and resistive sub-circuit that aids the solution of a MOS device internal node in the context of SPICE. This example also illustrates how such an RC sub-circuit model interacts with the core of a MOSFET model in circuit equation solution with the numerical iterative methods, such as the Newton-Raphson algorithm.

This chapter also answers this question: How does the distributed gate capacitance and channel resistance affect the transient property of a MOS transistor. The BSIM4 charge-deficit NQS model topology will be briefly reviewed first. This will then be followed by time point discretization, linearization of model equations, the loading of a device into a SPICE circuit matrix (*stamping* is the jargon). The forward and reverse bias scenarios and the source and drain swapping and stamping will be subsequently presented in detail. This chapter ends with discussions of the convergence check and simulation bypass algorithms for improving SPICE convergence and runtime. Throughout this chapter, the authors pay particular attention to the procedures and mathematical derivations required by the Newton-Raphson method. To allow for quick cross checking against the BSIM4 code implementation, the authors use the parameter and variable names of the BSIM4 implementation as much as possible.

## 10.2 Review of the Charge-Deficit Transient NQS Model

BSIM4 utilizes an auxiliary sub-circuit for NQS as shown in Fig. 10.1. (A similar approach has been employed in the modeling of the self-heating effects for SOI and high-voltage MOSFET and BJT SPICE models.) In this figure, the sub-circuit is composed of a non-linear resistor and a parallel capacitor to account for the NQS time constant. This time constant captures the first pole of MOSFET high-frequency operation. The quasi-static channel charge current is taken as the excitation stimulus that drives the resistor and capacitor. To solve for the value of  $v_{nqs}$ , which relates to the charge deficit  $Q_{def}$ , an internal NQS node is required for the Newton-Raphson iteration of the BSIM4 equations. In this case, the unknowns are  $v_{nqs}$  and the device terminal voltages  $V_d$ ,  $V_g$ ,  $V_s$ , and  $V_b$ .

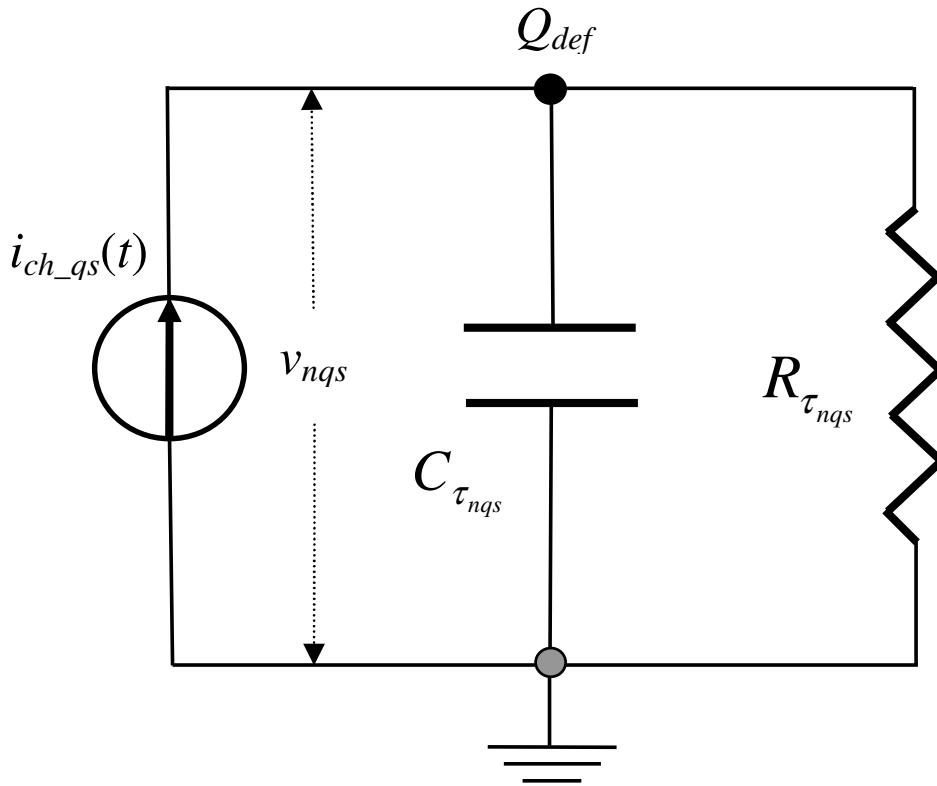


Fig. 10.1 The topology of the BSIM4 charge-deficit NQS sub-circuit model. BSIM4 chooses  $C_{\tau_{nqs}} = 1 \times 10^{-9}$  Farad to achieve fast simulation and good accuracy.

### 10.3 Time Discretization, Equation Linearization and Matrix Stamping

The BSIM4 charge-deficit NQS model presented above is developed for transient analysis. In the following, the numerical transient analysis techniques of SPICE will be reviewed [1 – 6]. This aims to provide the readers with necessary technical background to understand the BSIM4 SPICE implementation strategy and methodology that will be presented shortly.

Toward that end, the problem of this subject is defined as follows. In the time-domain transient analysis, the response of an electronic circuit over a specified time interval  $[0, T]$  is represented by a system of differential and algebraic equations (DAE) of the form

$$F(\mathbf{x}, \dot{\mathbf{x}}, t) = 0 \quad (10.1)$$

How do we find the solution  $\mathbf{x}$ ? Here the vector  $\mathbf{x}$  contains nodal voltages and branch currents,  $t$  is time,  $\dot{\mathbf{x}}$  is the time derivative of  $\mathbf{x}$ , and  $F$  is, in general, a nonlinear (differential) operator. It is nonlinear because for a CMOS circuit, the branch relations of a MOS device and other similar energy-storage devices have nonlinear DC transport currents and nonlinear capacitive charging displacement currents as well (refer to Chapters 3, 4, 5 and 6). For example, by applying Eq. (10.1) to the MOSFET drain node, the DAE equation is given by

$$I_{DC} + \frac{dQ_d(t)}{dt} = i_{external} \quad (10.2)$$

In this equation, the MOSFET branch relations  $I_{DC}$  and  $dQ_d(t)/dt$  are the DC current such as the channel current and the capacitive charging current, respectively, and  $i_{external}$  represents the current of any independent sources and/or the currents provided by the neighboring circuit elements that are connected to that drain node. [Note that in the time domain, the word DC of  $I_{DC}$  is not that exact as it really should be *quasi-DC* instead because these DC currents do vary with time.] In deriving Eq. (10.2) for the drain node, Kirchhoff's current law (KCL), which requires that the sum of all the inflow and outflow currents at any circuit node and time point must equal zero, was used together with the MOSFET branch relations.

The solution of Eqs. (10.1) or (10.2) is obtained by segmenting the time interval  $[0, T]$  into many discrete time points or smaller time intervals. At each point, numerical integrations are performed to replace these DAE equations with algebraic equations. For active, energy-storage devices such as MOS transistors, the resultant algebraic equations can be (highly) nonlinear. Again, take as an example the drain terminal charging current  $dQ_d(t)/dt$ , the second term on the left-hand side of Eq. (10.2). At the time point  $t_{n+1}$ ,  $Q_d$  can be approximated by

$$Q_{d,t_{n+1}} \approx Q_{d,t_n} + h \cdot \left. \frac{dQ_d(t)}{dt} \right|_{t_{n+1}} = Q_{d,t_n} + h \cdot i_{d,t_{n+1}} \quad (10.3)$$

In this example, the implicit Backward Euler numerical integration method is used in the time point discretization with the time step  $h = t_{n+1} - t_n$ . Here, the integration itself is to find the solution  $i_{d,t_{n+1}}$  at the time point  $t_{n+1}$  with the solutions of the previous time points obtained. The discretization reduces the solution of a DAE equation in the time domain to an easier problem of *quasi-DC* analyses.

With other integration algorithms, the discretization scheme changes. For Trapezoidal, Eq. (10.3) becomes

$$Q_{d,t_{n+1}} \approx Q_{d,t_n} + \frac{h}{2} \cdot \left( \left. \frac{dQ_d(t)}{dt} \right|_{t_{n+1}} + \left. \frac{dQ_d(t)}{dt} \right|_{t_n} \right) = Q_{d,t_n} + \frac{h}{2} \cdot (i_{d,t_{n+1}} + i_{d,t_n}) \quad (10.4)$$

Note here the effective time step is now half  $h$ .

Note also that Eqs. (10.3) and (10.4) are approximate because a constant time derivative of  $Q_d(t)$  was assumed over  $[t_n, t_{n+1}]$ :

$$Q_{d,t_{n+1}} = \int_0^{t_{n+1}} \frac{dQ_d(t)}{dt} \cdot dt = Q_{d,t_n} + \int_{t_n}^{t_{n+1}} \frac{dQ_d(t)}{dt} \cdot dt \approx Q_{d,t_n} + h \cdot i_{d,t_{n+1}} \quad (10.5)$$

According to the capacitance modeling theory given in Chapter 5, Eq. (10.5) provides good accuracy only when  $h$  is sufficiently small, or the time derivative of charges is really constant or both. Poor accuracy results when, for instance, the time step  $h$  becomes too large. This is typical of any numerical integration algorithms.

In the remainder of this chapter, only the Backward Euler method will be considered for the sake of ease of analysis. For other methods, similar analysis process applies equally.

In Eqs. (10.3) and (10.4),  $Q_{d,t_{n+1}}$  is a nonlinear function of MOSFET terminal voltages. These voltages must be first solved for in order to get  $i_{d,t_{n+1}}$ . In the context of a nonlinear solution algorithm such as the Newton-Raphson iteration method,  $Q_{d,t_{n+1}}$  can be linearized for the  $(k + 1)^{\text{th}}$  iteration by using the Taylor series expansion around the  $k^{\text{th}}$  iteration result  $Q_{d,t_n}$  and by keeping up to the first-order term. The expansion is performed as

$$Q_{d,t_{n+1}}^{k+1} = Q_{d,t_{n+1}}^k + \sum_j \left. \frac{\partial Q_d}{\partial V_j} \right|_{t_{n+1}}^k \cdot (V_{j,t_{n+1}}^{k+1} - V_{j,t_{n+1}}^k) \quad (10.6)$$

In this expression,  $V_j$  denotes any of the four MOSFET terminal voltages,  $V_d$ ,  $V_g$ ,  $V_s$  and  $V_b$  ( $j = d, g, s$ , and  $b$ ). The partial derivatives of  $Q_d$  with respect to  $V_j$  designate the trans-nodal capacitances of the transistor: For example, when  $j = g$ , that derivative gives from the BSIM4 model the drain-to-gate capacitance  $C_{dg}$ , which is known from the  $k^{\text{th}}$  iteration at  $t_{n+1}$ . Refer to Fig. 10.2 and Chapter 5 for details.

The above linearization process is repeated for every iteration and time point. Inserting Eq. (10.6) into Eq. (10.5) and then substituting  $i_{d,t_{n+1}}$  into the KCL equation, Eq. (10.2), for  $dQ_d(t)/dt$ , yield a linear algebraic equation to solve for such unknowns as  $V_{j,t_{n+1}}^{k+1}$ .

The non-linear time-domain DAE equation system Eq. (10.1) is now replaced by the system of *quasi-DC* linear algebraic equations of the general form  $\mathbf{J} \cdot \mathbf{x} = \mathbf{b}$ . Here  $\mathbf{J}$  is the coefficient matrix (interchangeably denoted by  $\mathbf{A}$  in publication). It is also known as the nodal admittance matrix, or simply the circuit Jacobian. Each entry of this matrix is an effective conductance in the form of  $(G_{ij} + C_{ij} / h)$ .  $G_{ij}$  and  $C_{ij}$  are device trans-nodal conductance and capacitance stamps, where  $i$  and  $j$  denote the nodal indices such as  $d, g, s$ , and  $b$ .  $\mathbf{x}$  is the vector of the unknowns of the circuit such as nodal voltages and/or branch currents.  $\mathbf{b}$  is the excitation vector or the equivalent current vector. It is often called RHS (right-hand side). The system of the linear algebraic equations is solved with such numerical techniques as Gaussian elimination or LU factorization.

It is worthwhile to point out that the linearization process introduces local truncation errors (LTE). LTE is the error in the solution of  $t_{n+1}$  assuming that the previous solutions are exact. It results from dropping the second-order term in the Taylor series expansion. For this reason, LTE is proportional to the size of time steps and can be accumulated as solution proceeds. In SPICE simulation, LTE tolerances are applied to control time steps to fulfill simulation accuracy requirement.

In the following sections, the discretization, linearization and stamping of the BSIM4 NQS model will be presented.

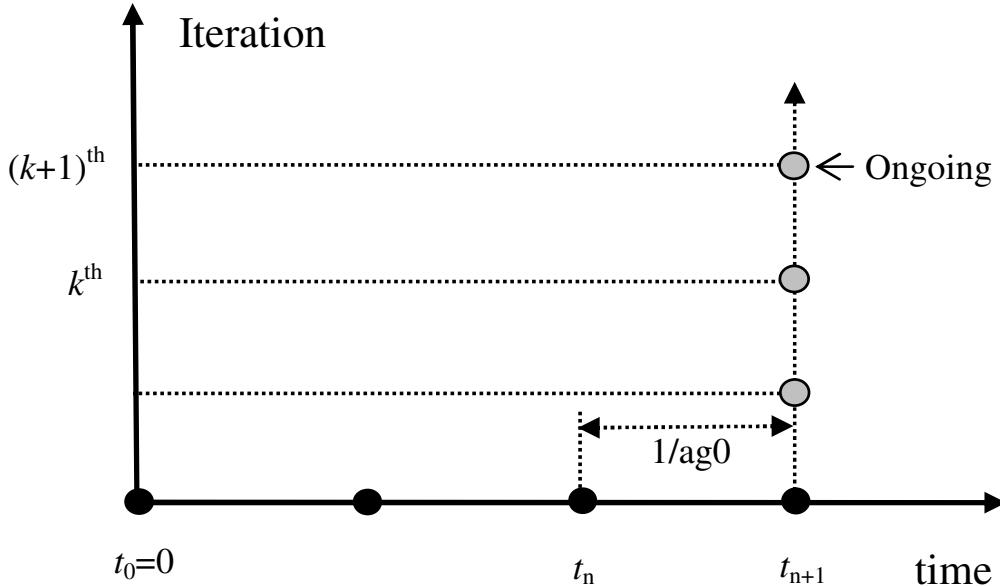


Fig. 10.2 Schematic illustration of numerical iterations versus time point discretization in time-domain simulation.

### 10.3.1 Discretization and Linearization of $i_{ch\_qs}(t)$

The charging current, namely the derivative of the device terminal charge with respect to time, can be approximated with a finite difference of this charge between two consecutive time points divided by the time interval between them (the time points), a treatment with Backward Euler. Take the channel charging current (i.e.,  $i_{ch\_qs}(t)$  of Fig. 10.1) as an example.

$$i_{ch\_qs}(t) = \frac{dQ_{ch\_qs}}{dt} \approx ag0 \cdot (Q_{ch\_qs,t_{n+1}}^{k+1} - Q_{ch\_qs,t_n}) \quad (10.7)$$

where  $ag0$ , in the SPICE3 terminology, is the inverse of the time step  $h$  for Backward Euler and  $2/h$  for Trapezoidal. Again,  $k$  and  $(k+1)$  denote the  $k^{\text{th}}$  and  $(k+1)^{\text{th}}$  iterations at  $t_{n+1}$ , with the  $(k+1)^{\text{th}}$  iteration being the ongoing iteration that SPICE uses to solve the circuit matrix for nodal voltages such as  $V_d^{k+1}$ ,  $V_g^{k+1}$ ,  $V_s^{k+1}$ ,  $V_b^{k+1}$ , and  $v_{nqs}^{k+1}$ . This is schematically illustrated in Fig. 10.2.

$Q_{ch\_qs,t_n}$  of Eq. (10.7) is already known from the solution at the past time point  $t_n$ . Thus, the major effort is to linearize  $Q_{ch\_qs,t_{n+1}}^{k+1}$  by performing a Taylor series expansion around  $Q_{ch\_qs,t_{n+1}}^k$  which is also known from the prior  $k^{\text{th}}$  iteration at  $t_{n+1}$ . Neglecting the high-order terms of the Taylor expansion leads to

$$i_{ch\_qs}(t) \approx ag0 \cdot \left\{ \begin{array}{l} Q_{ch\_qs,t_{n+1}}^k - Q_{ch\_qs,t_n} + \frac{\partial Q_{ch\_qs}}{\partial v_d} \Big|_{t_{n+1}}^k \cdot (v_{d,t_{n+1}}^{k+1} - v_{d,t_{n+1}}^k) \\ + \frac{\partial Q_{ch\_qs}}{\partial v_g} \Big|_{t_{n+1}}^k \cdot (v_{g,t_{n+1}}^{k+1} - v_{g,t_{n+1}}^k) \\ + \frac{\partial Q_{ch\_qs}}{\partial v_s} \Big|_{t_{n+1}}^k \cdot (v_{s,t_{n+1}}^{k+1} - v_{s,t_{n+1}}^k) \\ + \frac{\partial Q_{ch\_qs}}{\partial v_b} \Big|_{t_{n+1}}^k \cdot (v_{b,t_{n+1}}^{k+1} - v_{b,t_{n+1}}^k) \end{array} \right\} \quad (10.8)$$

Rearranging Eq. (10.8) in the SPICE3 BSIM4 coding terminology yields

$$i_{ch\_qs}(t) \approx ag0 \cdot \left\{ \begin{array}{l} Q_{ch\_qs,t_{n+1}}^k - Q_{ch\_qs,t_n} \\ - \left( \frac{\partial Q_{ch\_qs}}{\partial v_g} \Big|_{t_{n+1}}^k \cdot v_{gb,t_{n+1}}^k - \frac{\partial Q_{ch\_qs}}{\partial v_d} \Big|_{t_{n+1}}^k \cdot v_{bd,t_{n+1}}^k \right. \right. \\ \left. \left. - \frac{\partial Q_{ch\_qs}}{\partial v_s} \Big|_{t_{n+1}}^k \cdot v_{bs,t_{n+1}}^k \right. \right. \\ \left. + \frac{\partial Q_{ch\_qs}}{\partial v_d} \Big|_{t_{n+1}}^k \cdot v_{d,t_{n+1}}^{k+1} \right. \\ \left. + \frac{\partial Q_{ch\_qs}}{\partial v_g} \Big|_{t_{n+1}}^k \cdot v_{g,t_{n+1}}^{k+1} + \frac{\partial Q_{ch\_qs}}{\partial v_s} \Big|_{t_{n+1}}^k \cdot v_{s,t_{n+1}}^{k+1} \right. \\ \left. + \frac{\partial Q_{ch\_qs}}{\partial v_b} \Big|_{t_{n+1}}^k \cdot v_{b,t_{n+1}}^{k+1} \right) \end{array} \right\} \quad (10.9)$$

where the derivatives of  $Q_{ch\_qs}$  are given with respect to the physical (or actual) terminal voltages  $v_d$ ,  $v_g$ ,  $v_s$ , and  $v_b$  (all given in the lower case), such that Eq. (10.9) is valid for both the forward- and reverse-mode operations. These charge derivatives are related to the derivatives with respect to the electrical terminal voltages  $V_d$ ,  $V_g$ ,  $V_s$ , and  $V_b$  (in which a capital letter  $V$  is used to distinguish from the physical terminal voltages and the forward-bias scenario is always assumed) as

$$\left\{ \begin{array}{l} \frac{\partial Q_{ch\_qs}}{\partial v_d} \Big|_{t_{n+1}}^k = cqdb \equiv \frac{\partial Q_{ch\_qs}}{\partial V_d} \Big|_{t_{n+1}}^k \\ \frac{\partial Q_{ch\_qs}}{\partial v_g} \Big|_{t_{n+1}}^k = cqgb \equiv \frac{\partial Q_{ch\_qs}}{\partial V_g} \Big|_{t_{n+1}}^k \\ \frac{\partial Q_{ch\_qs}}{\partial v_s} \Big|_{t_{n+1}}^k = cqsb \equiv \frac{\partial Q_{ch\_qs}}{\partial V_s} \Big|_{t_{n+1}}^k \\ \frac{\partial Q_{ch\_qs}}{\partial v_b} \Big|_{t_{n+1}}^k = cqbb \equiv \frac{\partial Q_{ch\_qs}}{\partial V_b} \Big|_{t_{n+1}}^k \end{array} \right. \quad (10.9a)$$

for the forward mode ( $v_{ds} = V_{ds} \geq 0$ ); and for the reverse mode ( $v_{ds} = -V_{ds} < 0$ ),

$$\left\{ \begin{array}{l} \frac{\partial Q_{ch\_qs}}{\partial v_d} \Big|_{t_{n+1}}^k = cqsb \equiv \frac{\partial Q_{ch\_qs}}{\partial V_s} \Big|_{t_{n+1}}^k \\ \frac{\partial Q_{ch\_qs}}{\partial v_g} \Big|_{t_{n+1}}^k = cqgb \equiv \frac{\partial Q_{ch\_qs}}{\partial V_g} \Big|_{t_{n+1}}^k \\ \frac{\partial Q_{ch\_qs}}{\partial v_s} \Big|_{t_{n+1}}^k = cqdb \equiv \frac{\partial Q_{ch\_qs}}{\partial V_d} \Big|_{t_{n+1}}^k \\ \frac{\partial Q_{ch\_qs}}{\partial v_b} \Big|_{t_{n+1}}^k = cqbb \equiv \frac{\partial Q_{ch\_qs}}{\partial V_b} \Big|_{t_{n+1}}^k \end{array} \right. \quad (10.9b)$$

The only difference between Eq. (10.9a) and Eq. (10.9b) is that a source and drain swapping is applied as  $V_d$  and  $V_s$  are exchanged in deriving the capacitances  $cqsb$  and  $cqdb$ . Note that the capacitances  $cqdb$ ,  $cqgb$ ,  $cqsb$ , and  $cqbb$  are obtained from the BSIM4 charge and capacitance model, either  $CAPMOD = 0, 1$ , or  $2$ . They are expressed as functions of the electrical terminal voltages  $V_d$ ,  $V_g$ ,  $V_s$ , and  $V_b$ .

Note also that for conciseness and coding efficiency, the  $k^{\text{th}}$  individual nodal voltages have been collapsed into the voltages across the terminals in Eq. (10.9) by following the capacitance theory that the sum of all the capacitances in a row or column of a MOSFET capacitance matrix is zero; this is equivalent to stating that  $cqbb = -(cqdb + cqgb + cqsb)$ .

BSIM4 SPICE implementation derives a more compact form of Eq. (10.9)

$$\begin{aligned} i_{ch_qs}(t) \approx & i_{ch_qs\_RHS} + gcqdb \cdot v_{d,t_{n+1}}^{k+1} + gcqgb \cdot v_{g,t_{n+1}}^{k+1} \\ & + gcqs \cdot v_{s,t_{n+1}}^{k+1} + gcqbb \cdot v_{b,t_{n+1}}^{k+1} \end{aligned} \quad (10.10)$$

The first term on the right-hand side is the equivalent charging current:

$$\begin{aligned} i_{ch_qs\_RHS} = & ag0 \cdot (Q_{ch_qs,t_{n+1}}^k - Q_{ch_qs,t_n}) \\ & - (gcqgb \cdot v_{gb,t_{n+1}}^k - gcqdb \cdot v_{bd,t_{n+1}}^k - gcqs \cdot v_{bs,t_{n+1}}^k) \end{aligned} \quad (10.4a)$$

The equivalent conductances that will be stamped into the  $\mathbf{J}$  matrix are

$$\left\{ \begin{array}{l} gcqdb = cqdb \cdot ag0 \\ gcqgb = cqgb \cdot ag0 \\ gcqs = cqs \cdot ag0 \\ gcqbb = cqbb \cdot ag0 \end{array} \right. \quad (10.10b)$$

for the forward bias regime and

$$\left\{ \begin{array}{l} gcqdb = cqs \cdot ag0 \\ gcqgb = cqgb \cdot ag0 \\ gcqs = cqdb \cdot ag0 \\ gcqbb = cqbb \cdot ag0 \end{array} \right. \quad (10.10c)$$

for the reverse bias condition.

### 10.3.2 Stamping of $i_{ch\_qs}(t)$

The KCL law requires that the sum of all the current components that are associated with the NQS internal charge node in Fig. 10.1 be zero. This is to say

$$i_{ch\_qs}(t) - i_{C_{\tau_{nqs}}}(t) - i_{R_{\tau_{nqs}}}(t) \equiv 0 \quad (10.10d)$$

in which the minus sign is applied for the outflow currents,  $i_{C_{\tau_{nqs}}}(t)$  and  $i_{R_{\tau_{nqs}}}(t)$ . KCL together with device branch relations is the starting point for constructing a circuit element stamp and its equivalent circuit model for constructing the circuit matrix system. The element can be a simple bias-independent resistor or capacitor or a highly non-linear active transistor such as a BJT or a MOSFET. Applying this to the BSIM4 transient NQS model, one obtains the conductance and equivalent current stamps for  $i_{ch\_qs}(t)$  upon substituting Eq. (10.10) into Eq. (10.10d):

$$\begin{bmatrix} J & d & g & s & b & v_{nqs} \\ \hline d & 0 & 0 & 0 & 0 & 0 \\ g & 0 & 0 & 0 & 0 & 0 \\ s & 0 & 0 & 0 & 0 & 0 \\ b & 0 & 0 & 0 & 0 & 0 \\ v_{nqs} & -gcqdb & -gcqgb & -gcqsb & -gcqbb & 0 \end{bmatrix} \cdot \begin{bmatrix} x \\ v_{d,t_{n+1}}^{k+1} \\ v_{g,t_{n+1}}^{k+1} \\ v_{s,t_{n+1}}^{k+1} \\ v_{b,t_{n+1}}^{k+1} \\ v_{nqs,t_{n+1}}^{k+1} \end{bmatrix} = \begin{bmatrix} b \\ 0 \\ 0 \\ 0 \\ 0 \\ i_{ch\_qs\_RHS} \end{bmatrix} \quad (10.10e)$$

In this equation, the letters in the first row and column of the  $\mathbf{J}$  matrix represent the BSIM4 MOSFET topological node names. For example,  $g$  denotes the gate node if RGATEMOD = 0 and the internal gate node if RGATEMOD is set to a non-zero value.

### 10.3.3 Linearization of $i_{C_{\tau_{nqs}}}(t)$

$i_{C_{\tau_{nqs}}}(t)$  can be linearized in a similar way:

$$i_{C_{\tau_{nqs}}}(t) = \frac{\partial(v_{nqs} \cdot C_{\tau_{nqs}})}{\partial t} \approx ag0 \cdot C_{\tau_{nqs}} \cdot (v_{nqs,t_{n+1}}^{k+1} - v_{nqs,t_n}) \quad (10.11)$$

This equation can be transformed into Eq. (10.12) to take advantage of the charging current computation function `NlIntegrate()` of SPICE3:

$$i_{C_{\tau_{nqs}}}(t) \approx i_{C_{\tau_{nqs}}-RHS} + g_{C_{\tau_{nqs}}} \cdot v_{nqs,t_{n+1}}^{k+1} \quad (10.12)$$

where

$$i_{C_{\tau_{nqs}}-RHS} = ag0 \cdot [C_{\tau_{nqs}} \cdot (v_{nqs,t_{n+1}}^k - v_{nqs,t_n})] - g_{C_{\tau_{nqs}}} \cdot v_{nqs,t_{n+1}}^k \quad (10.12a)$$

The first term (square bracketed) on the right-hand side is returned from `NlIntegrate()`.

The effective conductance contributed by  $C_{\tau_{nqs}}$  in the time domain is given by

$$g_{C_{\tau_{nqs}}} = ag0 \cdot C_{\tau_{nqs}} \quad (10.12b)$$

### 10.3.4 Stamping of $i_{C_{\tau_{nqs}}}(t)$

Substituting Eq. (10.12) into the KCL equation Eq. (10.10d) for the NQS charge node permits writing the stamps for  $i_{C_{\tau_{nqs}}}(t)$  as

$$\begin{bmatrix} \mathbf{J} & | & \mathbf{d} & \mathbf{g} & \mathbf{s} & \mathbf{b} & v_{nqs} \\ \hline \mathbf{d} & | & 0 & 0 & 0 & 0 & 0 \\ \mathbf{g} & | & 0 & 0 & 0 & 0 & 0 \\ \mathbf{s} & | & 0 & 0 & 0 & 0 & 0 \\ \mathbf{b} & | & 0 & 0 & 0 & 0 & 0 \\ \hline v_{nqs} & | & 0 & 0 & 0 & 0 & g_{C_{\tau_{nqs}}} \end{bmatrix} \cdot \begin{bmatrix} \dots & x \\ v_{d,t_{n+1}}^{k+1} \\ v_{g,t_{n+1}}^{k+1} \\ v_{s,t_{n+1}}^{k+1} \\ v_{b,t_{n+1}}^{k+1} \\ v_{nqs,t_{n+1}}^{k+1} \end{bmatrix} = \begin{bmatrix} \dots & b \\ 0 \\ 0 \\ 0 \\ 0 \\ -i_{C_{\tau_{nqs}}} - RHS \end{bmatrix} \quad (10.12c)$$

### 10.3.5 Linearization of $i_{R_{\tau_{nqs}}}(t)$

The transport current  $i_{R_{\tau_{nqs}}}(t)$  that flows through  $R_{\tau_{nqs}}$  in Fig. 10.1 is equal to the voltage  $v_{nqs}$  multiplied by the inverse of  $R_{\tau_{nqs}}$

$$i_{R_{\tau_{nqs}}} = \frac{C_{\tau_{nqs}} \cdot gcrg}{C_{oxe} W_{effCV} L_{effCV} \cdot NF} \cdot v_{nqs} \quad (10.13)$$

The conductance  $gcrg$  is a non-linear function of the instantaneous voltages of the drain, gate, source and body nodes and it needs to be linearized by Taylor series expansions in a similar way to what was carried out above. This leads to Eq. (10.14):

$$i_{R_{\tau_{nqs}}} = \frac{C_{\tau_{nqs}}}{C_{oxe} W_{effCV} L_{effCV} \cdot NF} \cdot \left\{ \begin{array}{l} \left( gcrg \cdot v_{nqs} \right)_{t_{n+1}}^k \\ + gcrg_{t_{n+1}}^k \cdot \left( v_{nqs,t_{n+1}}^{k+1} - v_{nqs,t_{n+1}}^k \right) \\ \\ + v_{nqs,t_{n+1}}^k \cdot \left[ \begin{array}{l} \frac{\partial gcrg}{\partial v_d} \Big|_{t_{n+1}}^k \cdot \left( v_{d,t_{n+1}}^{k+1} - v_{d,t_{n+1}}^k \right) \\ + \frac{\partial gcrg}{\partial v_g} \Big|_{t_{n+1}}^k \cdot \left( v_{g,t_{n+1}}^{k+1} - v_{g,t_{n+1}}^k \right) \\ + \frac{\partial gcrg}{\partial v_s} \Big|_{t_{n+1}}^k \cdot \left( v_{s,t_{n+1}}^{k+1} - v_{s,t_{n+1}}^k \right) \\ + \frac{\partial gcrg}{\partial v_b} \Big|_{t_{n+1}}^k \cdot \left( v_{b,t_{n+1}}^{k+1} - v_{b,t_{n+1}}^k \right) \end{array} \right] \end{array} \right\} \quad (10.14)$$

Equation (10.14) is self explanatory and valid for both forward and reverse biases. The terms on the right side of Eq. (10.14) are regrouped into a more useful form for circuit matrix stamping

$$\begin{aligned} i_{R_{\tau_{nqs}}} &= i_{R_{\tau_{nqs}}-RHS} + ggtd \cdot v_{d,t_{n+1}}^{k+1} + ggtg \cdot v_{g,t_{n+1}}^{k+1} + ggts \cdot v_{s,t_{n+1}}^{k+1} \\ &+ ggtb \cdot v_{b,t_{n+1}}^{k+1} + gtau \cdot v_{nqs,t_{n+1}}^{k+1} \end{aligned} \quad (10.15)$$

Note that all the derivatives of  $gcrg$  with respect to the MOSFET terminal voltages sum up to be zero; in other words,  $ggtb = -(ggtd + ggtg + ggts)$ . Other symbols employed in Eq. (10.15) are defined by

$$\left\{ \begin{array}{l} i_{R_{\tau_{nqs}}-RHS} = ggt d \cdot v_{bd,t_{n+1}}^k - ggt g \cdot v_{gb,t_{n+1}}^k + ggt s \cdot v_{bs,t_{n+1}}^k \\ ggt d = \frac{C_{\tau_{nqs}} \cdot v_{nqs,t_{n+1}}^k}{C_{oxe} W_{effCV} L_{effCV} \cdot NF} \cdot \frac{\partial gcr g}{\partial V_d} \Big|_{t_{n+1}}^k \\ ggt g = \frac{C_{\tau_{nqs}} \cdot v_{nqs,t_{n+1}}^k}{C_{oxe} W_{effCV} L_{effCV} \cdot NF} \cdot \frac{\partial gcr g}{\partial V_g} \Big|_{t_{n+1}}^k \\ ggt s = \frac{C_{\tau_{nqs}} \cdot v_{nqs,t_{n+1}}^k}{C_{oxe} W_{effCV} L_{effCV} \cdot NF} \cdot \frac{\partial gcr g}{\partial V_s} \Big|_{t_{n+1}}^k \\ ggt b = \frac{C_{\tau_{nqs}} \cdot v_{nqs,t_{n+1}}^k}{C_{oxe} W_{effCV} L_{effCV} \cdot NF} \cdot \frac{\partial gcr g}{\partial V_b} \Big|_{t_{n+1}}^k \\ gtau u = \frac{C_{\tau_{nqs}} \cdot gcr g_{t_{n+1}}^k}{C_{oxe} W_{effCV} L_{effCV} \cdot NF} \end{array} \right. \quad (10.15a)$$

for the forward-mode operation ( $v_{ds} = V_{ds} \geq 0$ ). In the case of a reverse operation ( $v_{ds} = -V_{ds} < 0$ ), the following replacements need to be made into Eq. (10.15a) prior to loading into a circuit matrix:

$$\left\{ \begin{array}{l} ggt d = \frac{C_{\tau_{nqs}} \cdot v_{nqs,t_{n+1}}^k}{C_{oxe} W_{effCV} L_{effCV} \cdot NF} \cdot \frac{\partial gcr g}{\partial V_s} \Big|_{t_{n+1}}^k \\ ggt s = \frac{C_{\tau_{nqs}} \cdot v_{nqs,t_{n+1}}^k}{C_{oxe} W_{effCV} L_{effCV} \cdot NF} \cdot \frac{\partial gcr g}{\partial V_d} \Big|_{t_{n+1}}^k \end{array} \right. \quad (10.15b)$$

### 10.3.6 Stamping of $i_{R_{\tau_{nqs}}}(t)$

Substituting Eq. (10.15) into the KCL equation Eq. (10.10d) for the NQS charge node permits writing the stamps for  $i_{R_{\tau_{nqs}}}(t)$  as

$$\begin{bmatrix}
 J & d & g & s & b & v_{nqs} \\
 \hline
 d & 0 & 0 & 0 & 0 & 0 \\
 g & 0 & 0 & 0 & 0 & 0 \\
 s & 0 & 0 & 0 & 0 & 0 \\
 b & 0 & 0 & 0 & 0 & 0 \\
 v_{nqs} & ggt_d & ggt_g & ggt_s & ggt_b & gtau
 \end{bmatrix} \cdot \begin{bmatrix}
 x \\
 v_{d,t_{n+1}}^{k+1} \\
 v_{g,t_{n+1}}^{k+1} \\
 v_{s,t_{n+1}}^{k+1} \\
 v_{b,t_{n+1}}^{k+1} \\
 v_{nqs,t_{n+1}}^{k+1}
 \end{bmatrix} = \begin{bmatrix}
 b \\
 0 \\
 0 \\
 0 \\
 0 \\
 -i_{R_{\tau_{nqs}}} - RHS
 \end{bmatrix}$$

(10.15c)

### 10.3.7 Linearization of $i_g(t)$

The charging current that flows into the gate node as computed from the transient non-quasi-static model (Refer to Chapter 6 for more details) is repeated here for convenience of reference

$$i_g(t) = G_{xpart} \cdot \frac{v_{nqs} \cdot C_{\tau_{nqs}}}{\tau_{nqs}} = G_{xpart} \cdot \frac{v_{nqs} \cdot C_{\tau_{nqs}}}{C_{oxe} W_{effCV} L_{effCV} \cdot NF} \cdot gcrg$$

(10.16)

where together with  $v_{nqs}$ ,  $gcrg$ , a non-linear conductance as a function of the terminal voltages, is approximated with Taylor expansions

$$i_g(t) \approx \frac{G_{xpart} \cdot C_{\tau_{nqs}}}{C_{oxe} W_{effCV} L_{effCV} \cdot NF} \cdot \left\{ \begin{array}{l} \left( v_{nqs} \cdot gcrg \right)_{t_{n+1}}^k \\ + gcrg_{t_{n+1}}^k \cdot \left( v_{nqs,t_{n+1}}^{k+1} - v_{nqs,t_{n+1}}^k \right) \\ \\ + v_{nqs,t_{n+1}}^k \cdot \left[ \begin{array}{l} \frac{\partial gcrg}{\partial v_d} \Big|_{t_{n+1}}^k \cdot \left( v_{d,t_{n+1}}^{k+1} - v_{d,t_{n+1}}^k \right) \\ + \frac{\partial gcrg}{\partial v_g} \Big|_{t_{n+1}}^k \cdot \left( v_{g,t_{n+1}}^{k+1} - v_{g,t_{n+1}}^k \right) \\ + \frac{\partial gcrg}{\partial v_s} \Big|_{t_{n+1}}^k \cdot \left( v_{s,t_{n+1}}^{k+1} - v_{s,t_{n+1}}^k \right) \\ + \frac{\partial gcrg}{\partial v_b} \Big|_{t_{n+1}}^k \cdot \left( v_{b,t_{n+1}}^{k+1} - v_{b,t_{n+1}}^k \right) \end{array} \right] \end{array} \right\} \quad (10.17)$$

Regrouping and abbreviating the terms on the right-hand side of Eq. (10.17), one obtains

$$\begin{aligned} i_g(t) &\approx i_{g\_RHS} - ggt_d \cdot v_{d,t_{n+1}}^{k+1} - ggt_g \cdot v_{g,t_{n+1}}^{k+1} - ggts \cdot v_{s,t_{n+1}}^{k+1} \\ &\quad - ggt_b \cdot v_{b,t_{n+1}}^{k+1} - gtau \cdot v_{nqs,t_{n+1}}^{k+1} \end{aligned} \quad (10.18)$$

$G_{xpart} = -1$  was substituted in obtaining Eq. (10.18). The equivalent current is then found to be

$$i_{g\_RHS} = -ggt_d \cdot v_{bd,t_{n+1}}^k + ggt_g \cdot v_{gb,t_{n+1}}^k - ggts \cdot v_{bs,t_{n+1}}^k \quad (10.18a)$$

It is valid for both the forward and reverse operations provided Eqs. (10.15a) and (10.15b) are taken into account.

### 10.3.8 Stamping of $i_g(t)$

The KCL charging current equation for the gate node is written according to Fig. 10.1 as:

$$-i_g(t) + i_{g,external}(t) = 0 \quad (10.18b)$$

where  $i_{g,external}(t)$  is the charging current flowing into the gate node. Substituting Eq. (10.18) into Eq. (10.18b) permits writing the stamps for  $i_g(t)$  as

$$\begin{bmatrix} J & d & g & s & b & v_{nqs} \\ \hline d & 0 & 0 & 0 & 0 & 0 \\ g & -ggtd & -ggtg & -gcts & -ggtb & -gtau \\ s & 0 & 0 & 0 & 0 & 0 \\ b & 0 & 0 & 0 & 0 & 0 \\ v_{nqs} & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} x \\ v_{d,t_{n+1}}^{k+1} \\ v_{g,t_{n+1}}^{k+1} \\ v_{s,t_{n+1}}^{k+1} \\ v_{b,t_{n+1}}^{k+1} \\ v_{nqs,t_{n+1}}^{k+1} \end{bmatrix} = \begin{bmatrix} b \\ 0 \\ -i_{g,RHS} \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad (10.18c)$$

### 10.3.9 Linearization of $i_d(t)$

Linearization of the transient charge-based NQS drain current is similar to Eq. (10.16). The drain current due to NQS (refer to Chapter 6 for the formulation) is rewritten

$$i_d(t) = D_{xpart} \cdot \frac{v_{nqs} \cdot C_{\tau_{nqs}}}{C_{oxe} W_{effCV} L_{effCV} \cdot NF} \cdot gcrg \quad (10.19)$$

in which  $D_{xpart}$  is a non-linear function of the MOSFET terminal voltages and this is the only difference to consider in comparison with the linearization of Eq. (10.16) where  $G_{xpart} = -1$ . To be brief, Eq. (10.19) becomes approximately

$$i_d(t) \approx \frac{D_{xpart}^k \cdot C_{\tau_{nqs}}}{C_{oxe} W_{effCV} L_{effCV} \cdot NF} \cdot \left\{ \begin{array}{l} \left( v_{nqs} \cdot gcrg \right)_{t_{n+1}}^k + gcrg_{t_{n+1}}^k \cdot \left( v_{nqs,t_{n+1}}^{k+1} - v_{nqs,t_{n+1}}^k \right) \\ + v_{nqs,t_{n+1}}^k \cdot \left[ \begin{array}{l} \left[ \frac{\partial gcrg}{\partial v_d} \right]_{t_{n+1}}^k \cdot \left( v_{d,t_{n+1}}^{k+1} - v_{d,t_{n+1}}^k \right) \\ + \left[ \frac{\partial gcrg}{\partial v_g} \right]_{t_{n+1}}^k \cdot \left( v_{g,t_{n+1}}^{k+1} - v_{g,t_{n+1}}^k \right) \\ + \left[ \frac{\partial gcrg}{\partial v_s} \right]_{t_{n+1}}^k \cdot \left( v_{s,t_{n+1}}^{k+1} - v_{s,t_{n+1}}^k \right) \\ + \left[ \frac{\partial gcrg}{\partial v_b} \right]_{t_{n+1}}^k \cdot \left( v_{b,t_{n+1}}^{k+1} - v_{b,t_{n+1}}^k \right) \end{array} \right] \\ + v_{nqs,t_{n+1}}^k \cdot gtau \cdot \left[ \begin{array}{l} \left[ \frac{dD_{xpart}}{dv_d} \right]_{t_{n+1}}^k \cdot \left( v_{d,t_{n+1}}^{k+1} - v_{d,t_{n+1}}^k \right) \\ + \left[ \frac{dD_{xpart}}{dv_g} \right]_{t_{n+1}}^k \cdot \left( v_{g,t_{n+1}}^{k+1} - v_{g,t_{n+1}}^k \right) \\ + \left[ \frac{dD_{xpart}}{dv_s} \right]_{t_{n+1}}^k \cdot \left( v_{s,t_{n+1}}^{k+1} - v_{s,t_{n+1}}^k \right) \\ + \left[ \frac{dD_{xpart}}{dv_b} \right]_{t_{n+1}}^k \cdot \left( v_{b,t_{n+1}}^{k+1} - v_{b,t_{n+1}}^k \right) \end{array} \right] \end{array} \right\} \quad (10.20)$$

In this expression, the first term (including that curly bracketed) on the right-hand side is identical to that of Eq. (10.17) with  $G_{xpart}$  replaced by  $D_{xpart}$ . The rest is derived from the linearization of  $D_{xpart}$  with respect to the four terminal voltages, and  $gtau$  is taken from Eq. (10.15a). A compact form of Eq. (10.20) is obtained by grouping together similar variables and quantities

$$\begin{aligned}
 i_d(t) \approx & - \left[ D_{xpart}^k_{t_{n+1}} \cdot i_{g-RHS} + v_{nqs,t_{n+1}}^k \cdot gtau \cdot \begin{pmatrix} -dD_{xpart} dv_d \cdot v_{bd,t_{n+1}}^k \\ +dD_{xpart} dv_g \cdot v_{gb,t_{n+1}}^k \\ -dD_{xpart} dv_s \cdot v_{bs,t_{n+1}}^k \end{pmatrix} \right] \\
 & + \left( D_{xpart}^k_{t_{n+1}} \cdot ggtd + v_{nqs,t_{n+1}}^k \cdot gtau \cdot dD_{xpart} dv_d \right) \cdot v_{d,t_{n+1}}^{k+1} \\
 & + \left( D_{xpart}^k_{t_{n+1}} \cdot ggtg + v_{nqs,t_{n+1}}^k \cdot gtau \cdot dD_{xpart} dv_g \right) \cdot v_{g,t_{n+1}}^{k+1} \\
 & + \left( D_{xpart}^k_{t_{n+1}} \cdot ggts + v_{nqs,t_{n+1}}^k \cdot gtau \cdot dD_{xpart} dv_s \right) \cdot v_{s,t_{n+1}}^{k+1} \\
 & + \left( D_{xpart}^k_{t_{n+1}} \cdot ggtb + v_{nqs,t_{n+1}}^k \cdot gtau \cdot dD_{xpart} dv_b \right) \cdot v_{b,t_{n+1}}^{k+1} \\
 & + D_{xpart}^k_{t_{n+1}} \cdot gtau \cdot v_{nqs,t_{n+1}}^{k+1}
 \end{aligned} \tag{10.21}$$

where the voltage variables that are indexed  $(k+1)$  are unknowns to be solved for. Other variables in Eq. (10.21) are defined by

$$\left\{ \begin{array}{l} dD_{xpart-d} dv_d = \frac{dD_{xpart}}{dV_d} \Big|_{t_{n+1}}^k \\ dD_{xpart-d} dv_g = \frac{dD_{xpart}}{dV_g} \Big|_{t_{n+1}}^k \\ dD_{xpart-d} dv_s = \frac{dD_{xpart}}{dV_s} \Big|_{t_{n+1}}^k \\ dD_{xpart-d} dv_b = \frac{dD_{xpart}}{dV_b} \Big|_{t_{n+1}}^k \end{array} \right. \quad (10.21a)$$

for the forward-mode operation ( $v_{ds} = V_{ds} \geq 0$ ). They are given in Eq. (10.25d) of Section 10.3.11 once  $S_{xpart}$  has been computed for the reverse mode ( $v_{ds} = -V_{ds} < 0$ ).

### 10.3.10 Stamping of $i_d(t)$

The KCL charging current equation for the drain node is written according to Fig. 10.1 as:

$$-i_d(t) + i_{d,external}(t) = 0 \quad (10.22)$$

where  $i_{d,external}(t)$  is the charging current flowing into the drain node. Substituting Eq. (10.21) into Eq. (10.22) permits writing the stamps for  $i_d(t)$  as

$$\begin{array}{|c|c|c|c|c|c|} \hline
 J & d & g & s & b & v_{nqs} \\ \hline \hline
 d & gdnqs\_d & gdnqs\_g & gdnqs\_s & gdnqs\_b & gdnqs\_nqs \\ \hline
 g & 0 & 0 & 0 & 0 & 0 \\ \hline
 s & 0 & 0 & 0 & 0 & 0 \\ \hline
 b & 0 & 0 & 0 & 0 & 0 \\ \hline
 v_{nqs} & 0 & 0 & 0 & 0 & 0 \\ \hline
 \end{array}$$

$$\cdot \begin{bmatrix} x \\ v^{k+1}_{d,t_{n+1}} \\ v^{k+1}_{g,t_{n+1}} \\ v^{k+1}_{s,t_{n+1}} \\ v^{k+1}_{b,t_{n+1}} \\ v^{k+1}_{nqs,t_{n+1}} \end{bmatrix} = \begin{bmatrix} b \\ i_{d\_RHS} \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

(10.22a)

with the new symbols used to denote

$$\left\{ \begin{array}{l}
 gdnqs\_d = D_{xpart_{t_{n+1}}}^k \cdot ggt_d + v^k_{nqs,t_{n+1}} \cdot gtau \cdot dD_{xpart\_dv_d} \\
 gdnqs\_g = D_{xpart_{t_{n+1}}}^k \cdot ggt_g + v^k_{nqs,t_{n+1}} \cdot gtau \cdot dD_{xpart\_dv_g} \\
 gdnqs\_s = D_{xpart_{t_{n+1}}}^k \cdot ggt_s + v^k_{nqs,t_{n+1}} \cdot gtau \cdot dD_{xpart\_dv_s} \\
 gdnqs\_b = D_{xpart_{t_{n+1}}}^k \cdot ggt_b + v^k_{nqs,t_{n+1}} \cdot gtau \cdot dD_{xpart\_dv_b} \\
 gdnqs\_nqs = D_{xpart_{t_{n+1}}}^k \cdot gtau \\
 \\ 
 i_{d\_RHS} = D_{xpart_{t_{n+1}}}^k \cdot i_{g\_RHS} + v^k_{nqs,t_{n+1}} \cdot gtau \cdot \begin{pmatrix} -dD_{xpart\_dv_d} \cdot v^k_{bd,t_{n+1}} \\ +dD_{xpart\_dv_g} \cdot v^k_{gb,t_{n+1}} \\ -dD_{xpart\_dv_s} \cdot v^k_{bs,t_{n+1}} \end{pmatrix}
 \end{array} \right.$$

(10.22b)

### 10.3.11 Linearization of $i_s(t)$

In the event of the linearization of the transient charge-based NQS source terminal current  $i_s(t)$ , one may follow the same approach applied to  $i_d(t)$ . The  $D_{xpart}$  term need be replaced with  $S_{xpart}$  while the rest remains the same:

$$i_s(t) = S_{xpart} \cdot \frac{v_{nqs} \cdot C_{\tau_{nqs}}}{C_{oxe} W_{effCV} L_{effCV} \cdot NF} \cdot gcrg \quad (10.23)$$

Using Eq. (10.20) and substituting  $S_{xpart}$  for  $D_{xpart}$ , the linearization results in

$$i_s(t) \approx \frac{S_{xpart}^k \cdot C_{\tau_{nqs}}}{C_{oxe} W_{effCV} L_{effCV} \cdot NF} \cdot \left[ \begin{array}{l} \left( v_{nqs} \cdot gcrg \right)_{t_{n+1}}^k + gcrg_{t_{n+1}}^k \cdot \left( v_{nqs,t_{n+1}}^{k+1} - v_{nqs,t_{n+1}}^k \right) \\ + v_{nqs,t_{n+1}}^k \cdot \left[ \begin{array}{l} \left[ \frac{\partial gcrg}{\partial v_d} \right]_{t_{n+1}}^k \cdot \left( v_{d,t_{n+1}}^{k+1} - v_{d,t_{n+1}}^k \right) \\ + \left[ \frac{\partial gcrg}{\partial v_g} \right]_{t_{n+1}}^k \cdot \left( v_{g,t_{n+1}}^{k+1} - v_{g,t_{n+1}}^k \right) \\ + \left[ \frac{\partial gcrg}{\partial v_s} \right]_{t_{n+1}}^k \cdot \left( v_{s,t_{n+1}}^{k+1} - v_{s,t_{n+1}}^k \right) \\ + \left[ \frac{\partial gcrg}{\partial v_b} \right]_{t_{n+1}}^k \cdot \left( v_{b,t_{n+1}}^{k+1} - v_{b,t_{n+1}}^k \right) \end{array} \right] \\ + v_{nqs,t_{n+1}}^k \cdot gtau \cdot \left[ \begin{array}{l} \left[ \frac{dS_{xpart}}{dv_d} \right]_{t_{n+1}}^k \cdot \left( v_{d,t_{n+1}}^{k+1} - v_{d,t_{n+1}}^k \right) + \left[ \frac{dS_{xpart}}{dv_g} \right]_{t_{n+1}}^k \cdot \left( v_{g,t_{n+1}}^{k+1} - v_{g,t_{n+1}}^k \right) \\ + \left[ \frac{dS_{xpart}}{dv_s} \right]_{t_{n+1}}^k \cdot \left( v_{s,t_{n+1}}^{k+1} - v_{s,t_{n+1}}^k \right) + \left[ \frac{dS_{xpart}}{dv_b} \right]_{t_{n+1}}^k \cdot \left( v_{b,t_{n+1}}^{k+1} - v_{b,t_{n+1}}^k \right) \end{array} \right] \end{array} \right] \quad (10.24)$$

A compact version of Eq. (10.24) is written for the purpose of stamping

$$\begin{aligned}
i_s(t) \approx & - \left[ S_{xpart}^k \cdot i_{g-RHS} + v_{nqs,t_{n+1}}^k \cdot gtau \cdot \begin{pmatrix} -dS_{xpart} dv_d \cdot v_{bd,t_{n+1}}^k \\ +dS_{xpart} dv_g \cdot v_{gb,t_{n+1}}^k \\ -dS_{xpart} dv_s \cdot v_{bs,t_{n+1}}^k \end{pmatrix} \right] \\
& + \left( S_{xpart}^k \cdot ggtd + v_{nqs,t_{n+1}}^k \cdot gtau \cdot dS_{xpart} dv_d \right) \cdot v_{d,t_{n+1}}^{k+1} \\
& + \left( S_{xpart}^k \cdot ggtg + v_{nqs,t_{n+1}}^k \cdot gtau \cdot dS_{xpart} dv_g \right) \cdot v_{g,t_{n+1}}^{k+1} \\
& + \left( S_{xpart}^k \cdot ggts + v_{nqs,t_{n+1}}^k \cdot gtau \cdot dS_{xpart} dv_s \right) \cdot v_{s,t_{n+1}}^{k+1} \\
& + \left( S_{xpart}^k \cdot ggtb + v_{nqs,t_{n+1}}^k \cdot gtau \cdot dS_{xpart} dv_b \right) \cdot v_{b,t_{n+1}}^{k+1} \\
& + S_{xpart}^k \cdot gtau \cdot v_{nqs,t_{n+1}}^{k+1}
\end{aligned} \tag{10.25}$$

In the code implementation,  $S_{xpart}$  and its derivatives are computed from  $D_{xpart}$  for the forward-mode operation, since  $S_{xpart} = 1 - D_{xpart}$  and  $D_{xpart}$  has already been computed in this mode first. Therefore,

$$\begin{aligned}
dS_{xpart} dv_d &= -\frac{dD_{xpart}}{dV_d} \Big|_{t_{n+1}}^k \\
dS_{xpart} dv_g &= -\frac{dD_{xpart}}{dV_g} \Big|_{t_{n+1}}^k \\
dS_{xpart} dv_s &= -\frac{dD_{xpart}}{dV_s} \Big|_{t_{n+1}}^k \\
dS_{xpart} dv_b &= -\frac{dD_{xpart}}{dV_b} \Big|_{t_{n+1}}^k
\end{aligned} \tag{10.25a}$$

In the case of the reverse mode ( $v_{ds} = -V_{ds} < 0$ ),  $S_{xpart}$  instead of  $D_{xpart}$  is computed first for good code execution efficiency.  $S_{xpart}$  is defined as

$$S_{xpart} = -\frac{Q_d}{Q_g + Q_b} \tag{10.25b}$$

Here the denominator represents the quasi-static channel charge but with an opposite sign, namely  $Q_{ch\_qs} = -(Q_g + Q_b)$ . The intrinsic charge associated with the physical source terminal is in fact the intrinsic charge computed for the electrical drain of the model; hence the intrinsic electrical drain charge  $Q_d$  instead of  $Q_s$  of the source terminal, both as functions of the electrical terminal voltages  $V_d$ ,  $V_g$ ,  $V_s$ , and  $V_b$ , is present in Eq. (10.25b). Moreover, all the derivatives of the intrinsic charge components of Eq. (10.25b) with respect to the electrical terminal voltages are obtained from the quasi-static charge-capacitance model, CAPMOD = 0, 1, 2, or 3. With this in mind, one obtains for the reverse mode

$$\begin{aligned} dS_{xpart-dv_d} &= -\frac{dS_{xpart}}{dV_s} \Big|_{t_{n+1}}^k \\ dS_{xpart-dv_g} &= -\frac{dS_{xpart}}{dV_g} \Big|_{t_{n+1}}^k \\ dS_{xpart-dv_s} &= -\frac{dS_{xpart}}{dV_d} \Big|_{t_{n+1}}^k \\ dS_{xpart-dv_b} &= -\frac{dS_{xpart}}{dV_b} \Big|_{t_{n+1}}^k \end{aligned} \quad (10.25c)$$

Note again the electrical source and drain voltages above ( $V_s$  and  $V_d$ , respectively) have been swapped (interchanged) to obtain the derivatives of  $S_{xpart}$  with respect to the physical source and drain biases ( $v_s$  and  $v_d$ , respectively).

According to  $S_{xpart} = 1 - D_{xpart}$  again, once Eqs. (10.25b) and (10.25c) are computed,  $D_{xpart}$  and its derivatives are readily obtained for the reverse-mode operation

$$\begin{aligned} dD_{xpart-dv_d} &= -dS_{xpart-dv_d} \\ dD_{xpart-dv_g} &= -dS_{xpart-dv_g} \\ dD_{xpart-dv_s} &= -dS_{xpart-dv_s} \\ dD_{xpart-dv_b} &= -dS_{xpart-dv_d} \end{aligned} \quad (10.25d)$$

As has been mentioned in Section 10.3.9, this is useful for the linearization of  $i_d(t)$  and Eq. (10.21) for the reverse mode.

### 10.3.12 Stamping of $i_s(t)$

The KCL charging current equation for the source node is, according to Fig. 10.1,

$$-i_s(t) + i_{s,external}(t) = 0 \quad (10.26)$$

where  $i_{s,external}(t)$  is the charging current flowing into the source node from the circuit elements, devices, or sources that are connected to the source node. Substituting Eq. (10.25) into Eq. (10.26) permits writing the stamps for  $i_s(t)$  as

$$\begin{bmatrix} J & | & d & g & s & b & v_{nqs} \\ \hline d & 0 & 0 & 0 & 0 & 0 & 0 \\ g & 0 & 0 & 0 & 0 & 0 & 0 \\ s & gsnqs\_d & gsnqs\_g & gsnqs\_s & gsnqs\_b & gsnqs\_nqs & \\ b & 0 & 0 & 0 & 0 & 0 & 0 \\ v_{nqs} & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} x \\ v_{d,t_{n+1}}^{k+1} \\ v_{g,t_{n+1}}^{k+1} \\ v_{s,t_{n+1}}^{k+1} \\ v_{b,t_{n+1}}^{k+1} \\ v_{nqs,t_{n+1}}^{k+1} \end{bmatrix} = \begin{bmatrix} b \\ 0 \\ 0 \\ i_{s-RHS} \\ 0 \\ 0 \end{bmatrix} \quad (10.26a)$$

The new symbols are defined below.

$$\left\{ \begin{array}{l} gsnqs\_d = S_{xpart_{t_{n+1}}}^k \cdot ggt_d + v_{nqs,t_{n+1}}^k \cdot gtau \cdot dS_{xpart-d} dv_d \\ gsnqs\_g = S_{xpart_{t_{n+1}}}^k \cdot ggt_g + v_{nqs,t_{n+1}}^k \cdot gtau \cdot dS_{xpart-d} dv_g \\ gsnqs\_s = S_{xpart_{t_{n+1}}}^k \cdot ggt_s + v_{nqs,t_{n+1}}^k \cdot gtau \cdot dS_{xpart-d} dv_s \\ gsnqs\_b = S_{xpart_{t_{n+1}}}^k \cdot ggt_b + v_{nqs,t_{n+1}}^k \cdot gtau \cdot dS_{xpart-d} dv_b \\ gsnqs\_nqs = S_{xpart_{t_{n+1}}}^k \cdot gtau \\ i_{s-RHS} = S_{xpart_{t_{n+1}}}^k \cdot i_{g-RHS} \\ + v_{nqs,t_{n+1}}^k \cdot gtau \cdot \left( \begin{array}{l} -dS_{xpart-d} dv_d \cdot v_{bd,t_{n+1}}^k + dS_{xpart-d} dv_g \cdot v_{gb,t_{n+1}}^k \\ -dS_{xpart-d} dv_s \cdot v_{bs,t_{n+1}}^k \end{array} \right) \end{array} \right. \quad (10.26b)$$

Note that the denominator  $(Q_g + Q_b)$  of  $S_{xpart}$  and  $D_{xpart}$  of Eq. (10.25b) could take on an infinitesimally small value when the device operates away from inversion and approaches the flat-band condition. That would result in numerical overflows, divide-by-zero or “NaN” (not a number) violations in SPICE code executions, if the  $S_{xpart}$  and  $D_{xpart}$  equations were coded as is. To avoid such instabilities and make the SPICE implementation work simple, the following practical numerical treatment is applied in BSIM4 when the absolute value of the charge  $(Q_g + Q_b) \cdot 10^5$  is not greater than the product  $(C_{oxe} W_{effCV} L_{effCV} \cdot NF)$ :  $D_{xpart} = 0.4$  is chosen for the 40/60 charge partition if  $XPART < 0.5$ ;  $D_{xpart} = 0$  for 0/100 if  $XPART > 0.5$ ; and  $D_{xpart} = 0.5$  for 50/50 if  $XPART = 0.5$ , all for forward-mode operations.  $S_{xpart}$  is  $S_{xpart} = 1 - D_{xpart}$ . These choices are applied to  $S_{xpart}$  for reverse-mode operations and  $D_{xpart} = 1 - S_{xpart}$ .

The treatment is consistent with the quasi-static charge portioning. However, it introduces discontinuities in the derivatives of  $S_{xpart}$  and  $D_{xpart}$  with respect to voltages at the bias point where

$$|Q_g + Q_b| \cdot 10^5 = C_{oxe} W_{effCV} L_{effCV} \cdot NF \quad (10.27)$$

More numerically robust solutions are possible, an example of which would be let  $S_{xpart}$  and  $D_{xpart}$  be given

$$\sqrt{\frac{Q_d^2}{(Q_g + Q_b)^2 + 10^{-40}}}$$

which is smooth for all regions of operation (at a cost of additional CPU runtime on the square and square-root functions). The constant  $10^{-40}$  is a good choice which takes into account the possible numerical values that the MOSFET gate, channel, and body charges can take on.

## 10.4 Composite Stamps for Transient NQS Model

Once the individual stamps for each current components of the transient NQS model have been constructed, one can proceed to combine those individual stamps into a lump sum, called the composite stamps, by recalling and inserting Eqs. (10.10e), (10.12c), (10.15c), (10.18c), (10.22a), and (10.26a) and loading these composite stamps into the circuit matrix  $\mathbf{J} \cdot \mathbf{x} = \mathbf{b}$ . The composite stamps come in the form

$$\begin{array}{|c|cccc|c|} \hline
 J & d & g & s & b & v_{nqs} \\ \hline \hline
 d & gdnqs\_d & gdnqs\_g & gdnqs\_s & gdnqs\_b & gdnqs\_nqs \\ 
 g & -ggtd & -ggtg & -gcts & -ggtb & -gtau \\ 
 s & gsnqs\_d & gsnqs\_g & gsnqs\_s & gsnqs\_b & gsnqs\_nqs \\ 
 b & 0 & 0 & 0 & 0 & 0 \\ 
 v_{nqs} & ggtd - gcqdb & ggtg - gcqgb & ggts - gcqsb & ggtb - gcqbb & g_{C_t} + gtau \\ 
 \hline
 \end{array}$$

$$\begin{bmatrix} x \\ v_{d,t_{n+1}}^{k+1} \\ v_{g,t_{n+1}}^{k+1} \\ v_{s,t_{n+1}}^{k+1} \\ v_{b,t_{n+1}}^{k+1} \\ v_{nqs,t_{n+1}}^{k+1} \end{bmatrix} = \begin{bmatrix} b \\ i_{d-RHS} \\ -i_{g-RHS} \\ i_{s-RHS} \\ 0 \\ i_{ch\_qs\_RHS} - i_{C_t-RHS} - i_{R_t-RHS} \end{bmatrix} \quad (10.28)$$

Recall that the sum of each row or column of a correctly-built circuit  $\mathbf{J}$  matrix must be zero according to the charge neutrality requirements. It can be proved that this attribute holds in the case of Eq. (10.28) if only the  $4 \times 4$  conductance  $\mathbf{J}$  matrix that is associated with the MOSFET four terminals ( $d$ ,  $g$ ,  $s$ , and  $b$ ) is taken into account. For example, consider Column  $d$ . One can verify that

$$g_{dnqs-d} - ggtd + g_{snqs-d} \equiv 0 \quad (10.28a)$$

Including the effective trans-conductance at the  $v_{nqs}$  node with respect to the  $d$  node, namely  $(ggtd - gcqdb)$ , will however break the equality of Eq. (10.28a). This is to say

$$g_{dnqs-d} - ggtd + g_{snqs-d} + (ggtd - gcqdb) \neq 0 \quad (10.28b)$$

So is the case of the vector  $\mathbf{b}$  (namely,  $i_{d-RHS} - i_{g-RHS} + i_{s-RHS} \equiv 0$ ). This equality would no longer hold if the term  $(i_{ch-qs-RHS} - i_{Cl-RHS} - i_{Rl-RHS})$  should be inserted on its left-hand side. This is because the auxiliary NQS charge-deficit sub-circuit node as presented in Fig. 10.1, though coupled to the charging currents at the MOSFET four terminals to facilitate a boundary-value condition for the  $v_{nqs}$  iterations, does not constitute an additional real charge node that needs to satisfy the charge neutrality equation

$$Q_d + Q_g + Q_s + Q_b \equiv 0 \quad (10.28c)$$

of a MOS transistor. The MOS transistor itself is a complete and independent charge-neutral system under any circumstances, whether under the quasi-static assumption or not. This is analogous to the case of auxiliary self-heating sub-circuit for SOI modeling.

This attribute of an equivalent circuit matrix has many applications in analyzing and validating SPICE simulation, particularly in analog and RF IC designs where one often finds the need to verify, for instance, the consistency between the SPICE small-signal terminal resistances or gains and those computed (manually or by a software program) from the equivalent circuits. Violations of, for instance, Eq. (10.28a) often manifest themselves in the need for excessive SPICE iterations or to find false convergence or a convergence failure.

## 10.5 Bypass

Bypassing can significantly speed up simulations, especially for large circuits where sophisticated device models and stimuli such as square waveform voltage supplies are present. The speedup results from skipping the “unnecessary” charge, current, conductance and capacitance model computations needed by the  $\mathbf{J}$  and  $\mathbf{b}$  matrix stamping (as shown in Eq. (10.28)) to solve for the  $\mathbf{v}_{t_{n+1}}^{k+1}$  vector. The condition under which bypassing or skipping can take place is described in the following. For the terminal voltages, the absolute changes between the values  $\mathbf{v}_{t_{n+1}}^k$  just solved from the  $k^{\text{th}}$  iteration and the corresponding values  $\mathbf{v}_{t_{n+1}}^{k-1}$  from  $(k-1)^{\text{th}}$  (that is,  $|\mathbf{v}_{t_{n+1}}^k - \mathbf{v}_{t_{n+1}}^{k-1}|$ ) must be within a predefined range. For device DC currents, the changes in them, approximated by a linear projection based upon  $(\mathbf{v}_{t_{n+1}}^k - \mathbf{v}_{t_{n+1}}^{k-1})$ , must not exceed an analogous set of pre-defined range. As an example in SPICE3, the changes in  $v_{ds}$  and the DC channel current  $I_{ch}$  are

$$|\mathbf{v}_{ds,t_n+1}^k - \mathbf{v}_{ds,t_n+1}^{k-1}| < [\text{reltol} \cdot \max(|\mathbf{v}_{ds,t_n+1}^k|, |\mathbf{v}_{ds,t_n+1}^{k-1}|) + \text{volttol}] \quad (10.29a)$$

and

$$|I_{ch\_projected,t_n+1}^k - I_{ch,t_n+1}^{k-1}| < [\text{reltol} \cdot \max(|I_{ch\_projected,t_n+1}^k|, |I_{ch,t_n+1}^{k-1}|) + \text{abstol}] \quad (10.29b)$$

**reltol**, **volttol**, and **abstol** are the SPICE3 option statement parameters with default values of  $1 \times 10^{-3}$ ,  $1 \times 10^{-6}$  V, and  $1 \times 10^{-12}$  A, respectively, and  $\max(x, y)$  is the mathematic maximum function supplied by a typical C language compiler. It is equal to  $x$  if  $x \geq y$  and equal to  $y$  otherwise.

In Eq. (10.29b), a linear projection of the channel current  $I_{ch\_projected,t_n+1}^k$  based upon the voltage differences  $(\mathbf{v}_{t_{n+1}}^k - \mathbf{v}_{t_{n+1}}^{k-1})$  is used because the exact  $I_{ch,t_n+1}^k$  value is unavailable at this moment of time, which can only be known until after the model is evaluated. This linear

projection aids SPICE to make the decision whether the model equations need to be evaluated with  $v_{t_{n+1}}^k$  or just should be bypassed. A useful projection can be performed as follows

$$\begin{aligned} I_{ch\_projected,t_n+1}^k &\approx I_{ch,t_n+1}^{k-1} + \frac{\partial I_{ch}}{\partial v_d} \Big|_{t_{n+1}}^{k-1} \cdot (v_{d,t_{n+1}}^k - v_{d,t_{n+1}}^{k-1}) \\ &+ \frac{\partial I_{ch}}{\partial v_g} \Big|_{t_{n+1}}^{k-1} \cdot (v_{g,t_{n+1}}^k - v_{g,t_{n+1}}^{k-1}) \\ &+ \frac{\partial I_{ch}}{\partial v_s} \Big|_{t_{n+1}}^{k-1} \cdot (v_{s,t_{n+1}}^k - v_{s,t_{n+1}}^{k-1}) + \frac{\partial I_{ch}}{\partial v_b} \Big|_{t_{n+1}}^{k-1} \cdot (v_{b,t_{n+1}}^k - v_{b,t_{n+1}}^{k-1}) \end{aligned} \quad (10.30)$$

where  $(k-1) \geq 1$ . All currents and conductances will take on the values that have already been obtained from the  $(k-1)^{\text{th}}$  iteration at  $t_{n+1}$ .

By noting that

$$\frac{\partial I_{ch}}{\partial v_d} \Big|_{t_{n+1}}^{k-1} + \frac{\partial I_{ch}}{\partial v_g} \Big|_{t_{n+1}}^{k-1} + \frac{\partial I_{ch}}{\partial v_s} \Big|_{t_{n+1}}^{k-1} + \frac{\partial I_{ch}}{\partial v_b} \Big|_{t_{n+1}}^{k-1} \equiv 0 \quad (10.31)$$

Equation (10.30) reduces to

$$\begin{aligned} I_{ch\_projected,t_n+1}^k &\approx I_{ch,t_n+1}^{k-1} + G_{ds} \cdot (v_{ds,t_{n+1}}^k - v_{ds,t_{n+1}}^{k-1}) \\ &+ G_m \cdot (v_{gs,t_{n+1}}^k - v_{gs,t_{n+1}}^{k-1}) + G_{mbs} \cdot (v_{bs,t_{n+1}}^k - v_{bs,t_{n+1}}^{k-1}) \end{aligned} \quad (10.32)$$

with

$$\left\{ \begin{array}{l} \frac{\partial I_{ch}}{\partial v_d} \Big|_{t_{n+1}}^{k-1} = \frac{\partial I_{ch}}{\partial V_d} \Big|_{t_{n+1}}^{k-1} = G_{ds} \\ \frac{\partial I_{ch}}{\partial v_g} \Big|_{t_{n+1}}^{k-1} = \frac{\partial I_{ch}}{\partial V_g} \Big|_{t_{n+1}}^{k-1} = G_m \\ \frac{\partial I_{ch}}{\partial v_b} \Big|_{t_{n+1}}^{k-1} = \frac{\partial I_{ch}}{\partial V_b} \Big|_{t_{n+1}}^{k-1} = G_{mbs} \end{array} \right. \quad (10.32a)$$

for the forward bias mode ( $v_{ds} = V_{ds} \geq 0$ ), and to

$$I_{ch\_projected,t_n+1}^k \approx I_{ch,t_n+1}^{k-1} - G_{ds} \cdot (v_{ds,t_{n+1}}^k - v_{ds,t_{n+1}}^{k-1}) \\ + G_m \cdot (v_{gd,t_{n+1}}^k - v_{gd,t_{n+1}}^{k-1}) + G_{mbs} \cdot (v_{bs,t_{n+1}}^k - v_{bs,t_{n+1}}^{k-1}) \quad (10.32b)$$

for the reverse mode ( $v_{ds} = -V_{ds} < 0$ ) by recognizing the need for an electrical source and drain terminal swapping. Hence

$$\begin{cases} \frac{\partial I_{ch}}{\partial v_s} \Big|_{t_{n+1}}^{k-1} = \frac{\partial I_{ch}}{\partial V_d} \Big|_{t_{n+1}}^{k-1} = G_{ds} \\ \frac{\partial I_{ch}}{\partial v_g} \Big|_{t_{n+1}}^{k-1} = \frac{\partial I_{ch}}{\partial V_g} \Big|_{t_{n+1}}^{k-1} = G_m \\ \frac{\partial I_{ch}}{\partial v_b} \Big|_{t_{n+1}}^{k-1} = \frac{\partial I_{ch}}{\partial V_b} \Big|_{t_{n+1}}^{k-1} = G_{mbs} \end{cases} \quad (10.32c)$$

where  $G_{ds}$ ,  $G_m$ , and  $G_{mbs}$  have their usual intended meanings.

Subtracting  $I_{ch,t_n+1}^{k-1}$  from both sides of Eqs. (10.32) and (10.32b) and taking the absolute values for them yield the left-hand side of Eq. (10.29b). Once it is determined that Eqs. (10.29a) and (10.29b) are both satisfied and thus bypassing is to be in effect for this particular transistor, then all its DC currents, charges, conductances and capacitances will take on the values that are labeled with (k-1) above without the need for evaluating the model equations again in order to solve for  $v_{t_{n+1}}^{k+1}$ . They will then be stamped and loaded directly into Eq. (10.28), where substitutions of  $v_{t_{n+1}}^{k-1}$  for  $v_{t_{n+1}}^k$  also have to be applied. This SPICE3 bypass treatment or its variants have been implemented in many SPICE simulators as well. To have a graphical view, refer to Fig. 10.3, in which a typical SPICE and device model iteration flow chart is exemplified with the emphasis placed on the bypassing of the BSIM4 charge-deficit transient NQS model [Detailed explanation of this flow is given at the end of this section].

Upon examining Fig. 10.3, one realizes that the NQS charge node voltage  $v_{nqs}$  plays no role in the model evaluations until the stage of the conductance and capacitance swapping and the right-hand side equivalent current calculation. Therefore, Eqs. (10.29a) and (10.29b) can be implemented to skip the DC and capacitance model evaluations prior to that stage without requiring the changes in the NQS charge node voltage  $v_{nqs}$  and its nodal currents to be evaluated.

Furthermore, it has been verified with many SPICE simulation experiments that even if the MOSFET terminal voltages and DC currents can satisfy the criteria of Eqs. (10.29a) and (10.29b), the charge node voltage  $v_{nqs}$  in many occasions cannot. The reason is readily appreciated:  $v_{nqs}$  is mathematically an independent state variable and can be physically forced to have a significant change in its magnitude within a very short period of fast transients. Therefore, any attempts to substitute  $v_{nqs,t_{n+1}}^{k-1}$  for  $v_{nqs,t_{n+1}}^k$  at any steps of Fig. 10.3 are irrational. By following the line of this thought, one fix associated with  $v_{nqs}$  was made by commenting out the following line from the BSIM4 bypass code section (or Step 2-2 of Fig. 10.3)

```
qdef = * (ckt->CKTstate0 + here->BSIM4qdef);
```

in which  $qdef$  denotes  $v_{nqs,t_{n+1}}^k$  while the right side signifies  $v_{nqs,t_{n+1}}^{k-1}$  in the nomenclature used in this chapter. (Note that an offset in the memory address pointer such as  $(ckt->CKTstate0 + here->BSIM4qdef)$  stores the numeric values of terminal/nodal voltages and charges from the  $(k-1)^{th}$  iteration. These offsets are symbolized by  $V_{state0}$  for convenient reference in Fig. 10.3. More detailed descriptions of  $V_{state0}$  are given at the end of this section.

It is worthwhile to point out that the bypass criteria, Eqs. (10.29a) and (10.29b), although designed with heuristics, work very well. However, the bypass implementation method, such as the one just mentioned above has limitations — accuracy loss or even false convergence can be observed for applications with low-voltage and fast transient stimuli, where the changes in terminal voltages between several consecutive time points can be so small that bypassing, once kicked in, will continue with the same terminal bias values. The consequence is then no updates are to be made to the circuit matrix, for many subsequent time points. This abnormal bypassing results in piece-wise or “saw”-shaped waveforms as exemplified in Fig. 10.4, where the quasi-static gate charges  $Q_g$  (Fig. 10.4 (c)) are plotted.

It has been proven that a fix to this problem is possible. A hint is that the circuit matrix does have to be updated even when bypassing is taking place, so that the simulator will not get stuck with the same matrix for the subsequent time points.

Now turn back to Fig. 10.3 for more discussions of the simulation flow. The ongoing  $(k+1)^{\text{th}}$  iteration for  $v_{t_{n+1}}^{k+1}$  uses the vectors  $\mathbf{v}_{t_{n+1}}^k$ ,  $\mathbf{v}_{nqs,t_{n+1}}^k$  and  $\mathbf{V}_{\text{state}0}$  as the input terminal/nodal voltages (plus  $v_{nqs}$  if the transient charge deficit/surplus NQS model is invoked). A few observations are worth making. First, since the quasi-static charge, current, conductance, and capacitance model equations are independent of the NQS charge node voltage  $v_{nqs}$ , whether the quasi-static model evaluation should be bypassed or not is not contingent upon whether  $v_{nqs}$

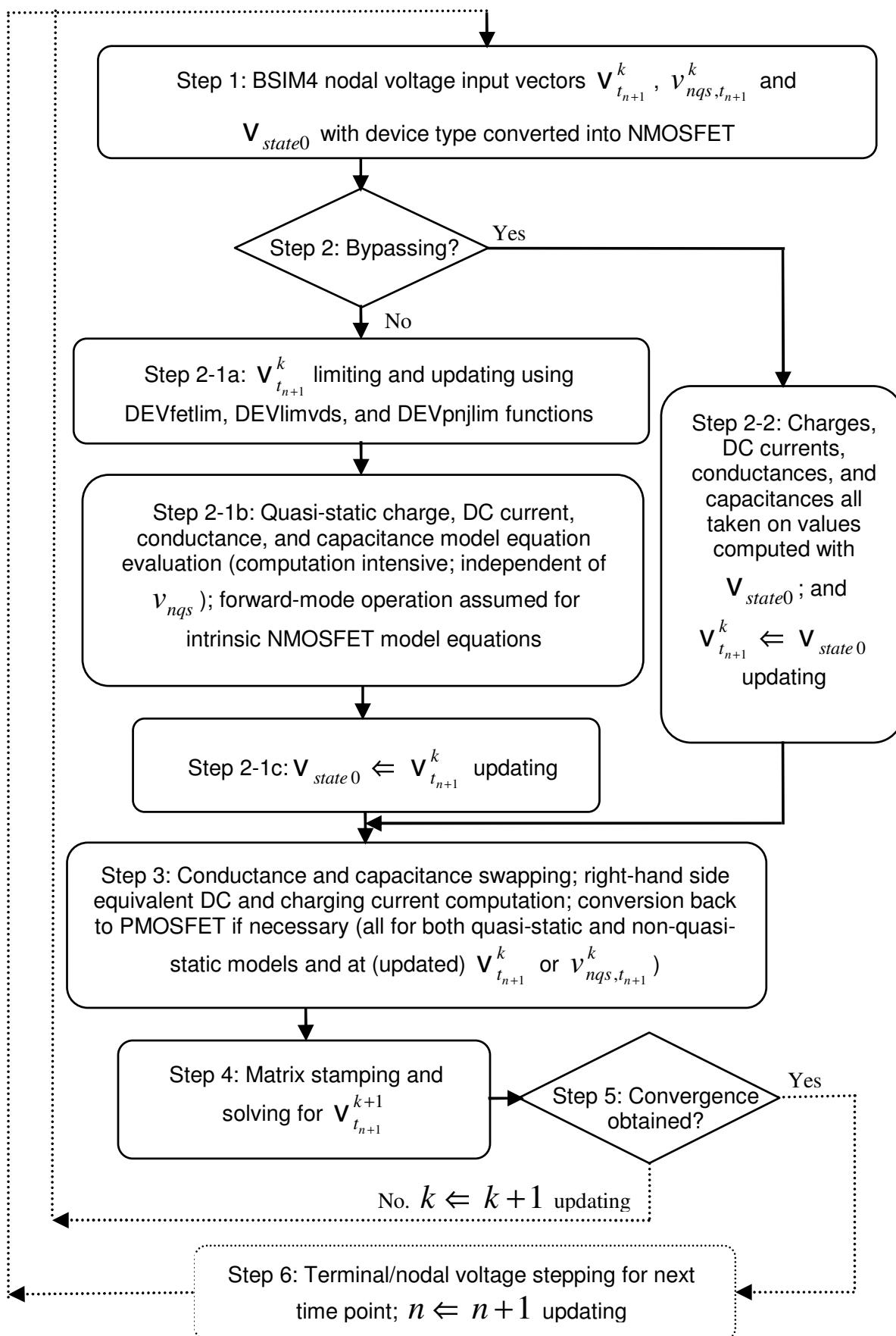


Fig. 10.3 BSIM4 iteration flow within SPICE in the time domain. LTE is not included.

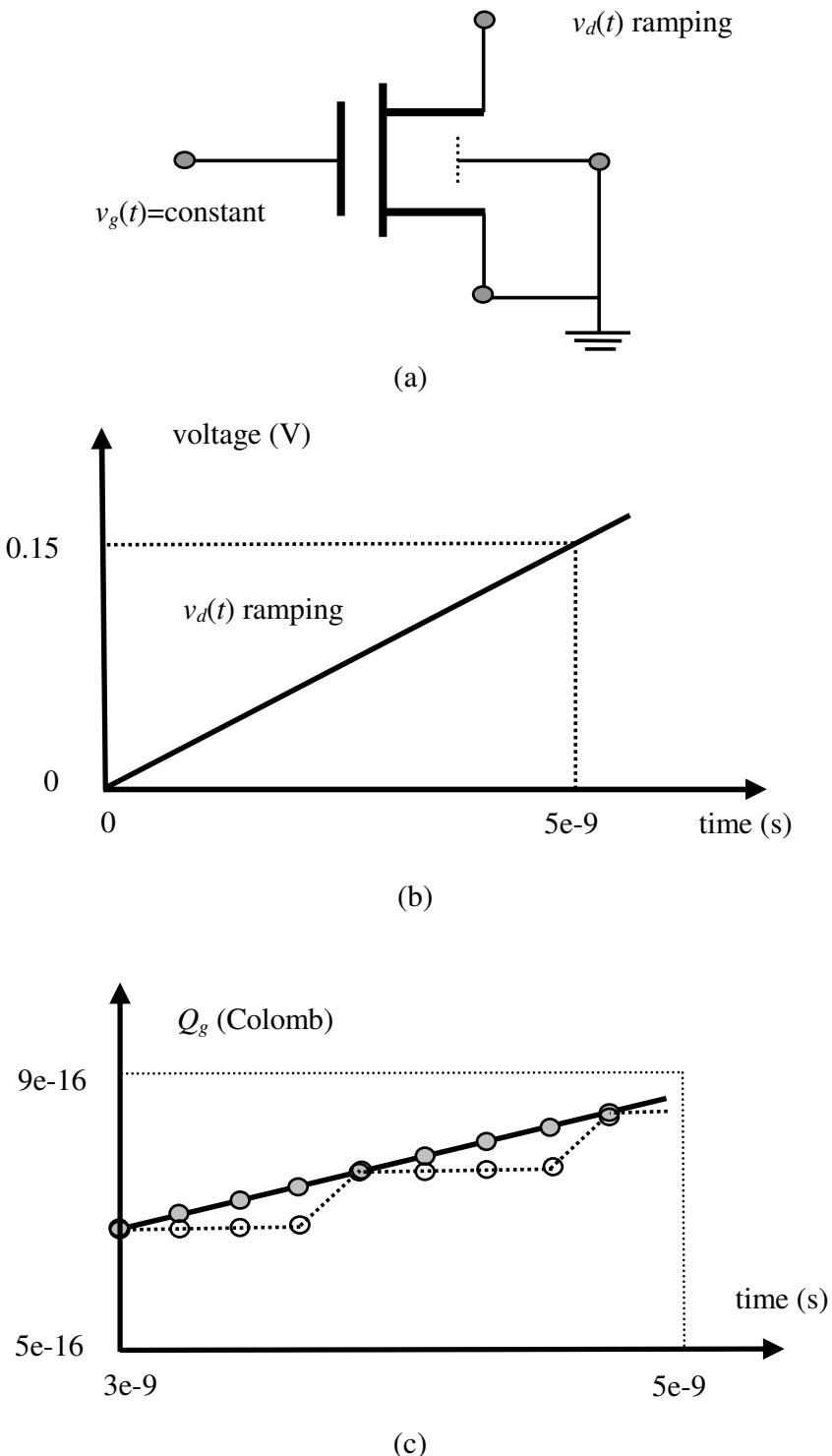


Fig. 10.4 A  $Q_g$  accuracy and false convergence problem (c) induced by using the simple SPICE3 bypassing algorithm for a single-transistor circuit (a) with a fast-ramping low-voltage stimulus (b). In (c), the solid line with solid circles is obtained in the accuracy mode while the dotted line with open circles results from using this simple bypass.

can satisfy Eqs. (10.29a) and (10.29b) in Step 2 of Fig. 10.3. Secondly,  $\mathbf{V}_{state0}$  is defined as the voltage vector that is created to store the voltage vectors,  $\mathbf{v}_{t_{n+1}}^{k-1}$  or possibly updated  $\mathbf{v}_{t_{n+1}}^{k-1}$  first, and then  $\mathbf{v}_{t_{n+1}}^k$  or possibly updated  $\mathbf{v}_{t_{n+1}}^k$ .  $\mathbf{V}_{state0}$  does not include  $v_{nqs}$ . The need for updating  $\mathbf{v}_{t_{n+1}}^{k-1}$  and  $\mathbf{v}_{t_{n+1}}^k$  arises from the voltage limiting performed at Step 2-1a. A heuristic measure is often needed to pull back any terminal/nodal voltages that may be far away from their anticipated values. Thirdly, the bypass checking performed at Step 2 is based upon the changes of the terminal/nodal voltages and the DC currents between  $\mathbf{v}_{t_{n+1}}^k$  and  $\mathbf{V}_{state0}$ . Convergence is accomplished at Step 5 if the changes of the DC currents between  $\mathbf{v}_{t_{n+1}}^{k+1}$  and  $\mathbf{V}_{state0}$  fulfill the convergence criteria. Finally, once bypassing has been determined at Step 2 to take place subsequently, all the DC currents, charges, conductances, and capacitances that are to be stamped into the circuit matrix will take on their respective previous values.

## 10.6 Convergence Checking

The discussion of convergence checking and implementation will be presented in this section.

The linearization using the Taylor series expansion is a necessary approximation to the non-linear charge and DC current equations in order to stamp them into the linear circuit matrix. Numerical iterations to solve the matrix are required at each time point to attain a matrix solution that can satisfy the non-linear equations. To minimize any possible resulting errors, a number of iterations are performed so that the simulation results meet a pre-defined accuracy threshold. This process is known as convergence checking.

Almost with no exception for any SPICE simulators, one convergence criterion is to use the maximum number of iterations

allowed for each time point. A convergence failure will be reported once that maximum number is reached and the errors have not been reduced below the accuracy tolerance threshold.

The magnitude of the error is defined as the absolute difference of a DC current value between the current from the linear projection with  $v_{t_{n+1}}^{k+1}$  and the current computed with  $v_{t_{n+1}}^k$ . (Rigorously speaking, here  $v_{t_{n+1}}^k$  should be  $v_{state0}$  because  $v_{t_{n+1}}^k$  may be updated as a result of voltage limiting from the inception of the  $(k+1)^{th}$  iteration.) Take the channel current  $I_{ch}$  as an example, a convergence for the channel current is said to be achieved for the  $(k+1)^{th}$  iteration only if

$$\left| I_{ch\_linearized,t_{n+1}}^{k+1} - I_{ch,t_{n+1}}^k \right| < [\text{reitol} \cdot \max(|I_{ch\_linearized,t_{n+1}}^k|, |I_{ch,t_{n+1}}^{k-1}|) + \text{abstol}] \quad (10.33)$$

is satisfied, an analog of Eq. (10.29b) with the same interpretations for **reitol** and **abstol**. The left-hand side of Eq. (10.33) accounts for iteration errors and the right-hand side denotes the pre-defined convergence tolerance threshold. How to formulate  $I_{ch\_linearized,t_{n+1}}^{k+1}$  in code implementation? The answer is

$$\begin{aligned} I_{ch\_linearized,t_{n+1}}^{k+1} \approx I_{ch,t_{n+1}}^k &+ \frac{\partial I_{ch}}{\partial V_d} \Big|_{t_{n+1}}^k \cdot (v_{ds,t_{n+1}}^{k+1} - v_{ds,t_{n+1}}^k) \\ &+ \frac{\partial I_{ch}}{\partial V_g} \Big|_{t_{n+1}}^k \cdot (v_{gs,t_{n+1}}^{k+1} - v_{gs,t_{n+1}}^k) + \frac{\partial I_{ch}}{\partial V_b} \Big|_{t_{n+1}}^k \cdot (v_{bs,t_{n+1}}^{k+1} - v_{bs,t_{n+1}}^k) \end{aligned} \quad (10.34a)$$

for the forward mode operation ( $v_{ds} = V_{ds} \geq 0$ ) and

$$\begin{aligned} I_{ch\_linearized,t_{n+1}}^{k+1} \approx I_{ch,t_{n+1}}^k &- \frac{\partial I_{ch}}{\partial V_d} \Big|_{t_{n+1}}^k \cdot (v_{ds,t_{n+1}}^{k+1} - v_{ds,t_{n+1}}^k) \\ &+ \frac{\partial I_{ch}}{\partial V_g} \Big|_{t_{n+1}}^k \cdot (v_{gd,t_{n+1}}^{k+1} - v_{gd,t_{n+1}}^k) + \frac{\partial I_{ch}}{\partial V_b} \Big|_{t_{n+1}}^k \cdot (v_{bd,t_{n+1}}^{k+1} - v_{bd,t_{n+1}}^k) \end{aligned} \quad (10.34b)$$

for the reverse mode ( $v_{ds} = -V_{ds} < 0$ ). The voltage derivatives of Eqs. (10.34a) and (10.34b) are known for both the forward and reverse modes

$$\left\{ \begin{array}{l} \frac{\partial I_{ch}}{\partial V_d} \Big|_{t_{n+1}}^k = G_{ds} \\ \frac{\partial I_{ch}}{\partial V_g} \Big|_{t_{n+1}}^k = G_m \\ \frac{\partial I_{ch}}{\partial V_b} \Big|_{t_{n+1}}^k = G_{mbs} \end{array} \right. \quad (10.35)$$

All the terms given in the above three equations are already known assuming that the  $(k+1)^{th}$  solution of the circuit matrix at the time instant  $t_{n+1}$  is available. It should be noted that  $I_{ch,t_{n+1}}^k$  is computed exactly from the (non-linear)  $I_{ch}$  model equation itself, not from a linear projection approximation.

The same implementation of Eq. (10.33) through Eq. (10.35) needs to be carried out for all other DC current model components of a device such as those of the junction diode, impact ionization, GIDL, and gate direct tunneling currents. The same convergence checking is also required for the current through the source and drain diffusion resistances (bias independent) and the LDD resistance (bias dependent) when they are present between the internal and external source/drain nodes. This requirement also holds for the gate resistance (if RGATEMOD is set to be non zero and the gate direct tunneling current component is turned on). Resistances are usually treated as (voltage-controlled) current sources in SPICE modeling.

Only when the criteria for all these components are fulfilled, can a convergence be declared for that BSIM4 device at the time point  $t_{n+1}$ . Furthermore, convergence can then be declared for the entire circuit when all the devices meet the convergence criteria simultaneously.

A rigorous implementation of convergence checking would require the checking of every individual DC current component. However, some current components such as the impact ionization, GIDL, gate-to-drain direct tunneling, or junction diode DC currents are often insignificant in magnitude in the triode and saturation operation regions as compared to that of the channel current. In this case, one can exempt these small currents when applying convergence checking. Furthermore, in order to improve the performance of the convergence checking, one can consider doing so for the terminal lump sum of all components rather than each individual component.

Convergence checking on the nodal/terminal voltages is also required in order to avoid potential false convergence that may occur, for instance, in the saturation bias regime. In this region, a significant voltage change in the drain-source voltage may induce a change in  $I_{ch}$  that is still deemed to meet the current convergence criteria. In order to amend this shortcoming, perform the voltage checking by following what is similar to the voltage bypassing criterion given in Eq. (10.29a):

$$\left| v_{ds,t_{n+1}}^{k+1} - v_{ds,t_{n+1}}^k \right| < \left[ \text{reltol} \cdot \max\left(\left| v_{ds,t_{n+1}}^{k+1} \right|, \left| v_{ds,t_{n+1}}^k \right| \right) + \text{volttol} \right] \quad (10.36)$$

Convergence checking for other terminal voltages can be done similarly.

## 10.7 Chapter Summary

This chapter presented and discussed the theory, methodology and techniques of implementing a semiconductor device compact model into a SPICE simulator for accurate, efficient and robust IC simulation. The presented implementation issues include numerical integration, time discretization, equation linearization, model stamping/loading into circuit

matrices, model evaluation bypassing, as well as simulation convergence checking. The example of BSIM4 transient NQS model implementation is employed throughout this chapter for illustration purpose. The author would like to remark that implementation using a computer language such as C enables better customization and optimization of SPICE device model than behavioral programming languages such as Verilog-A. Future improvements in behavioral language compliers are needed to narrow the difference.

## References

- [1] L. W. Nagel, "SPICE2: A Computer Program to Simulate Semiconductor Circuits," Electronics Research Laboratory, College of Engineering, University of California, Berkeley, Memo No. UCB/ERL-M520, May 1975.
- [2] Thomas L. Quarles, "The SPICE3 Implementation Guide," Electronics Research Laboratory, College of Engineering, University of California, Berkeley, Memo No. UCB/ERL-M89/44, April 24, 1989.
- [3] Thomas L. Quarles, "Analysis of Performance and Convergence Issues for Circuit Simulation," Electronics Research Laboratory, College of Engineering, University of California, Berkeley, Memo No. UCB/ERL-M89/42, April 1989.
- [4] Thomas L. Quarles, "Adding Devices to SPICE3," Electronics Research Laboratory, College of Engineering, University of California, Berkeley, Memo No. UCB/ERL-M89/45, April 24, 1989.
- [5] V. Litovski and M. Zwolinski, "VLSI Circuit Simulation and Optimization," Chapman & Hall, 1997.
- [6] Andrew Yang, "Computation Methods for Circuit Analysis and Simulation," University of Washington, EE 537 course notes, 1992.

# **Chapter 11**

## **Multi-Gate Transistor Model**

### **11.1 Introduction and Chapter Objectives**

BSIM4 has served the MOSFET compact modeling well. However, the traditional *single-gate, planar* CMOS transistor structure is undergoing a replacement/transition to multiple gates. The scaling of planar CMOS to 14nm is increasingly difficult owing to subthreshold leakage, sensitivity to gate length variations, random doping fluctuations, and ionized impurity scattering [1-2]. These challenges are also correlated among themselves and are complex to solve with the existing planar MOSFET device structure. Multi-gate FETs such as FinFETs provide a better solution and can extend the scaling into the 14nm node and later below 10nm [3-4]. The strong electrostatic control over the channel originating from multiple gates and ultra-thin body overwhelms the capacitive coupling between source/drain and the channel. Therefore the multi-gate transistor can be scaled to a much smaller gate length than bulk planar CMOS for a given dielectric thickness. Many efforts are underway for readying the production of multi-gate FET ICs and in developing production-worthy compact models. At the same time, circuit designers have embarked on multi-gate FET circuit designs.

This chapter will present and discuss briefly the BSIM-CMG model. CMG stands for common multiple gates. BSIM-CMG is not part of BSIM4. It has a different set of I-V and C-V model formulations and parameters as well as a different nomenclature from those of BSIM4.

## 11.2 Advantages of FinFETs Over Planar CMOS

What makes a field-effect transistor (FET) different from a resistor is that the gate, not the drain, controls the conduction of its channel. The gate exerts its control through capacitive (electrostatic) coupling to the channel. When shrinking the MOSFET gate to a smaller size, the drain is pulled closer to the middle of the channel and the source. That increases the capacitive coupling between the drain and the channel and between the drain and the source as well (see Fig. 11.1). When the channel becomes shorter, the drain gets increasingly larger control of the channel. As a consequence, the drain-induce barrier lowering (DIBL) and  $V_{th}$  roll-off becomes worse and the transistor becomes leakier. This has been the problem for the planar CMOS. Heavy channel doping has been the ad-hoc remedy in the past, which produces significant side effects such as low carrier mobility owing to carrier scattering, and large channel depletion and junction capacitances. Threshold voltage  $V_{th}$  variations caused by random dopant fluctuations are also a significant side effect.

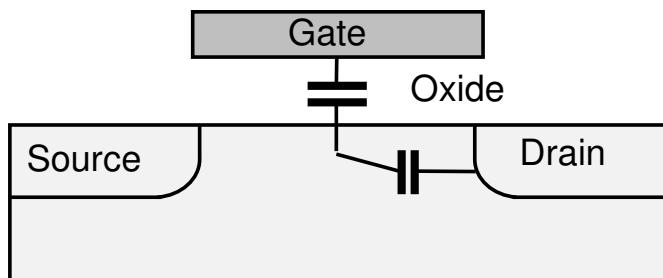


Fig. 11.1 When the gate length is reduced, the drain-to-channel capacitance becomes significant relative to the gate capacitance and the gate loses its absolute ability to shut off the channel leakage current.

The gates of multi-gate MOSFETs such as FinFETs [5] control the channel from two or more sides (see Fig. 11.2). That increases the gate control and permits the gate length to be further scaled down. The FinFET device structure holds the record for the smallest demonstrated gate length of 3nm. It has several other important merits that improve the CMOS technology scaling ability. It can reduce the sensitivity of  $V_{th}$  to the gate length, thus partially alleviating the serious manufacturing variation problem. It has a steeper subthreshold slope of the channel current versus the gate voltage (i.e., a smaller sub-threshold swing  $S$ -

factor in mV/decade). The channel carrier mobility is higher because of lower required channel doping and hence reduced impurity scattering. In particular, it can also reduce the surface roughness scattering because of lower transverse electric fields (perpendicular to the gate stack).

A common multi-gate MOSFET transistor can be built on a bulk or an SOI silicon substrate (refer to Fig. 11.2). All the gates are physically and electrically connected. BSIM-CMG supports both of these device technologies.

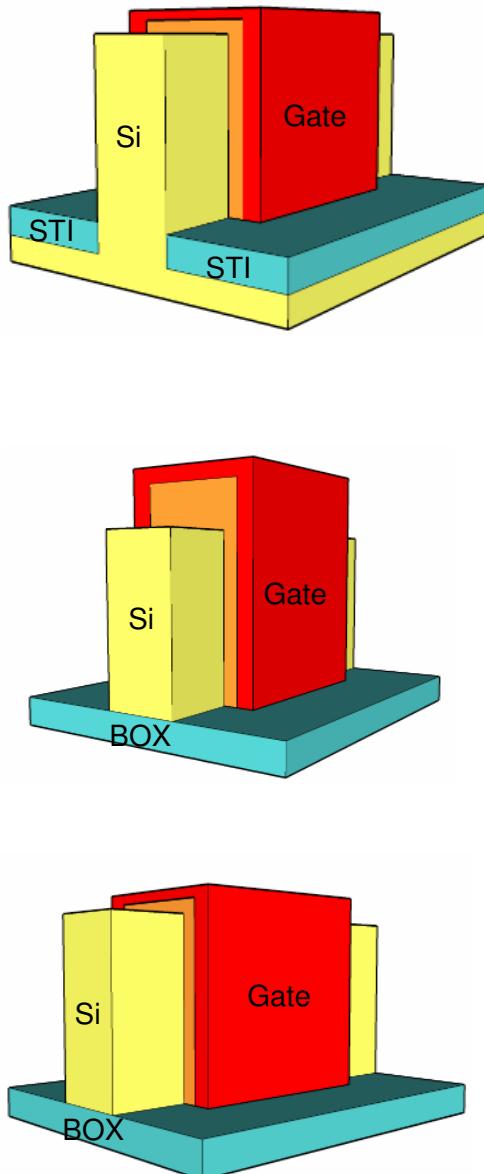


Fig. 11.2 Common multi-gate MOSFET device structures: A triple-gate FinFET on bulk Si (on the top); double-gate FinFET on SOI (in the middle); and triple-gate FinFET on SOI (at the bottom). STI stands for shallow-trench isolation and BOX for buried oxide of an SOI substrate.

## 11.3 BSIM-CMG

BSIM-CMG is a surface potential based model. In this approach, all the device terminal currents, conductances, charges and capacitances are expressed as functions of the surface potentials. These model equations are all  $C_\infty$ -continuous, which means that any-order current and charge derivatives are continuous with respect to device terminal voltages. The calculation of the surface potentials is the starting point of this model. BSIM-CMG considers any channel doping levels. This is in contrast to other models in which the channel is assumed to be undoped, intrinsic silicon [6-7]. BSIM-CMG models all the modern device physical effects. They include short-channel effects, mobility degradation, velocity saturation, quantum mechanical effects, poly-silicon gate depletion, and gate direct-tunneling. The concept and the model formulations for many of these physical effects originate from BSIM4, which was presented and discussed in the preceding chapters of this book.

In the following, the core or basic BSIM-CMG formulations [8] is presented and discussed first. This is followed by the modeling of the real device physical effects. The model comparison and validation against measured data [9] are then presented. The SPICE implementation of this model follows the same methodology detailed in Chapter 10. This model was released in Verilog-A rather than the C language.

### 11.3.1 *The Core Model: Surface Potential Modeling*

A long-channel symmetric double-gate (DG) MOSFET is employed to develop the core BSIM-CMG model. Fig. 11.3 shows the schematic of the symmetric and common-gate DG-FET device structure. The convention for the axes and the symbols used are defined in this figure. The electric potential in the body is obtained by solving Poisson's equation. For a long-channel transistor, the gradual channel approximation (GCA) is applied, which states that the horizontal electric field is much smaller than the vertical. This leads to a quasi 2-D Poisson's equation (in the vertical dimension) to solve. It includes both the inversion carriers and the bulk charges in the silicon body. This equation is written as

$$\frac{\partial^2 \psi(x, y)}{\partial x^2} = \frac{q n_i}{\epsilon_{Si}} \cdot e^{\frac{q(\psi(x, y) - \phi_B - V_{ch}(y))}{kT}} + \frac{q N_A}{\epsilon_{Si}} \quad (11.1)$$

where  $\psi(x, y)$  is the electric potential in the body,  $V_{ch}(y)$  is the channel potential ( $V_{ch}(0) = 0$ ) and ( $V_{ch}(L) = V_{ds}$ ),  $N_A$  is the body doping concentration, and the familiar Fermi potential is given by

$$\phi_B = \frac{kT}{q} \cdot \ln \left( \frac{N_A}{n_i} \right) \quad (11.2)$$

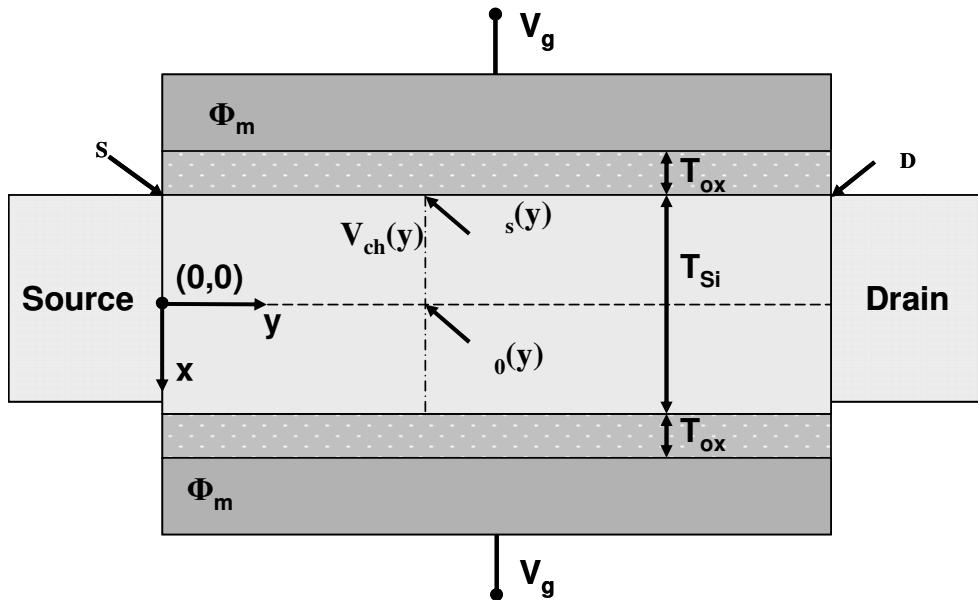


Fig. 11.3 Schematic of a symmetric common-gate DG-MOSFET device structure.  $\Phi_m$  is the gate work function.

For a lightly-doped Si channel, the doping can be neglected. Eq. (11.1) is solved by applying Gauss' law as the boundary condition. For a moderate to heavy body doping, the doping concentration cannot be neglected and that makes the calculation of the surface potential a complex task because Eq. (11.1) cannot be solved in a closed form. Instead, a perturbation approach is used [8]. The body potential now becomes

$$\psi(x, y) = \psi_1(x, y) + \psi_2(x, y) \quad (11.3)$$

Here the first term,  $\psi_1(x,y)$ , is the potential attributed to the inversion carrier term of Eq. (11.1). The second term,  $\psi_2(x,y)$ , is the perturbation of the potential introduced by the body doping. It is known that the body region can be either fully depleted or partially depleted depending on the gate voltage  $V_{gs}$ , the body doping  $N_A$  and the silicon channel thickness  $T_{Si}$ . The perturbation method yields the body potential in both the full-depletion and partial-depletion regimes.

In the full-depletion regime, the inversion carriers are spread through the entire body. The contribution of inversion carriers to the potential,  $\psi_1(x,y)$ , is calculated by dropping the term associated with the body dopants in Eq. (11.1).

$$\frac{\partial^2 \psi_1(x,y)}{\partial x^2} = \frac{qn_i}{\epsilon_{Si}} \cdot e^{\frac{q(\psi_1(x,y)-\phi_B-V_{ch}(y))}{kT}} \quad (11.4)$$

which can then be integrated to obtain a closed-form  $\psi_1(x,y)$ .

$$\psi_1(x,y) = \psi_0(y) - \frac{2kT}{q} \cdot \ln \left( \cos \left( \sqrt{\frac{q^2}{2\epsilon_{Si}kT} \frac{n_i^2}{N_A}} e^{\frac{q(\psi_0(y)-V_{ch}(y))}{kT}} \cdot \frac{x}{2} \right) \right) \quad (11.5)$$

$\psi_0(y)$  is the potential at the center of the body as shown in Fig. 11.3. Substituting Eqs. (11.3) and (11.4) into Eq. (11.1) yields a second-order partial differential equation of  $\psi_2(x,y)$

$$\frac{\partial^2 \psi_2(x,y)}{\partial x^2} = \frac{qn_i}{\epsilon_{Si}} \cdot e^{\frac{q(\psi_1(x,y)-\phi_B-V_{ch}(y))}{kT}} \cdot \left( e^{\frac{q\psi_2(x,y)}{kT}} - 1 \right) + \frac{qN_A}{\epsilon_{Si}} \quad (11.6)$$

It follows that the contribution by the body dopant charge to the potential and electric field at the mid-plane ( $x = 0$ ) is such that the following boundary condition holds for solving for  $\psi_2(x,y)$

$$\psi_2(x=0, y) = 0 \quad \text{and} \quad \left. \frac{d\psi_2}{dx} \right|_{(x=0, y)} = 0 \quad (11.7)$$

With Eqs. (11.5) and (11.7), Eq. (11.6) can be solved to obtain  $\psi_2(x,y)$

$$\psi_2(x, y) = \frac{2qn_i}{\epsilon_{Si}} \cdot \frac{e^{\frac{q\phi_B}{kT}}}{a(x, y)} \cdot \left( \frac{e^{\frac{x\sqrt{a(x, y)}}{2}} - 1}{2e^{\frac{x\sqrt{a(x, y)}}{2}}} \right)^2 \quad (11.8)$$

here

$$a(x, y) = \frac{q^2 n_i}{\epsilon_{Si} kT} \cdot e^{\frac{q(\psi_1(x, y) - V_{ch}(y) - \phi_B)}{kT}} \quad (11.9)$$

The potential at  $y$  along the surface is the sum of  $\psi_1(x, y)$  and  $\psi_2(x, y)$ , both evaluated at the surface

$$\psi_s(y) = \psi_1\left(\frac{T_{Si}}{2}, y\right) + \psi_2\left(\frac{T_{Si}}{2}, y\right) \quad (11.10)$$

Note that  $\psi_2(x, y)$  is a function of  $\psi_1(x, y)$  (see Eq. (11.8)) and  $\psi_1(x, y)$  is a function of  $\psi_0(y)$  (refer to Eq. (11.5)). As a result,  $\psi_s(y)$  is a function of only  $\psi_0(y)$ , the potential at the center of the body.

The electric field at the surface is obtained by integrating Eq. (11.1). Applying Gauss' law at the surface gives

$$V_g = V_{fb} + \psi_s(y) + \frac{\epsilon_{Si}}{C_{ox}} \cdot \sqrt{\frac{2qn_i}{\epsilon_{Si}} \cdot \left( \frac{e^{\frac{q\psi_s(y)}{kT}} - e^{\frac{q\psi_0(y)}{kT}}}{q/kT} \cdot e^{\frac{-q(V_{ch}(y) + \phi_B)}{kT}} + e^{\frac{q\phi_B}{kT}} \cdot (\psi_s(y) - \psi_0(y)) \right)} \quad (11.11)$$

In the partial-depletion regime, the depletion width (thickness) is bias dependent. At the edge of the depletion width,  $x_{dep}$ , the electric potential is zero and hence  $\psi_1|(x=x_{dep}) = 0$ . The surface potential can now be re-derived in the same method as that for the full-depletion case. The contribution to the surface potential by the inversion carriers is

$$\psi_1\left(\frac{T_{Si}}{2}, y\right) = -\frac{2kT}{q} \cdot \ln \left( \cos \left( \sqrt{\frac{q^2}{2\epsilon_{Si}kT} \frac{n_i^2}{N_A} e^{\frac{-qV_{ch}(y)}{kT}}} \cdot \frac{x_{dep}}{2} \right) \right) \quad (11.12)$$

and the perturbation term due to body dopant charge is

$$\psi_2\left(\frac{T_{Si}}{2}, y\right) = \frac{2qn_i}{\epsilon_{Si}} \cdot \frac{e^{\frac{q\phi_B}{kT}}}{a} \cdot \left( \frac{e^{x_{dep}\frac{\sqrt{a}}{2}} - 1}{2e^{x_{dep}\frac{\sqrt{a}}{2}}} \right)^2 \quad (11.13)$$

here

$$a = \frac{q^2 n_i}{\epsilon_{Si} kT} \cdot e^{\frac{q(\psi_1(\frac{T_{Si}}{2}, y) - V_{ch}(y) - \phi_B)}{kT}} \quad (11.14)$$

In the case of partial depletion, Gauss' law gives

$$V_g = V_{fb} + \psi_s(y) + \frac{\epsilon_{Si}}{C_{ox}} \cdot \sqrt{\frac{2qn_i}{\epsilon_{Si}} \cdot \left( \frac{e^{\frac{q\psi_s(y)}{kT}} - 1 \cdot e^{\frac{-q(V_{ch}(y) + \phi_B)}{kT}}}{q/kT} + e^{\frac{q\phi_B}{kT}} \cdot \psi_s(y) \right)} \quad (11.15)$$

The only unknown variable in the surface potential solution for the partial-depletion regime is  $x_{dep}$ , which is obtained by solving Eq. (11.15) numerically. Once  $x_{dep}$  is determined, the surface potential is calculated from Eq. (11.10) with the assistance of Eqs. (11.12) through (11.14).

In order to obtain continuous expressions for terminal currents and charges, it is necessary to capture the transition between the full-depletion and partial-depletion operation modes smoothly (this transition region is usually termed as the dynamic depletion). Also, in order to improve the model evaluation efficiency, a simplified expression for  $\psi_2(x,y)$  is derived (denoted  $\psi_{pert}$ ). This new expression is continuous in the transition between depletion regimes. Therefore, a single continuous equation can be obtained for the surface potential by defining

$$\beta = \frac{T_{Si}}{2} \sqrt{\frac{q^2}{2\epsilon_{Si} kT} \frac{n_i^2}{N_A} e^{\frac{q(\psi_0(y) - V_{ch}(y))}{kT}}} \quad (11.16)$$

The unified surface potential  $\psi_s$  equation used in the core model for BSIM-CMG is given by

$$f_0(\beta) = \ln(\beta) - \ln(\cos(\beta)) - \frac{V_g - V_{fb} - V_{ch}}{2\frac{kT}{q}} + \ln\left(\frac{2}{T_{Si}} \sqrt{\frac{2\epsilon_{Si}kTN_A}{qn_i^2}}\right) + \frac{2\epsilon_{Si}}{T_{Si}C_{ox}} \cdot \sqrt{\beta^2 \left( \frac{e^{\frac{q\psi_{pert}}{kT}}}{\cos^2(\beta)} - 1 \right) + \frac{\psi_{pert}}{\left(\frac{kT}{q}\right)^2} \left( \psi_{pert} - 2\frac{kT}{q} \ln(\cos(\beta)) \right)} = 0 \quad (11.17)$$

$\beta$  is the only unknown variable. In BSIM-CMG, the transcendental  $\psi_s$  equation (Eq. (11.17)) is solved for  $\beta$  using an analytical approximation instead of an iterative method to make the model numerically robust and efficient. This is another significant advantage of the simplified perturbation term  $\psi_{pert}$ .

The surface potential is a function of  $\beta$  and is given by

$$\psi_s = 2\frac{kT}{q} \left( \ln(\beta) - \ln(\cos(\beta)) + \ln\left(\frac{2}{T_{Si}} \sqrt{\frac{2\epsilon_{Si}kTN_A}{qn_i^2}}\right) \right) + \psi_{pert} \quad (11.18)$$

The surface potential at source ( $\psi_s$ ) is obtained by solving Eq. (11.17) at the source end, i.e.,  $V_{ch} = 0$ . Similarly, the surface potential at the drain end ( $\psi_D$ ) is calculated by solving Eq. (11.17) with  $V_{ch} = V_{ds}$ .

Fig. 11.4(a) compares the accuracy of the surface potential model against TCAD simulations. All the TCAD simulations use the gate materials with mid-gap work functions and assume constant carrier mobility. The surface potential is calculated as a function of the gate voltage for a wide range of body dopings ranging from a light doping of  $1 \times 10^{15} \text{ cm}^{-3}$  to a heavy doping of  $5 \times 10^{18} \text{ cm}^{-3}$ . The transition from the partial-depletion regime to the full-depletion regime with increasing gate voltage is clearly visible in the heavily doped DG-FET. The absolute error in the analytical approximation of  $\psi_s$  is within a few nano-volts as shown in Fig. 11.4(b), which is usually a challenge for an analytical surface potential approach to meet.

Eq. (11.17) yields the surface potential for both the light and heavy body dopings as shown in Fig. 11.4(a). However, the inclusion of bulk dopant charges in the analysis of a lightly doped body is redundant and it would lead to significant overhead in model equation evaluation. Therefore, for a lightly-doped DG-FET, Eq. (11.17) is now simplified to

$$\ln \beta - \ln(\cos \beta) - \frac{V_g - V_{fb} - V_{ch}}{kT} + \ln\left(\frac{2}{T_{Si}} \sqrt{\frac{2\epsilon_{Si} k T N_A}{qn_i^2}}\right) + \frac{2\epsilon_{Si}}{T_{Si} C_{ox}} \beta \tan \beta = 0 \quad (11.19)$$

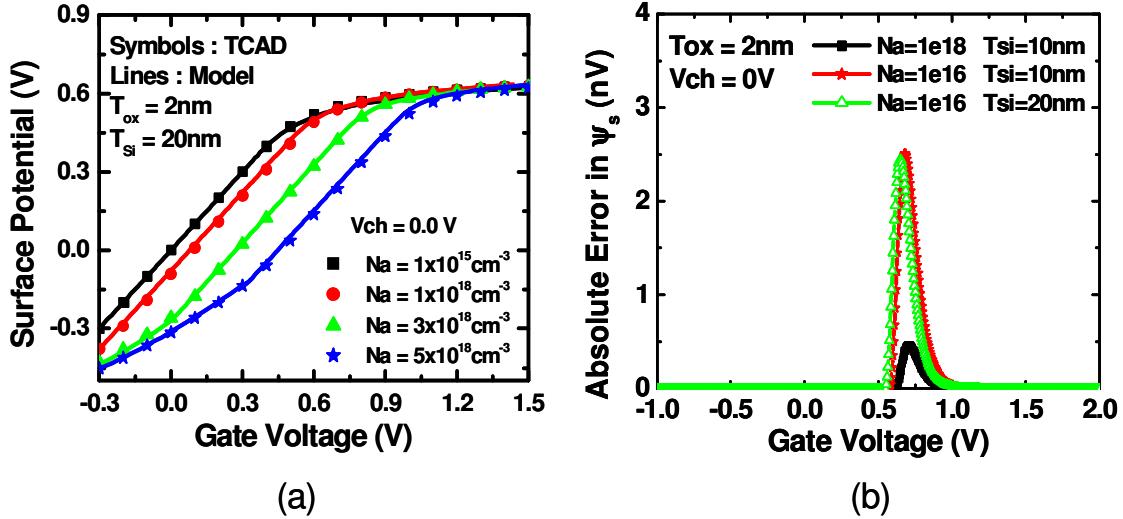


Fig. 11.4 Comparison of the surface potential model against TCAD for several body dopings in (a). Absolute errors of the analytical model of  $\psi_s$  relative to numerical simulations in (b).

### 11.3.2 Channel I-V Model

The BSIM-CMG channel DC I-V model is developed using the drift-diffusion theory without any charge-sheet approximation [10]. The current of a DG-FET is given by

$$I_d = 2 \cdot \mu \cdot W \cdot Q_{inv}(y) \frac{dV_{ch}}{dy} \quad (11.20)$$

where  $\mu$  is the carrier mobility,  $W$  is the channel width,  $Q_{inv}(y)$  is the inversion charge density and the factor 2 accounts for the front and back channel currents in a symmetric common-gate DG-FET.

The inversion charge density  $Q_{inv}$  is the difference between the densities of the total charge in the body and the bulk charge:

$$Q_{inv}(y) = Q_{total}(y) - Q_{bulk}(y) \quad (11.21)$$

The bulk charge  $Q_{bulk}$  can be obtained from the potential perturbation  $\psi_{pert}$

$$Q_{bulk} = \sqrt{2q\epsilon_{Si}N_A\psi_{pert}} \quad (11.22)$$

The total charge in the body is found from Gauss's law

$$Q_{total}(y) = C_{ox} \cdot (V_g - V_{fb} - \psi_s(y)) \quad (11.23)$$

The drain current equation can be obtained by integrating Eq. (11.20) from the source to the drain by substituting Eqs. (11.21) through (11.23). The integration is analogous to that presented in Chapter 3. The resulting drain current has two terms, associated with the source and the drain:

$$I_d = 2 \cdot \mu \cdot \frac{W}{L} \cdot (f(\psi_s) - f(\psi_d)) \quad (11.24)$$

Here the function  $f(\psi_s)$  is

$$f(\psi_s) = \frac{Q_{inv}^2}{2C_{ox}} + 2\frac{kT}{q}Q_{inv} - \frac{kT}{q} \left( 5\frac{\epsilon_{Si}kT}{qT_{Si}} + Q_{bulk} \right) \cdot \ln \left( 5\frac{\epsilon_{Si}kT}{qT_{Si}} + Q_{bulk} + Q_{inv} \right) \quad (11.25)$$

$f(\psi_d)$  has a similar form. Eqs. (11.24) and (11.25) give the drain current model for a symmetric DG-FET. Please refer to [8] and [9] for the details of the derivation.

The accuracy of the I-V model is demonstrated against TCAD simulations. Fig. 11.5 shows the comparison for a heavily doped DG-FET ( $N_A = 3 \times 10^{18} \text{ cm}^{-3}$ ). BSIM-CMG agrees over all regions of operation. Fig. 11.6 shows the accuracy comparison over a wide range of body doping. The model can predict the drain current in both full-depletion and partial-depletion regimes.

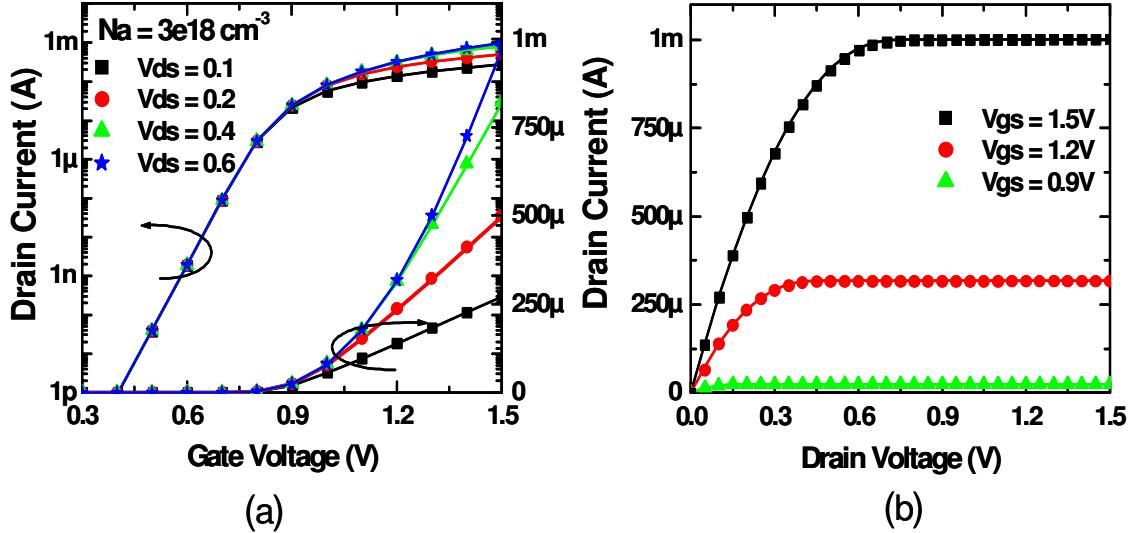


Fig. 11.5 (a)  $I_d$ - $V_g$  and (b)  $I_d$ - $V_d$  characteristics from the BSIM-CMG I-V model (lines) and TCAD simulations (symbols). Device process parameters are  $N_A = 3 \times 10^{18} \text{ cm}^{-3}$ ,  $T_{Si} = 20\text{nm}$ , and  $T_{ox} = 2\text{nm}$ .

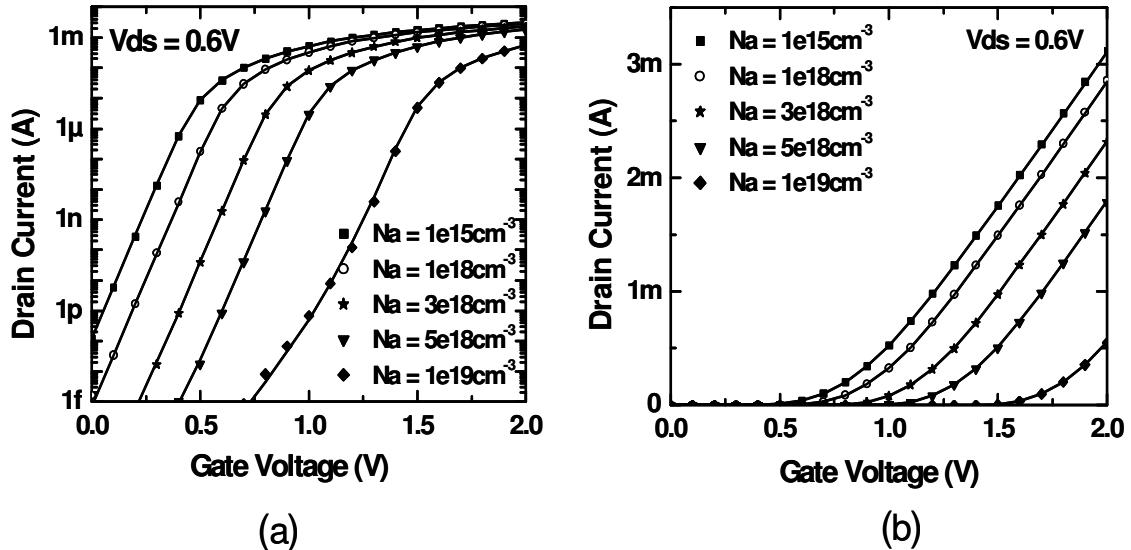


Fig. 11.6  $I_d$ - $V_g$  characteristics for different body doping concentrations calculated from the I-V model (lines) and TCAD (symbols). Device process parameters are  $T_{Si} = 20\text{nm}$  and  $T_{ox} = 2\text{nm}$ .

A lightly-doped DG-FET with a thin body is subject to the volume inversion in the sub-threshold region, which results from low bulk charge densities in the body and the limited number of mobile carriers. Therefore, there is negligible potential drop between the surface and the center of the body. Fig. 11.7(a) confirms a virtually flat potential profile in this case from both BSIM-CMG and TCAD simulations. Furthermore,

the potential in the body has a weak dependence on body thickness. Any small increase in the gate voltage in the sub-threshold region increases the potential throughout the entire body volume, which causes the inversion in the entire body. This phenomenon is called bulk inversion or volume inversion. Note that since the electric potential is virtually independent of body thickness, the amount of inversion carriers in the body is linearly proportional to the body thickness. As a result, the sub-threshold current in a lightly-doped DG-FET is a linear function of the body thickness. The I-V model is able to predict this trend correctly as shown in Fig. 11.7(b), where the  $I_d$  for a 20nm thick body is ~4 times as large as the current of a device with a 5nm body thickness in the sub-threshold regime.

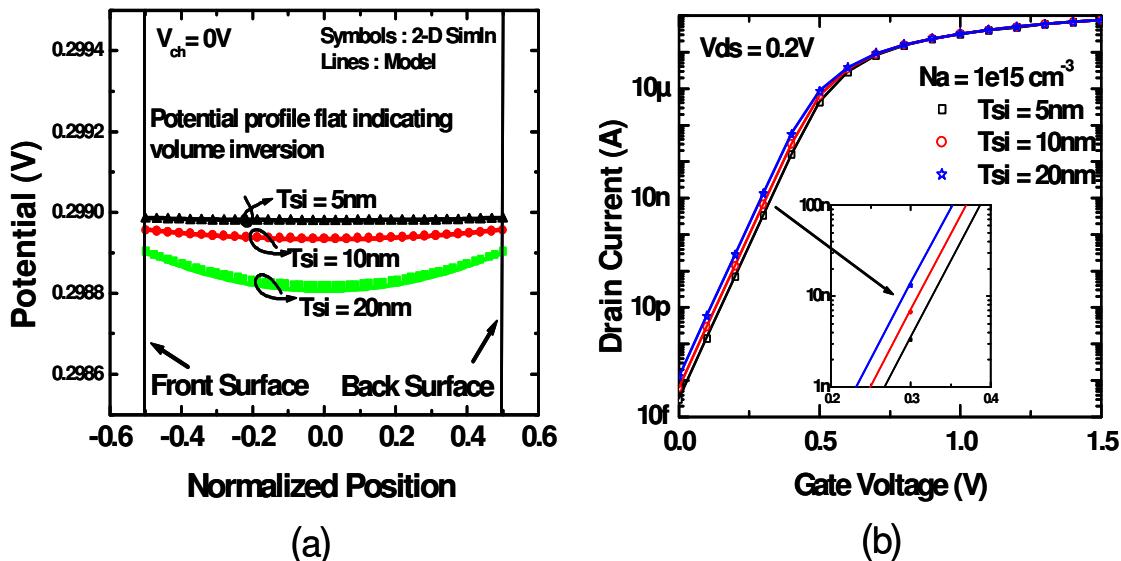


Fig. 11.7 Volume inversion in lightly-doped DG-FETs. (a) Potential distribution in the body. (b) Sub-threshold  $I_d$ - $V_g$  characteristics for different body thicknesses showing volume inversion. In both (a) and (b) lines represent model and symbols represent TCAD. Device process parameters used are  $N_A = 1 \times 10^{15} \text{ cm}^{-3}$ , and  $T_{ox} = 2 \text{ nm}$ .

### 11.3.3 Charge and Capacitance Models

The BSIM-CMG intrinsic charge and capacitance model are derived in the following. According to charge neutrality, the sum of the charge on the top and bottom gates is equal to the total charge in the body. The total charge is computed by integrating the charge along the channel. Since the two gates are electrically connected, the gate charge is given by

$$Q_s = 2WC_{ox} \int_0^L (V_g - V_{fb} - \psi_s(y)) \cdot dy \quad (11.26)$$

The channel charge needs to be partitioned into the source and drain nodes (refer to the 40/60 charge partition scheme in Chapter 5 and the Ward-Dutton partition model in [11]). The charge into the source  $Q_s$  is

$$Q_s = -2WC_{ox} \int_0^L \left(1 - \frac{y}{L}\right) \cdot \left(V_g - V_{fb} - \psi_s(y) - \frac{Q_{bulk}}{C_{ox}}\right) \cdot dy \quad (11.27)$$

The charge into the drain is given by

$$Q_d = -2WC_{ox} \int_0^L \frac{y}{L} \cdot \left(V_g - V_{fb} - \psi_s(y) - \frac{Q_{bulk}}{C_{ox}}\right) \cdot dy \quad (11.28)$$

The surface potential can be obtained from the current continuity equation

$$I_d(L) = I_d(y) \quad \text{where } 0 \leq y \leq L \quad (11.29)$$

However, the expression of the drain current in Eqs. (11.24) and (11.25) is too complex to consider for the charge model in practice, for the sake of computational cost. Thus, a simplified version of I-V model is utilized to obtain the surface potential

$$I_d(y) = 2 \cdot \mu \cdot \frac{W}{y} \cdot (g(\psi_s) - g(\psi_s(y))) \quad (11.30)$$

where the function  $g(\psi_s(y))$  is defined as

$$g(\psi_s) = \frac{Q_{inv}^2}{2C_{ox}} + 2 \frac{kT}{q} Q_{inv} \quad (11.31)$$

Note Eqs. (11.30) and (11.31) retain good accuracy in strong inversion but it can overestimate the drain current in sub-threshold region.  $\psi_s(y)$  is related to  $\psi_S$  and  $\psi_D$  by

$$(11.32)$$

$$\frac{y}{L} \cdot (B - \psi_s - \psi_d)(\psi_d - \psi_s) = (B - \psi_s - \psi_s(y))(\psi_s(y) - \psi_s)$$

here

$$B = 2 \left( V_g - V_{fb} - \frac{Q_{bulk}}{C_{ox}} + \frac{2kT}{q} \right) \quad (11.33)$$

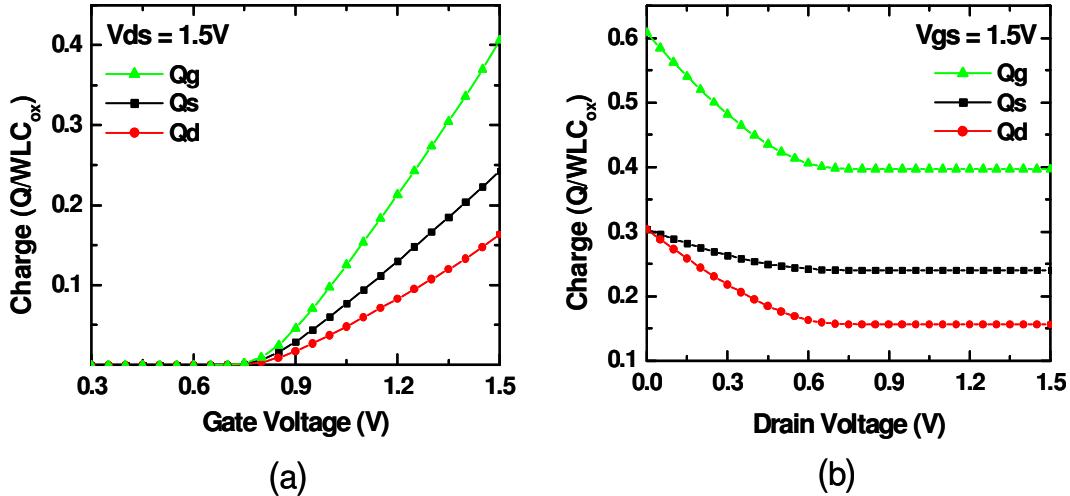


Fig. 11.8 Device terminal charges using Eq. (11.33) as a function of (a)  $V_{gs}$  and (b)  $V_{ds}$ . Device process parameters:  $N_A = 3 \times 10^{18} \text{ cm}^{-3}$ ,  $T_{ox} = 2 \text{ nm}$ , and  $T_{Si} = 20 \text{ nm}$ .

The terminal charges are obtained by substituting  $\psi_s(y)$  into Eqs. (11.26) through (11.28) and integrating from source to drain.

$$Q_g = 2WLC_{ox} \left( V_{gs} - V_{fb} - \frac{\psi_s + \psi_d}{2} + \frac{(\psi_d - \psi_s)^2}{6(B - \psi_d - \psi_s)} \right)$$

$$Q_d = -2WLC_{ox} \left( \frac{V_{gs} - V_{fb} - \frac{Q_{bulk}}{C_{ox}}}{2} - \frac{\psi_s + \psi_d}{4} + \frac{(\psi_d - \psi_s)^2}{60(B - \psi_d - \psi_s)} + \frac{(5B - 4\psi_d - 6\psi_s)(B - 2\psi_d)(\psi_s - \psi_d)}{60(B - \psi_d - \psi_s)^2} \right) \quad (11.34)$$

$$Q_s = -(Q_{fg} + Q_{bg} + Q_{bulk} + Q_d)$$

They are continuous over all regions of operation. Fig. 11.8 plots the terminal charges calculated from Eq. (11.34) as a function of  $V_{ds}$  and  $V_{gs}$ . As can be seen, the ratio of the drain node charge to the source node charge is 40/60 in the saturation region, which agrees to the Ward-Dutton charge partition under the quasi-static condition.

The BSIM-CMG capacitances are derived from Eq. (11.34). They are

$$C_{ij} = \delta_{ij} \cdot \frac{\partial Q_i}{\partial V_j} \quad (11.35)$$

where  $i$  and  $j$  denote the multi-gate FET terminals (refer to Chapter 5 for details). Note that  $C_{ij}$  must satisfy the charge neutrality requirement

$$\sum_i C_{ij} = \sum_j C_{ij} = 0 \quad (11.36)$$

The C-V model is also verified by TCAD simulations. The capacitances are plotted as a function of gate voltage and drain voltage in Figs. 11.9 and 11.10. At  $V_{ds} = 0V$ ,  $C_{sg}=C_{dg}$  and  $C_{gs}=C_{gd}$ . This demonstrates the symmetry of the model.

The surface potential model together with the I-V and C-V model for the DG-FET transistor constitutes the core model for BSIM-CMG.

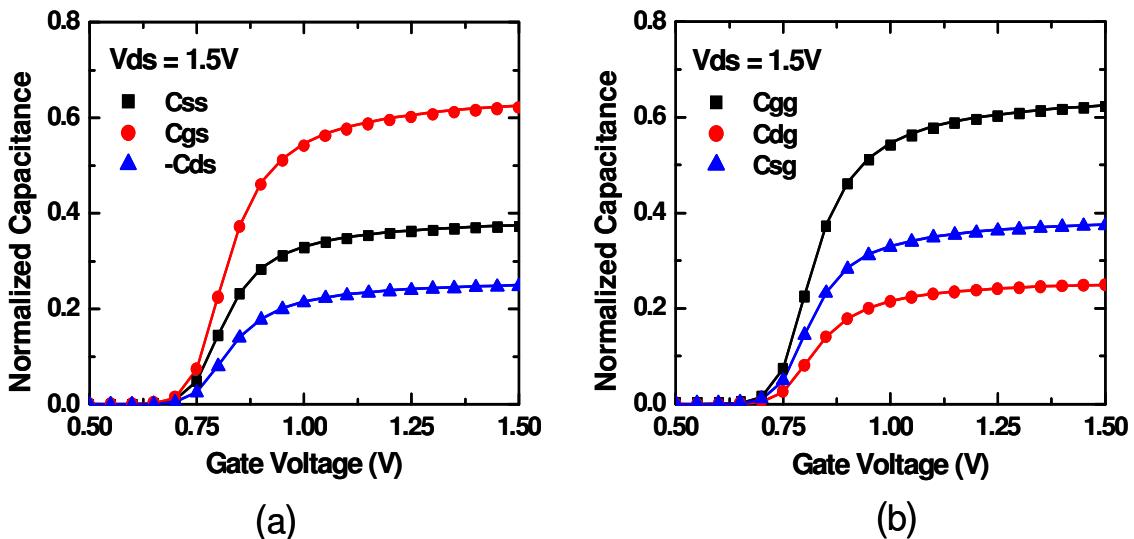


Fig. 11.9 Capacitances (normalized to  $2WLC_{ox}$ ) calculated from the C-V model (lines) and TCAD (symbols) as a function of  $V_{gs}$ . Device process parameters:  $N_A = 3 \times 10^{18} \text{ cm}^{-3}$ ,  $T_{ox} = 2\text{nm}$ , and  $T_{Si} = 20\text{nm}$ .

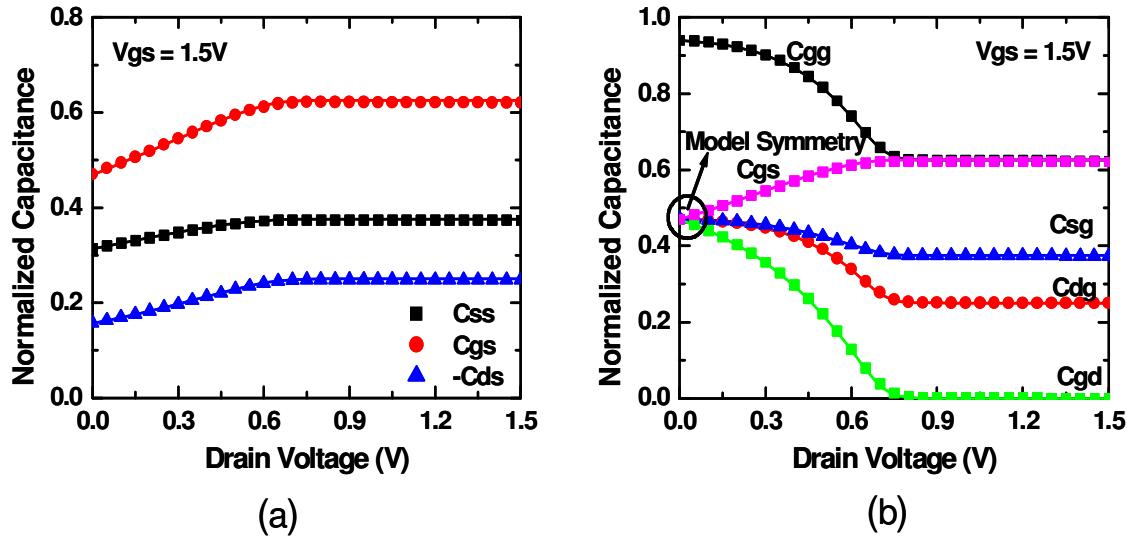


Fig. 11.10 Capacitances (normalized to  $2WLC_{ox}$ ) calculated from the C-V model (lines) and TCAD (symbols) as a function of  $V_{ds}$ . Device process parameters:  $N_A = 3 \times 10^{18} \text{ cm}^{-3}$ ,  $T_{ox} = 2\text{nm}$ , and  $T_{Si} = 20\text{nm}$ .

### 11.3.4 Modeling of Advanced Physical Effects

On top of the core model described in the preceding subsections, BSIM-CMG models numerous physical phenomena that are important for multi-gate MOSFET process technologies [8]. For example, using a thin body leads to significant quantum mechanical effects. Short-channel effects such as drain-induced barrier lowering (DIBL), threshold voltage ( $V_{th}$ ) roll-off and sub-threshold slope degradation are critical as well. In particular, the poly-silicon gate depletion must also be accounted for in the consideration that FinFETs with poly-silicon gates can be a competitive candidate for low-cost high-density memory arrays. Table 11.1 summarizes the physical effects that are modeled by BSIM-CMG. In the following, examples of the model fitting results are illustrated.

Table 11.1 The physical effects modeled in BSIM-CMG

- 
- 1) Quantum mechanical effects
  - 2) Short channel effects
    - a)  $V_{th}$  roll-off
    - b) Drain-induced barrier lowering (DIBL)
    - c) Sub-threshold slope degradation
    - d) Channel length modulation
  - 3) Poly-silicon gate depletion
  - 4) Source and drain series resistances
  - 5) Mobility degradation
  - 6) Velocity saturation
  - 7) Velocity overshoot/Source-end velocity limit
  - 8) Gate-induced drain/source leakage currents (GIDL/GISL)
  - 9) Impact ionization
  - 10) Source and drain junction leakage currents
  - 11) Gate direct-tunneling currents
  - 12) Parasitic capacitances
- 

Figure 11.11 demonstrates the BSIM-CMG channel leakage current on channel lengths of double-gate FETs in comparison with TCAD simulations [8, 9]. Fig. 11.12 illustrates that a triple-gate FET is more effective in suppressing the channel leakage current than a double-gate FET. Fig. 11.13 shows the BSIM-CMG  $V_{th}$  roll-off model for triple-gate FET as verified with TCAD simulations.

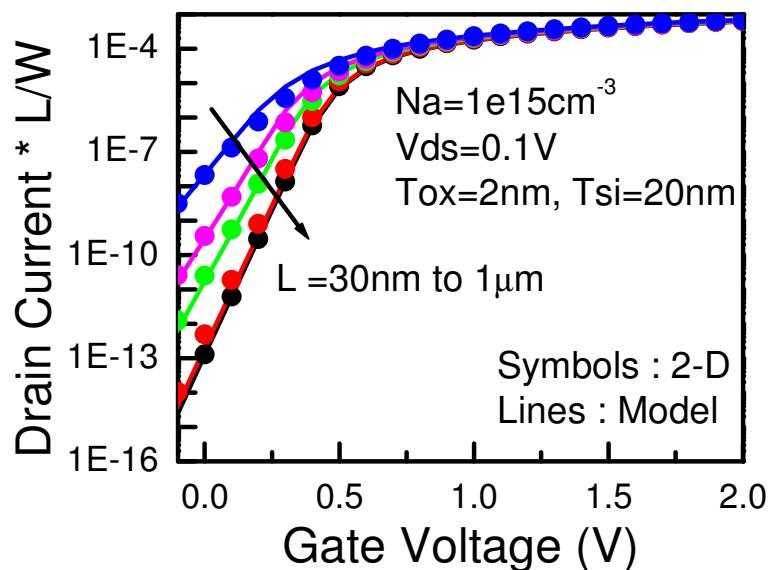


Fig. 11.11 Model-predicted channel-length dependence of  $I_d$ - $V_g$  characteristics against TCAD simulations.

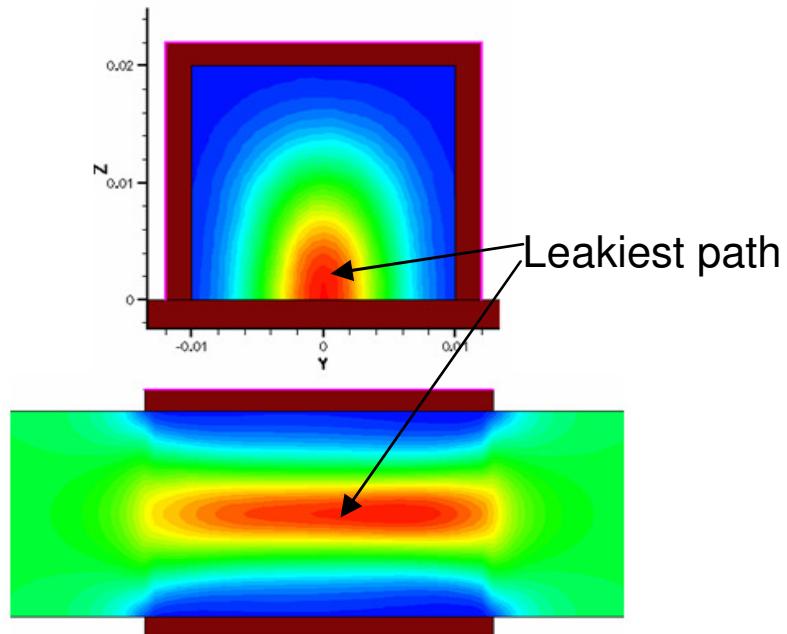


Fig. 11.12 Leakage current path is different in triple-gate FETs due to 3-D effects. The leakiest path is located at the bottom center of the fin.

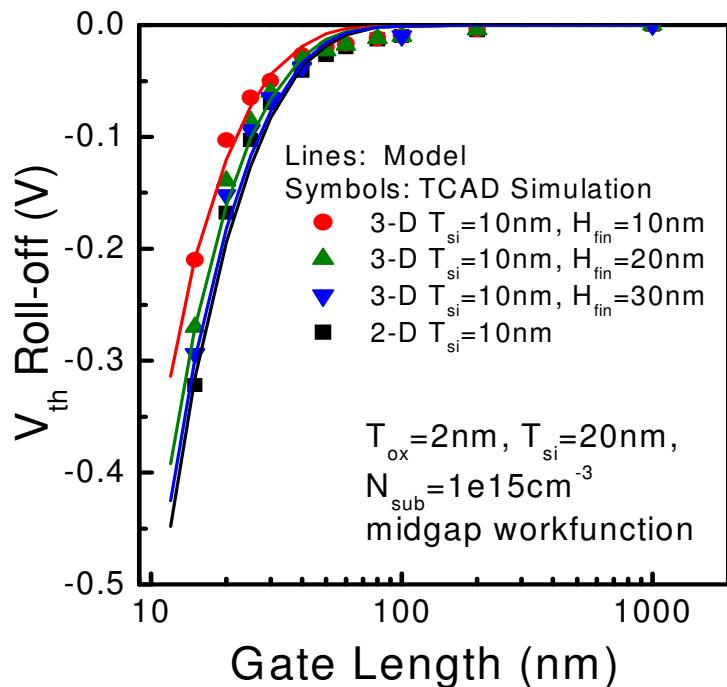


Fig. 11.13  $V_{th}$  roll-off of the BSIM-CMG short-channel effect model and TCAD simulation. A smaller fin height of triple-gate FETs has better  $V_{th}$  roll-off. (Symbols: TCAD, Lines: Model)

## 11.4 Model Validation

In this section, the experimental validation of BSIM-CMG is presented with two different FinFET technologies – SOI FinFETs and bulk FinFETs [9]. The validation is conducted on the modeling of the drain current and its derivatives, the gate trans-conductance ( $g_m$ ) and output conductance ( $g_{ds}$ ), for both long channel and short channel devices.

The SOI FinFETs were fabricated on a lightly doped 60nm thick film with a 2nm  $\text{SiO}_2$  dielectric and a strained TiSiN gate for enhanced channel carrier mobility. Each device has 20 parallel fins, each fin being 22nm thick. Fig. 11.14 shows the model fitting to  $I_d$  and its derivatives for a short-channel  $L_g = 90\text{nm}$  device. Accurate modeling of the physical phenomena such as DIBL, mobility degradation, channel length modulation and GIDL is achieved. The model also gives good accuracy in the modeling of device parameters for analog applications. This is demonstrated in Fig. 11.15, where the trans-conductance efficiency (i.e.,  $g_m/I_d$ ) and output conductance ( $g_{ds}$ ) of a long channel SOI FinFET ( $L_g = 1\text{ m}$ ) are well reproduced by the model.

The model has also been validated with bulk FinFET measurements. These devices with moderate channel doping were fabricated with a TiN gate. They have 25nm thick fins and an equivalent oxide thickness (EOT) of 1.95nm. Fig. 11.16 shows the measured short-channel ( $L_g = 50\text{nm}$ ) characteristics and the model fitting results. Compared to its SOI version, the topology of the BSIM-CMG bulk model adds a body node, which the impact ionization body current flows into. The measured body current of a short-channel FinFET together with the model fitting is shown in Fig. 11.16(b) with excellent agreement. The model and measurement comparisons of the trans-conductances of the drain current,  $g_m$  and  $g_{ds}$ , are shown for a long-channel bulk FinFET device in Fig. 11.17. Note that the measured output channel conductance  $g_{ds}$  can be noisy and hard to fit when it is very small.

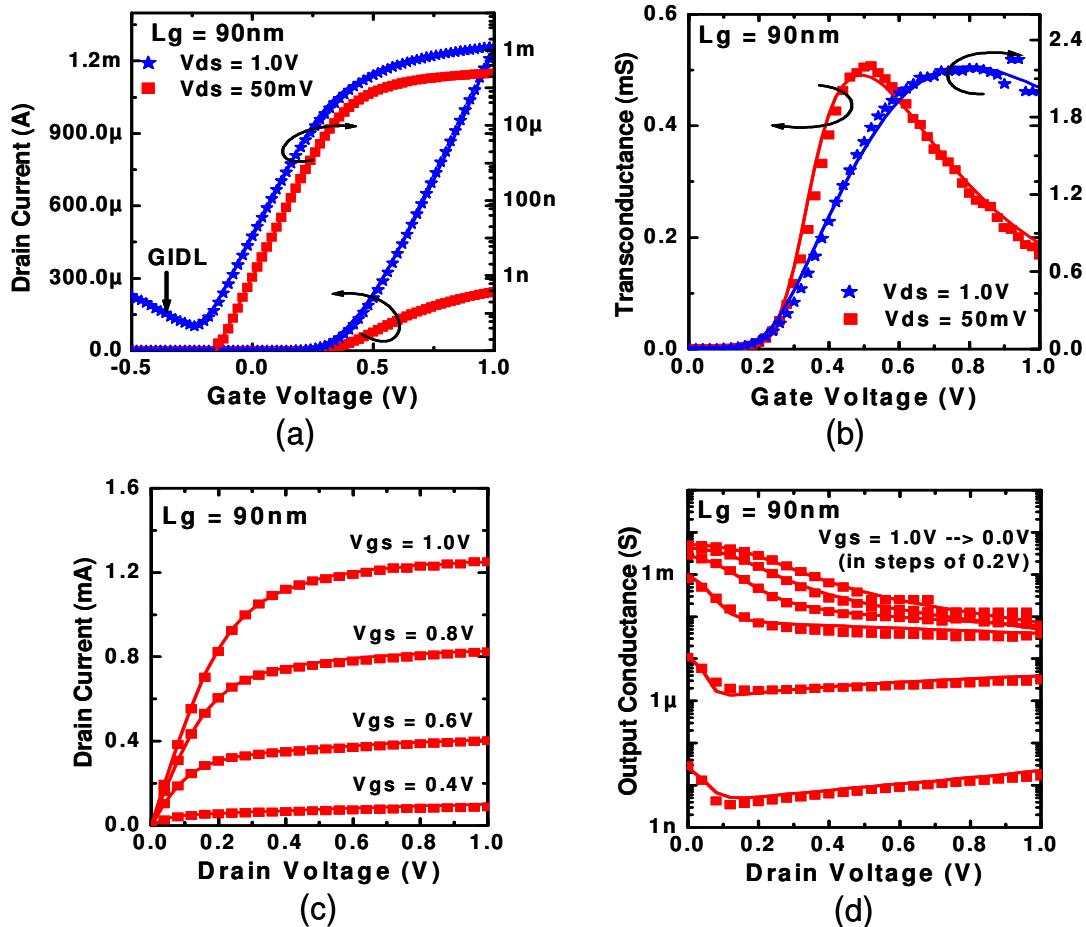


Fig. 11.14 BSIM-CMG model comparison with short-channel SOI FinFETs (gate length = 90nm). Symbols represent the measured data and lines indicate model fitting results.

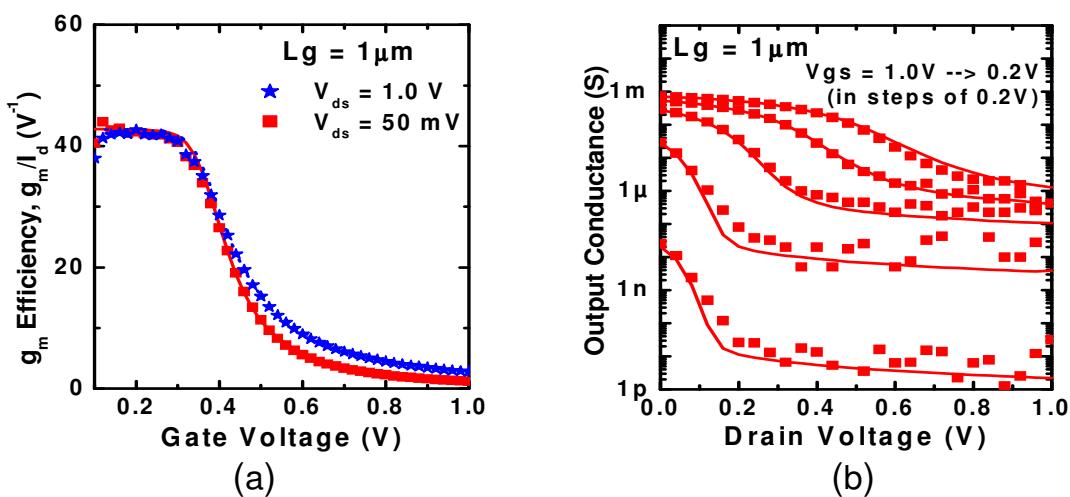


Fig. 11.15 BSIM-CMG model comparison with analog circuit parameters (a)  $g_m/I_d$  and (b)  $g_{ds}$  for long-channel SOI FinFET (gate length = 1 m). Symbols represent the measured data and lines indicate model fitting results.

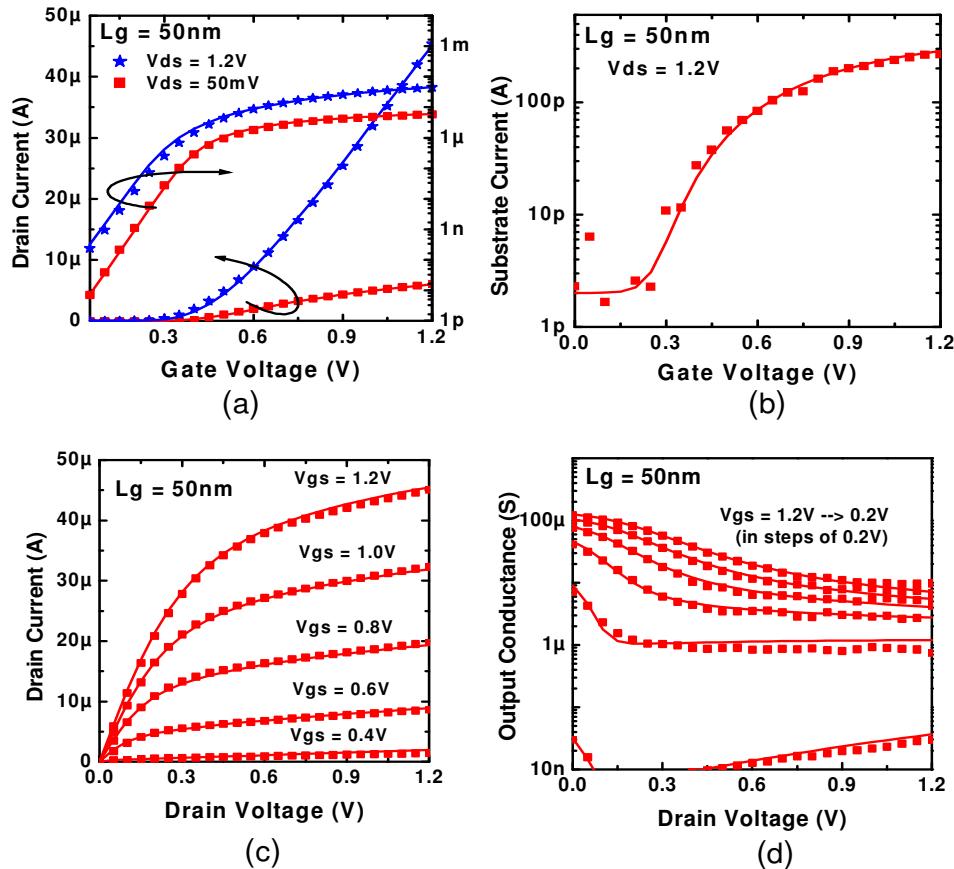


Fig. 11.16 BSIM-CMG model comparison with short-channel bulk FinFETs (gate length =  $50\text{nm}$ ). Symbols represent the measured data and lines indicate model fitting results.

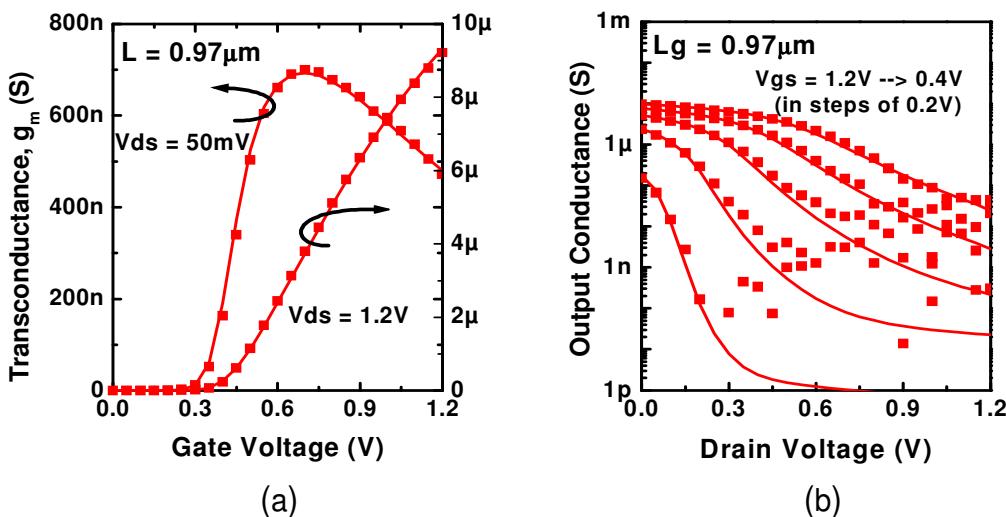


Fig. 11.17 BSIM-CMG model comparison with analog design metrics (a) transconductance  $g_m$  and (b) output conductance  $g_{ds}$  for a long-channel bulk FinFET (gate length =  $0.97\mu\text{m}$ ). Symbols represent the measured data and lines indicate the BSIM-CMG model fitting results.

The comparisons show that BSIM-CMG accurately captures the characteristics of advanced multi-gate FETs. The validation with triple-gate FETs demonstrates the ability of the model to capture phenomena such as gate corner effects which are unique to tri-gate and quadruple-gate FETs. The model is able to describe both SOI and bulk silicon based multi-gate FET technologies.

## 11.5 Chapter Summary

Compact models of semiconductor devices are the bridge between IC manufacturing and design. BSIM4, as the industry standard MOSFET compact model, has served the IC industry worldwide successfully from the 130nm down to today's 20nm technology node as well as from the logic and memory to the analog/RF and mixed-signal IC applications. Over the past decade, BSIM4 has contributed to the IC industry worth untold billions of dollars. It has brought gigantic benefits to the economy and the quality of life around the world.

BSIM-CMG is developed for the multi-gate MOSFETs/FinFETs, a technology that allows CMOS size reduction to continue for years to come.

## References

- [1] Y. Taur, D. A. Buchanan, Wei Chen, D. J. Frank, K. E. Ismail, Shih-Hsien Lo, G. A. Sai-Halasz, R. G. Viswanathan, H.-J. C. Wann, S. J. Wind, and Hon-Sum Wong, "CMOS scaling into the nanometer regime," *Proceedings of IEEE*, vol. 85, pp. 486-504, April 1997.
- [2] D. J. Frank, R. H. Dennard, E. Nowak, P. M. Solomon, Y. Taur, and H.-S.P. Wong, "Device scaling limits of Si MOSFETs and their application dependencies," *Proceedings of IEEE*, vol. 89, pp. 259-288, March 2001.
- [3] D. Hisamoto, Wen-Chin Lee, J. Kedzierski, H. Takeuchi, K. Asano, C. Kuo, E. Anderson, Tsu-Jae King, J. Bokor, and Chenming Hu, "FinFET – A self-aligned double gate MOSFET scalable to 20nm," *IEEE Trans. Electron Devices*, vol. 47, pp. 2320-2325, December 2000.
- [4] H.-S. P. Wong, D. J. Frank, and P. M. Solomon, "Device design considerations for double-gate, ground plane and single-gate ultra-thin SOI MOSFETs at the 25nm channel length generation," *IEDM Tech. Dig.*, pp. 407-410, San Francisco, December 1998.

## 410 BSIM4 AND MOSFET MODELING FOR IC SIMULATION

By Weidong Liu and Chenming Hu

- [5] Xuejue Huang, Wen-Chin Lee, Charles Kuo, D. Hisamoto, Leland Chang, J. Kedzierski, E. Anderson, H. Takeuchi, Yang-Kyu Choi, K. Asano, K. V. Subramanian, Tsu-Jae King, J. Bokor, and Chenming Hu, "Sub-50 nm FinFET: PMOS", Tech. Dig. of IEDM, pp. 67-70, Washington D. C., December 1999.
- [6] Y. Taur, "Analytic solutions of charge and capacitance in symmetric and asymmetric double-gate MOSFETs," IEEE Trans. Electron Devices, vol. 48, pp. 2861-2869, December 2001.
- [7] G. D. J. Smit, A. J. Scholten, N. Serra, R. M. T. Pijper, R. van Langevelde, A. Mercha, G. Gildenblat, and D. B. M. Klassen, "PSP-based compact FinFET model describing dc and RF measurements," IEDM Tech. Dig., pp. 175-178, San Francisco, December 2006.
- [8] M. V. Dunga, C-H. Lin, A. Niknejad, and C. Hu, "BSIM-CMG: A Compact Model for Multi-Gate Transistors," Chapter 3 in FinFETs and Other Multi-Gate Transistors, J. P. Colinge, Edited, Springer Science, New York, NY , pp.113-153, 2007.
- [9] Mohan V. Dunga, Chung-Hsun Lin, Darsen D. Lu, Weize Xiong, C. R. Cleavelin, P. Patruno, Jiunn-Ren Hwang, Fu-Liang Yang, Ali M. Niknejad, and Chenming Hu, "BSIM-MG: A versatile multi-gate FET model for mixed-signal design," Proceedings of the VLSI Technology Symposium, pp. 80-81, 2007.
- [10] J. R. Brews, "A charge sheet model of the MOSFET," Solid-State Electronics, vol. 21, pp. 345-355, 1978.
- [11] D. Ward, and R. Dutton, "A charge-oriented model for MOS transistor capacitances," IEEE J. Solid State Circuits, vol. SSC-13, pp. 703-708, October 1978.

# Index

## A

AC currents, 155  
auxiliary sub-circuit, 346

## B

Backward Euler numerical integration method, 349  
band-to-band tunneling, 116, 143  
binned model, 17  
binning, 6, 13, 17, 61  
body charge, 23  
body depletion region, 23  
body resistance network, 214  
body-bias coefficient, 23, 34  
Boltzmann constant, 21, 122  
breakdown coefficient, 313  
breakdown factor, 312  
BSIM (Berkeley Short-channel IGFET Model), 1  
built-in potential, 325, 332  
bulk inversion, 399  
bulk-charge coefficient, 45  
bulk-charge effects, 45, 168  
buried oxide, 389

## C

capacitive currents, 155  
carrier mobility, 93, 389, 396  
carrier scattering, 250, 388  
channel “pinch-off”, 100, 258  
channel charge density, 89  
channel length and width, 15, 16  
channel length modulation, 100, 253

Channel Thermal Noise, 254  
channel voltage, 24  
characteristic drain-field length, 102  
characteristic length, 28  
charge-deficit AC NQS, 211, 213  
charge-deficit transient NQS, 201, 205  
charge-thickness model, 168, 169  
circuit Jacobian, 350  
CMOS (Complementary Metal-Oxide-Semiconductor), 1  
CMOS IC, 226  
CMOS, 1, 5, 6, 8, 14, 18, 22, 31, 64, 65, 87, 107, 203, 235, 253, 269, 345, 348, 387, 388  
common multiple gates, 387  
Compact Model Council, 6, 115, 269  
compact models, 1  
conduction band, 305  
constant noise figure circle, 245  
convergence checking, 381  
Coulombic scattering, 53, 54, 56, 246  
critical electric field, 94, 95  
critical vertical electric field, 53  
current continuity equation, 128  
cut-off frequency, 213

## D

depletion layer capacitance, 89, 93  
depletion width, 393  
differential and algebraic equations, 347  
diffusion resistances, 46

diode breakdown, 309  
 diode charge and capacitance, 322  
 divide-by-zero errors, 129, 135  
 drain-induced barrier lowering, 403  
 drain-induced threshold voltage shift, 104  
 drift-diffusion theory, 396  
 dynamic depletion, 394

**E**

Early voltage, 99  
 effective body-bias voltage, 42  
 effective channel doping concentration, 34  
 effective drain-to-source voltage, 196  
 effective gate bias, 89  
 effective gate oxide capacitance, 20  
 effective junction perimeter and area, 277  
 effective mobility, 53, 54, 94  
 effective time step, 349  
 electrical gate capacitance, 89  
 electrical gate oxide capacitance, 206  
 electrical gate oxide thickness, 121, 206  
 electron affinity, 35, 51, 55  
 energy band gap, 55  
 energy barrier, 25  
 energy relaxation time, 96  
 energy-band gap, 21  
 equivalent diode noise current, 241  
 equivalent noise resistance, 242  
 equivalent noise temperature, 242  
 equivalent oxide thickness, 20, 406  
 equivalent saturated diode noise current, 242

**F**

fast SPICE, 2  
 Fermi level, 247, 328  
 Fermi potential, 23, 121, 391

finite charge-thickness, 51  
 flat-band voltage, 21, 23, 50, 125, 171, 172  
 flicker noise, 239, 247

**G**

gate charge, 23  
 gate direct-tunneling current, 116, 205  
 gate electrode resistance, 190  
 gate work function, 55  
 gate-and-channel tunneling current, 116  
 gate-body tunneling current  $I_{gb}$ , 121  
 gate-channel tunneling, 125, 126  
 gate-edge perimeter, 278  
 gate-induced source and drain leakage, 143  
 Gauss' law, 26, 391, 393, 397  
 Gaussian elimination, 350  
 Gaussian equations, 140  
 global model, 17

**H**

high-k metal gate, 19, 51, 53, 56  
 high-performance, 22  
 holistic thermal noise model, 256

**I**

impact ionization, 105, 139, 406  
 induced gate noise, 256, 264  
 internal source and drain nodes, 48  
 intrinsic carrier concentration, 21  
 intrinsic-input gate resistance, 193, 226  
 inversion charge, 396  
 ionized dopant atoms, 53  
 isolation-sidewall perimeter, 278

**J**

junction breakdown voltage, 313  
 junction depletion capacitance, 322  
 junction diffusion capacitance, 322  
 junction grading coefficient, 323, 324

junction leakage currents, 278  
 junction saturation current, 304, 329

**K**

Kirchhoff's current law, 348

**L**

layout parasitics extraction, 47  
 layout-dependent effects, 58  
 LDD resistances, 46  
 length of oxide definition, 60  
 LOCOS, 31  
 long-channel zero body-bias threshold voltage, 27  
 longitudinal electric field, 94  
 low-field mobility, 55, 60  
 low-noise amplification, 243

**M**

matching impedance, 245  
 material model, 20, 43  
 mechanical stress effects, 56  
 minimum noise figure, 244  
 mobility fluctuation, 250  
 model card selection, 17  
 moderate inversion, 91  
 momentum relaxation time, 96  
 MOSFETs (Metal-Oxide-Semiconductor Field Effect Transistors), 1  
 multi-finger, 61

**N**

Newton-Raphson iteration, 346, 349  
 nodal admittance matrix, 350  
 noise figure contours, 245  
 noise figure, 242  
 noise matching network., 245  
 noise parameters, 242  
 noise partitioning, 263  
 noise portioning coefficient, 259

noise reference temperature, 243  
 noise source reflection, 243  
 noise voltages, 236  
 non-ideality factor, 306  
 non-Si substrate material, 20  
 non-uniform substrate doping, 32  
 non-uniform vertical doping, 33, 34  
 normalized equivalent noise resistance, 244  
 NQS, 213  
 number of fingers, 15  
 numerical integrations, 348  
 Nyquist's theorem, 241

**O**

optical phonons, 96  
 optimum complex reflection coefficient, 244  
 output resistance, 98

**P**

parameter binning, 18  
 parameter extraction, 199, 214, 226, 244  
 parasitic junction diode series resistance, 304  
 physical gate oxide capacitance, 19  
 point contact, 293  
 Poisson equation, 23, 36, 175, 177, 390  
 poly-silicon gate depletion, 42, 89, 146  
 poly-silicon sheet resistance, 192  
 power spectral intensity, 236  
 process technology, 22

**Q**

quantum mechanical effect, 13, 19, 390  
 quasi two-dimensional approximations, 26  
 quasi-Fermi potential, 88, 197

**R**

random phase, 264  
RC time constant, 192, 207  
relative dielectric constant, 19  
relaxation-time approximation, 203  
remote soft-phonon, 56  
round-off errors, 45  
Rout degradation factor, 103

**S**

saturation current densities, 291  
saturation velocity, 62  
saturation voltage, 96  
scattering centers, 53  
Schrodinger equation, 51  
shallow-trench isolation, 59, 193  
shared source and drain, 272  
sheet resistance, 47  
Shockley-Read-Hall (SRH) Generation and Recombination, 305  
shot noise, 240  
signal-to-noise power ratio, 242  
small-signal equivalent circuit, 238  
smoothing functions, 44  
source-end velocity limit, 107  
S-parameters, 199, 226  
SPICE, 13, 16, 22, 29, 39, 60, , 89, 107, 119, 124, 138, 147, 190, 193, 198, 201, 226, 235, 238, 245, 254, 273, 304, 313, 315, 322, 345, 350, 390  
strained-silicon technology, 59  
stress memorization technique, 60  
substrate permittivity, 20  
substrate-current induced body effect, 142  
subthreshold swing, 92  
surface potential, 13, 21, 34, 88  
surface roughness scattering, 56  
surface scattering, 53  
susceptance, 240

**T**

target oxide thickness, 37  
Taylor series expansion, 349  
telegraph signal, 247  
temperature-dependence, 21, 39, 57  
tensile stress, 59  
thermal noise, 239  
thermal voltage, 122  
thermionic emission, 107  
threshold voltage, 13, 24, 119, 126  
threshold voltage, 4, 89, 99, 106  
time-domain transient analysis, 347  
trans-conductance amplification coefficient, 258  
trans-conductance amplification, 257, 260  
trans-conductance efficiency, 406  
Transient noise, 238  
trans-nodal capacitances, 350  
transport currents, 348  
trap-assisted tunneling, 307, 322, 330  
tunneling barrier height, 123  
tunneling mechanisms, 118  
two-port active network, 240

**U**

universal mobility-field relationship, 53

**V**

valence band, 306  
velocity overshoot, 96  
velocity saturation, 93, 95  
voltage-controlled current source, 260

**W**

well-proximity effects, 56, 64  
wide contact, 293

**Y**

Y parameter, 198, 214