UNIVERSITY OF CALIFORNIA SAN DIEGO

Non-GCA Modeling for Double-Gate and Ground-Plane MOSFETs

A Dissertation submitted in partial satisfaction of the requirements
for the degree Doctor of Philosophy

in

Electrical Engineering (Applied Physics)

by

Mei-Hua Su

Committee in charge:

Professor Yuan Taur, Chair
Professor Chung-Kuan Cheng
Professor Kenji Nomura
Professor Paul K. L. Yu

2024

The Dissertation of Mei-Hua Su is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2024

DEDICATION

This dissertation is dedicated to my parents for their great support and unbounded love.

TABLE OF CONTENTS

DISSERTATION APPROVAL PAGE iii

DEDICATION iv

TABLE OF CONTENTS v

LIST OF FIGURES vii

ACKNOWLEDGEMENTS xii

VITA xiii

ABSTRACT OF THE DISSERTATION xiv

CHAPTER 1 INTRODUCTION 1

CHAPTER 2 PROBLEMS WITH GCA MODELING OF THE MOSFET SATRURATION REGION 3

2.1 GCA MODEL UNDER CONSTANT MOBILITY 3
2.2 GCA MODEL UNDER VELOCITY SATURATION 11

CHAPTER 3 HISTORY OF MODELING THE MOSFET SATURATION REGION 19

3.1 REDDI AND SAH'S CONCEPT OF PINCH-OFF IN METAL-OXIDE-SEMICONDUCTOR TRANSISTOR (MOST) 19
3.2 EL-MANSY AND BOOTHROYD'S TWO-DIMENSIONAL MODEL IN THE SATURATION REGION 20
3.3 PING KO'S PHD THESIS 23

CHAPTER 4 NON-GCA MODEL FOR DG MOSFETS 28

4.1 TCAD SIMULATIONS 28
4.2 CONSTANT MOBILITY 36
4.3 *N*=1 VELOCITY SATURATION 43
4.4 *N*=2 VELOCITY SATURATION 47
4.5 EXPLICIT SOLUTION BY REGIONAL APPROXIMATION 54
4.6 NUMERICAL SOLUTION METHODS: FORWARD EULAER VERSUS BACKWARD EULER 57

CHAPTER 5 NON-GCA MODEL FOR BULK MOSFETS 63

5.1 UNIFORM DOPING 63
5.2 GROUND-PLANE MOSFETS 70

CHAPTER 6 SCE OF ET-SOI MOSFETS 91

6.1 SHORT-CHANNEL SOI MOSFETS 91
6.2 EFRECTS OF BOX THICKNESS, SILICON THICKNESS, AND BACKGATE BIAS ON SCE 95

CHAPTER 7 CONCLUSION 104

LIST OF FIGURES

Figure 2.1: A schematic MOSFET cross section. 3

Figure 2.2: Inversion charge density as a function of the quasi-Fermi potential V. 7

Figure 2.3: I_{ds}-V_{ds} curves (solid) solved. 8

Figure 2.4: Schematic diagram of a double-gate MOSFET. 9

Figure 2.5: GCA model generated I_{ds}-V_{ds} characteristics. 12

Figure 2.6: I_{ds}-V_{ds} characteristics generated by the GCA models under the n = 1 velocity saturation model. 15

Figure 2.7: I_{ds}-V_{ds} characteristics generated by the GCA and under the n = 2 velocity saturation relation. 16

Figure 2.8: dV/dy at the drain ($y = L$) versus V_{ds} for the case for n=1. 17

Figure 2.9: dV/dy at the drain ($y = L$) versus V_{ds} for n=2. 18

Figure 3.1: A schematic cross section of an IGFET to illustrate source and drain section and axes. 20

Figure 3.2: Definition of the drain section boundaries. 22

Figure 3.3: A closer look into the cross section of a MOSFET near drain. 23

Figure 4.1: Double-gate (DG) MOSFET structure. 28

Figure 4.2: GCA model generated I_{ds}-V_{ds} characteristics compared to TCAD. 28

Figure 4.3: From TCAD at bias point A on the V_{gs} = 1.5 V curve in Fig. 4.2. 30

Figure 4.4: From TCAD at bias point B on the V_{gs} = 1.5 V curve in Fig. 4.2. 30

Figure 4.5: From TCAD at bias point B in Fig. 2, V_{gs} = 1.5 V, V_{ds} = 1.8 V. 32

Figure 4.6: From TCAD: Potential and Fermi potential along a cut through the center. 32

Figure 4.7: Comparison of $Q_i(V)$ from TCAD, GCA model, and Eq. (4.19). 35

Figure 4.8: Potential contour plot from TCAD. The bias point is V_{gs} = 0.9 V, V_{ds} = 1.2 V, under the n = 2 velocity saturation model. 36

Figure 4.9: (a) Solution to Eq. (4.9) with I_{ds} set at 6% over the peak (I_{dsat}). (b) Agreement between the I_{ds}-V_{ds} computed point by point and that by multiplying (I_{ds0}/L) to the $y(V)$ curve. .40

Figure 4.10: Figure caption example. 42

Figure 4.11: I_{ds}-V_{ds} characteristics generated by the GCA and non-GCA models under the n = 1 velocity saturation model. 45

Figure 4.12: dV/dy at the drain ($y = L$) versus V_{ds} for the case in Fig. 4.11. 46

Figure 4.13: I_{ds}-V_{ds} characteristics (n = 1 velocity saturation) generated by the continuous non-GCA model compared with TCAD. 47

Figure 4.14: Comparison of $g_{dc} \equiv dI_{ds}/dV_{ds}$ versus V_{ds} (n = 1 vel. sat.) between TCAD and the non-GCA model. 47

Figure 4.15: I_{ds}-V_{ds} characteristics generated by the GCA and non-GCA models under the n = 2 velocity saturation model. μ_0 = 200 cm^2/V-s, $v_{sat} = 10^7$ cm/s. C_{inv} is taken to be ε_i/t_i.....50

Figure 4.16: dV/dy at the drain ($y = L$) versus V_{ds} for the case in Fig. 4.12. Labels above the non-GCA curve indicate the carrier velocity at those bias points. 51

Figure 4.17: I_{ds}-V_{ds} characteristics (n = 2 velocity saturation) generated by the continuous non-GCA model compared with TCAD. 52

Figure 4.18: Comparison of $g_{dc} \equiv dI_{ds}/dV_{ds}$ versus V_{ds} (n = 2 vel. sat.) between TCAD and the non-GCA model. Open squares are calculated from the explicit Eq. (4.54) valid for $V_{ds} > V_{dsat}$. 52

Figure 4.19: Comparison of the rigorous all region model, Eqs. (4.1), (4.2) to the C_{inv} model, Eq. (4.19) at V_{gs} 70-270 mV above V_t. 53

Figure 4.20: Model validity versus channel length. n = 1 velocity saturation model is assumed in both model and TCAD. 57

Figure 4.21: Solutions for Backward Euler and Averaged Euler methods with different step sizes dy.

.....60

Figure 4.22: Solutions from Forward Euler and Averaged Euler methods with step sizes dy of 0.1 nm.

.....61

Figure 5.1: Plots from TCAD simulations. (a) Potential $\psi(x)$ and electron density $n(x)$ (right scale) along three vertical cuts (b) Electron density versus depth in silicon along five vertical cuts between the saturation point and the drain (y = 500 nm).

.....65

Figure 5.2: $y(V)$ solution to Eq. (5.5) for two values of I_{ds}: I_{ds1} is 3% over I_{dsat}, I_{ds2} is 6% over I_{dsat}. 69

Figure 5.3: I_{ds}-V_{ds} curves (solid) solved from Eq. (5.5) for the device described in the caption to Fig. 5.2. The dashed curves are from the GCA model for which currents saturate at I_{dsat}. 69

Figure 5.4: A schematic diagram showing the low-high (retrograde) step doping profile. 73

Figure 5.5: Band diagram and charge distribution of an extreme retrograde-doped or ground-plane nMOSFET at the threshold condition. 75

Figure 5.6: A schematic cross-section of ground-plane MOSFETs. Shown on the right is the depth profile of body doping along a vertical cutline. 76

Figure 5.7: Band diagram of a ground-plane nMOSFET biased near the threshold. The p^+ ground plane is grounded to the n^+ source. 77

Figure 5.8: Mobile charge density per area at a point in the channel versus electron quasi-Fermi potential for a given gate voltage. 81

Figure 5.9: I_{ds}-V_{gs} characteristics generated by the model in both linear and log scales compared to TCAD. 82

Figure 5.10: I_{ds}-V_{ds} characteristics generated by the model compared to TCAD. The squares are from the GCA model discussed in this section. 82

Figure 5.11: I_{ds}-V_{ds} characteristics generated by the n = 1 non-GCA model compared to TCAD. 85

Figure 5.12: I_{ds}-V_{ds} characteristics generated by the n = 2 non-GCA model compared to TCAD. 86

Figure 5.13: I_{ds}-V_{ds} characteristics generated by the n = 1 non-GCA model compared to TCAD with different d_{si}. 87

Figure 5.14: I_{ds}-V_{ds} characteristics generated by the n = 1 velocity saturation model compared to the published data of 20 nm MOSFETs. (a) No source and drain resistance. (b) With source and drain resistance (values given in the main text) added to the model. 89

Figure 6.1: A schematic cross-section of SOI CMOS, with shallow trench isolation, dual polysilicon gates, and self-aligned silicide. 91

Figure 6.2: 2-D constant potential contours of (a) bulk and (b) SOI MOSFETs. 93

Figure 6.3: Short-channel V_t roll-off of the bulk and SOI MOSFETs in Fig. 6.2. 94

Figure 6.4: $\Delta\psi_{s,min}$, the minimum surface potential between the source and drain of a short channel device with respect to that of the long channel device for the SOI and bulk MOSFETs in Fig. 6.5.....95

Figure 6.5: Cross-section of ET-SOI MOSFET investigated in this work. 96

Figure 6.6: Short-channel V_t roll-off versus BOX thickness. 98

Figure 6.7: Short-channel V_t roll-off versus silicon thickness. 99

Figure 6.8: Comparison of V_t roll-off of nMOS with respect to substrate doping type and concentration.....100

Figure 6.9: Comparison of SCE for different gate work function and backgate bias. 101

Figure 6.10: Potential versus depth for the cases of Vbg = 3 V and Vbg = -3 V in Fig. 6.8. 102

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, Prof. Yuan Taur. Without his tremendous help and enlightening guidance during the past 5 years, I am unable to reach this milestone in my life. He is a great expert with 50 years' research experience and knowledge in the field of semiconductor physics. His deep insight and vast knowledge benefit me throughout the whole PhD study. The way he thinks and the attitude he has in pursuing the truth significantly influences me. I can take great advantage of all I learn from him for my whole life.

Tons of thanks should be given to my committee members, Prof. Paul K. L. Yu, Prof. Kenji Nomura, Prof. Prabhakar Bandaru and Prof. Chung-Kuan Cheng for taking their time to serve as committee members to review my dissertation and give me valuable comments.

Many thanks to my friends in UCSD, in particular, Chuyang Hong, Zhongjie Ren, Ruoman Yang, Chi-Hsin Huang, Hao-ping Lin, Hsuan Chang and Ben Qiu. Those friendships have meant a lot to me.

Finally, and most importantly, I want to express my sincere gratitude to my parents for their endless love and support. This dissertation is dedicated to them.

Chapter 4, in full, is a reprint of the material as it appears in Yuan Taur, Woojin Choi, Jianing Zhang, and Meihua Su, "A Non-GCA DG MOSFET Model Continuous into the Velocity Saturation Region", *IEEE Trans. Electron Device*, pp. 1160-1166, Mar. 2019. The dissertation author was an investigator and author of this paper.

Chapter 5, in full, is a reprint of the material as it appears in M.-H. Su, C. Hong, and Y. Taur, "A Non-GCA Model for Ground-Plane MOSFETs*", Solid-State Electronics*, vol. 209, p. 108754, Nov. 2023. The dissertation author was the primary investigator and author of this paper.

Chapter 6, in full, is a reprint of the material as it appears in M.-H. Su, C. Hong, S. Cristoloveanu and Y. Taur, "Effects of BOX Thickness, Silicon Thickness, and Backgate Bias on SCE of ET-SOI MOSFETs," Microelectronic Engineering, 238, 111506, Jan. 2021. The dissertation author was the primary investigator and author of this paper.

VITA

2017 Bachelor of Electrophysics, National Chiao Tung University

2019 Master of Science in Electrical Engineering (Applied Physics), University of California San Diego

2024 Doctor of Philosophy in Electrical Engineering (Applied Physics), University of California San Diego

PUBLICATIONS

[1] **"**A Non-GCA DG MOSFET Model Continuous into the Velocity Saturation Region**",** Yuan Taur, Woojin Choi, Jianing Zhang, and Meihua Su, IEEE TRANSACTIONS ON ELECTRON DEVICES, MARCH 2019
[2] E. M. Su, D. C. Hong and Y. Taur, "Effects of BOX thickness, silicon thickness, and backgate bias on SCE of ET-SOI MOSFETs", Microelectronic Engineering, vol. 238, pp. 111506, Jan. 2021.
[3] M.-H. Su, C. Hong, and Y. Taur, "A Non-GCA Model for Ground-Plane MOSFETs", Solid State Electronics, vol. 209, pp. 108754, Aug. 2023

FIELD OF STUDY

Major Field: Electrical Engineering
Focused Field: Applied Physics/Engineering Physics

ABSTRACT OF THE DISSERTATION

Non-GCA Modeling for Double-Gate and Ground-Plane MOSFETs

by

Mei-Hua Su

Doctor of Philosophy in Electrical Engineering (Applied Physics)

University of California San Diego, 2024

Professor Yuan Taur, Chair

In this dissertation, non-GCA models are developed for both DG (Double-Gate) MOSFETs and ground-plane bulk MOSGETs. It is widely known that MOSFET velocity saturation region is beyond the framework of GCA first invoked by Shockley in 1952, the bedrock of virtually all MOSFET models. A few papers in the literature have dealt with the 2-D nature of the field pattern in the saturation region of bulk or DG. In general, such models are unable to generate Ids-Vds curves continuous from the triode region into the velocity saturation region.

A DG MOSFET model that goes beyond the *gradual channel approximation* is developed by incorporating the effect of lateral field gradient on carrier density. It is shown that while the oxide field crosses zero at the point of saturation and becomes negative beyond it, the channel is not *pinched off* of charge carriers. The model generates I_{ds}-V_{ds} characteristics continuous into the saturation region with finite output conductance consistent with TCAD. An explicit expression is derived for the output conductance in saturation in terms of basic device parameters.

The continuous model is later extending MOSFET I-V characteristics into the velocity saturation region with finite output conductance. Both the n = 1 and n = 2 models have been employed. It is shown that the standard relation of *channel length modulation* (CLM) for constant mobility must be modified for velocity saturation because the drain current is not simply inversely proportional to the channel length. Regional approximations are applied to derive explicit expressions for the output conductance in the velocity saturation region in terms of basic device parameters.

In the following section, a non-GCA (Gradual Channel Approximation) model continuous into the velocity saturation region is developed for ground-plane bulk MOSFETs. The I_{ds}-V_{ds} characteristics generated by both the n = 1 and the n = 2 models are consistent with 2-D simulations. By incorporating source and drain series resistance into the model, it is shown that the model can reproduce the I_{ds}-V_{ds} data of 20 nm bulk MOSFETs published in the literature.

CHAPTER 1 INTRODUCTION

In a field-effect transistor (FET), the current in the channel between the source and drain is modulated by the voltage applied to the gate. Under most bias conditions, the field in the gate direction is much stronger than the field in the source-drain direction. Modeling of an FET is much simplified under the *gradual channel approximation* (GCA), which assumes in the Poisson's equation that the field in the source-drain direction is negligible compared to the field in the gate direction. Virtually all FET models stemmed from the framework of GCA first invoked by Shockley in 1952 [1][2]. While the application of GCA led to analytic models for the linear, parabolic, and subthreshold regions, it renders either no solution or a negative slope in the saturation region, i.e., in the I_{ds}-V_{ds} characteristics at drain voltages beyond the value where the current saturates.

It has been recognized early on that the field pattern in the velocity saturation region is of a 2-D nature. A few papers in the literature [3][4] have dealt with the 2-D nature of the field pattern in the saturation region of MOSFETs. Their common approach is to divide the device into two sections. In the section on the source side, the GCA holds. In the section on the drain side, 2-D Gauss' law is applied to obtain the length of the section, known as channel length modulation, as a function of drain voltage. However, the critical lateral field at the start of the "velocity saturation" section cannot be unambiguously defined. In general, such models are unable to generate I_{ds}-V_{ds} curves continuous from the triode region into the velocity saturation region.

In this dissertation, a non-GCA model is formulated by adding a source-drain field term to the gate-induced mobile charge density of the GCA model. It generates I_{ds}-V_{ds} curves continuous

from the triode region into the velocity saturation region for double-gate (DG) and ground-plane bulk MOSFETs. All have been verified by TCAD simulations.

The outline of the dissertation is as follows. Chapter 2 details the problems encountered with GCA modeling of the MOSFET saturation region. Chapter 3 reviews the previous modeling of the MOSFET saturation region in the literature. Chapters 4 describes the formulation of the proposed non-GCA model and its application to DG MOSFETs. Chapter 5 applies the non-GCA model to bulk MOSFETs, including the uniformly-doped and ground-plane devices. Chapter 6 is also part of the research work during the PhD study, but on a different topic: SCE (Short Channel Effect) on ET-SOI (Extreme Thin Silicon on Insulator) MOSFETs. Chapter 7 is the conclusion.

References:

[1] W. Shockley, "A unipolar field-effect transistor," *Proc. IRE*, vol. 40, pp. 1365-1376, Nov. 1952.

[2] C. T. Sah, "Characteristics of the metal-oxide-semiconductor transistors," *IEEE Trans. Electron Device*, pp. 324-345, July 1964.

[3] Y. El-Mansy and A. Boothroyd, "A simple two-dimensional model for IGFET operation in the saturation region," *IEEE Trans. Electron Devices*, pp. 254-262, Mar. 1977.

[4] P. K. Ko, R. S. Muller, and C. Hu, "A unified model for hot electron currents in MOSFETs," *1981 IEDM Technical Digest*, pp. 600-603.

CHAPTER 2 PROBLEMS WITH GCA MODELING OF THE MOSFET SATURATION REGION

2.1 GCA Model under Constant Mobility

Figure 2.1 shows the schematic cross section of an n-channel MOSFET in which the source is the n^+ region on the left, and the drain is the n^+ region on the right. A thin oxide film separates the gate from the channel region between the source and drain. The *x*-axis is perpendicular to the gate electrode and is pointing into the p-type substrate with $x = 0$ at the silicon surface. The *y*-axis is parallel to the channel or the current flow direction, with $y = 0$ at the source and $y = L$ at the drain. The MOSFET is assumed to be uniform along the *z*-axis over a distance called the *channel width*, *W*, determined by the boundaries of the thick field oxide.

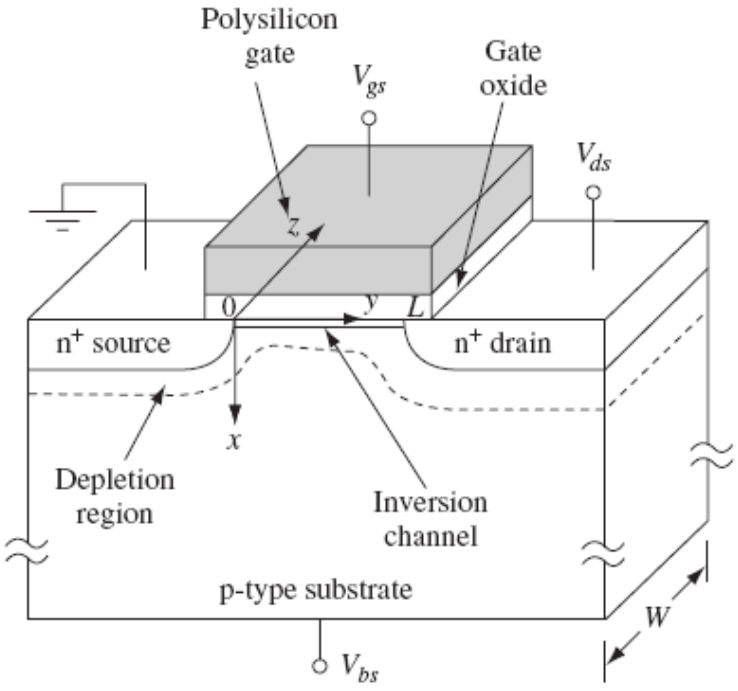


Figure 2.1 A schematic MOSFET cross section, showing the axes of coordinates and the bias voltages at the four terminals for the drain-current model.

Conventionally, the source voltage is defined as the ground potential. The drain voltage is V_{ds}, the gate voltage is V_{gs}, and the p-type substrate is biased at V_{bs}. We assume $V_{bs} = 0$, i.e., the substrate contact is grounded to the source potential. The p-type substrate is assumed to be uniformly doped with an acceptor concentration N_a.

2.1.1 Bulk MOSFETs, Constant Mobility

Gradual Channel Approximation

A major assumption in any 1-D MOSFET model is the ***gradual channel approximation (GCA), which assumes that the variation of the electric field in the y-direction (along the channel) is much less than the corresponding variation in the x-direction (perpendicular to the channel)*** (Pao and Sah, 1966). This allows us to reduce the 2-D Poisson equation to 1-D slices (x-component only).

$\psi(x, y)$ is the band bending, or intrinsic potential, at (x, y) with respect to the intrinsic potential of the bulk substrate. We further assume that $V(y)$ is the electron quasi-Fermi potential at a point y along the channel with respect to the Fermi potential of the n^+ source. The assumption that V is independent of x in the direction perpendicular to the surface is justified by the consideration that current is proportional to the gradient of the quasi-Fermi potential and that MOSFET current flows predominantly in the source-to-drain, or y-direction. At the source end of the channel, $V(y = 0) = 0$. At the drain end of the channel, $V(y = L) = V_{ds}$. The electron quasi-Fermi potential at a point in the channel is essentially flat in the vertical direction across the n-type inversion layer. The effect of V is to multiply the electron density by $e^{-qV/kT} \propto e^{-q\phi_n/kT} = e^{E_{fn}/kT}$ over its $V = 0$ value.

$$\frac{\partial^2\psi}{\partial x^2}+\frac{\partial^2\psi}{\partial y^2}=\frac{q}{\varepsilon_{si}}\left[N_a+n_i e^{q(\psi-V)/kT}\right] \tag{2.1}$$

Assume $\frac{\partial^2\psi}{\partial x^2} >> \frac{\partial^2\psi}{\partial y^2}$

$$\frac{d^2\psi}{dx^2}=\frac{q}{\varepsilon_{si}}\left[N_a+n_i e^{q(\psi-V)/kT}\right] \tag{2.2}$$

Coupled to the current continuity eq.: (for constant mobility)

$$I_{ds}=\mu_{eff}\frac{W}{L}\int_0^{V_{ds}}\left(-Q_i(V)\right)dV \tag{2.4}$$

Its integration over the inversion layer gives the inversion charge per unit gate area, Q_i:

$$Q_i(y)=-q\int_0^{x_i} n(x,y)dx. \tag{2.5}$$

$$I_{ds}=q\mu_{eff}\frac{W}{L}\int_0^{V_{ds}}\left(\int_{\delta}^{\psi_s}\frac{n_i e^{q(\psi-V)/kT}}{\mathscr{E}(\psi,V)}d\psi\right)dV. \tag{2.6}$$

This is referred to as *Pao and Sah's double integral* (Pao and Sah, 1966). The boundary value ψ_s is determined by two coupled equations: $V_g-V_{fb}=V_{ox}+\psi_s=\frac{-Q_s}{C_{ox}}+\psi_s$ and $Q_s=-\varepsilon_{si}\mathscr{E}_s(\psi_s)$ or Gauss's law, where $\mathscr{E}_s(\psi_s)$ is obtained by letting $\psi=\psi_s$ in the equation:

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

In depletion and inversion where $q\psi_s/kT \gg 1$, only two of the terms in the above equation are significant and need to be kept. The merged equation is then

$$V_{gs}=V_{fb}+\psi_s-\frac{Q_s}{C_{ox}}=V_{fb}+\psi_s+\frac{\sqrt{2\varepsilon_{si}kTN_a}}{C_{ox}}\left[\frac{q\psi_s}{kT}+\frac{n_i^2}{N_a^2}e^{q(\psi_s-V)/kT}\right]^{1/2}, \tag{2.7}$$

which is an implicit equation for $\psi_s(V)$. Equations (2.6) and (2.7) can only be solved numerically.

Charge Sheet Model

Pao and Sah's double integral can be simplified to a single integral if the inversion charge density Q_i can be expressed as a function of ψ_s. This is the approach taken by the *charge-sheet model* (Brews, 1978). It is based on the fact that the inversion layer is located very close to the silicon surface like a thin sheet of charge. There is a sharp increase of the field (spatial integration of the volume charge density) across the thin inversion layer, but very little change of the potential (spatial integration of the field). The central assumption of the charge-sheet model is that Eq. (2.8) for the depletion charge density,

$$Q_d = -qN_aW_d = -\sqrt{2\varepsilon_{si}qN_a\psi_s}, \tag{2.8}$$

can be extended to strong inversion and beyond. Since the total silicon charge density Q_s is given by Eq. (2.7) or $V_g - V_{fb} = V_{ox} + \psi_s = \frac{-Q_s}{C_{ox}} + \psi_s$, Eq. (2.8) allows the inversion charge density to be expressed as

$$Q_i = Q_s - Q_d = -C_{ox}(V_{gs} - V_{fb} - \psi_s) + \sqrt{2\varepsilon_{si}qN_a\psi_s}. \tag{2.9}$$

The above is plotted in Fig. 2.2 for a fixed V_{gs}. Note from Eq. (2.4) that the drain current is proportional to the area under the $|Q_i(V)|$ curve between $V = 0$ and V_{ds}. When V_{ds} is small (linear region), the inversion charge density at the drain end of the channel is only slightly lower than that at the source end. As the drain voltage increases (for a fixed gate voltage), the area or current increases, but the inversion charge density at the drain decreases until finally it goes to zero when $V_{ds} = V_{dsat} = (V_{gs} - V_t)/m$. At this point, I_{ds} reaches its maximum value, I_{dsat} of $I_{ds} = I_{dsat} = \mu_{eff}C_{ox}\frac{W}{L}\frac{(V_{gs}-V_t)^2}{2m}$.

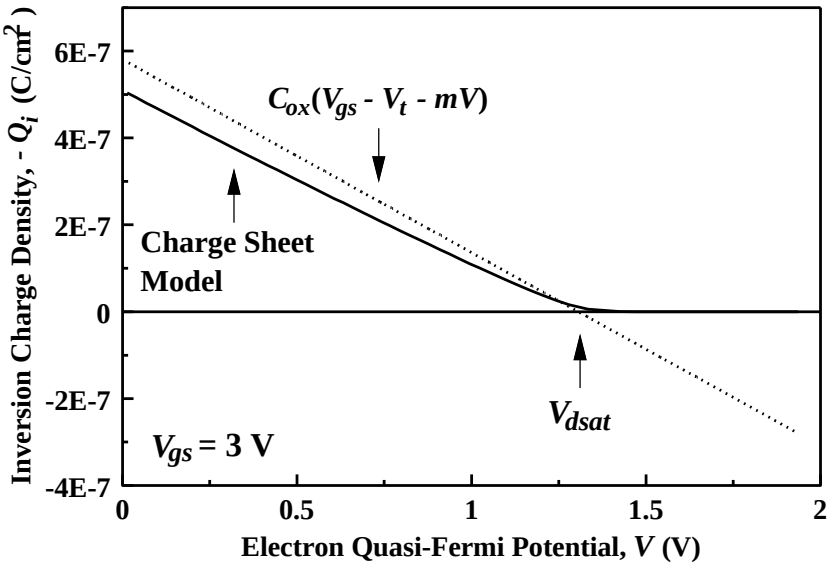


Figure 2.2 Inversion charge density as a function of the quasi-Fermi potential V. The solid curve is generated from the charge sheet model.

Also plotted in Fig. 2.2 is the continuous $-Q_i(V)$ curve of the charge sheet model generated by numerically solving the implicit Eq. (2.7) for $\psi_s(V)$, then calculating $Q_i(\psi_s)$ from Eq. (2.9). At $V = 0$, $-Q_i$ is slightly lower than $C_{ox}(V_{gs} - V_t)$ due to the inversion layer capacitance effect discussed in the last subsection. Instead of $-Q_i = 0$ at $V = V_{dsat}$ then going negative as in the piecewise model, $-Q_i$ of the charge sheet model approaches 0 continuously as $V \rightarrow \infty$. This means that I_{ds}, proportional to the area under the charge sheet $-Q_i(V)$, converges continuously to the saturation value as V_{ds} becomes >> V_{dsat}.

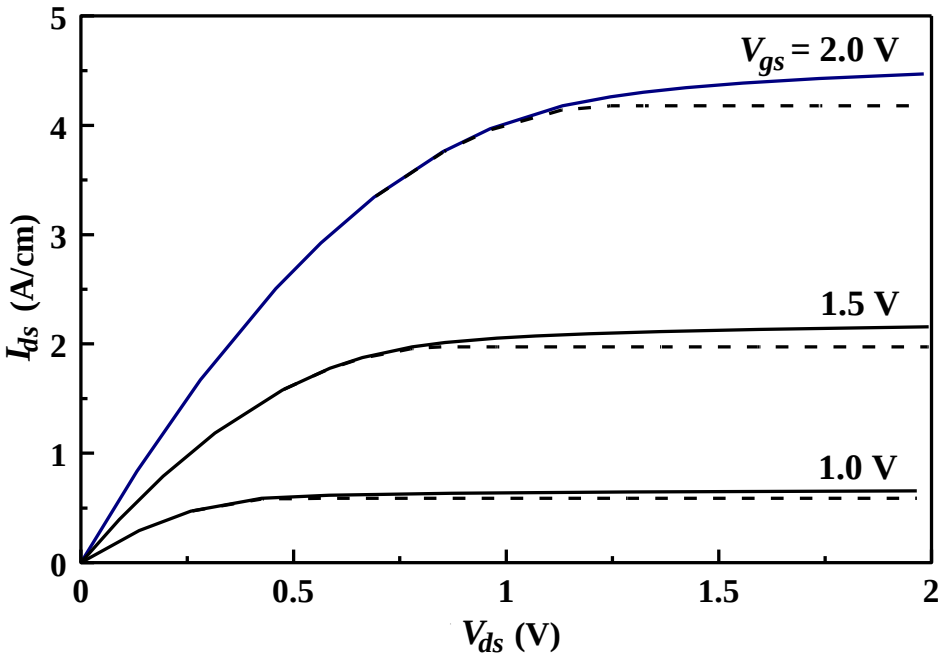


Figure 2.3 I_{ds}-V_{ds} curves (solid) solved from $\frac{I_{ds}}{\mu_{eff}W}y = C_{inv}\left[(V_{gs}-V_t)V - \frac{m}{2}V^2\right] + \frac{\varepsilon_{si}d_{si}}{2}\left[\left(\frac{dV}{dy}\right)^2 - E_0^2\right]$ for the device. The dashed curves are from the GCA model for which currents saturate at I_{dsat}.

2.1.2 DG MOSFETs, Constant Mobility

$$\frac{\partial^2\psi}{\partial x^2} + \frac{\partial^2\psi}{\partial y^2} = \frac{q}{\varepsilon_{si}}n(x,y) = \frac{q}{\varepsilon_{si}}n_i e^{q(\psi-V)/kT} \tag{2.10}$$

and the current continuity eq.,

$$\frac{\partial J_x}{\partial x} + \frac{\partial J_y}{\partial y} = \frac{\partial}{\partial x}\left(qn\mu\frac{\partial V}{\partial x}\right) + \frac{\partial}{\partial y}\left(qn\mu\frac{\partial V}{\partial y}\right) = 0\,, \tag{2.11}$$

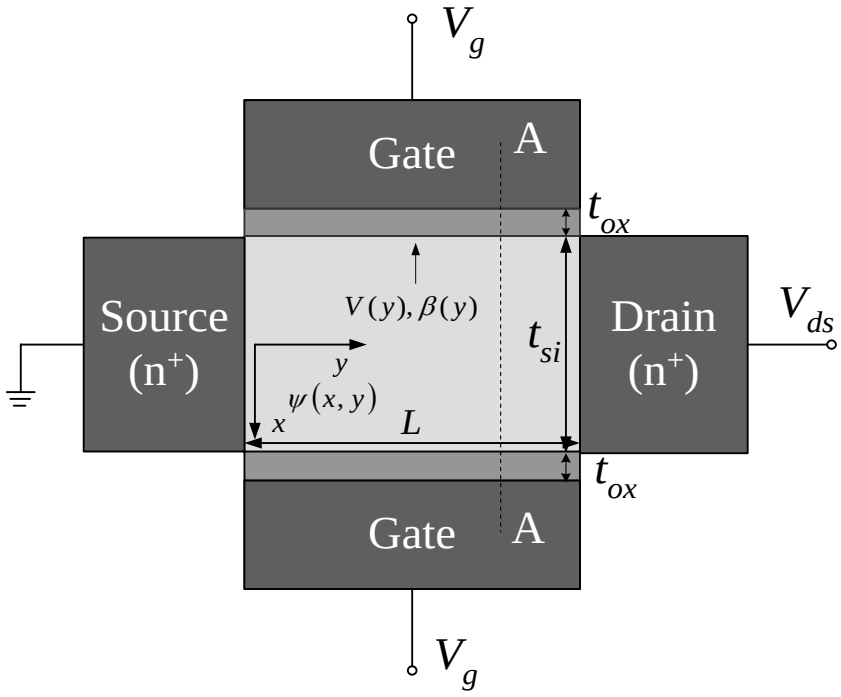


Figure 2.4 Schematic diagram of a double-gate MOSFET. $V(y)$ is the quasi-Fermi potential at a point in the channel. β is a function of V.

where ψ is the potential, n is the carrier density, n_i the intrinsic carrier concentration, V the Fermi potential, and J_x, J_y are the current densities. Fig. 1 shows the geometry of a symmetric double-gate MOSFET. For a lightly doped body, the fixed charge density is negligible. Since the current is predominantly in the source-to-drain or y-direction, V is essentially a function of y only, independent of x. Eq. (2) is then simplified to

$$J_y = qn\mu \frac{dV}{dy} = \text{constant}, \tag{2.12}$$

independent of y. J_y can be integrated in the x-direction to yield the total source-to-drain current:

$$I_{ds} = \mu W Q_i \frac{dV}{dy}. \tag{2.13}$$

Here, μ is the mobility, W is the device width, and Q_i is the mobile charge density per area equal to the integral of $q \times n$ over thickness (x).

In order to solve the coupled Poisson's eq. and the current continuity eq. in 1-D slices in the *x*-direction, GCA is invoked that assumes $\partial^2\psi/\partial y^2 << \partial^2\psi/\partial x^2$ so that Eq. (1) is reduced to

$$\frac{\partial^2\psi}{\partial x^2} = \frac{q}{\varepsilon_{si}} n_i e^{q(\psi - V)/kT} . \tag{2.14}$$

With the condition $\partial\psi/\partial x = 0$ at $x = 0$ for symmetric DG MOSFETs, the solution to Eq. (5) takes the general form [8]:

$$\psi(x) = V + \frac{2kT}{q} \ln\left[\sqrt{\frac{8\varepsilon_{si} kT}{q^2 n_i t_{si}^2}} \frac{\beta}{\cos(2\beta\, x / t_{si})} \right], \tag{2.15}$$

where β is a constant of *x*, but a function of *y*. For every $y \in (0, L)$, ψ satisfies the condition

$$\varepsilon_i \frac{V_{gs} - (\phi_m - \chi - E_g / 2q) - \psi(x = \pm t_{si} / 2)}{t_i} = \pm\varepsilon_{si} \left.\frac{\partial\psi}{\partial x}\right|_{x=\pm t_{si}/2} \tag{2.16}$$

at the silicon-oxide interface. Here, ϕ_m is the gate work function and χ is the electron affinity of silicon. Substitution of Eq. (6) in Eq. (7) yields a relation between *V* and β,

$$V_{gs} - V_t - V = \frac{2kT}{q} \left[\ln\beta - \ln(\cos\beta) + 2\frac{\varepsilon_{si} t_i}{\varepsilon_i t_{si}} \beta \tan\beta \right], \tag{2.17}$$

where

$$V_t \equiv \phi_m - \chi - \frac{E_g}{2q} + \frac{2kT}{q} \ln\sqrt{\frac{8\varepsilon_{si} kT}{q^2 n_i t_{si}^2}} . \tag{2.18}$$

Both *V* and β are functions of *y*. Eq. (8) gives their one-to-one correspondence for a fixed V_{gs}. The use of the intermediary parameter β allows explicit expressions of charge, potential, and field at any point in the channel. For example,

$$Q_i = 2\varepsilon_{si} \left.\frac{\partial\psi}{\partial x}\right|_{x=\pm t_{si}/2} = 8\frac{kT}{q}\frac{\varepsilon_{si}}{t_{si}} \beta\tan\beta , \tag{2.19}$$

and

$$\psi(x) = V_{gs} - V_t + \frac{2kT}{q}\left[ln\left(\sqrt{\frac{8\varepsilon_{si}kT}{q^2 n_i t_{si}^2}}\frac{cos\,\beta}{cos(2\beta\, x/t_{si})}\right) - \frac{2\varepsilon_{si}t_i}{\varepsilon_i t_{si}}\beta\, tan\,\beta\right] \quad (2.20)$$

The current continuity Eq. (4) can then be integrated with respect to β to obtain the source-drain current:

$$I_{ds} = \mu\frac{W}{L}\int_0^{V_{ds}} Q_i dV = \mu\frac{W}{L}\int_{\beta_s}^{\beta_d} Q_i(\beta)\frac{dV}{d\beta}d\beta = \mu\frac{W}{L}\frac{4\varepsilon_{si}}{t_{si}}\left(\frac{2kT}{q}\right)^2\left[\beta\tan\beta - \frac{\beta^2}{2} + \frac{\varepsilon_{si}t_i}{\varepsilon_i t_{si}}\beta^2\tan^2\beta\right]\Bigg|_{\beta_d}^{\beta_s}. \quad (2.21)$$

β_s is the solution to Eq. (8) for $V = 0$, and β_d is the solution to Eq. (8) for $V = V_{ds}$. Fig. 2 shows the I_{ds}-V_{ds} characteristics generated by this model compared to TCAD simulation.

2.2 GCA Model under Velocity Saturation

2.2.1 Bulk MOSFETs, velocity saturation

n = 1 Velocity Saturation

The GCA model approach is discussed first. We replace the low-field drift velocity, μ_{eff} dV/dy, in $I_{ds}(y) = -\mu_{eff}W\frac{dV}{dy}Q_i(y) = -\mu_{eff}W\frac{dV}{dy}Q_i(V)$ with $v = \frac{\mu_{eff}dV/dy}{1+(\mu_{eff}/v_{sat})dV/dy}$ to obtain:

$$I_{ds} = -WQ_i(V)\frac{\mu_{eff}\,dV/dy}{1+(\mu_{eff}/v_{sat})dV/dy}. \quad (2.22)$$

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

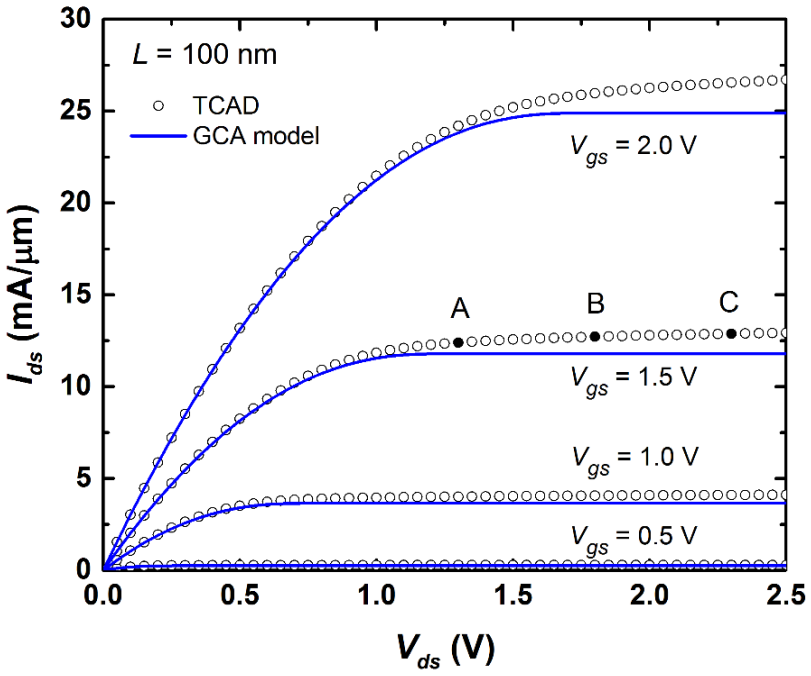


Figure 2.5 GCA model generated I_{ds}-V_{ds} characteristics compared to TCAD (μ = 200 cm^2/V-s). The MOSFET structural parameters are shown in Fig. 1. SCE is negligible in this case.

Equation (2.22) can be rearranged to yield

$$I_{ds} = -\left(\mu_{eff} W Q_i(V) + \frac{\mu_{eff} I_{ds}}{v_{sat}} \right) \frac{dV}{dy}. \tag{2.23}$$

After multiplying dy to the LHS, the above can be integrated from $y = 0$ to L and from $V = 0$ to V_{ds} to solve I_{ds}:

$$I_{ds} = \frac{-\mu_{eff} (W/L) \int_0^{V_{ds}} Q_i(V) dV}{1 + (\mu_{eff} V_{ds} / v_{sat} L)}. \tag{2.24}$$

The numerator is simply the long-channel current, Eq. (2.4), without velocity saturation. It is clear that if the "average" field along the channel, V_{ds}/L, is much less than the critical field $\mathcal{E}_c = v_{sat}/\mu_{eff}$, the drain current is hardly affected by velocity saturation. When V_{ds}/L becomes comparable to or

greater than $\mathcal{E}_c$, however, the drain current is significantly reduced. A convenient, approximate expression for $Q_i(V)$ is Eq. (2.25):

$$-Q_i(V) = C_{inv}(V_{gs} - V_t - mV), \tag{2.25}$$

where $C_{inv}(V_{gs} - V_t)$, is given by Eq. (2.9) of the charge sheet model with $\psi_s = \psi_{s,s}$ for $V = 0$. The integration in Eq. (2.24) can then be carried out to yield

$$I_{ds} = \frac{\mu_{eff} C_{inv}(W/L)[(V_{gs} - V_t)V_{ds} - (m/2)V_{ds}^2]}{1 + (\mu_{eff} V_{ds} / v_{sat} L)}. \tag{2.26}$$

For a given V_{gs}, I_{ds} increases with V_{ds} until a maximum current is reached. The saturation voltage, V_{dsat}, is found by solving $dI_{ds}/dV_{ds} = 0$. To compact the equations, a dimensionless parameter

$$z \equiv \frac{2\mu_{eff}(V_{gs} - V_t)}{m v_{sat} L} \tag{2.27}$$

is introduced. It is a measure of the severity of velocity saturation. Then,

$$V_{dsat} = \frac{2(V_{gs} - V_t)/m}{1 + \sqrt{1 + 2\mu_{eff}(V_{gs} - V_t)/(m v_{sat} L)}} \equiv \frac{L v_{sat}}{\mu_{eff}} \left(\sqrt{1+z} - 1 \right) \tag{2.28}$$

This expression is always less than the long-channel saturation voltage, $(V_{gs} - V_t)/m$. Substituting Eq. (2.28) into Eq. (2.26), we find the saturation current,

$$I_{dsat} = C_{inv} W v_{sat} (V_{gs} - V_t) \frac{\sqrt{1+z} - 1}{\sqrt{1+z} + 1}. \tag{2.29}$$

For $z << 1$, Eq. (2.29) is reduced to the long-channel saturation current,

$$I_{dsat} = \mu_{eff} C_{inv} \frac{W}{L} \frac{(V_{gs} - V_t)^2}{2m}. \tag{2.30}$$

For $z >> 1$, Eq. (2.29) becomes the velocity-saturation-limited current,

$$I_{dsat} = C_{inv} W v_{sat} (V_{gs} - V_t). \tag{2.31}$$

Note that in this limit, I_{dsat} is independent of channel length L and varies linearly with $V_{gs} - V_t$ instead of quadratically as in the long-channel case.

At the saturation point, $V(y = L) = V_{dsat}$. It can be shown that

$$I_{dsat} = C_{inv} W v_{sat} (V_{gs} - V_t - m V_{dsat}) = -W v_{sat} Q_i (y = L). \tag{2.32}$$

In other words, carriers at the drain are traveling at the saturation velocity, which means $dV/dy \rightarrow \infty$ in $v = \frac{\mu_{eff} dV/dy}{1+(\mu_{eff}/v_{sat})dV/dy}$. Note that $-Q_i$ of Eq. (2.25) is positive at this point. The commonality between the current saturation in the case of constant mobility and in the case of velocity saturation is therefore not $-Q_i \rightarrow 0$, but the divergence of dV/dy under the GCA model.

For $V_{ds} > V_{dsat}$, the GCA model breaks down.

n = 2 Velocity Saturation

It has been known that the $n = 1$ velocity saturation model has a discontinuity problem with the 2nd order derivative around $V_{ds} = 0$ because the dV/dy factor in the denominator of Eq. (6.41) should in fact be $|dV/dy|$ to keep it always positive (Joardar *et al.*, 1998).

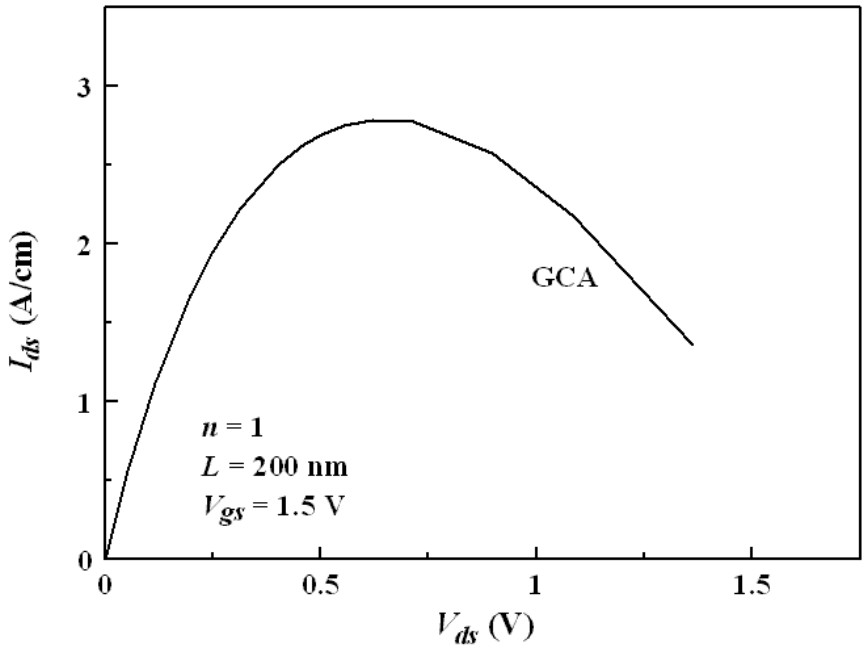


Figure 2.6 I_{ds}-V_{ds} characteristics generated by the GCA models under the n = 1 velocity saturation model. The MOSFET parameters are t_{inv} = 3.3 nm, N_a = 10^{18} cm^{-3} (uniform), n$^+$ silicon gate work function, so V_t = 0.4 V and m = 1.28. Other parameters are μ_{eff} = 200 cm^2/V-s, v_{sat} = 10^7 cm/s, and d_{si} = 20 nm.

To satisfy the continuity requirement, n needs to be an even integer. The least of which is 2. For the GCA model with n = 2 velocity saturation, Eq. (2.22) becomes

$$I_{ds} = -WQ_i(V)\frac{\mu_{eff}\,(dV/dy)}{\sqrt{1+(\mu_{eff}/v_{sat})^2(dV/dy)^2}}. \tag{2.33}$$

It can be re-arranged to yield an integral equation between I_{ds} and V_{ds} for a given V_{gs},

$$LI_{ds} = \mu_{eff}\int_0^{V_{ds}} \sqrt{[WQ_i(V)]^2 - (I_{ds}/v_{sat})^2}\,dV. \tag{2.34}$$

With $Q_i(V)$ of Eq. (2.25), the above integral can be carried out by transforming V to an intermediary variable u,

$$WC_{inv}(V_{gs} - V_t - mV) = (I_{ds}/v_{sat})\cosh u. \tag{2.35}$$

Then,

$$L = \frac{\mu_{eff} I_{ds}}{2mWC_{inv} v_{sat}^{\,2}} \left[\sinh u \cosh u - u \right]_{u_d}^{u_s}, \tag{2.36}$$

where u_s and u_d are given by $WC_{inv}(V_{gs} - V_t - mV) = (I_{ds}/v_{sat})\, cosh\, u$, with V = 0 and V_{ds}, respectively.

The I_{ds}-V_{ds} curve generated for a fixed V_{gs} is shown in Fig. 2.7. There is a maximum V_{ds} = V_{dsat} where I_{ds} reaches a peak value I_{dsat} beyond which no solution exists. This corresponds to u_d = 0 where the factor in the square root of Eq. (2.34) is zero, meaning carriers are traveling at v_{sat} and $dV/dy \rightarrow \infty$. The peak current is

$$I_{dsat} = WC_{inv}(V_{gs} - V_t) v_{sat} / \cosh u_s, \tag{2.37}$$

where u_s (for the peak point) is solved by the implicit equation,

$$L = \frac{\mu_{eff}(V_{gs} - V_t)}{2mv_{sat}} \left[\sinh u_s - \frac{u_s}{\cosh u_s} \right]. \tag{2.38}$$

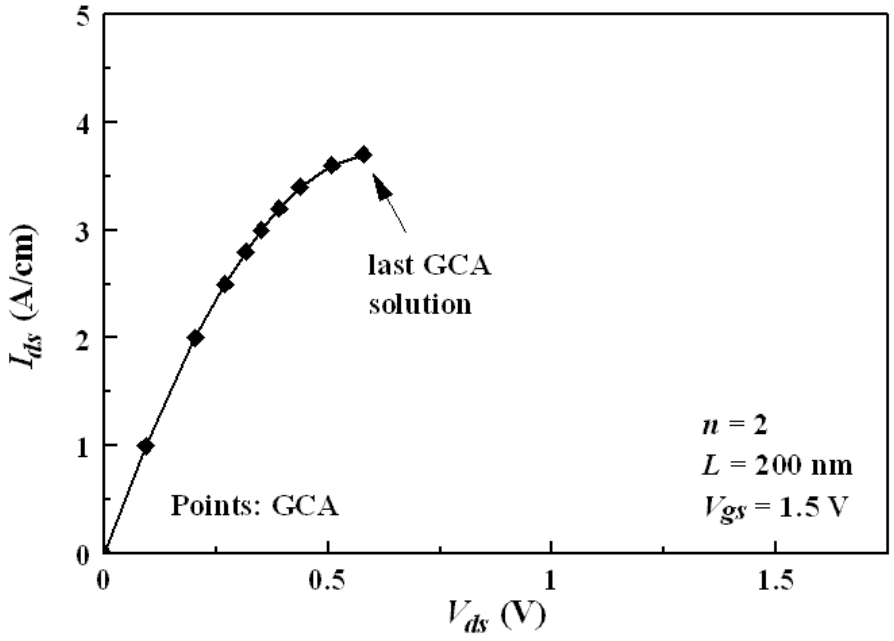


Figure 2.7 I_{ds}-V_{ds} characteristics generated by the GCA and under the n = 2 velocity saturation relation. The device parameters are the same as those described in the caption to Fig. 2.6.

2.2.2 DG MOSFETs, velocity saturation

Fig. 2.8 and 2.9 plot the gradient of Fermi potential dV/dy for n=1 and n=2 cases at $y = L$, i.e., the drain end versus V_{ds}. At the current peak in the GCA model, $dV/dy \rightarrow \infty$ and $v = v_{sat}$. Past the peak, $dV/dy < 0$, clearly unphysical.

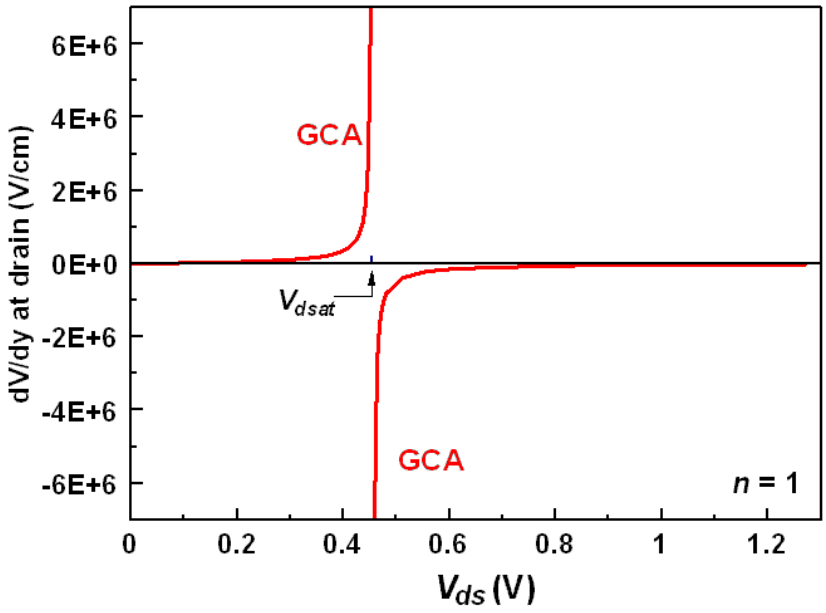


Figure 2.8 dV/dy at the drain ($y = L$) versus V_{ds} for the case in Fig. 2.8. Labels above the non-GCA curve indicate the carrier velocity at those bias points.

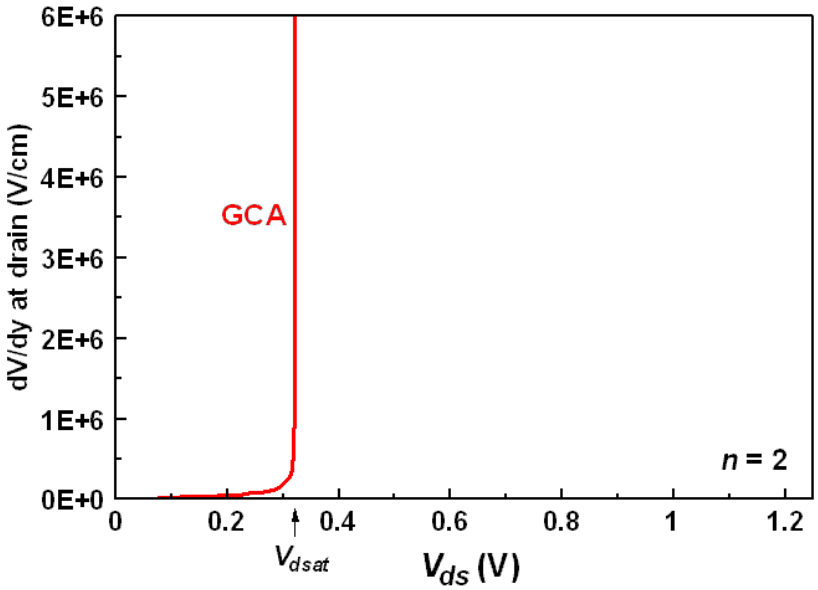


Figure 2.9 *dV*/*dy* at the drain (*y* = *L*) versus V_{ds} for the case in Fig. 2.8. Labels above the non-GCA curve indicate the carrier velocity at those bias points.

References:

[1] Yuan Taur and Huang-Hsuan Lin, "Modeling of DG MOSFET I-V Characteristics", *IEEE Trans. Electron Device*, pp. 1714-1720, April 2018

[2] Yuan Taur, Tak H Ning, "Fundamentals of modern VLSI devices", *Cambridge university press*, December 2021

CHAPTER 3 HISTORY OF MODELING THE MOSFET SATURATION REGION

3.1 Reddi and Sah's Concept of Pinch-off in Metal-Oxide-Semiconductor Transistor (MOST)

As the drain voltage is increased beyond V_{DS}, the length of the pinch-off region will widen resulting in a decrease of the effective channel length; this in turn will cause the drain current to increase. This is one of the causes for finite source to drain incremental resistance for $V_D > V_{DS}$. This effect in a way is analagous to the Early effect in bipolar transistors.

The channel shrinkage ***(ΔL)*** can be approximated by

$$\Delta L = [\,2\varepsilon(V_d - V_{ds})/qN_A]^{1/2} \tag{3.1}$$

Early effect in BJT is a 1-D effect, not the 2-D effect with MOSFET saturation.

The channel region of MOSFET has mobile charge, unlike the depletion region of a p-n junction.

Thus, for $V_D > V_{DS}$, I_D can be expressed as

$$I_{D\,V_D>V_{DS}} = \frac{I_{DS}}{(1-\frac{\Delta L}{L})} = \frac{LI_{DS}}{L-\left\{\frac{2\varepsilon}{qN_A}(V_d-V_{ds})\right\}^{1/2}} \tag{3.2}$$

In a DG MOSFET, there is no doping hence $N_A = 0 \rightarrow$ clearly does not work.

It is worth noting that in some works, e.g., [11], a CLM of $\Delta L = \sqrt{2\varepsilon_s (V_{ds} - V_{dsat})/qN_a}$ is used, derived from the widening of the space charge region due to drain voltage. The physics is analogous to the finite output conductance in the forward active region of a bipolar junction transistor (BJT), or Early effect. This clearly does not apply to saturation in MOSFETs. Early effect in BJT is due to encroachment of the base-collector depletion region into the neutral base region. It is a 1-D phenomenon at moderate fields involving the fixed dopant charge in the base region. MOSFET saturation, on the other hand, has to do with the 2-D nature of the device. It happens at high fields and involves only the mobile charge. One factor in common between BJT and MOSFET is that the output conductance in the active or the saturation region goes up with thinner base width or shorter channel length (before SCE kicks in).

3.2 El-Mansy and Boothroyd's Two-Dimensional Model in the Saturation Region

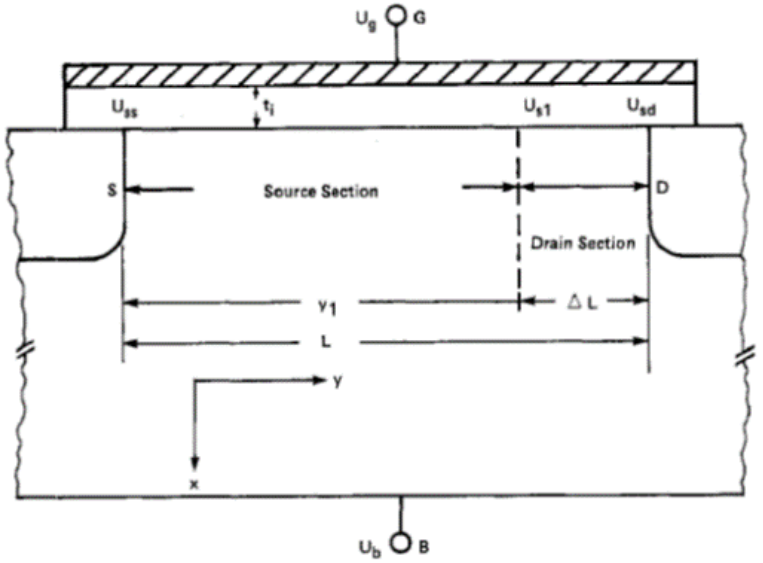


Figure 3.1 A schematic cross section of an IGFET to illustrate source and drain section and axes.

THE MODEL

The model developed here is based on the second approach in the previous discussion, i.e., dividing the space-charge region along the channel into two sections. In the source section the GCA is valid and any of the various models available [6], [9], [12] for this section can be used. In the drain section the two-dimensional nature of the potential distribution is accounted for. Condition for the validity of the GCA is

$$\frac{\partial^2 U}{\partial x^2} \Big/ \frac{\partial^2 U}{\partial y^2} \geq K \tag{3.3}$$

where *K* is a large number (note that exact validity of the GCA corresponds to an infinite value for *K)*.

Use an empirical criterion to divide the MOSFET channel into a source section (where GCA works) and a drain section which they analyze later.

The expression

$$U_{s1} = U_{ss} + \frac{U_{sp} - U_{ss}}{1 + F \cdot (t_i/L)} \tag{3.4}$$

was found to yield results close to those obtained from numerically solving (la) for a wide range of device parameters and applied voltages. The factor *F* is, in general, a slowly increasing function of gate voltage, and different values for it may be needed for different voltage ranges.

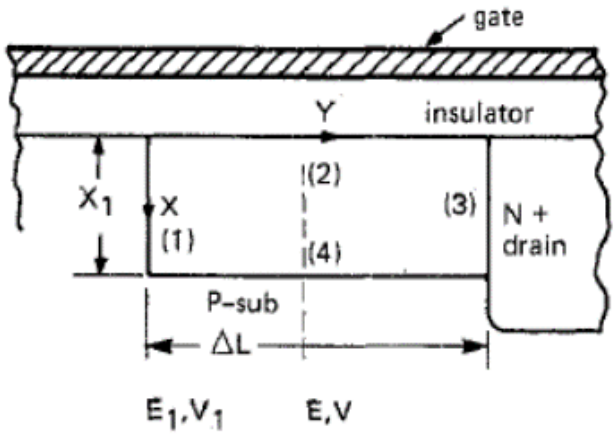


Figure 3.2 Definition of the drain section boundaries.

Boundary 1): This is the boundary separating the source and drain sections. The potential and field at this boundary are defined from the solution in the source section, while the physical location of that boundary along the surface is defined from the solution in the drains section.

Boundary 2): This is the semiconductor-insulator interface.

Boundary 3): This is the drain metallurgical junction.

Boundary 4): This is located at a distance x_1 from the surface. The potential and electric field at this boundary are assumed to be zero.

$$\frac{dI_d}{dU_{sd}} = \frac{1}{L_e} \cdot \frac{1}{(-E_{sd})} \tag{3.5}$$

has to be solved numerically. This is a first order differential equation, and any of the standard methods can be used to solve it. Of course the boundary condition for solving the equation is $\boldsymbol{I_d = I_1}$ when $\boldsymbol{U_{sd} = U_{S1}}$.

3.3 Ping Ko's PhD Thesis

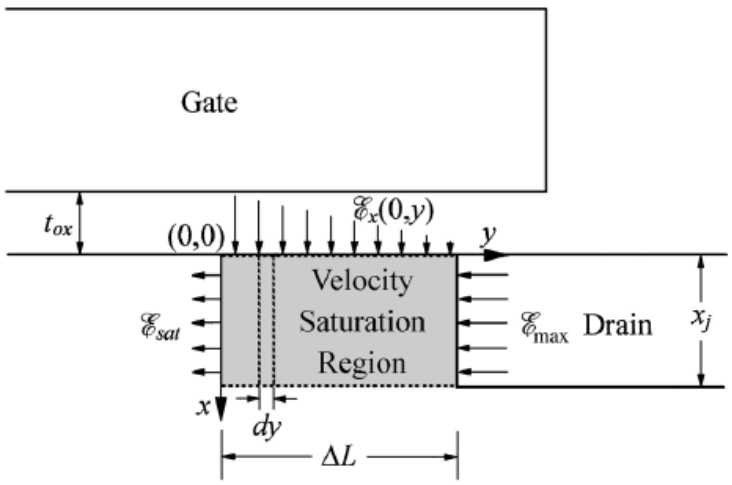


Figure 3.3 A closer look into the cross section of a MOSFET near drain

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

$$V_g - V_{fb} = V_{ox}(y) + \psi_s(y) = V_{ox}(y) + 2\psi_B + V(y) \tag{3.6}$$

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

$$\mathcal{E}_x(0,y) = \frac{\varepsilon_{ox}}{\varepsilon_{si}}\mathcal{E}_x(y) = \frac{\varepsilon_{ox}}{\varepsilon_{si}}\frac{V_{ox}(y)}{t_{ox}} \tag{3.7}$$

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

$$\mathcal{E}_x(0,0) = \frac{qN_a x_j + Q_i(y=0)}{\varepsilon_{si}}, \tag{3.8}$$

where Q_i (> 0) is the mobile (electron) charge density per unit area. It is assumed here that the junction depth x_j is comparable to the depletion width W_{dm}.

Similar to El-Mansy and Boothroyd, also apply 2-D Gauss' law to the velocity saturation region near the drain.

Carriers are already traveling at velocity such that $I_{ds} = WQ_i v_{sat}$, the mobile charge density,

$$Q_i(y) = q\int_0^{x_j} n(x,y)dx, \tag{3.9}$$

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

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

$$\mathcal{E}_x(0,y)dy - \mathcal{E}_y(y+dy)x_j + \mathcal{E}_y(y)x_j = \frac{qN_a x_j dy + Q_i(y)dy}{\varepsilon_{si}} \tag{3.10}$$

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

$$-x_j\frac{d\mathcal{E}_y}{dy}=\mathcal{E}_x(0,0)-\mathcal{E}_x(0,y). \tag{3.11}$$

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

$$\mathcal{E}_x(0,0)-\mathcal{E}_x(0,y)=\frac{\varepsilon_{ox}}{\varepsilon_{si}t_{ox}}[V_{ox}(0)-V_{ox}(y)]=\frac{\varepsilon_{ox}}{\varepsilon_{si}t_{ox}}[V(y)\ -V(0)\ ]. \tag{3.12}$$

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

$$\frac{d^2V}{dy^2}=\frac{\varepsilon_{ox}}{\varepsilon_{si}t_{ox}x_j}[V(y)-V_{dsat}], \tag{3.13}$$

or

$$\frac{d^2V}{dy^2}=\frac{V(y)-V_{dsat}}{l^2}, \tag{3.14}$$

Where the characteristic length l is given by

$$l=\sqrt{\frac{\varepsilon_{si}}{\varepsilon_{ox}}t_{ox}x_j}\approx\sqrt{3t_{ox}x_j}. \tag{3.15}$$

Did not realize that the vertical field in oxide goes through zero and becomes negative close to the drain.

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

$$V(y) = V_{dsat} + l\mathcal{E}_{sat} sinh\left(\frac{y}{l}\right). \tag{3.16}$$

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

Peak Field at the Drain

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

$$\Delta L = lln\left[\frac{V_{ds} - V_{dsat}}{l\mathcal{E}_{sat}} + \sqrt{\left(\frac{V_{ds} - V_{dsat}}{l\mathcal{E}_{sat}}\right)^2 + 1}\right]. \tag{3.17}$$

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

$$\mathcal{E}_y(y) = -\frac{dV}{dy} = -\mathcal{E}_{sat} cosh(\tfrac{y}{l}), \tag{3.18}$$

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

$$\mathcal{E}_{max} \equiv \mathcal{E}_y(y = \Delta L) = -\sqrt{\left(\frac{V_{ds}-V_{dsat}}{l}\right)^2 + \mathcal{E}_{sat}{}^2}. \quad (3.19)$$

This field can be as high as mid-10^5 to 10^6 V/cm and is responsible for a variety of hot carrier effects such as impact ionization, substrate current, and oxide degradation. In general, all models that partitioned MOSFET into two sections are not continuous from the triode (GCA) region to the saturation region. They cannot predict where the partition point (e.g., E_{sat} above) is.

References:

[1] C. T. Sah, "Characteristics of the metal-oxide-semiconductor transistors," *IEEE Trans. Electron Device*, pp. 324-345, July 1964.

[2] Y. El-Mansy and A. Boothroyd, "A simple two-dimensional model for IGFET operation in the saturation region," *IEEE Trans. Electron Devices*, pp. 254-262, Mar. 1977.

[3] P. K. Ko, R. S. Muller, and C. Hu, "A unified model for hot electron currents in MOSFETs," *1981 IEDM Technical Digest*, pp. 600-603.

[4] Yuan Taur, Tak H Ning, "Fundamentals of modern VLSI devices", *Cambridge university press*, December 2021

CHAPTER 4 NON-GCA MODEL FOR DG MOSFETS

4.1 TCAD Simulations

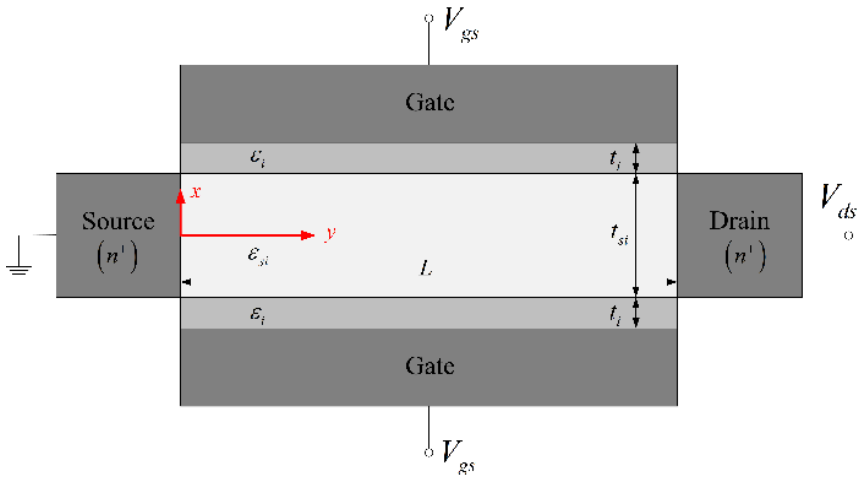


Figure 4.1 Double-gate (DG) MOSFET structure assumed in this work. t_{si} = 4 nm, t_i = 2 nm, ε_{si} = ε_i = 11.8ε_0. The gate work function is such that V_t = 0.33 V.

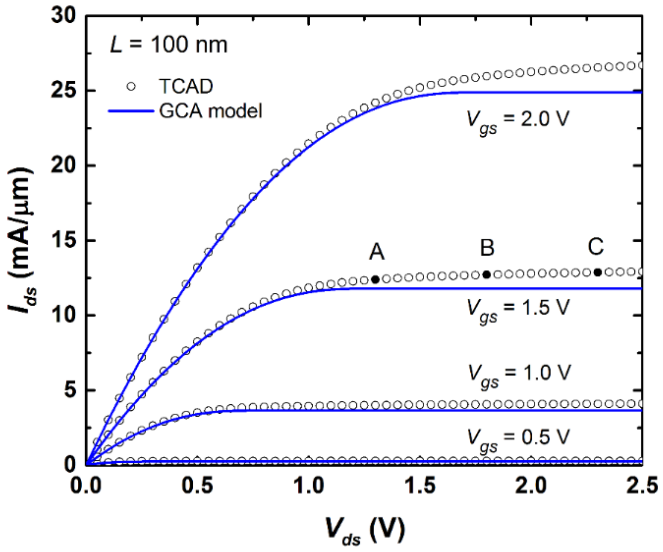


Figure 4.2 GCA model generated I_{ds}-V_{ds} characteristics compared to TCAD (μ = 200 cm^2/V-s). The MOSFET structural parameters are shown in Fig. 4.1. SCE is negligible in this case.

To gain a deeper understanding of the physical picture in the saturation region, we dissect in detail the TCAD simulation case in Fig. 4.2. Drain-induced barrier lowering (DIBL) due to SCE is negligible in this example. The point labeled A at V_{ds} = 1.3 V on the V_{gs} = 1.5 V curve is slightly beyond the saturation voltage of $V_{dsat} = V_{gs} - V_t \approx 1.17$ V. The electron density n near the drain is shown in Fig. 4.3 along several lateral cuts at various depths. It is clear that there is no "*pinchoff*" of channel depicted in the textbooks based on the GCA. Even along the surface (x = 2 nm), the electron density never falls below 10^{19} cm^{-3}. This fact was also pointed out in a 2012 publication with TCAD simulations. It clearly demonstrates the failure of GCA.

Fig. 4.4 goes further by looking at point B in Fig. 4.2, where V_{ds} = 1.8 V on the V_{gs} = 1.5 V curve. Here, we plot the potential ψ versus depth (x) between the gates along three vertical cuts near the drain, at y = 93.2, 95.2, and 97.2 nm. There is a change of sign of the vertical field, $\mathcal{E}_x = -\partial\psi/\partial x$, at y = 95.2 nm. On the source side of 95.2 nm, E_x is such that electrons are attracted toward the gates. This is the normal direction of the field effect that gives rise to "inversion" or turns the device on. However, on the drain side of 95.2 nm, $\mathcal{E}_x$ is such that electrons are repelled from the gates. Thus the so-called "*pinchoff*" point should be interpreted as the point where $\mathcal{E}_x$ changes sign or where $\mathcal{E}_x = 0$. The Fermi potential V at the point of zero oxide field is $V_{dsat} \approx 1.17$ V. It is at y = 95.2 nm in this case while $V(y = 100 \text{ nm}) = V_{ds} = 1.8$ V. However, the channel is not pinched off when $\mathcal{E}_x = 0$. The electron density, also plotted in Fig. 4.4, in the y = 95.2 nm case is above 10^{19} cm^{-3} at every depth. *Channel length modulation* should then be interpreted as the movement of the point of zero oxide field toward the source as the drain voltage goes beyond saturation.

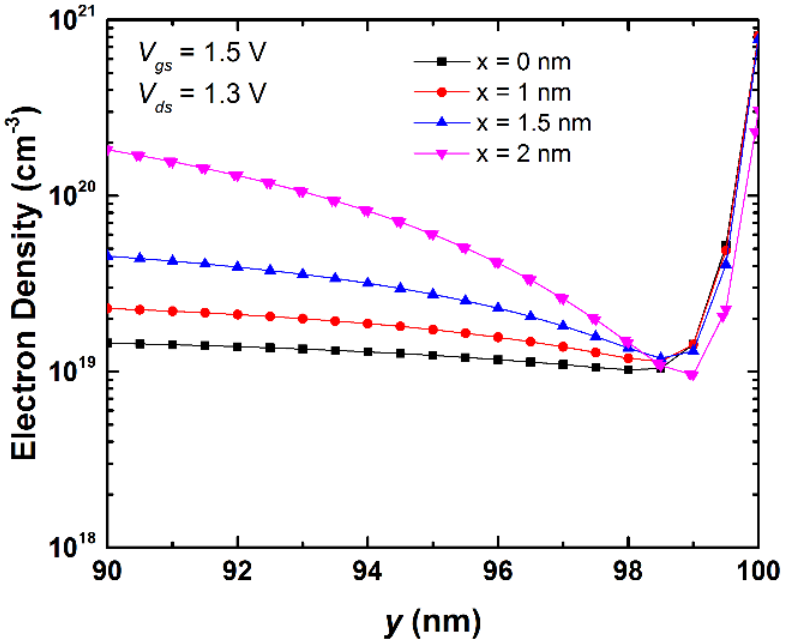


Figure 4.3 From TCAD at bias point A on the V_{gs} = 1.5 V curve in Fig. 4.2: Electron concentration near the drain (y = 100 nm) along several lateral cuts from the surface to the center. The source-drain doping level is 10^{21} cm^{-3}.

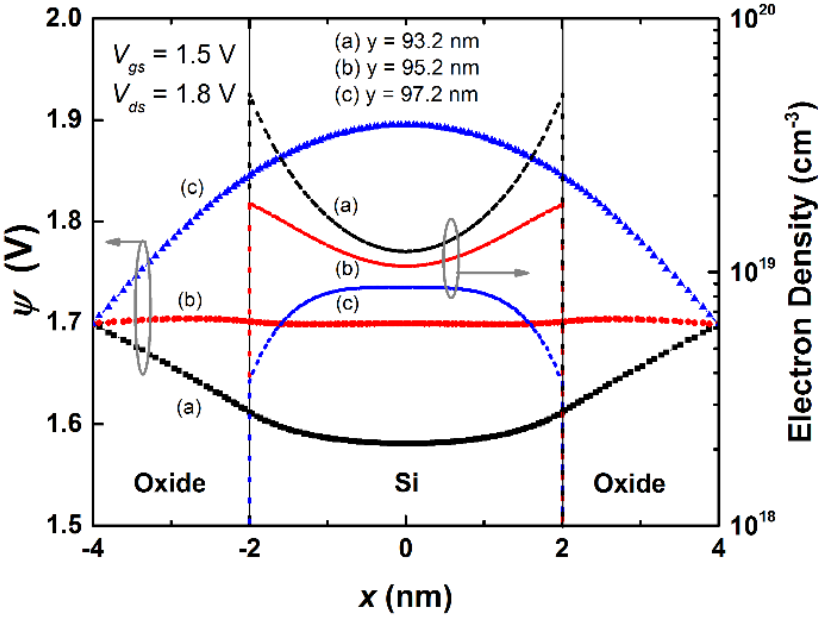


Figure 4.4 From TCAD at bias point B on the V_{gs} = 1.5 V curve in Fig. 4.2: Potential (solid) versus depth along 3 cuts, before the point of saturation (y = 93.2 nm), at the point of saturation (y = 95.2 nm), and after saturation (y = 97.2 nm). Electron density (dotted) in each case is shown using the scale to the right.

When $\mathcal{E}_x = 0$, $\partial^2\psi/\partial x^2$ is also 0 (at y = 95.2 nm). When $\mathcal{E}_x < 0$ (on the negative x side), $\partial^2\psi/\partial x^2$ is also < 0. This clearly contradicts Eq. (2.14) of GCA. The key factor is, of course, that the $\partial^2\psi/\partial y^2$ term in the full 2-D Eq. (2.10) cannot be neglected when biased near and beyond the saturation point.

The current continuity Eq. (2.13) is based on the assumption that the Fermi potential V varies predominantly in the direction of current flow, namely, the y-direction. Fig. 4.5 verifies that this is still a good approximation in the saturation region. The condition of current continuity constrains the product of Q_i and dV/dy to be a constant, independent of y. When biased near or beyond V_{dsat}, Q_i plummets as y moves toward the drain. GCA says that $Q_i \rightarrow 0$ (pinchoff) and $dV/dy \rightarrow \infty$ at the point of saturation. However, when dV/dy increases sharply with y, d^2V/dy^2 also increases and becomes appreciable. It is shown in Fig. 4.6 that dV/dy and $\partial\psi/\partial y$ tend to track each other owing to the fact that the current in this bias region is predominantly a drift current. The $\partial\psi^2/\partial y^2$ term in Eq. (2.10) then makes the electron density n nonzero and positive even though $\partial\psi^2/\partial x^2$ is zero or negative. From this picture, pinchoff never happens. When Q_i is diminishing, current continuity forces dV/dy to go up, which in turn causes $\partial\psi^2/\partial y^2$ to go up and replenishes Q_i. This picture is consistent with the TCAD revelations in Figs. 4.3 and 4.4

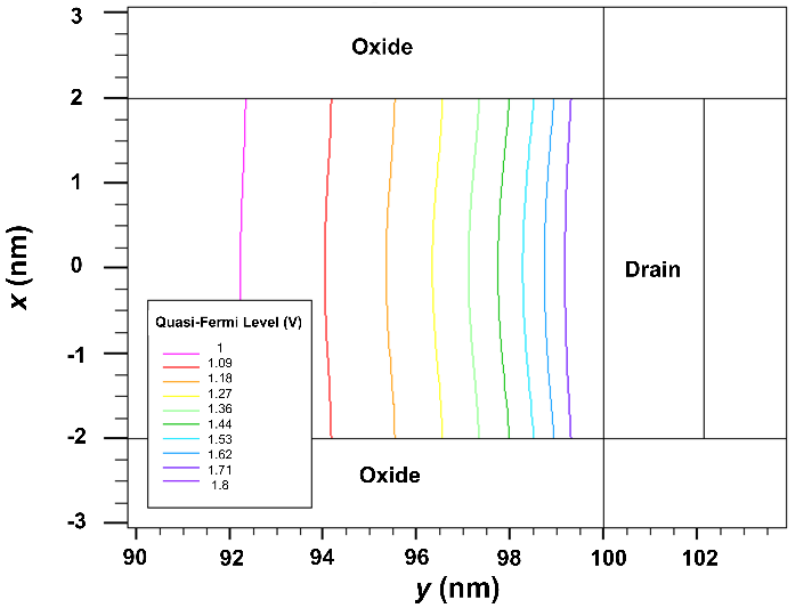


Figure 4.5 From TCAD at bias point B in Fig. 2, V_{gs} = 1.5 V, V_{ds} = 1.8 V: Constant Fermi potential contours near the drain. The most sloped angle between the gradient of $V(x, y)$ and y-axis is 5°, meaning J_y is cos 5° = 0.996 of the total magnitude of J.

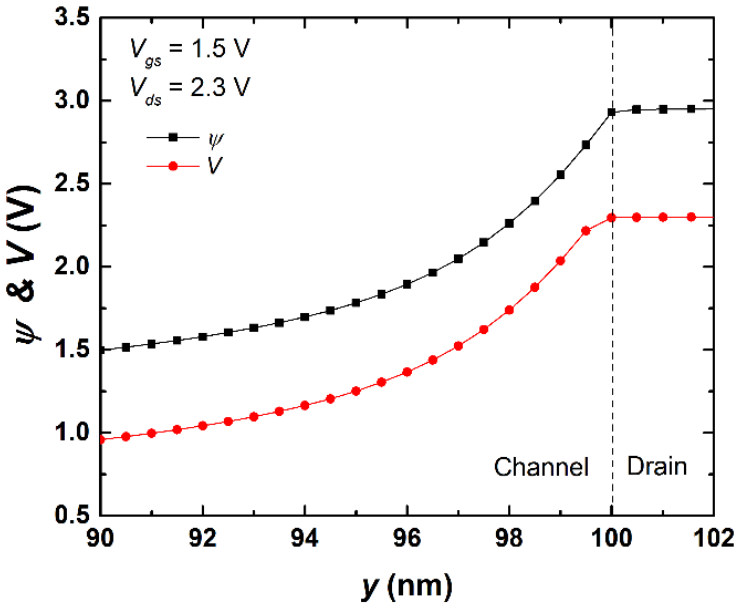


Figure 4.6 From TCAD: Potential and Fermi potential along a cut through the center of silicon. The bias point is labeled C in Fig. 4.2.

Justification of $Q_i(V)$ and C_{inv} Determination

Q_i as a function of V, or Eq. (4.19), is closely examined in this section with the help of TCAD. Based on the analytic potential model for DG MOSFETs, a GCA model, Q_i is given by

$$Q_i = 8\frac{kT}{q}\frac{\varepsilon_{si}}{t_{si}}\beta\tan\beta\,, \tag{4.1}$$

where the intermediary parameter $\beta \in (0, \pi/2)$ is related to V_{gs} and V through an implicit equation,

$$V_{gs} - V_t - V = \frac{2kT}{q}\left[\ln\beta - \ln(\cos\beta) + 2\frac{\varepsilon_{si}t_i}{\varepsilon_i t_{si}}\beta\tan\beta\right], \tag{4.2}$$

with

$$V_t \equiv \phi_m - \chi - \frac{E_g}{2q} + \frac{2kT}{q}\ln\sqrt{\frac{8\varepsilon_{si}kT}{q^2 n_i t_{si}^2}}\,. \tag{4.3}$$

In the above, ϕ_m is the gate work function, χ and n_i are the electron affinity and intrinsic carrier concentration of silicon. An example of Q_i versus V of the above model is shown in Fig. 4.7. Note that when V exceeds $V_{gs} - V_t$ and $\to \infty$, the LHS of Eq. (4.2) is negative and β becomes < 1 or << 1. Q_i exhibits subthreshold behavior, i.e., exponentially approaching zero but staying positive, much like the case when $V_{gs} - V_t < 0$. This is a common fallacy of all GCA models, including the charge sheet model for bulk MOSFETs. While the subthreshold behavior is correct when $V_{gs} - V_t < 0$, it is incorrect in saturation when $V_{gs} - V_t > 0$ but < V.

Stemmed from the $\partial\psi^2/\partial x^2$ term in 2D Poisson's equation, Q_i is the charge induced in the channel by the gate, directly related to the field in the oxide perpendicular to the channel, $\mathcal{E}_x$. As shown in the potential contour plot from TCAD in Fig. 4.8, this field changes sign along the channel: positive between the source and the point where $V \approx V_{gs} - V_t$, and negative beyond it. While $\mathcal{E}_x$ therefore $Q_i < 0$ is not allowed in GCA models, it is perfectly fine with the non-GCA

model in which ΔQ_i of Eq. (4.26) ensures that the total mobile charge density, $Q_i + \Delta Q_i$, stays positive (for current continuity) even when Q_i is negative. Proper modeling of the negative Q_i behavior, however, is important because it causes ΔQ_i to go higher, thereby lowering the output conductance.

The TCAD $Q_i(V)$ curve in Fig. 4.6 is extracted from $\mathscr{E}_x$ close to the gate which by Gauss's law gives the total charge density in the two gate electrodes, $Q_i = 2\varepsilon_i \mathscr{E}_x$. Near and beyond the point where $\mathscr{E}_x$ changes sign, $\mathscr{E}_x$ close to silicon deviates from that close to the gate because the effect of the lateral component, $\mathscr{E}_y$, becomes appreciable. It is worth noting in Fig. 4.7 that the point of $V = V_{dsat}$, beyond which GCA stops working, is unremarkable. This shows that the transition from the GCA region to the velocity saturation region is rather gradual, which suggests that there is no clear cut division of the channel into two distinct regions.

The most elementary form of Q_i,

$$Q_i = 2C_{ox}(V_{gs} - V_t - V) \tag{4.4}$$

where $C_{ox} = \varepsilon_i/t_i$, does capture the negative going behavior for V beyond $V_{gs} - V_t$, consistent with TCAD. But its value at $V = 0$, indicated in Fig. 4.7, is over estimated because the semiconductor capacitance is not taken into account. This Q_i value at the source is of critical importance as the GCA current is directly proportional to it. Here then lies the rationale behind Eq. (4.19): insert a correction factor, C_{inv}/C_{ox}, given by the ratio of Eq. (4.1) to Eq. (4.2) for $\beta = \beta_s$ at $V = 0$, namely,

$$\frac{C_{inv}}{C_{ox}} = \frac{2(\varepsilon_{si}t_i / \varepsilon_i t_{si})\beta_s \tan\beta_s}{\ln\beta_s - \ln(\cos\beta_s) + 2(\varepsilon_{si}t_i / \varepsilon_i t_{si})\beta_s \tan\beta_s}. \tag{4.5}$$

This factor has a slight dependence on V_{gs}. It varies from 0.842 at $V_{gs} = 1.2$ V to 0.734 at $V_{gs} = 0.6$ V in our device example. As can be seen in Fig. 6, $Q_i(V)$ of Eq. (4.19) does not exactly match the

TCAD curve. But it turns out, as shown in Figs. 4.13-14 and Figs 4.17-18 below, the deviation has little or no effect on the I_{ds}-V_{ds} or output conductance characteristics.

While the TCAD $Q_i(V)$ curve in Fig. 4.7 is extracted from the device under the n = 2 velocity saturation model, additional examination reveals similar results under the n = 1 model. It appears that $Q_i(V)$ characteristic is transport independent.

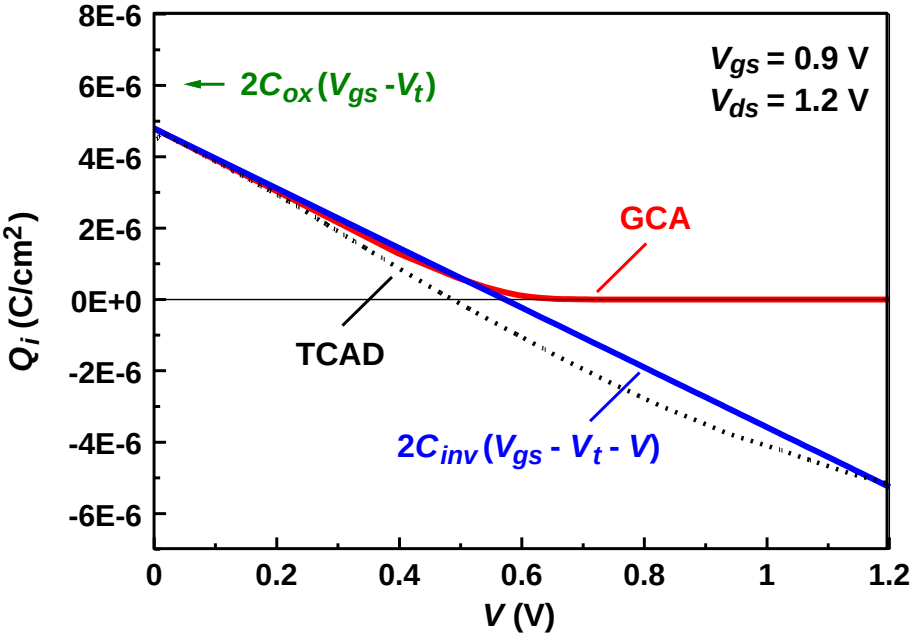


Figure 4.7 Comparison of $Q_i(V)$ from TCAD, GCA model, and Eq. (4.19).

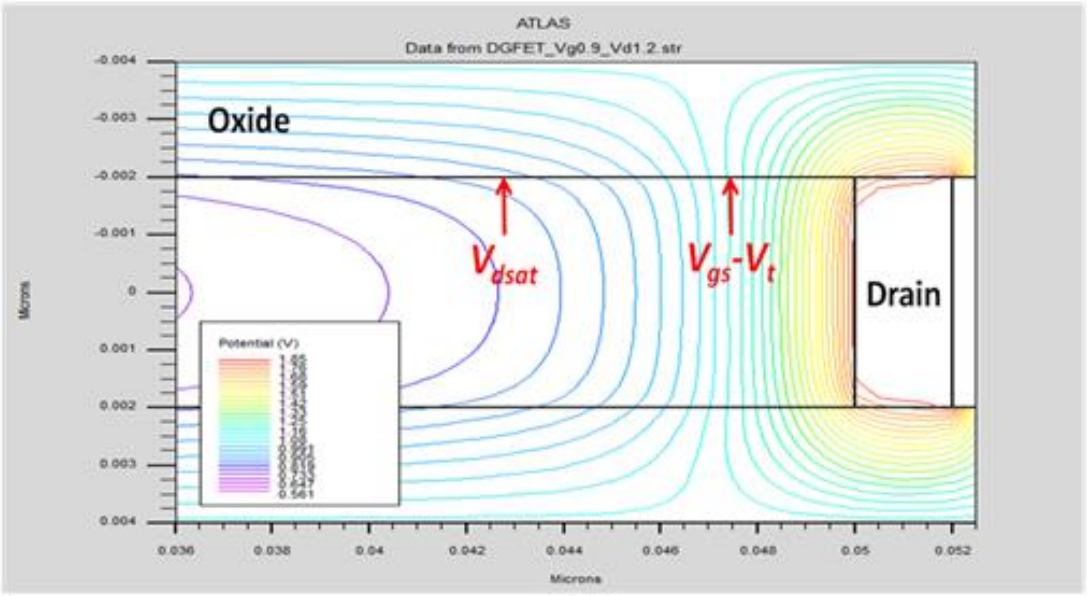


Figure 4.8 Potential contour plot from TCAD. The bias point is V_{gs} = 0.9 V, V_{ds} = 1.2 V, under the n = 2 velocity saturation model. The gates (top and bottom of the plotting window) are at a potential of 1.1 V. The x- and y-label units are in μm, with L = 50 nm. The arrows point to the location in channel where the quasi-Fermi potentials are $V_{gs} - V_t$ = 0.57 V, and V_{dsat} = 0.26 V, respectively.

4.2 Constant Mobility

It is abundantly clear that the key factor missing in GCA is the effect of $\partial\psi^2/\partial y^2$ on the mobile charge density. To construct a continuous model that extends into saturation region, we begin with the textbook expression of inversion charge density as a function of the Fermi potential V in the channel,

$$Q_i = 2C_{ox}(V_{gs} - V_t - V)\,. \tag{4.6}$$

This equation over simplifies the inversion charge to a delta function of zero depth, hence over estimates the current. But it serves to bring forth the key concept of the approach. Instead of using this expression as the only Q_i in the current continuity Eq. (2.13) as in a GCA model, we add a ΔQ_i due to $\partial\psi^2/\partial y^2$:

$$\Delta Q_i = (q\Delta n)t_{si} = \varepsilon_{si} t_{si} \frac{\partial^2 \psi}{\partial y^2} \approx \varepsilon_{si} t_{si} \frac{d^2 V}{dy^2} . \tag{4.7}$$

Fig. 4.6 justifies the approximation $\partial \psi^2/\partial y^2 \approx d^2V/dy^2$. Also, Δn is taken to be uniform over t_{si}, in view of Fig. 4.5. With ΔQ_i added to Q_i of Eq. (4.6), the current continuity Eq. (2.13) becomes

$$I_{ds} = \mu W \left[2C_{ox}(V_{gs} - V_t - V) + \varepsilon_{si} t_{si} \frac{d^2 V}{dy^2} \right] \frac{dV}{dy} . \tag{4.8}$$

This equation can be integrated once to yield

$$\frac{I_{ds}}{\mu W} y = 2C_{ox} \left[(V_{gs} - V_t)V - \frac{V^2}{2} \right] + \frac{\varepsilon_{si} t_{si}}{2} \left[\left(\frac{dV}{dy} \right)^2 - \mathsf{E}_0^2 \right], \tag{4.9}$$

where $\mathcal{E}_0 = (dV/dy)|_{y=0}$ at the source. Since d^2V/dy^2 is negligible at the source, setting $V = 0$ in Eq. (4.8) gives

$$\mathsf{E}_0 = \frac{I_{ds}}{2\mu W C_{ox}(V_{gs} - V_t)} . \tag{4.10}$$

For a given I_{ds}, Eq. (4.9) can be solved for $y(V)$ or $V(y)$ with the initial condition $V(0) = 0$. Then V_{ds} is given by the value of V where y reaches L. In other words, the model constructs I_{ds}-V_{ds} characteristics by finding V_{ds} for given I_{ds} rather than the more conventional way of solving I_{ds} given V_{ds}. Needless to say, further efforts are needed to turn it into a SPICE-like model.

To generate a continuous solution $y(V)$, a repetitive numerical procedure should be followed with good accuracy whether $(dV/dy)^2$ is negligible or not. The method we practiced is to go from a point of (V, y) to the next point, $(V + \delta V, y + \delta y)$, by solving δy from

$$\frac{I_{ds}}{\mu W}(y + \delta y) = 2C_{ox} \left[(V_{gs} - V_t)(V + \delta V) - \frac{(V + \delta V)^2}{2} \right] + \frac{\varepsilon_{si} t_{si}}{2} \left[\left(\frac{\delta V}{\delta y} \right)^2 - \mathsf{E}_0^2 \right] \tag{4.11}$$

for a given incremental δV. The above can be re-organized into a cubic equation of unknown δy with explicit solutions. [Note: While in this specific case, it is easier to solve a quadratic equation

in δV for a given δy, the cubic equation approach is more general and applicable to the model discussed later with more complex 1st term (due to Q_i).]

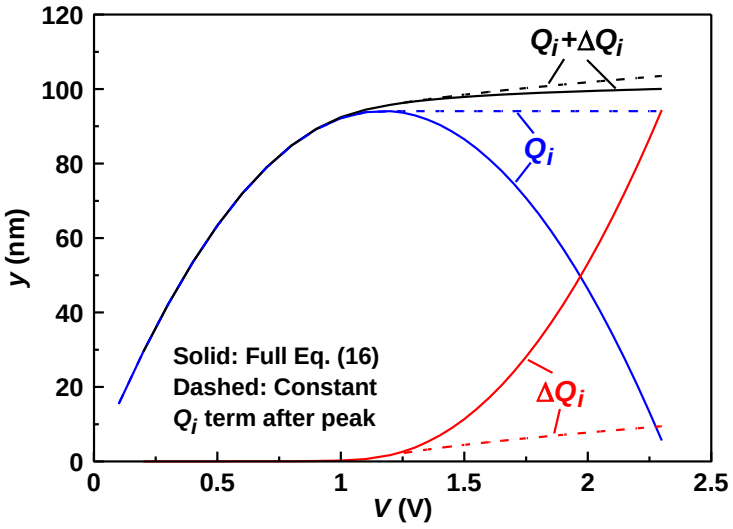
Without the 2nd term (due to ΔQ_i), Eq. (4.9) is just the standard GCA result that has a peak value of $C_{ox}(V_{gs} - V_t)^2$ at $V = V_{gs} - V_t$, as illustrated in Fig. 4.9(a) (solid curve labeled Q_i). It means that I_{ds} cannot exceed the saturation current, $I_{dsat} = \mu(W/L)C_{ox}(V_{gs} - V_t)^2$. Past the peak, Q_i of Eq. (4.6) becomes negative, hence forbidden by the GCA model. It is often regarded as unphysical in the textbooks. However, as revealed in Fig. 4.4, the vertical field ($\mathcal{E}_x$) does change sign at the point where $V = V_{gs} - V_t$. What is unphysical then is not $\partial \psi^2/\partial x^2 < 0$. Rather, what is unphysical is the GCA itself past the point of saturation. Note that when $y(V)$ approaches the peak with the 1st term dominating, $dy/dV \rightarrow 0$, which turns on the square of the reciprocal, $(dV/dy)^2$ in the 2nd term ($\mathcal{E}_0^2$ is negligible), thereby removing the peak. Even when the 1st term (due to Q_i) decreases past $V = V_{gs} - V_t$, the 2nd term (due to ΔQ_i) just picks up the slack and ensures that the sum (y) keeps on increasing with V, albeit at a lower rate [Fig. 4.9(a), solid curve labeled $Q_i + \Delta Q_i$].

To investigate the effect of the negative oxide field, Eq. (4.9) is also solved with the 1st term (due to Q_i) set at a constant equal to the peak value for $V > V_{gs} - V_t$, i.e., past the peak. The results are shown as dashed lines in Fig. 4.9(a). It is noteworthy that while the Q_i terms are dramatically different in the two cases, the total $y(V)$, labeled as $Q_i + \Delta Q_i$ in Fig 4.9(a), differ only slightly. This means that the 2nd term (ΔQ_i in Fig. 4.9(a)) adjusts to the 1st term to make the total $y(V)$ slope slightly positive beyond the point of saturation. Mathematically, dy/dV can decrease with V indefinitely but can never reach or cross zero, because that would mean the reciprocal, dV/dy, in Eq. (16) goes to infinity.

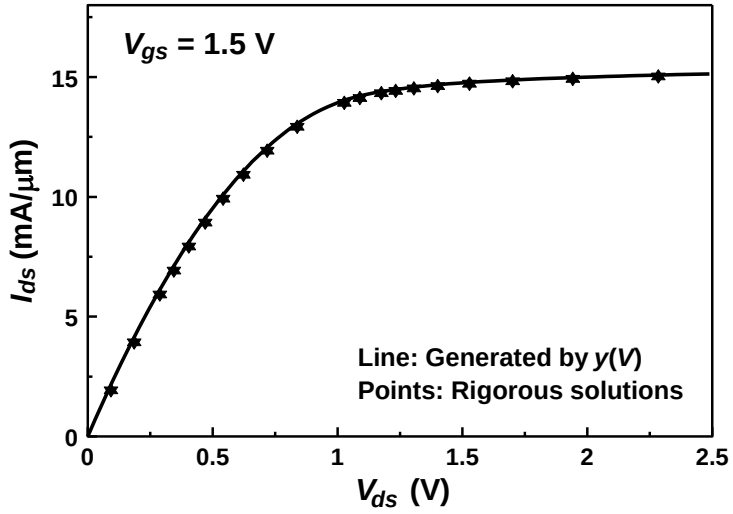
The $y(V)$ curve in Fig. 4.9(a) was obtained with a presumed I_{ds} value (6% over I_{dsat}). It gives the V_{ds} value (2.3 V) when y reaches L (100 nm). In principle, this process needs be repeated by varying I_{ds} over an array of values to generate an entire I_{ds}-V_{ds} curve point by point. A much simpler approximation is to turn the $y(V)$ curve generated with one fixed I_{ds0} into an I_{ds}-V_{ds} curve by multiplying y with (I_{ds0}/L). To justify it mathematically, we rewrite Eq. (4.9) by introducing $z \equiv y\, I_{ds0}$:

$$\frac{z}{\mu W} = 2C_{ox}\left[(V_{gs} - V_t)V - \frac{V^2}{2}\right] + \frac{\varepsilon_{si} t_{si}}{2}\left[I_{ds0}^2\left(\frac{dV}{dz}\right)^2 - \mathsf{E}_0^2\right] \tag{4.12}$$

Since the $(dV/dz)^2$ term is only significant in saturation where $I_{ds} \approx I_{dsat}$ or slightly higher, a choice of the factor in front: $I_{ds0}^2 \approx I_{dsat}^2$ will give it the right magnitude. In the triode region, only the 1st term on the RHS of Eq. (4.12) is important, and z/L gives the I_{ds} in that region as a function of $V = V_{ds}$. At the point of $I_{ds} = I_{ds0}$ (in saturation), the solution is exact because at the voltage $V = V_{ds}$ where $y(V) = L$, $z/L = I_{ds0}$. For I_{ds} slightly below or above I_{ds0}, e.g., $I_{ds} = (1 + \delta)I_{ds0}$, $z/L = I_{ds}$ if y is taken to $(1 + \delta)L$. Fig. 4.9(b) compares the I_{ds}-V_{ds} from the rigorous point-by-point solution of Eq. (4.9) to that generated by multiplying the $y(V)$ curve with (I_{ds0}/L). While the two are not exactly identical, the latter is an excellent approximation to the former.



(a)



(b)

Figure 4.9 (a) Solution to Eq. (4.9) with I_{ds} set at 6% over the peak (I_{dsat}). The curve labeled Q_i is the contribution of the 1st term to y. ΔQ_i is from the 2nd term. The dashed curves are the solution with the Q_i term set to the peak value after the peak. (b) Agreement between the I_{ds}-V_{ds} computed point by point and that by multiplying (I_{ds0}/L) to the $y(V)$ curve.

Regional approximation

In saturation, the 1st term of

$$\frac{I_{ds}}{\mu W}y = \frac{4\varepsilon_{si}}{t_{si}}\left(\frac{2kT}{q}\right)^2\left[\beta\tan\beta - \frac{\beta^2}{2} + \frac{\varepsilon_{si}t_i}{\varepsilon_i t_{si}}\beta^2\tan^2\beta\right]_{\beta}^{\beta_s} - \frac{C_{ox}}{4}\left[\left|V_{gs}-V_t-V\right| - (V_{gs}-V_t-V)\right]^2 + \frac{\varepsilon_{si}t_{si}}{2}\left[\left(\frac{dV}{dy}\right)^2 - \mathsf{E}_0^2\right]$$ is

essentially a constant, equal to $I_{dsat}(L/\mu W)$. The 2nd term is simply given by

$\int_{V_{gs}-V_t}^{V} 2C_{ox}(V_{gs}-V_t-V)dV = -C_{ox}(V_{gs}-V_t-V)^2$. Thus,

$$\frac{I_{ds}}{\mu W}y - \frac{I_{dsat}}{\mu W}L + C_{ox}(V-V_{dsat})^2 = \frac{\varepsilon_{si}t_{si}}{2}\left[\left(\frac{dV}{dy}\right)^2 - \mathsf{E}_0^2\right] \tag{4.13}$$

As shown in Fig. 4.9(b), there is an equivalence of I_{ds} to $(y/L)I_{dsat}$. Therefore, $dy/L = dI_{ds}/I_{dsat}$, which simply restates the very concept of CLM. With that, $(dV/dy)^{-1}$ is converted to the output conductance and solved from Eq. (4.13):

$$\frac{dI_{ds}}{dV_{ds}} = \frac{I_{dsat}}{L}\sqrt{\frac{\varepsilon_{si}t_{si}}{2}}\left[\frac{(I_{ds}-I_{dsat})L}{\mu W} + C_{ox}(V_{ds}-V_{dsat})^2 + \frac{\varepsilon_{si}t_{si}}{2}\mathsf{E}_0^2\right]^{-1/2} \tag{4.14}$$

where $V = V_{ds}$ at $y = L$. If V_{ds} is not too close to V_{dsat}, the 1st and the 3rd terms in the square bracket are negligible compared to the 2nd term. An approximate expression for the output conductance in the saturation region is then:

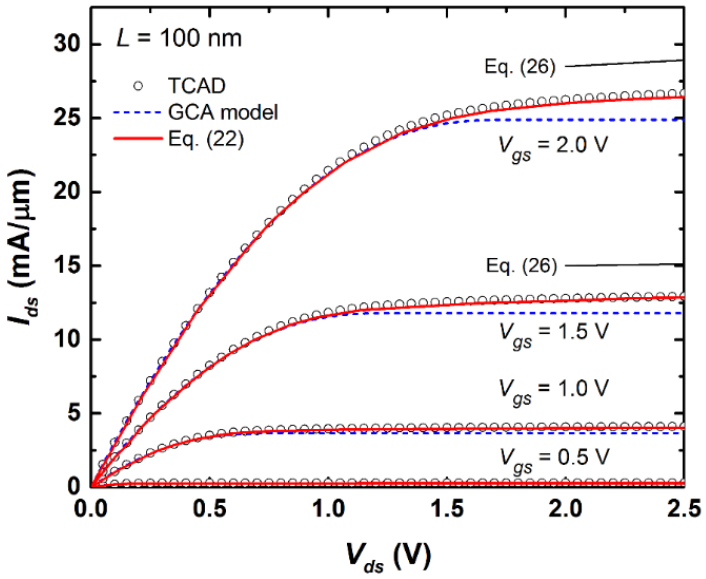
$$\frac{dI_{ds}}{dV_{ds}} \approx \sqrt{\frac{\varepsilon_{si}t_{si}t_i}{2\varepsilon_i L^2}}\,\frac{I_{dsat}}{V_{ds}-V_{dsat}}. \tag{4.15}$$

The output conductance decreases with increasing V_{ds} bias, as depicted in Fig. 4.10(b). Note that $\Delta L(\text{CLM}) \sim \log(V_{ds} - V_{dsat})$.

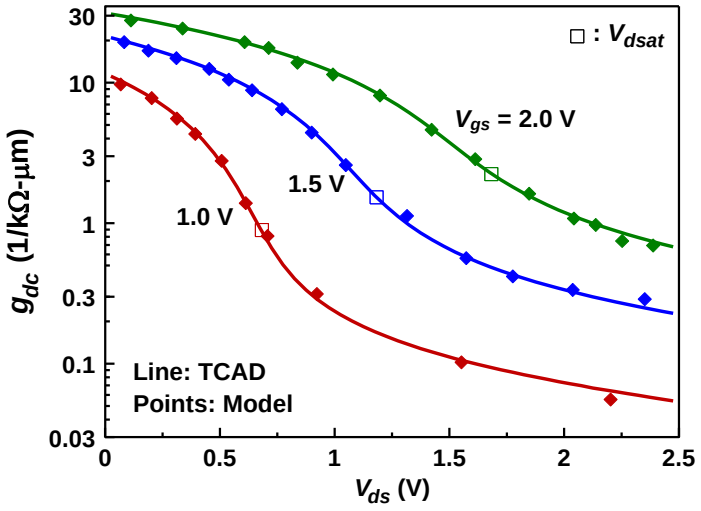
The dimensionless factor, $t_{si}t_i/L^2$, in the square root of Eq. (4.15) indicates that the saturation region characteristics are scalable with respect to the *x*- and *y*-dimensions of the device. For our example of $\varepsilon_{si} = \varepsilon_{ox}$, $t_{si} = 4$ nm, $t_{ox} = 2$ nm, and $L = 100$ nm,

$$\frac{dI_{ds}}{dV_{ds}} \approx \frac{1}{50}\,\frac{I_{dsat}}{V_{ds}-V_{dsat}}. \tag{4.16}$$

The slopes given by the above equation are indicated in Fig. 4.10(a) above the V_{gs} = 1.5 V and 2.0 V curves.



(a)



(b)

Figure 4.10 (a) I_{ds}-V_{ds} curves solved compared to TCAD. L = 100 nm. (b) Output conductance versus V_{ds}. Open squares in each curve indicate where V_{dsat} is for that V_{gs}.

4.3 *n* =1 Velocity Saturation

With the n = 1 velocity saturation model, the current continuity eq. takes the form

$$I_{ds} = \mu W Q_i \frac{dV}{dy} = \frac{\mu_0 W Q_i}{1 + (\mu_0 / v_{sat})(dV/dy)} \frac{dV}{dy}, \tag{4.17}$$

where I_{ds} is the source-drain current independent of y, μ_0 is the low-field mobility, W is the device width, Q_i is the mobile charge density per unit area, and $V(y)$ is the electron quasi-Fermi potential at a point y in the channel. Here, the driving force is taken to be dV/dy so that I_{ds} will not exceed WQ_iv_{sat}. By multiplying the denominator on both sides and integrating from $V(0) = 0$ to $V(L) = V_{ds}$, Eq. (4.17) yields

$$I_{ds} = \frac{\mu_0 W}{L + (\mu_0 / v_{sat}) V_{ds}} \int_0^{V_{ds}} Q_i(V) dV \cdot \tag{4.18}$$

The simplest expression for $Q_i(V)$ under GCA is

$$Q_i = 2C_{inv}(V_{gs} - V_t - V), \tag{4.19}$$

where C_{inv} is the inversion layer capacitance per unit area and V_t is the threshold voltage. While here the non-GCA model is applied to DG MOSFETs, the only factor pertaining to DG MOSFETs is $2C_{inv}$. The same model can be easily adopted for bulk MOSFETs by using a different C_{inv} appropriate for bulk MOSFETs.

Eq. (4.18) is then easily integrated to give

$$I_{ds} = \frac{\mu_0 W C_{inv} [2(V_{gs} - V_t) V_{ds} - {V_{ds}}^2]}{L + (\mu_0 / v_{sat}) V_{ds}}. \tag{4.20}$$

The I_{ds}-V_{ds} characteristics are plotted in Fig. 4.11 in which we see the problem with GCA for V_{ds} beyond V_{dsat} where I_{ds} reaches its peak, I_{dsat}. V_{dsat} can be solved from the condition, $dI_{ds}/dV_{ds} = 0$,

$$V_{dsat} = \frac{Lv_{sat}}{\mu_0}\left(\sqrt{1+z}-1\right), \tag{4.21}$$

where z is a dimensionless parameter defined as

$$z \equiv \frac{2\mu_0 (V_{gs} - V_t)}{v_{sat} L}. \tag{4.22}$$

Substituting V_{dsat} back in Eq. (4.20) gives

$$I_{dsat} = 2C_{inv} W v_{sat} (V_{gs} - V_t) \frac{\sqrt{1+z}-1}{\sqrt{1+z}+1}. \tag{4.23}$$

It can also be shown from the above that

$$I_{dsat} = 2C_{inv} W v_{sat} (V_{gs} - V_t - V_{dsat}), \tag{4.24}$$

namely, carriers move at v_{sat} at the drain end under the peak current condition.

Note from Eq. (4.23) that I_{dsat} is not simply $\propto 1/L$ as in the constant mobility case. This means that the conventional relation for *channel length modulation* (CLM), $\delta I_{ds}/I_{dsat} = \delta L/L$ where $\delta L/L$ is the fractional reduction of the GCA channel length, needs to be modified for the velocity saturation case. From Eq. (4.23),

$$\frac{\delta I_{ds}}{I_{dsat}} = \frac{1}{\sqrt{1+z}} \frac{\delta z}{z} = \frac{1}{\sqrt{1+z}} \frac{\delta L}{L}. \tag{4.25}$$

Note that if $z = 0$, it reduces to the familiar form of CLM for the constant mobility case. But if $z >> 1$, I_{ds} can be independent of L if fully velocity saturated (at the source). The factor on CLM under $n = 1$ velocity saturation, $(\delta I_{dsat}/I_{dsat})/(\delta L/L)$, for the $L = 50$ nm device at $V_{gs} = 1.2$ V is ~ 0.35, meaning only 3.5% increase of current for 10% modulation of channel length. This relation will be applied to derive an output conductance for the $n = 1$ velocity saturation case.

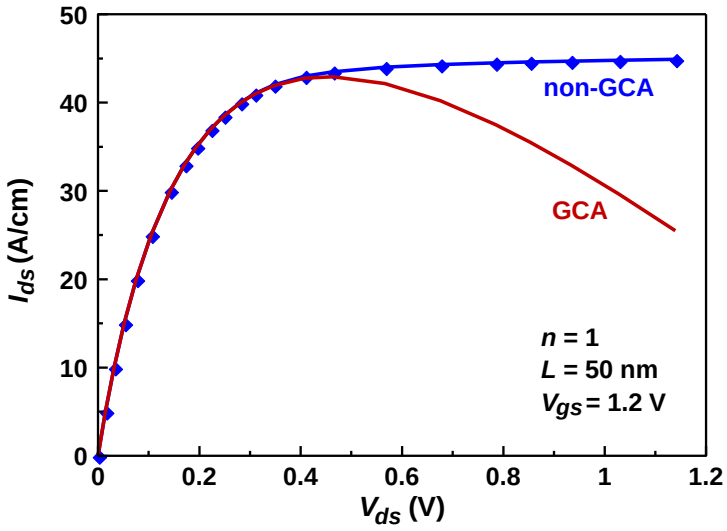


Figure 4.11 I_{ds}-V_{ds} characteristics generated by the GCA and non-GCA models under the n = 1 velocity saturation model. μ_0 = 200 cm^2/V-s, v_{sat} = 10^7 cm/s. C_{inv} is taken to be ε_i/t_i.

The problem of negative slope is solved by taking the effect of lateral field gradient, d^2V/dy^2, on mobile charge density into account. Adding

$$\Delta Q_i = (q\Delta n)t_{si} = \varepsilon_{si}t_{si}\frac{\partial^2\psi}{\partial y^2} \approx \varepsilon_{si}t_{si}\frac{d^2V}{dy^2} \tag{4.26}$$

to Q_i of Eq. (4.7) yields

$$I_{ds} = \frac{\mu_0}{1+(\mu_0/v_{sat})(dV/dy)}W\left(Q_i + \varepsilon_{si}t_{si}\frac{d^2V}{dy^2}\right)\frac{dV}{dy}. \tag{4.27}$$

By multiplying the denominator to the LHS, it can be integrated once:

$$\frac{I_{ds}}{\mu_0 W}y + \frac{I_{ds}}{v_{sat}W}V = 2C_{inv}\left[(V_{gs}-V_t)V - \frac{V^2}{2}\right] + \frac{\varepsilon_{si}t_{si}}{2}\left[\left(\frac{dV}{dy}\right)^2 - \mathcal{E}_0^{\,2}\right], \tag{4.28}$$

where $\mathcal{E}_0 = (dV/dy)|_{y=0}$ at the source. Since d^2V/dy^2 is negligible at the source, setting V = 0 in Eq. (4.17) gives

$$\mathcal{E}_0 = \frac{I_{ds}}{\mu_0 W Q_i(V=0) - (\mu_0/v_{sat})I_{ds}}. \tag{4.29}$$

For a given I_{ds}, Eq. (4.28) is a 1st order ordinary differential equation that can be solved numerically for $V(y)$, and therefore $V_{ds} = V(L)$. This can be done in small steps of, e.g., δy = 0.5 nm, with a general-purpose mathematical tool like matlab, or even with a spread sheet . The continuous I_{ds}-V_{ds} characteristics generated are also shown in Fig. 4.11.

Fig. 4.12 plots the gradient of Fermi potential dV/dy at $y = L$, i.e., the drain end versus V_{ds}. At the current peak in the GCA model, $dV/dy \rightarrow \infty$ and $v = v_{sat}$. Past the peak, $dV/dy < 0$, clearly unphysical. The key effect of the $(dV/dy)^2$ term in Eq. (4.28) is to remove the singularity and keep dV/dy finite and positive. Carrier velocity approaches v_{sat}, but never reaches v_{sat}. Past V_{dsat}, dV/dy is approximately a linear function of V_{ds} with an intercept $\approx V_{dsat}$. This turns out to be a general behavior regardless of n = 1 or n = 2 velocity saturation models.

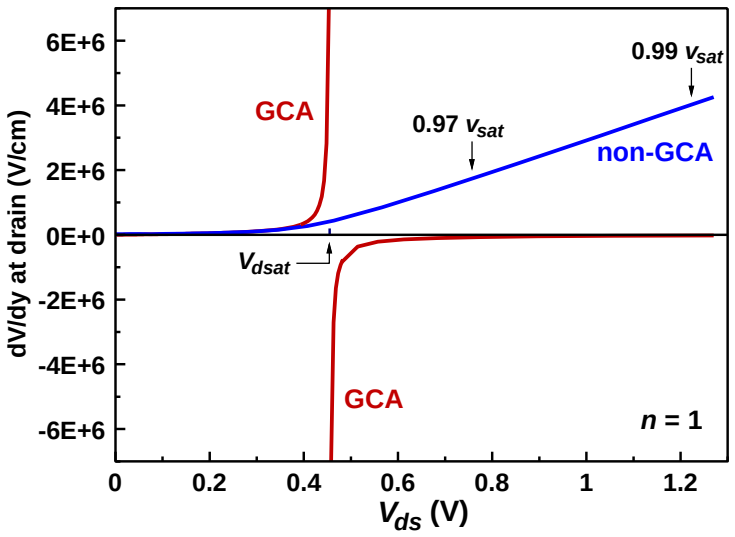


Figure 4.12 dV/dy at the drain ($y = L$) versus V_{ds} for the case in Fig. 4.11. Labels above the non-GCA curve indicate the carrier velocity at those bias points.

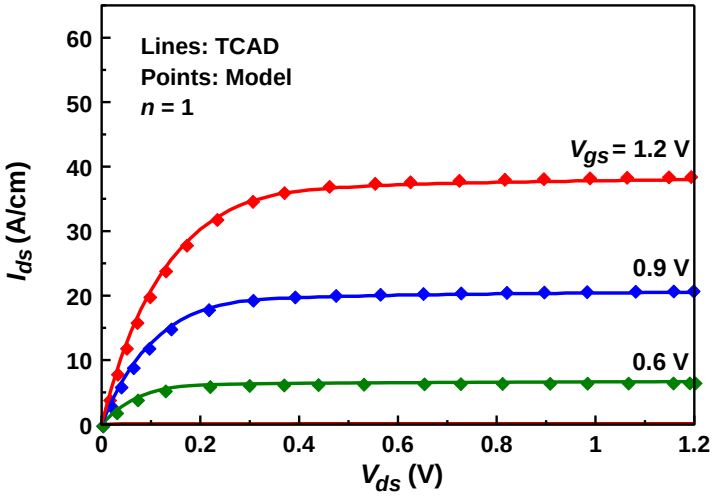


Figure 4.13 I_{ds}-V_{ds} characteristics (n = 1 velocity saturation) generated by the continuous non-GCA model compared with TCAD.

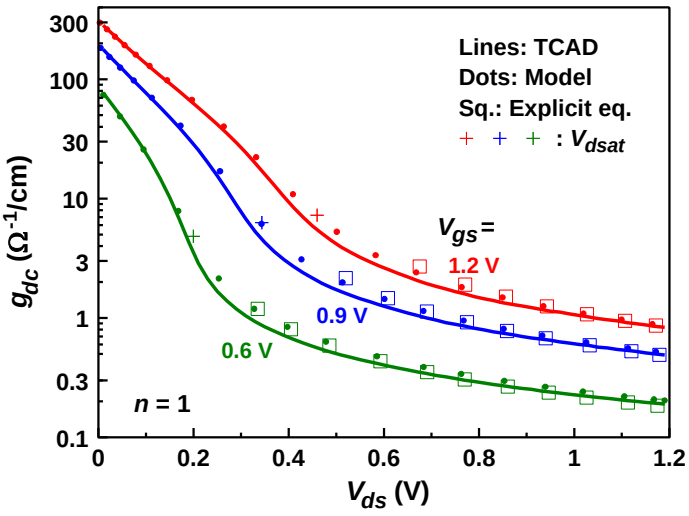


Figure 4.14 Comparison of $g_{dc} \equiv dI_{ds}/dV_{ds}$ versus V_{ds} (n = 1 vel. sat.) between TCAD and the non-GCA model. Open squares are calculated from the explicit Eq. (42) valid for $V_{ds} > V_{dsat}$.

4.4 n =2 Velocity Saturation

It has been known that the n = 1 velocity saturation models have discontinuity in their 2nd derivative around V_{ds} = 0. While this is all right for modeling of digital circuits, it may cause

problems with modeling of analog circuits like mixers. Here we extend the non-GCA model described above to the n = 2 velocity saturation case at the expense of further mathematical complexity.

For n = 2, Eq. (4.17) becomes

$$I_{ds} = \frac{\mu_0 W Q_i}{\sqrt{1+(\mu_0 / v_{sat})^2 (dV/dy)^2}} \frac{dV}{dy} . \tag{4.30}$$

For the GCA part of the model, the above can be re-arranged to yield an integral equation between I_{ds} and V_{ds}:

$$LI_{ds} = \int_0^{V_{ds}} \sqrt{W^2 Q_i^2 \mu_0^2 - I_{ds}^2 (\mu_0 / v_{sat})^2}\, dV . \tag{4.31}$$

With $Q_i(V)$ given by the simple relation, Eq. (4.19), the integral can be carried out by introducing an intermediary parameter u,

$$L = \frac{\mu_0 I_{ds}}{4WC_{inv} v_{sat}^2} \left[\sinh u \cosh u - u\right]_{u_d}^{u_s} , \tag{4.32}$$

where u_s and u_d satisfy

$$2WC_{inv}(V_{gs} - V_t) = (I_{ds} / v_{sat}) \cosh u_s \tag{4.33}$$

and

$$2WC_{inv}(V_{gs} - V_t - V_{ds}) = (I_{ds} / v_{sat}) \cosh u_d . \tag{4.34}$$

For a given I_{ds}, u_s is given explicitly by Eq. (4.33). Then Eq. (4.32) is an implicit equation that solves for u_d, which in turn is used to determine V_{ds} in Eq. (4.34). The I_{ds}-V_{ds} curve generated for V_{gs} = 1.2 V is shown in Fig. 4.15. There is a maximum $V_{ds} = V_{dsat}$ where I_{ds} reaches a peak value I_{dsat} beyond which no solution exists. This corresponds to u_d = 0 where the factor in the square root of Eq. (4.31) is zero. The same Eq. (4.24) also holds for the n = 2 case. At saturation, u_s is the solution to the implicit equation

$$L = \frac{\mu_0 (V_{gs} - V_t)}{2 v_{sat}} \left[\sinh u_s - \frac{u_s}{\cosh u_s} \right] . \tag{4.35}$$

And I_{dsat} is given by

$$I_{dsat} = 2WC_{inv}(V_{gs} - V_t) v_{sat} / \cosh u_s . \tag{4.36}$$

The above equations give the channel length modulation for the $n = 2$ case:

$$\frac{\delta I_{ds}}{I_{dsat}} = \frac{\delta L}{L} \frac{\sinh u_s \cosh u_s - u_s}{\sinh u_s \cosh u_s + u_s} . \tag{4.37}$$

The factor on CLM, $(\delta I_{dsat}/I_{dsat})/(\delta L/L)$, under $n = 2$ velocity saturation is ~ 0.3, for the $L = 50$ nm device at $V_{gs} = 1.2$ V.

To continue the solution beyond the current peak, ΔQ_i of Eq. (4.26) is added to Q_i as in the $n = 1$ case:

$$I_{ds} = W \left(Q_i + \varepsilon_{si} t_{si} \frac{d^2 V}{dy^2} \right) \frac{\mu_0}{\sqrt{1 + (\mu_0 / v_{sat})^2 (dV/dy)^2}} \frac{dV}{dy} . \tag{4.38}$$

This equation cannot be integrated like Eq. (4.27). Instead, we convert the 2nd derivative to

$$\frac{d^2 V}{dy^2} = \frac{dV}{dy} \frac{d}{dV} \left(\frac{dV}{dy} \right) = \frac{1}{2} \frac{d}{dV} \left(\frac{dV}{dy} \right)^2 . \tag{4.39}$$

By squaring Eq. (4.38) and defining

$$g(V) = \left(\frac{dV}{dy} \right)^2 , \tag{4.40}$$

a 1st order differential equation is obtained:

$$I_{ds}{}^2 \left[1 + \left(\frac{\mu_0}{v_{sat}} \right)^2 g \right] = W^2 \mu_0{}^2 g \left[Q_i(V) + \frac{\varepsilon_{si} t_{si}}{2} \frac{dg}{dV} \right]^2 \tag{4.41}$$

With $Q_i(V)$ given by Eq. (4.19), this equation is numerically solved for $g(V)$. After that, $g^{-1/2} = dy/dV$ is readily integrated from $V = 0$ to V_{ds} where $y = L$ is reached. The continuous solution of

I_{ds}-V_{ds} is plotted in Fig. 4.15. Fig. 4.16 plots dV/dy at the drain ($y = L$) versus V_{ds}. The same linear behavior as in the $n = 1$ case is observed. We derive the general expression based on regional approximation in the velocity saturation region.

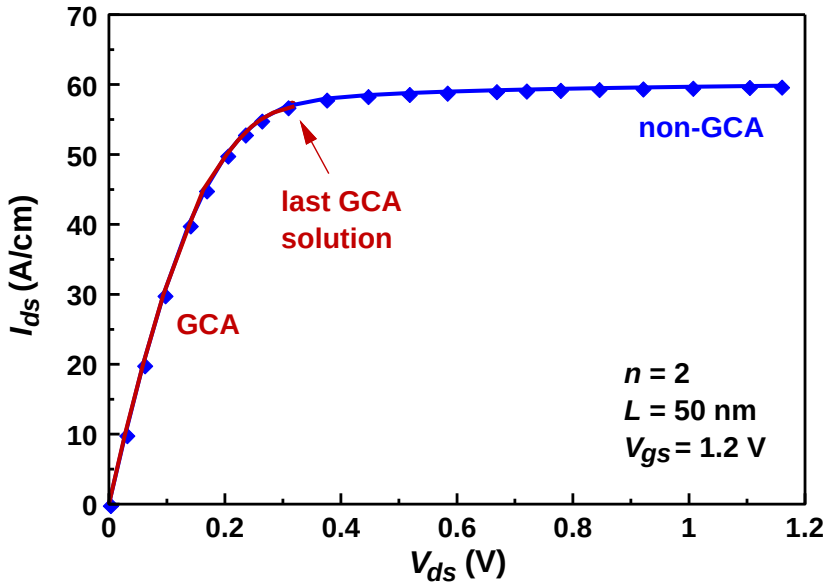


Figure 4.15 I_{ds}-V_{ds} characteristics generated by the GCA and non-GCA models under the $n = 2$ velocity saturation model. $\mu_0 = 200$ cm^2/V-s, $v_{sat} = 10^7$ cm/s. C_{inv} is taken to be ε_i/t_i.

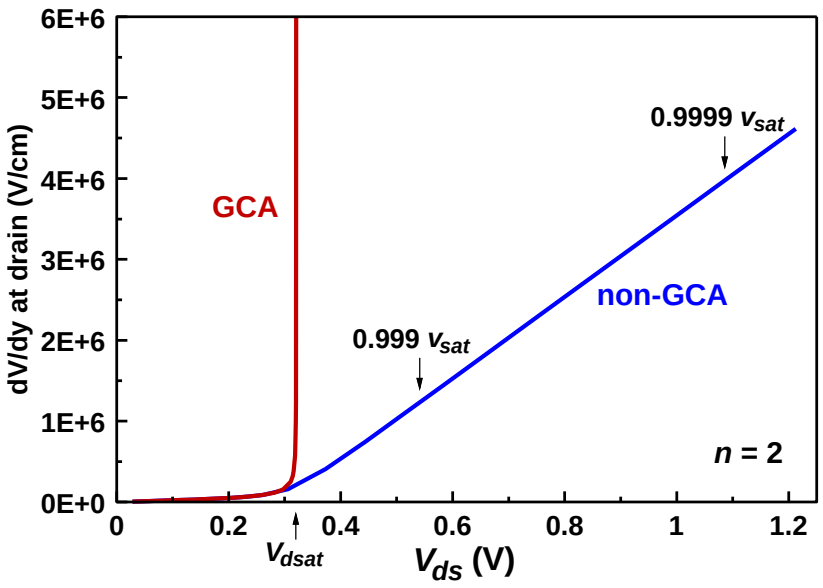


Figure 4.16 *dV*/*dy* at the drain (*y* = *L*) versus V_{ds} for the case in Fig. 4.12. Labels above the non-GCA curve indicate the carrier velocity at those bias points.

Figs. 4.13 and 4.17 show that the I_{ds}-V_{ds} characteristics generated by the non-GCA model with the above factor are in close agreement with TCAD for both the n = 1 and n = 2 cases. Further examination of the output conductance, $g_{dc} \equiv dI_{ds}/dV_{ds}$, in Figs. 4.14 and 4.18 again shows reasonably close agreements between the model and TCAD over the entire range of V_{ds}.

Although the simple Eq. (4.19) for $Q_i(V)$ works fine for $V_{gs} - V_t > 250$ mV ($\approx 10\ kT/q$), it loses its accuracy when V_{gs} is within 100-200 mV of V_t. In that case, the relation between Q_i and $V_{gs} - V_t - V$ is more accurately expressed by the rigorous, all region model of Eqs. (4.1) and (4.2) through the intermediary parameter β.

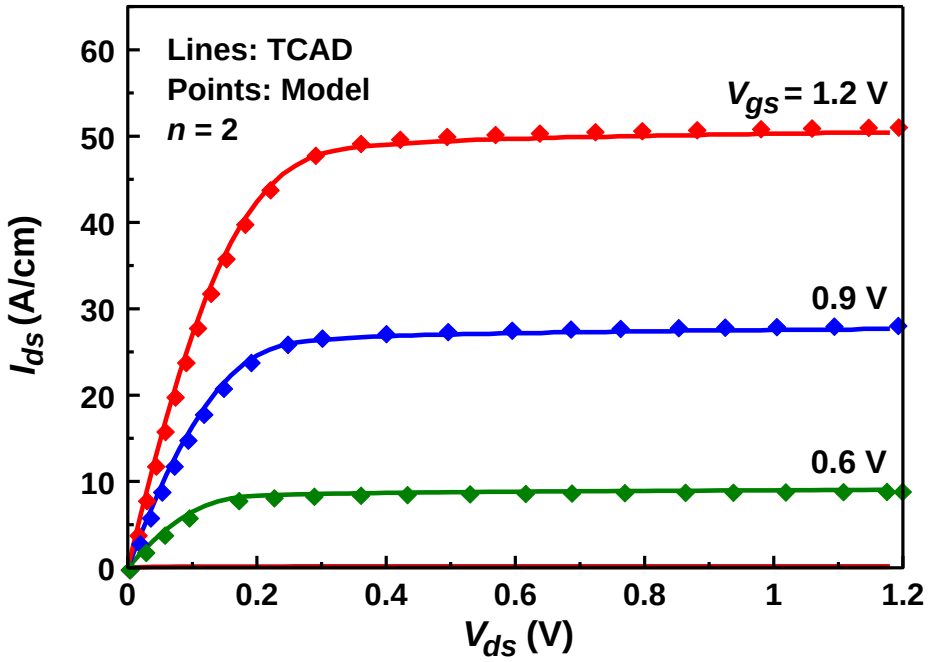


Figure 4.17 I_{ds}-V_{ds} characteristics (n = 2 velocity saturation) generated by the continuous non-GCA model compared with TCAD.

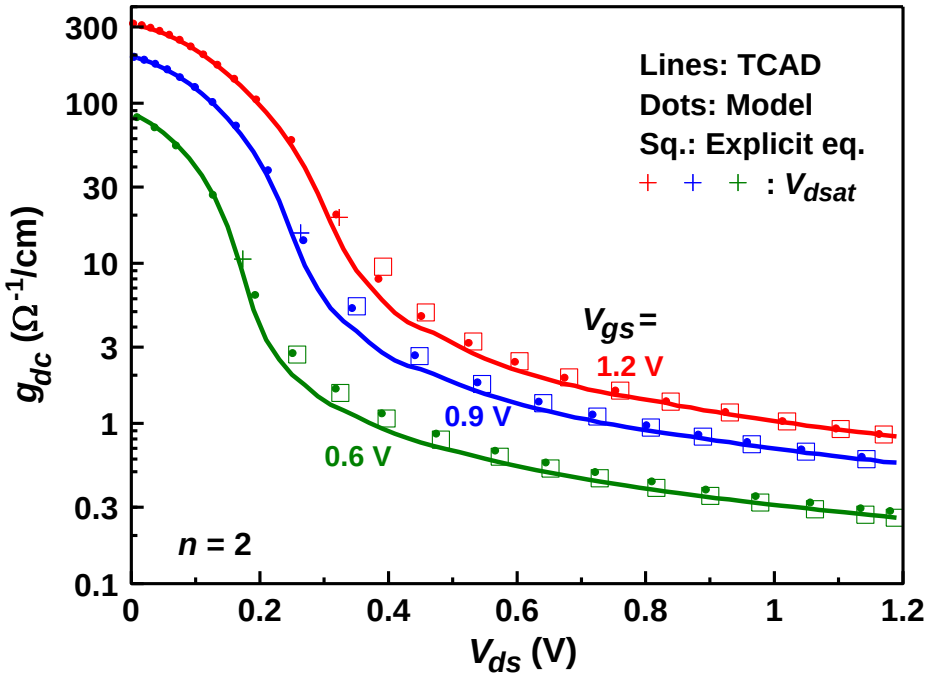


Figure 4.18 Comparison of $g_{dc} \equiv dI_{ds}/dV_{ds}$ versus V_{ds} (n = 2 vel. sat.) between TCAD and the non-GCA model. Open squares are calculated from the explicit Eq. (4.54) valid for $V_{ds} > V_{dsat}$.

This is depicted in Fig. 4.19, where the discrepancy starts to show at V_{gs} = 0.5 V and becomes worse at V_{gs} = 0.4 V—only 3kT/q above V_t. One fix is to generalize Eq. (4.19) to

$$Q_i = 2C_{inv}(V_{gs} - V_t - mV), \tag{4.42}$$

by introducing a parameter m (< 1) to describe the decreased slope of Q_i versus V when $V_{gs} - V_t$ is only a few kT/q. m can be determined from the all region model, Eqs. (4.1), (4.2). For example, m ≈ 0.7 when V_{gs} = 0.4 V. Far above V_t, m ≈ 1. The non-GCA model can be modified in a straightforward way to accommodate this additional parameter.

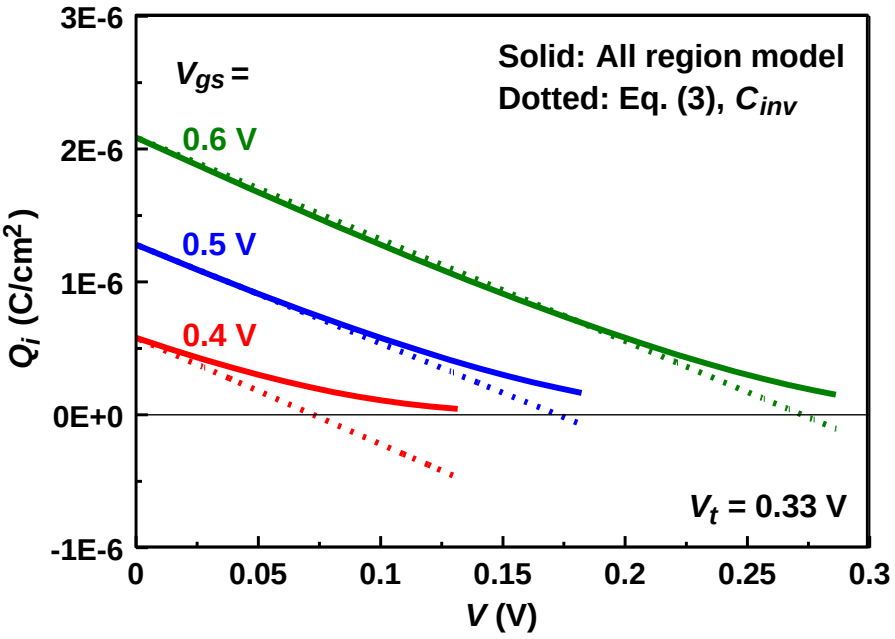


Figure 4.19 Comparison of the rigorous all region model, Eqs. (4.1), (4.2) to the C_{inv} model, Eq. (4.19) at V_{gs} 70-270 mV above V_t.

4.5 Explicit Solution by Regional Approximation

The results that $dV/dy\big|_{y=L}$ is a linear function of V_{ds} for both $n = 1$ (Fig. 4.12) and $n = 2$ (Fig. 4.16) clearly indicate that it is more general than the specific velocity saturation model. This function is derived analytically below following a regional approximation. In the velocity saturation region, $(\mu_0/v_{sat})(dV/dy) >> 1$ such that carrier velocity $\approx v_{sat}$. Both Eqs. (4.27) and (4.38) can then be simplified to

$$I_{ds} = Wv_{sat}\left[2C_{inv}(V_{gs} - V_t - V) + \varepsilon_{si}t_{si}\frac{d^2V}{dy^2}\right], \tag{4.43}$$

with Q_i given by Eq. (4.19). By applying Eq. (4.39), the above equation becomes

$$\frac{I_{ds}}{Wv_{sat}} - 2C_{inv}(V_{gs} - V_t - V) = \frac{\varepsilon_{si}t_{si}}{2}\frac{d}{dV}\left(\frac{dV}{dy}\right)^2. \tag{4.44}$$

Integrating the above from V_{dsat} to V, and making use of Eq. (4.24) for I_{dsat}, it can be shown that

$$\frac{I_{ds} - I_{dsat}}{Wv_{sat}}(V - V_{dsat}) + C_{inv}(V - V_{dsat})^2 = \frac{\varepsilon_{si}t_{si}}{2}\left[\left(\frac{dV}{dy}\right)^2 - \left(\frac{dV}{dy}\right)^2_{V_{dsat}}\right]. \tag{4.45}$$

The above can be written as

$$\left(\frac{dV}{dy}\right)^2 = \frac{2C_{inv}}{\varepsilon_{si}t_{si}}\left[(V - V_{dsat} + a)^2 + b^2\right] \tag{4.46}$$

where

$$a \equiv \frac{I_{ds} - I_{dsat}}{2Wv_{sat}C_{inv}} = \frac{I_{ds} - I_{dsat}}{I_{dsat}}(V_{gs} - V_t - V_{dsat}) \tag{4.47}$$

and

$$a^2 + b^2 = \frac{\varepsilon_{si}t_{si}}{2C_{inv}}\left(\frac{dV}{dy}\right)^2_{V_{dsat}}. \tag{4.48}$$

Since $I_{ds} \approx I_{dsat}$, a is ~ 0.03 V or less for the device being considered. If V_{ds} is not too close to V_{dsat}, $dV/dy|_{Vds}$ at the drain >> $dV/dy|_{Vdsat}$. Therefore,

$$\left.\frac{dV}{dy}\right|_{y=L} = \left.\frac{dV}{dy}\right|_{V_{ds}} \approx \sqrt{\frac{2C_{inv}}{\varepsilon_{si}t_{si}}}(V_{ds} - V_{dsat} + a) \cdot \tag{4.49}$$

This agrees well with the straight lines in Figs. 4.12 and 4.16.

Eq. (4.46) indicates that V is an exponential function of y beyond the point of saturation. Integration with the condition $V(y = L) = V_{ds}$ yields

$$V - V_{dsat} + a + \sqrt{(V - V_{dsat} + a)^2 + b^2} \approx 2(V_{ds} - V_{dsat}) \exp\left[\sqrt{\frac{2C_{inv}}{\varepsilon_{si}t_{si}}}(y - L)\right], \tag{4.50}$$

under the assumption that V_{ds} is not too close to V_{dsat}. In terms of CLM, the point $y = L - \Delta L$ where $V = V_{dsat}$ moves toward the source as V_{ds} increases:

$$\Delta L = \sqrt{\frac{\varepsilon_{si}t_{si}}{2C_{inv}}} \ln\left(\frac{2(V_{ds} - V_{dsat})}{a + \sqrt{a^2 + b^2}}\right). \tag{4.51}$$

Similar exponential expressions of $V(y)$ have been derived in for bulk MOSFETs where $2C_{inv}$ becomes C_{ox} and t_{si} is replaced by x_j, the source-drain junction depth. ΔL of Eq. (4.51) is weakly dependent on $dV/dy|_{Vdsat}$ which goes into $a^2 + b^2$ per Eq. (4.48). $dV/dy|_{Vdsat}$ cannot be determined analytically because it is at the transition point between the GCA model and the fully velocity saturated model, Eq. (4.43). Numerically, $dV/dy|_{Vdsat}$ depends on V_{gs}, as well as on whether the v_{sat} model is $n = 1$ or $n = 2$. For the device considered, $dV/dy|_{Vdsat}$ goes from $3\times(v_{sat}/\mu_0)$ to $9\times(v_{sat}/\mu_0)$.

The factor $a + \sqrt{a^2 + b^2}$ in Eq. (4.51) then ranges from 0.04 V to 0.11 V, meaning a log factor as large as ln(50) ~ 4 and ΔL of ~ 9 nm.

To derive an explicit expression for the output conductance in the velocity saturation region, we use Eq. (4.25) for the $n = 1$ case and note from Eq. (4.50) that for an incremental δV_{ds}, the GCA channel length is further shortened by

$$\delta L = \sqrt{\frac{\varepsilon_{si} t_{si}}{2C_{inv}}} \frac{\delta V_{ds}}{V_{ds} - V_{dsat}} \tag{4.52}$$

Therefore,

$$\frac{dI_{ds}}{dV_{ds}} = \frac{\delta I_{ds}}{\delta V_{ds}} = \sqrt{\frac{\varepsilon_{si} t_{si}}{2C_{inv}(1+z)(L-\Delta L)^2}} \frac{I_{dsat}}{V_{ds} - V_{dsat}}. \tag{4.53}$$

Here, for better accuracy, L in Eq. (4.25) is replaced by the GCA channel length, $L - \Delta L$. It can make as much as 20% difference on the conductance result. For the $n = 2$ case, Eq. (4.37) is used:

$$\frac{dI_{ds}}{dV_{ds}} = \sqrt{\frac{\varepsilon_{si} t_{si}}{2C_{inv}(L-\Delta L)^2}} \left(\frac{\sinh u_s \cosh u_s - u_s}{\sinh u_s \cosh u_s + u_s} \right) \frac{I_{dsat}}{V_{ds} - V_{dsat}}. \tag{4.54}$$

Because of the factors due to modified CLM, the output conductance in the velocity saturation region is lower than that in the saturation region of the constant mobility case. The output conductance calculated from the analytic Eqs. (4.53), (4.54) is also shown in Figs. 4.14 and 4.18 over the range of $V_{ds} > V_{dsat}$ for each V_{gs} bias. They agree well with the numerical model and TCAD results.

SCE is negligible at $L = 50$ nm. It is worthwhile to push the model-TCAD comparison to shorter L and find out at what channel length SCE starts to have non-negligible effect on the output conductance in the saturation region. Fig. 4.20 shows that the model is accurate down to $L = 20$ nm. Below that SCE sets in, having a stronger influence on g_{dc} at $V_{gs} = 0.6$ V than 1.2 V because the closer V_{gs} is to V_t, the more sensitive is Q_i to V_t reduction due to DIBL. The onset is generally comparable to the slope of $\exp(-\pi L/2\lambda)$ from the scale length model where $\lambda = t_{si} + 2\ t_i = 8$ nm.

This can be generalized to state that the range of model validity is $L \geq 2\lambda$, similar to the common criterion for tolerable SCEs based on the subthreshold leakage current.

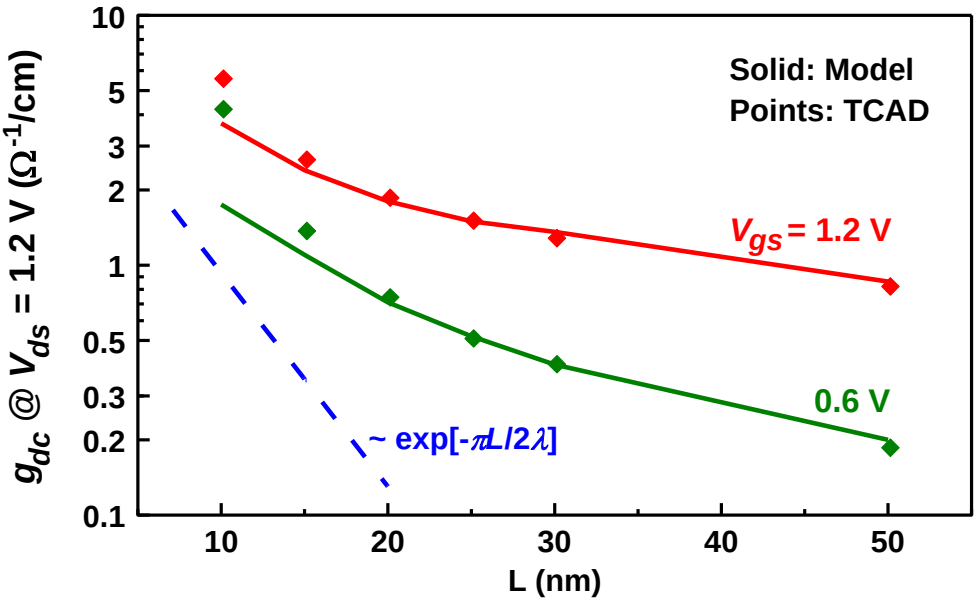


Figure 4.20 Model validity versus channel length. n = 1 velocity saturation model is assumed in both model and TCAD.

4.6 Numerical Solution Methods: Forward Euler versus Backward Euler

In numerical analysis and scientific computing, the backward Euler method (or implicit Euler method) is one of the most basic numerical methods for the solution of ordinary differential equations. It is similar to the (standard) Euler method, but differs in that it is an implicit method. The backward Euler method has error of order one in time.

Consider the ordinary differential equation $\frac{dy}{dt} = f(t, y)$ with initial value $y(t_0) = y_0$ Here the function f and the initial data t_0 and y_0 are known; the function y depends on the real variable t and is unknown. A numerical method produces a sequence $y_0, y_1, y_2, \ldots$ such that y_k approximates $y(t_0 + kh)$, where h is called the step size.

The backward Euler method computes the approximations using $y_{k+1} = y_k + hf(t_{k+1}, y_{k+1})$. This differs from the (forward) Euler method in that the forward method uses $f(t_k, y_k)$ in place of $f(t_{k+1}, y_{k+1})$.

The backward Euler method is an implicit method: the new approximation y_{k+1} appears on both sides of the equation, and thus the method needs to solve an algebraic equation for the unknown y_{k+1}. For non-stiff problems, this can be done with fixed-point iteration:

$$y_{k+1}^{[0]} = y_k, \quad y_{k+1}^{[i+1]} = y_k + hf\left(t_{k+1}, y_{k+1}^{[i]}\right).$$

If this sequence converges (within a given tolerance), then the method takes its limit as the new approximation y_{k+1}.

For a given I_{ds}, solve the following 1st order differential eq. for $V(y)$ with the boundary condition $V(y = 0) = 0$. Use a constant step size of, for example, $\delta y = 0.5$ nm. Then, $V_{ds} = V(y = L)$, i.e., the value of V when y reaches 50 nm.

$$\frac{I_{ds}}{\mu_0 W} y + \frac{I_{ds}}{v_{sat} W} V = 2C_{ox}\left[(V_g - V_0)V - \frac{1}{2}V^2\right] + \frac{\varepsilon_{si} t_{si}}{2}\left[\left(\frac{dV}{dy}\right)^2 - \mathrm{E}_0^{\,2}\right] \tag{4.55}$$

and

$$E_0 = \frac{I_{ds}}{2\mu_0 W C_{ox}(V_g - V_0) - (\mu_0 / v_{sat}) I_{ds}}. \tag{4.56}$$

Going from *V* at a point *y* to the next point *y* + δy *with* $dV/dy|_V$, at point n+1, y goes to y + δy *and V goes to* V + δV. We would like to find the value of $dV/dy|_{V+\delta V}$. The value of dV/dy is in eq. 4.57:

$$\frac{dV}{dy} = \sqrt{E_0^{\,2} + \frac{2}{\varepsilon_{si} t_{si}} \left\{ \frac{I_{ds}}{\mu_0 W} y + \frac{I_{ds}}{v_{sat} W} V - 2C_{ox} \left[(V_g - V_0)V - \frac{1}{2} V^2 \right] \right\}} \equiv f(y, V) \tag{4.57}$$

Different Euler methods were tried to solve the above equations. Forward Euler is defined as $dV = dy \times dV/dy|_V$ where $dV/dy|_V$ is the dV/dy evaluated at the point (y, V). Backward Euler method is defined as $dV = dy \times dV/dy|_{V+dV}$ where $dV/dy|_{V+dV}$ is the dV/dy evaluated at the point (y + dy, V + dV).

i.e.,

$$\frac{I_{ds}}{\mu W}(y + \delta y) + \frac{I_{ds}}{v_{sat} W}(V + \delta V) = 2C_{ox} \left[(V_{gs} - V_t)(V + \delta V) - \frac{(V + \delta V)^2}{2} \right] + \frac{\varepsilon_{si} t_{si}}{2} \left[\left(\frac{\delta V}{\delta y} \right)^2 - E_0^2 \right], \tag{4.58}$$

Then δV *is* solved from the above eq.

In Averaged Euler,

dV is *dy* × ½{$dV/dy|_{V+dV}$ + $dV/dy|_V$}, therefore $dV/dy|_{V+dV} = 2 \times dV/dy - dV/dy|_V$}.

i.e.,

$$\frac{I_{ds}}{\mu W}(y + \delta y) + \frac{I_{ds}}{v_{sat} W}(V + \delta V) = 2C_{ox} \left[(V_{gs} - V_t)(V + \delta V) - \frac{(V + \delta V)^2}{2} \right] + \frac{\varepsilon_{si} t_{si}}{2} \left[\left(2\frac{\delta V}{\delta y} - \left. \frac{dV}{dy} \right|_V \right)^2 - E_0^2 \right]. \tag{4.59}$$

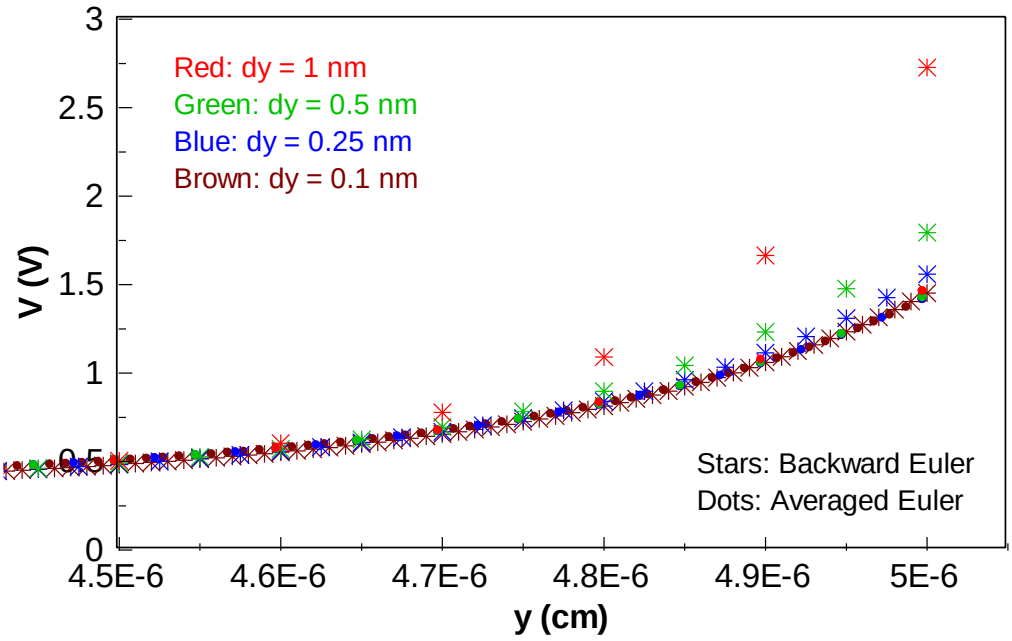


Figure 4.21 Solutions for Backward Euler and Averaged Euler methods with different step sizes *dy*.

Backward Euler has *V(y)* near the drain is step-size dependent until *dy* = 0.1 nm while in, averaged Euler, *V(y)* has little step-size dependence. *dy* = 1 nm is good enough. The plot below shows an example that the forward Euler method may not converge even with a small *dy* of 0.1 nm. The averaged and backward Euler methods always converge with a *dy* = 1 nm.

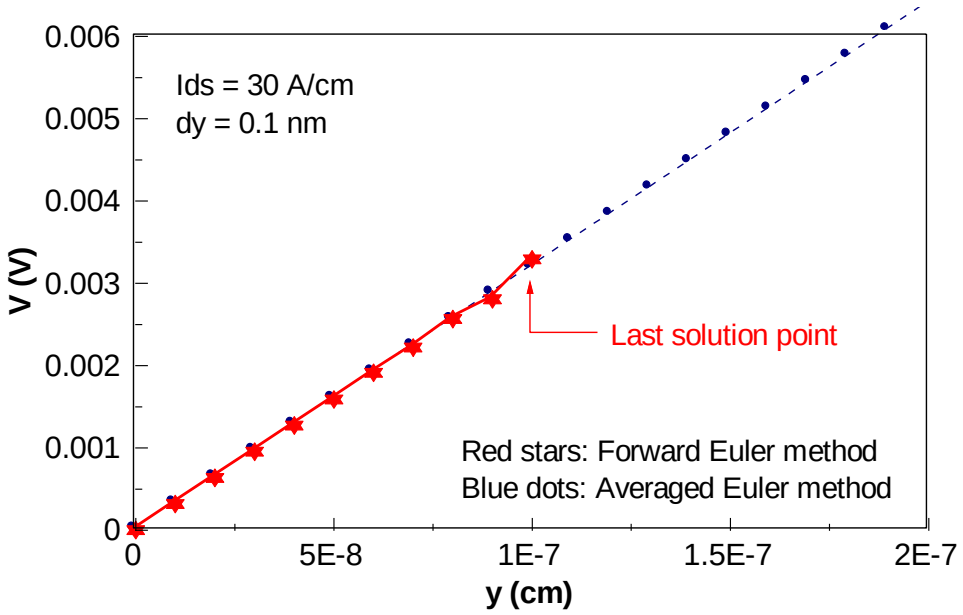


Figure 4.22 Solutions from Forward Euler and Averaged Euler methods with step sizes *dy* of 0.1 nm.

References:

[1] Yuan Taur and Huang-Hsuan Lin, "Modeling of DG MOSFET I-V Characteristics", *IEEE Trans. Electron Device,* pp. 1714-1720, April 2018

[2] Yuan Taur, Woojin Choi, Jianing Zhang, and Meihua Su, "A Non-GCA DG MOSFET Model Continuous into the Velocity Saturation Region", *IEEE Trans. Electron Device*, pp. 1160-1166, Mar. 2019

[3] Yuan Taur, Tak H Ning, "Fundamentals of modern VLSI devices", *Cambridge university press*, December 2021

Acknowledgments

Chapter 4, in full, is a reprint of the material as it appears in Yuan Taur, Woojin Choi, Jianing Zhang, and Meihua Su, "A Non-GCA DG MOSFET Model Continuous into the Velocity Saturation Region", *IEEE Trans. Electron Device*, pp. 1160-1166, Mar. 2019. The dissertation author was an investigator and author of this paper.

CHAPTER 5 NON-GCA MODEL FOR BULK MOSFETS

5.1 Uniform Doping

Non-GCA Model for the Saturation Region

The MOSFET current model covered thus far has been developed under the framework of Gradual Channel Approximation (GCA). It assumes that the field gradient in the *y*-direction or the channel direction is negligible compared to the field gradient in the *x*-direction or the gate direction so 2-D Poisson's equation,

$$\frac{\partial^2\psi}{\partial x^2}+\frac{\partial^2\psi}{\partial y^2}=-\frac{q}{\varepsilon_{si}}[p(x)-n(x)+N_d^+(x)-N_a^-(x)], \tag{5.1}$$

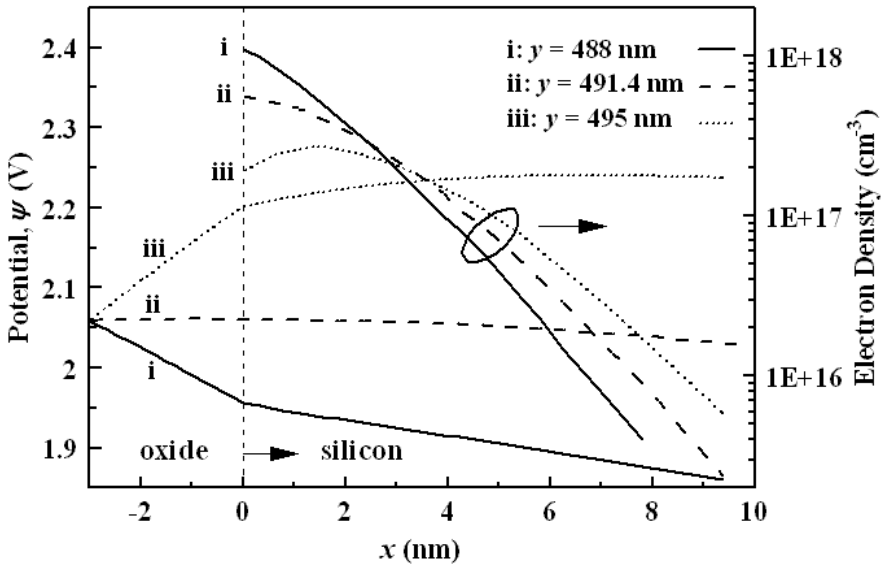
is reduced to the 1-D MOS equation of $\frac{d^2\psi}{dx^2}=-\frac{d\mathscr{E}}{dx}=-\frac{q}{\varepsilon_{si}}[p(x)-n(x)+N_d^+(x)-N_a^-(x)]$. The GCA model works fine in the linear, parabolic, and subthreshold regions, but fails in the saturation region when $V_{ds} > V_{dsat}$. However, the current continuity equation, $I_{ds}(y)=-\mu_{eff}W\frac{dV}{dy}Q_i(y)=-\mu_{eff}W\frac{dV}{dy}Q_i(V)$:

$$I_{ds}=-\mu_{eff}W\frac{dV}{dy}Q_i(V) \tag{5.2}$$

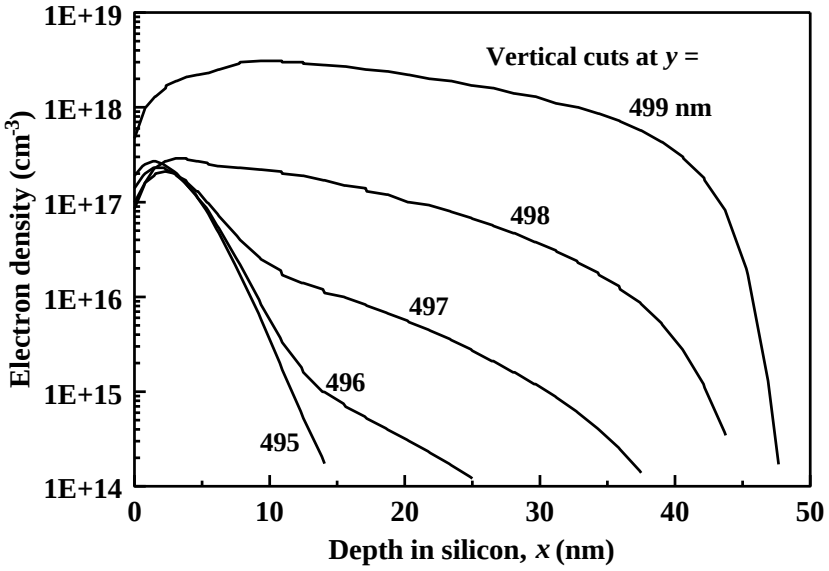
demands that the product $(-Q_i)\times dV/dy$ be a constant throughout the channel. When $-Q_i \to 0$, $dV/dy \to \infty$ thus invalidating the GCA.

In most standard texts, this is called the "*pinch-off*" condition. Pinch-off is a term originally applied to JFETs (Junction Field-Effect Transistor) in the early days of transistor development (Shockley, 1952). It describes how a p- or n-type conducting path is squeezed to zero by the encroaching depletion regions of reverse-biased p-n junctions on both sides of the path. It is rather misleading to use "*pinch-off*" to describe the point of current saturation in

MOSFETs because what goes to zero at $V = V_{dsat}$ is the vertical field, $\mathsf{E}_x = -(\partial\psi / \partial x)\big|_{x=0}$, or the gate induced charge density, not the entire mobile charge density. As a matter of fact, ***both $\partial^2\psi/\partial x^2$ and E_x become negative beyond $V = V_{dsat}$,*** as seen in Fig. 5.1(a) from 2-D numerical simulations. This shows that the above V_{dsat} behavior of the charge-sheet curve is a consequence of the GCA model not allowing E_x to go negative, rather than being physically correct. Also shown in Fig. 5.1(a) is that ***the electron density is never zero whether $(\partial\psi/\partial x)|_{x=0}$ is positive or negative***. From the 2-D Eq. (5.1) perspective, when $\partial^2\psi / \partial x^2$ is negative, the $\partial^2\psi / \partial y^2$ term becomes more positive to overcome the negative $\partial^2\psi / \partial x^2$, thus making the total sum positive. In this regard, "pinch-off" never happens; $\partial^2\psi / \partial y^2$ and therefore the lateral field increase sharply while the vertical field takes on negative values when $V_{ds} > V_{dsat}$.



(a)



(b)

Figure 5.1 Plots from TCAD simulations. (a) Potential $\psi(x)$ and electron density $n(x)$ (right scale) along three vertical cuts: (i) before the saturation point, (ii) at the saturation point, (iii) beyond the saturation point. For this plot, ψ is defined as the intrinsic potential with respect to the Fermi potential of the source. The MOSFET parameters are L = 500 nm, t_{inv} = 3.3 nm, $N_a = 10^{18}$ cm^{-3} (uniform), V_{gs} = 1.5 V, V_{ds} = 2.0 V. The gate work function is that of n^+ silicon. (b) Electron density versus depth in silicon along five vertical cuts between the saturation point and the drain (y = 500 nm). The junction depth is x_j = 50 nm in this case.

A Continuous Non-GCA Model into the Saturation Region

To construct a non-GCA model, a $\partial^2\psi / \partial y^2$ term is added to $-Q_i$ in the current continuity equation (Taur and Lin, 2018):

$$I_{ds} = \mu_{eff} W \left[-Q_i(V) + \varepsilon_{si} d_{si} \frac{d^2\psi}{dy^2} \right] \frac{dV}{dy}. \tag{5.3}$$

Here, d_{si} is an effective depth in silicon to convert the per volume charge density, $\varepsilon_{si} d^2\psi / dy^2$, to an area charge density. For double-gate MOSFETs with thin silicon film, the clear choice for d_{si} is the silicon thickness. For bulk MOSFETs, d_{si} is some fraction of the junction depth x_j. This can be seen in the TCAD plot in Fig. 5.1(b) of the depth distribution of the electron density beyond the point of saturation. When the vertical cut moves closer to the drain junction, the electron density spreads deeper towards the junction depth, x_j = 50 nm, indicating a similar spread of the current density. In this regard, d_{si} is an effective or average depth rather than a physical depth. For this example, a depth parameter of d_{si} = 20 nm serves as a good approximation. Also seen in Fig. 5.1(b) is that the electron density per area, i.e., $n(x)$ integrated over x, which has been decreasing before the saturation point ($y \approx$ 491.4 nm), keeps on decreasing through the saturation point until a point of minimum at $y \approx$ 497 nm very close to the drain junction edge.

For the expression of $-Q_i(V)$ in $I_{ds} = \mu_{eff} W \left[-Q_i(V) + \varepsilon_{si} d_{si} \frac{d^2\psi}{dy^2} \right] \frac{dV}{dy}$, we choose Eq. (5.4):

$$-Q_i(V) = C_{inv}(V_{gs} - V_t - mV), \tag{5.4}$$

which does go negative beyond $V = V_{dsat} = (V_{gs} - V_t)/m$ (see the dotted line in Fig. 2.2). Here, C_{inv} is used in place of C_{ox} to take the inversion layer capacitance into account. At the source, $C_{inv}(V_{gs} - V_t)$ is given by $Q_i = Q_s - Q_d = -C_{ox}(V_{gs} - V_{fb} - \psi_s) + \sqrt{2\varepsilon_{si} q N_a \psi_s}$ of the charge

sheet model with $\psi_s = \psi_{s,s}$ for $V = 0$. The linear slope of $Q_i(V)$ is a reasonable approximation for $V_{gs} - V_t$ larger than several kT/q, e.g., $V_{gs} - V_t > 0.2$ V. For near-threshold bias conditions, the decrease of $|Q_i|$ with V is much softer due to inversion layer capacitance effects (Ren and Taur, 2020).

To make Eq. (5.3) easier to solve, an approximation, $d^2\psi / dy^2 \approx d^2V / dy^2$, is made on the grounds that near the drain where the $\varepsilon_{si} d^2\psi / dy^2$ term is important, the current is mostly drift, i.e., $d\psi / dy \approx dV / dy$. With that substitution, Eq. (5.3) can be integrated once to yield:

$$\frac{I_{ds}}{\mu_{eff} W} y = C_{inv}\left[(V_{gs} - V_t)V - \frac{m}{2}V^2\right] + \frac{\varepsilon_{si} d_{si}}{2}\left[\left(\frac{dV}{dy}\right)^2 - \mathcal{E}_0^2\right], \tag{5.5}$$

where $\mathcal{E}_0$ is dV/dy at $y = 0$. Since the non-GCA term in Eq. (5.3) is negligible at $y = 0$, we have

$$\mathcal{E}_0 = \frac{I_{ds}}{\mu_{eff} W C_{inv}(V_{gs} - V_t)}. \tag{5.6}$$

Equation (5.5) is a 1st-order ordinary differential equation valid for all regions above threshold, both before and after saturation. For a given I_{ds}, it solves for $V(y)$ numerically from $y = 0$ to $y = L$, yielding $V_{ds} = V(L)$ as the result. The standard method of evaluating $dV/dy = f(y, V)$ and applying it to get to the next point runs into the trouble of magnifying the numerical imprecision in the region of $V << V_{dsat}$ where $y(V)$ is simply given by the 1st term (GCA) on the RHS of Eq. (5.5) with the $(dV / dy)^2$ term (non-GCA) negligible. Instead, to go from a point (y, V) to the next point $(y + \delta y, V + \delta V)$, the following difference equation is used:

$$\frac{I_{ds}}{\mu_{eff} W}\delta y = C_{inv}\left\{(V_{gs} - V_t)\delta V - \frac{m}{2}\left[(2V\delta V + (\delta V)^2\right]\right\} + \frac{\varepsilon_{si} d_{si}}{2}\left\{\left[2\left(\frac{\delta V}{\delta y}\right) - \left.\frac{dV}{dy}\right|_{y,V}\right]^2 - \left[\left.\frac{dV}{dy}\right|_{y,V}\right]^2\right\}, \tag{5.7}$$

where $(dV/dy)|_{y,V}$ is the value of dV/dy at (y, V). For a given δy, the above can be re-grouped into a quadratic equation for δV with standard solutions. This procedure can be repeated for a large number of steps on a spread sheet to produce a continuous transition from the GCA dominated region to the non-GCA region.

Examples of the solution $V(y)$ for two different values of I_{ds}, both slightly over I_{dsat}, are plotted in Fig. 5.2 as y versus V so that y can be decomposed into its two components: the 1st term on the RHS of Eq. (5.5) stemming from $-Q_i$ (labeled GCA) and the 2nd term from $(dV/dy)^2$ (labeled non-GCA). Consider first the GCA curve. It has a peak value of $y = (I_{dsat}/I_{ds})L$ at $V = V_{dsat} = (V_{gs} - V_t)/m$, then decreases toward zero. This would be unphysical, like the downturn of I_{ds} past V_{dsat}, were the $-Q_i$ component solely responsible for the current. In the non-GCA model, the additional component from $(dV/dy)^2$, while negligible for $V < V_{dsat}$, increases sharply beyond V_{dsat} so the sum y (solid curves) continues to increase towards L, as seen in Fig. 5.2. The slope dy/dV is, of course, never negative although is much reduced in the saturation region than before saturation.

The notion of *Channel Length Modulation* (CLM) is based on the fact that the peak y-value of the GCA curve, $(I_{dsat}/I_{ds})L$, at $V = V_{dsat}$ becomes $< L$ if $I_{ds} > I_{dsat}$. If we let this y value to be $L - \Delta L$, we obtain $I_{ds} = I_{dsat}/(1 - \Delta L/L)$. In view of the full non-GCA model, CLM only serves as an approximation as the y value at $V = V_{dsat}$ on the solid curve in Fig. 5.2 is slightly higher than the y value at $V = V_{dsat}$ on the dashed (GCA) curve.

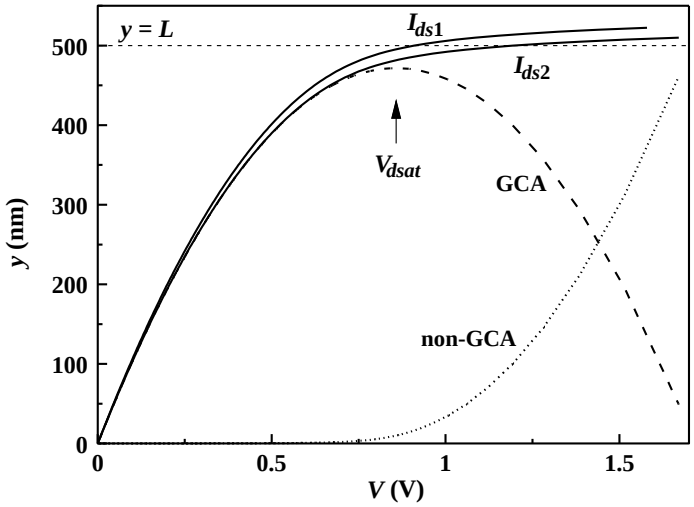


Figure 5.2 $y(V)$ solution to Eq. (5.5) for two values of I_{ds}: I_{ds1} is 3% over I_{dsat}, I_{ds2} is 6% over I_{dsat}. The device is the same as that of Fig. 5.1, with V_t = 0.4 V, m = 1.28, biased at V_{gs} = 1.5 V so V_{dsat} = 0.86 V and I_{dsat} = 2.0 A/cm. d_{si} is chosen to be 20 nm. The crossover with the $y = L$ line gives the V_{ds} solution for the particular I_{ds}. The I_{ds2} result is further partitioned into two curves, according to the two terms on the RHS of Eq. (5.5). The dashed curve labeled GCA is the 1st term divided by ($I_{ds}/\mu_{eff}W$). The dotted curve labeled non-GCA is the 2nd term divided by the same.

Figure 5.3 shows the I_{ds}-V_{ds} curves generated from this model. They are continuous from the linear and parabolic regions into the saturation region.

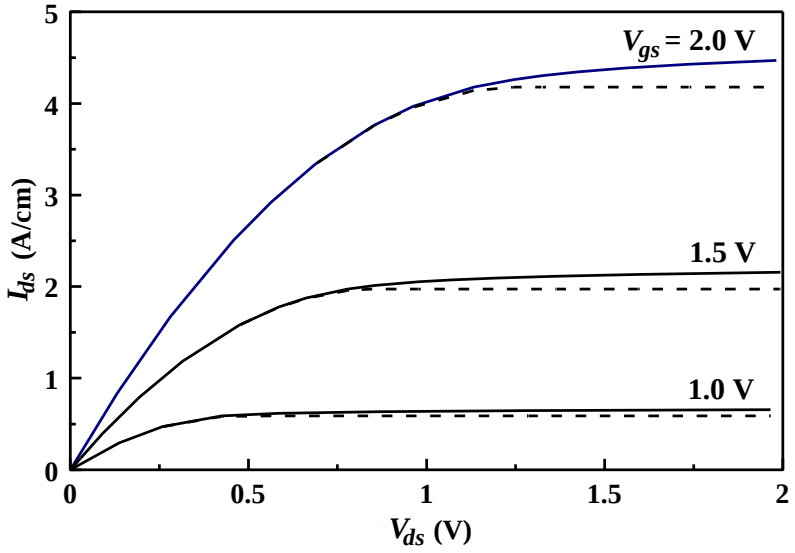


Figure 5.3 I_{ds}-V_{ds} curves (solid) solved from Eq. (5.5) for the device described in the caption to Fig. 5.2. The dashed curves are from the GCA model for which currents saturate at I_{dsat}.

Regional Approximation for the Saturation Region

Equation (5.5) can be greatly simplified in the saturation region where $y \approx L$ as is evident in Fig. 5.2. The E_0^2 term can also be dropped. Further rearrangement yields

$$\frac{L}{\mu_{eff} W}(I_{ds} - I_{dsat}) + \frac{m}{2} C_{inv}(V - V_{dsat})^2 = \frac{\varepsilon_{si} d_{si}}{2}\left(\frac{dV}{dy}\right)^2, \tag{5.8}$$

where V_{dsat} and I_{dsat} are given by $V_{ds} = V_{dsat} = \frac{V_{gs}-V_t}{m}$ and $I_{ds} = I_{dsat} = \mu_{eff} C_{ox} \frac{W}{L} \frac{(V_{gs}-V_t)^2}{2m}$. If V_{ds} is not too close to V_{dsat}, the first term on the LHS is much smaller than the second term. ***It then follows that in the saturation region, dV/dy increases linearly with $V - V_{dsat}$. Further integration gives V(y) as an exponential function of y***, $\propto \exp\left[y / \sqrt{\varepsilon_{si} d_{si} / (m C_{inv})} \right]$. The correlation between the characteristic lateral length $\sqrt{\varepsilon_{si} d_{si} / (m C_{inv})}$ and the vertical dimensions reflects the 2-D nature of the non-GCA effect (Ko *et al*., 1981).

Based on the CLM picture, there is a correspondence of Δy with ΔI_{ds}. Specifically, $\Delta y / L = \Delta I_{ds} / I_{dsat}$. Equation (5.8) then gives the output conductance in the saturation region:

$$\frac{dI_{ds}}{dV_{ds}} = \frac{I_{dsat}}{L}\left(\frac{dV}{dy}\right)^{-1} = \sqrt{\frac{\varepsilon_{si} d_{si}}{m C_{inv} L^2}} \frac{I_{dsat}}{V_{ds} - V_{dsat}}. \tag{5.9}$$

For not too short channel devices, the dimensionless square root factor is << 1, e.g., ~ 1/40 for the device in Fig. 5.3. The slope in the saturation region increases with V_{gs} through the I_{dsat} factor and decreases with V_{ds} for a given V_{gs}.

5.2 Ground-Plane MOSFETs

Nonuniform Channel Doping

In this section, we consider the threshold voltage and the maximum depletion width of a nonuniformly doped MOSFET. Specific examples include high-low and low-high doping profiles.

By employing the depletion approximation in subthreshold, the electric field, surface potential, and threshold voltage can be solved for an arbitrary p-type doping profile $N(x)$. The electric field is obtained by integrating Poisson's equation once (neglecting mobile carriers in the depletion region):

$$\mathsf{E}(x) = \frac{q}{\varepsilon_{si}} \int_x^{W_d} N(x)dx, \tag{5.10}$$

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

$$\psi_s = \frac{q}{\varepsilon_{si}} \int_0^{W_d} \int_x^{W_d} N(x')dx'dx \tag{5.11}$$

Using integration by parts, one can show that the above is equivalent to (Brews, 1979)

$$\psi_s = \frac{q}{\varepsilon_{si}} \int_0^{W_d} xN(x)dx. \tag{5.12}$$

The integral of $xN(x)$ equals the center of mass of $N(x)$ within $(0, W_d)$ times the integral of $N(x)$.

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

Retrograde (Low–High) Channel Profile

When the channel length is scaled to 0.25 μm and below, higher doping concentration is needed in the channel to reduce W_{dm} and control the short-channel effect. If a uniform profile were used, the threshold voltage $V_t = V_{fb} + 2\psi_B + \frac{\sqrt{4\varepsilon_{si} q N_a \psi_B}}{C_{ox}}$ would be too high even with dual polysilicon gates. The problem is further aggravated by quantum effects, which, can add another 0.1–0.2 V to the threshold voltage because of the increasing fields (van Dort *et al.*, 1994).

To reduce the threshold voltage without significantly increasing the gate depletion width, a retrograde channel profile, i.e., a low–high doping profile as shown schematically in Fig. 5.4, is required (Sun *et al*., 1987; Shahidi *et al*., 1989). Such a profile is formed using higher-energy implants that peak below the surface. It is assumed that the maximum gate depletion width extends into the higher-doped region. All the equations in the previous section remain valid for $N_s < N_a$. For simplicity, we assume an ideal retrograde channel profile for which $N_s = 0$. Equation $V_t = V_{fb} + 2\psi_B + \frac{1}{C_{ox}}\sqrt{2\varepsilon_{si} q N_a \left(2\psi_B - \frac{q(N_s - N_a)x_s^2}{2\varepsilon_{si}}\right)} + \frac{q(N_s - N_a)x_s}{C_{ox}}$ then becomes

$$V_t = V_{fb} + 2\psi_B + \frac{qN_a}{C_{ox}}\sqrt{\frac{4\varepsilon_{si}\psi_B}{qN_a} + x_s^2} - \frac{qN_a x_s}{C_{ox}}. \tag{5.13}$$

Similarly, $W_{dm} = \sqrt{\frac{2\varepsilon_{si}}{qN_a}\left(2\psi_B - \frac{q(N_s - N_a)x_s^2}{2\varepsilon_{si}}\right)}$ with $N_s = 0$ gives the maximum depletion width,

$$W_{dm} = \sqrt{\frac{4\varepsilon_{si}\psi_B}{qN_a} + x_s^2}. \tag{5.14}$$

The net effect of low–high doping is that the threshold voltage is reduced, but the depletion width has increased, just opposite to that of high-low doping. All other expressions, such as those

for the subthreshold swing and the substrate sensitivity, in the previous subsection apply with W_{dm} replaced by Eq. (5.14).

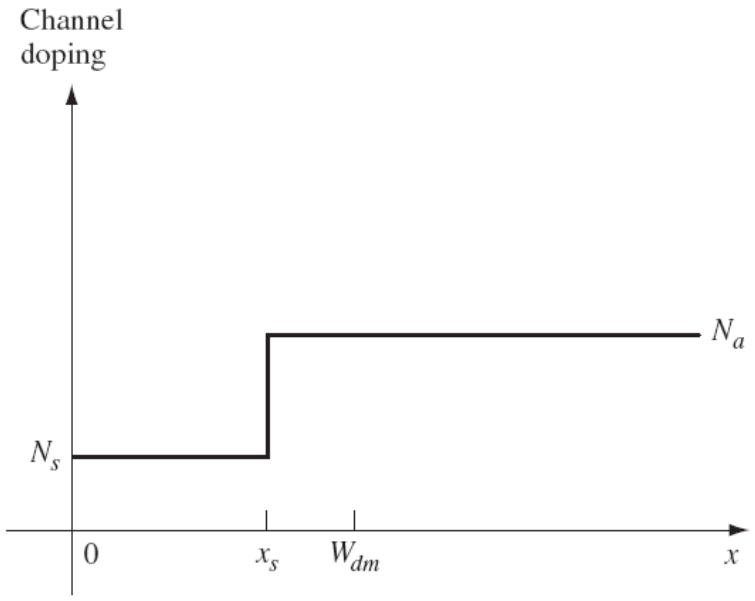


Figure 5.4 A schematic diagram showing the low-high (retrograde) step doping profile. $x = 0$ denotes the silicon–oxide interface.

Extreme Retrograde Profile and Ground-Plane MOSFET

Two limiting cases are worth discussing. If $\boldsymbol{x_s \ll (4\varepsilon_{si}\psi_B/qN_a)^{1/2}}$, then W_{dm} remains essentially unchanged from the uniformly doped value [Eq. (5.14)], while V_t is lowered by a net amount equal to qN_ax_s/C_{ox} [Eq. (5.13)]. In the other limit, N_a is sufficiently high that $x_s \gg (4\varepsilon_{si}\psi_B/qN_a)^{1/2}$. In that case, $W_{dm} \approx x_s$, and the entire depletion region is undoped. All the depletion charge is located at the edge of the depletion region. The square root term in Eq. (5.13) can be expanded into a power series to yield

$$V_t = V_{fb} + 2\psi_B + \frac{\varepsilon_{si}\left(2\psi_B / x_s\right)}{C_{ox}}. \tag{5.15}$$

The last term, due to the depletion charge density in silicon, $\varepsilon_{si}(2\psi_B/x_s)$, can also be derived from Gauss's law by considering that the field in the undoped region is constant and equals $2\psi_B/x_s$ at threshold. Note that the work function difference that goes into V_{fb} is between the gate and the p^+ silicon at the edge of the depletion region. Using $m = 1 + 3t_{ox}/W_{dm} = 1 + 3t_{ox}/x_s$, we can rewrite Eq. (5.15) as

$$V_t = V_{fb} + 2\psi_B + (m-1)2\psi_B. \qquad 5.16)$$

Comparison with $V_t = -\frac{E_g}{2q} + \psi_B + \frac{4\varepsilon_{si}\psi_B}{W_{dm}}\frac{t_{ox}}{\varepsilon_{ox}} = -\frac{E_g}{2q} + \psi_B + 2(m-1)(2\psi_B)$ shows that, with the extreme retrograde profile, the depletion charge (the third) term of V_t is reduced to half of the uniformly doped value.

All the essential device characteristics, such as SCE (W_{dm}), subthreshold slope (m), and threshold voltage, are determined by the depth of the undoped layer, x_s. ***The limiting case of retrograde channel profile therefore degenerates into a ground-plane MOSFET*** (Yan *et al.*, 1991). The band diagram and charge distribution of such a device at the threshold condition are shown schematically in Fig.5.5. Note that the field is constant (no curvature in potential) in the undoped region between the surface and x_s. There is an abrupt change of field at $x = x_s$, where a delta function of depletion charge (area = $2\varepsilon_{si}\psi_B/x_s$) resides. Beyond x_s, the bands are essentially flat. It is desirable not to extend the p^+ region under the source and drain junctions, since that will increase the parasitic junction capacitance. The ideal channel doping profile is then that of a low–high–low type, in which the narrow p^+ region is used only to confine the gate depletion width. Such a profile is also referred to as *pulse-shaped doping* or *delta doping* in the literature.

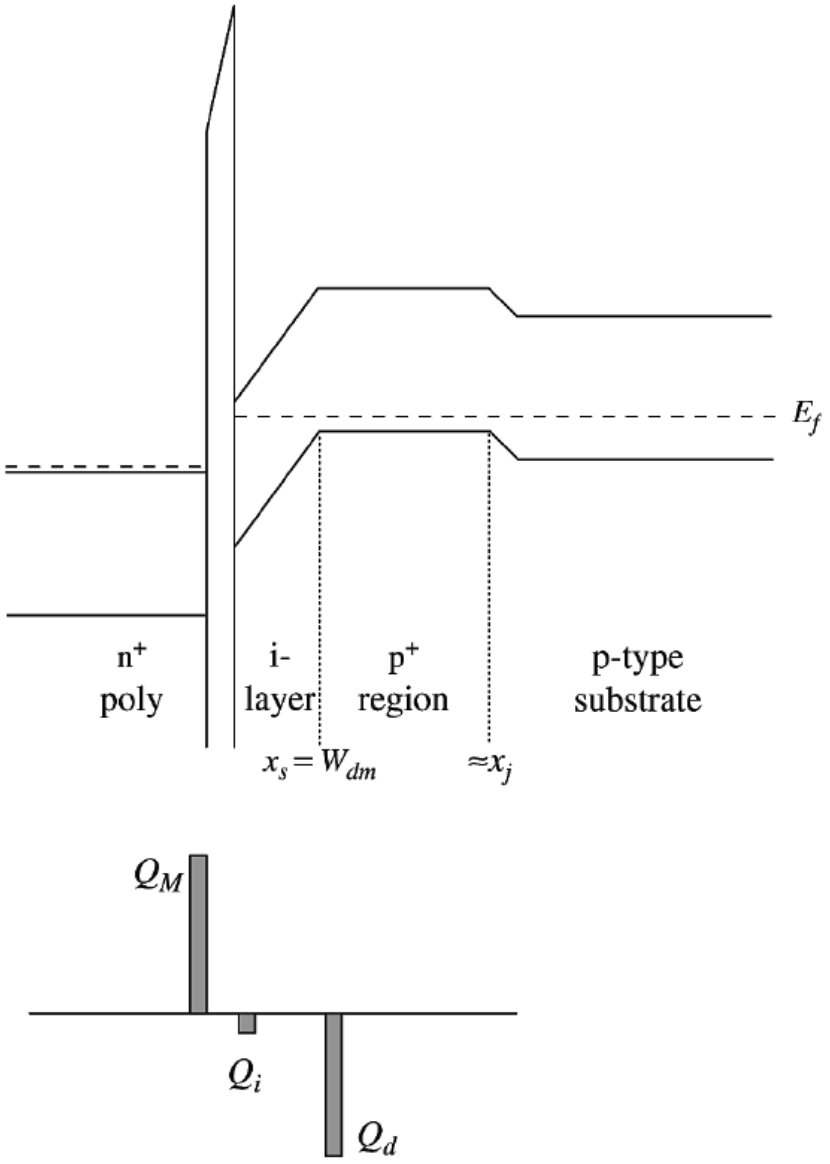


Figure 5.5 Band diagram and charge distribution of an extreme retrograde-doped or ground-plane nMOSFET at the threshold condition.

Near the limits of bulk MOSFET scaling, the body needs to be doped above 10^{19} cm^{-3} to constrain the gate depletion depth to ~ 10 nm for control of the short-channel effect. The threshold voltage, on the other hand, needs to be scaled down below ~ 0.3 V for a supply voltage of ~ 1.0 V. This is accomplished by employing low-high (retrograde) body doping which, for a given gate

depletion width, has a reduced depletion charge density. The extreme limit of low-high doping is a ground-plane MOSFET (also known as super-steep retrograde) shown schematically in Fig. 5.6. It consists of a lightly doped or essentially undoped surface layer of depth x_s on top of a highly doped (N_a) p^+ body (for nMOSFETs). The gate depletion width is essentially x_s with all the depletion charge located at the step where the body doping changes abruptly from 0 to N_a.

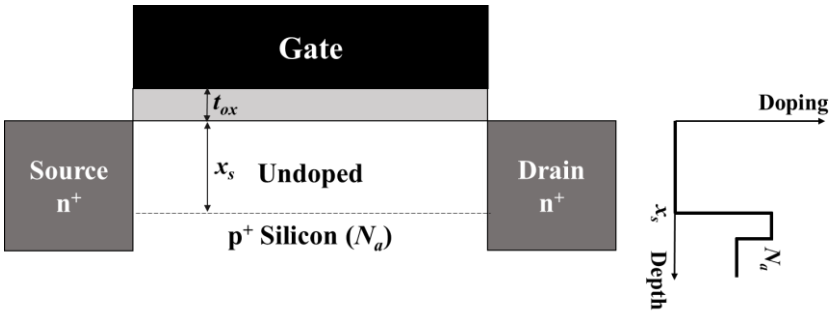


Figure 5.6 A schematic cross-section of ground-plane MOSFETs. Shown on the right is the depth profile of body doping along a vertical cutline.

A Non-GCA Model for Ground-Plane MOSFETs

In this work, we develop a model for ground-plane MOSFETs by first deriving an analytic solution under the Gradual Channel Approximation (GCA). The GCA model works for long channel MOSFETs with constant mobility. For shorter length MOSFETs in which velocity saturation occurs, it is necessary to implement a non-GCA model incorporating the gate-induced mobile charge density from the GCA model. The I_{ds}-V_{ds} characteristics generated from both the n =1 and n = 2 velocity saturation models are verified by TCAD simulations. For comparison with the published hardware data of 20 nm bulk MOSFETs, source-drain series resistances are added to the model.

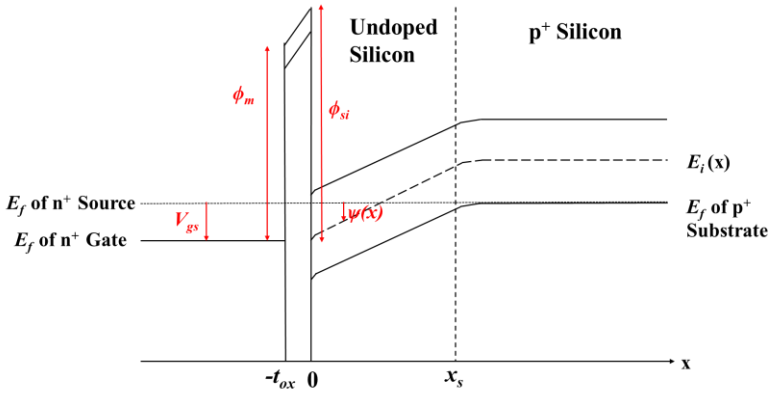


Figure 5.7 Band diagram of a ground-plane nMOSFET biased near the threshold. The p^+ ground plane is grounded to the n^+ source.

5.2.1 Long-Channel GCA Model

The band diagram of a ground-plane nMOSFET is shown schematically in Fig. 5.7. In the undoped silicon region between $x = 0$ and $x = x_s$, there is only the inversion charge (electrons) density hence Poisson's equation takes the form

$$\frac{d^2\psi}{dx^2} = \frac{q}{\varepsilon_{si}} n_i e^{q(\psi - V)/kT} \tag{5.17}$$

where $\psi(x) \equiv -[E_i(x) - E_{fs}]/q$ is the potential function defined in Fig. 5.7 referenced to the source Fermi level E_{fs}. In the above equation, ε_{si} is the permittivity of silicon, n_i is the intrinsic carrier concentration, and V is the electron quasi-Fermi potential at a point in the channel. V is independent of x but varies along the channel from 0 at the source to V_{ds} at the drain.

The hole density is neglected in Eq. (5.17). This is valid for most of the undoped region where the valence band edge is well below the E_f of the p^+ substrate (Fig. 5.7). The highest hole

density in the undoped region occurs right below $x = x_s$. The justification for neglecting the hole density there is given later after discussing depletion of the ground plane.

By multiplying $d\psi/dx$ to both sides, Eq. (5.7) can be integrated once to obtain

$$\frac{d\psi}{dx} = -\sqrt{\frac{2kTn_i}{\varepsilon_{si}} e^{q(\psi-V)/kT} + E_0^2} \tag{5.18}$$

where E_0 is a constant of integration. The general solution to Eq. (5.17) includes the possibility of a negative constant, or $E_0^2 < 0$. As far as the ground-plane MOSFET is concerned, only the $E_0^2 > 0$ solution is needed. Equation (5.18) has the closed form solution:

$$\psi(x) = V + \frac{2kT}{q} \ln\left\{\frac{\sqrt{\frac{\varepsilon_{si}}{2kTn_i}} E_0}{\sinh\left[\frac{qE_0 x}{2kT} + z_0\right]}\right\} \tag{5.19}$$

where z_0 is the second integration constant.

The constants E_0 and z_0 are determined by the boundary conditions at $x = 0$ and $x = x_s$. At $x = 0$, the continuity of displacement at the Si-SiO_2 interface yields

$$\varepsilon_{ox} \frac{V_{gs} - \phi_{mi} - \psi(0)}{t_{ox}} = -\varepsilon_{si} \left.\frac{d\psi}{dx}\right|_{x=0} \tag{5.20}$$

Here, ε_{ox} is the permittivity of the gate oxide with a thickness t_{ox}, V_{gs} is the applied gate to source voltage, and $\phi_{mi} \equiv (\phi_m - \phi_{si})$ is the work function difference between the gate and intrinsic silicon. At $x = x_s$, the simple boundary condition is $\psi(x_s) = -E_g/2q$ if there were no depletion in the p^+ ground plane. But the depletion effect is nonnegligible on the nanometer scale even if the p^+ is doped as high as $N_a = 10^{20}$ cm^{-3}. To account for that, we note that for $\psi'(x)$ in the p^+ region ($x \geq x_s$),

$$\frac{d^2\psi'}{dx^2} = \frac{q}{\varepsilon_{si}}\left[N_a - n_i e^{-q\psi'/kT}\right]. \tag{5.21}$$

Integrate once with the condition $d\psi'/dx = 0$ at $\psi' = -(kT/q)\,ln(N_a/n_i)$,

$$\left(\frac{d\psi'}{dx}\right)^2 = \frac{2kTN_a}{\varepsilon_{si}}\left[\frac{n_i}{N_a}e^{-q\psi'/kT} + \frac{q\psi'}{kT} + ln\left(\frac{N_a}{n_i}\right) - 1\right]. \tag{5.22}$$

The boundary condition at $x = x_s$ for ψ and $d\psi/dx$ is then

$$\left(\left.\frac{\mathrm{d}\psi}{\mathrm{dx}}\right|_{\mathrm{x=x_s}}\right)^2 = \frac{2kTN_a}{\varepsilon_{si}}\left[\frac{n_i}{N_a}e^{-q\psi(x_s)/kT} + \frac{q\psi(x_s)}{kT} + ln\left(\frac{N_a}{n_i}\right) - 1\right] \tag{5.23}$$

for matching ψ' and $d\psi'/dx$ in the p^+ region.

Applying Eq. (5.18) to Eq. (5.23) yields

$$\mathrm{E_0^2} = \frac{2\mathrm{kTN_a}}{\varepsilon_{\mathrm{si}}}\left[\frac{\mathrm{n_i}}{\mathrm{N_a}}\mathrm{e}^{-\mathrm{q}\psi(\mathrm{x_s})/\mathrm{kT}} + \ln\left(\frac{\mathrm{N_a}}{\mathrm{n_i}}\mathrm{e}^{\mathrm{q}\psi(\mathrm{x_s})/\mathrm{kT}}\right) - 1\right] - \frac{2kTn_i}{\varepsilon_{si}}e^{q[\psi(x_s)-V]/kT} \tag{5.24}$$

Let $x = x_s$ in Eq. (5.19) and approximate $\sinh\left[\frac{\mathrm{qE_0x_s}}{2\mathrm{kT}} + \mathrm{z_0}\right]$ as $\frac{1}{2}\exp\left[\frac{\mathrm{qE_0x_s}}{2\mathrm{kT}} + \mathrm{z_0}\right]$, we obtain

$$\frac{2\mathrm{kT}}{\mathrm{q}}\mathrm{z_0} = \mathrm{V} + \frac{2\mathrm{kT}}{\mathrm{q}}\ln\left\{2\sqrt{\frac{\varepsilon_{\mathrm{si}}}{2\mathrm{kTn_i}}}\mathrm{E_0}\right\} - \mathrm{E_0x_s} - \psi(\mathrm{x_s}) \tag{5.25}$$

Let $x = 0$ in Eqs. (5.18), (5.19) and substitute $\psi(0)$ and $(d\psi/dx)|_{x=0}$ in Eq. (5.20):

$$V_{gs} - \phi_{mi} = V + \frac{2kT}{q}ln\left\{\sqrt{\frac{\varepsilon_{si}}{2kTn_i}}\mathcal{E}_0\right\} - \frac{2kT}{q}ln[sinh\,z_0] + \frac{\varepsilon_{si}}{\varepsilon_{ox}}t_{ox}\mathcal{E}_0\,coth\,z_0 \tag{5.26}$$

An implicit equation for a single unknown $\psi(x_s)$ is obtained by expressing z_0 in Eq. (5.26) with Eq. (5.25), then replacing all E_0 in terms of $\psi(x_s)$ using Eq. (5.24).

To justify that the hole density is negligible in Eq. (5.17), we first note that the average field in the undoped region (0, x_s) is $E_{av} \approx (E_g/q)/x_s$ for gate voltages close to the condition where

the front surface inverts, like that shown in Fig. 5.7. For x_s = 10 nm in our case, $E_{av} \approx 1$ MV/cm. For this field, Eq. (5.23) with $N_a = 10^{20}$ cm^{-3} gives $\psi(x_s) \sim -0.54$ V thus $p(x_s) = n_i \exp[-q\psi(x_s)/kT]$ $\sim 10^{19}$ cm^{-3}. For $x < x_s$, the effect of $p(x) = n_i \exp[-q\psi(x)/kT]$ on the field ΔE can be estimated from Poisson's equation as follows:

$$\Delta E = \frac{q}{\varepsilon_{si}} \int_0^{x_s} p(x)dx = \frac{q}{\varepsilon_{si}} n_i \int_{\psi(0)}^{\psi(x_s)} \frac{e^{-q\psi/kT}}{-E} d\psi \approx \frac{kT}{\varepsilon_{si} E_{av}} n_i e^{-q\psi(x_s)/kT}$$

Plugging in the numbers above, $\Delta E \approx (kT/\varepsilon_{si}E_{av})p(x_s) \sim 0.04$ MV/cm << E_{av}. This shows that the hole density in the undoped region has a negligible effect on $\psi(x)$ compared to the existing field E_{av} in the undoped region.

Once E_0 and z_0 are solved for given V_{gs} and V, the mobile charge density Q_i (taken as positive) can be evaluated from Gauss' law:

$$Q_i(V_{gs}, V) = -\varepsilon_{si}\left[\left.\frac{d\psi}{dx}\right|_{x=0} - \left.\frac{d\psi}{dx}\right|_{x=x_s}\right] = \varepsilon_{si} E_0 \left\{\coth z_0 - \coth\left[\frac{qE_0 x_s}{2kT} + z_0\right]\right\} \qquad (5.27)$$

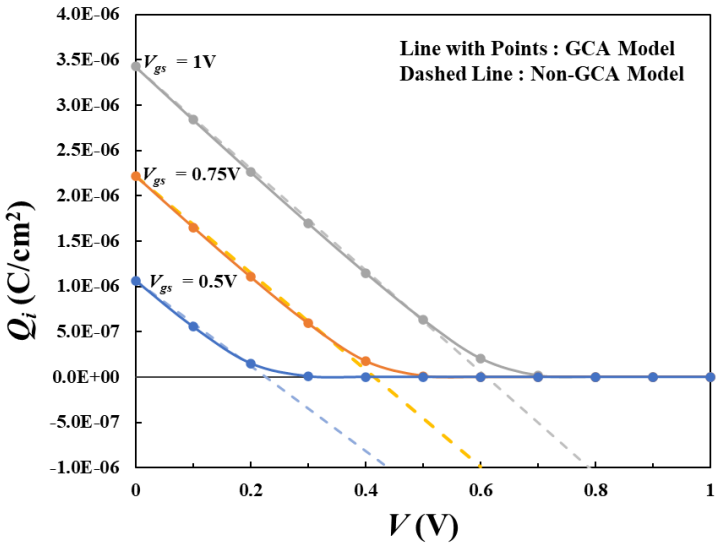


Figure 5.8 Mobile charge density per area at a point in the channel versus electron quasi-Fermi potential for a given gate voltage. The solid lines are solved from Eq. (5.27). The dashed lines are the linear approximation of Eq. (5.32).

Figure 5.8 shows an example of Q_i versus V plots solved from the model for several values of V_{gs}. For the case of constant mobility (μ_0), the long channel MOSFET current is simply given by the integral of Q_i with respect to V from 0 to V_{ds}, the source-to-drain voltage:

$$I_{ds}(V_{gs}, V_{ds}) = \mu_0 \frac{W}{L} \int_0^{V_{ds}} Q_i(V_{gs}, V) dV. \tag{5.28}$$

Here, W and L are the width and length of the MOSFET. Model generated I_{ds}-V_{gs} plots for L = 1 μm are shown in Fig. 5.9. They are consistent with TCAD simulations. The I_{ds}-V_{ds} plots are shown in Fig. 5.10. The model currents in saturation are slightly (~ 3%) below those of TCAD. The slight discrepancy is resolved by applying the non-GCA model described in the next section to this case.

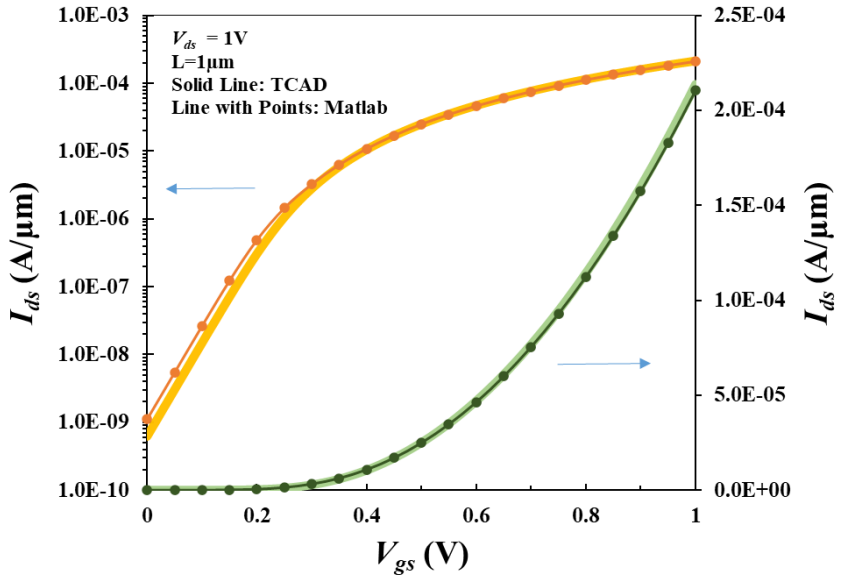


Figure 5.9 I_{ds}-V_{gs} characteristics generated by the model in both linear and log scales compared to TCAD. The device parameters assumed are: t_{ox} = 2 nm, $\varepsilon_{ox} = \varepsilon_{si}$, x_s = 10 nm, $N_a = 10^{20}$ cm^{-3}, and μ_0 = 200 cm^2/V-s. The gate work function is that of n$^+$ silicon.

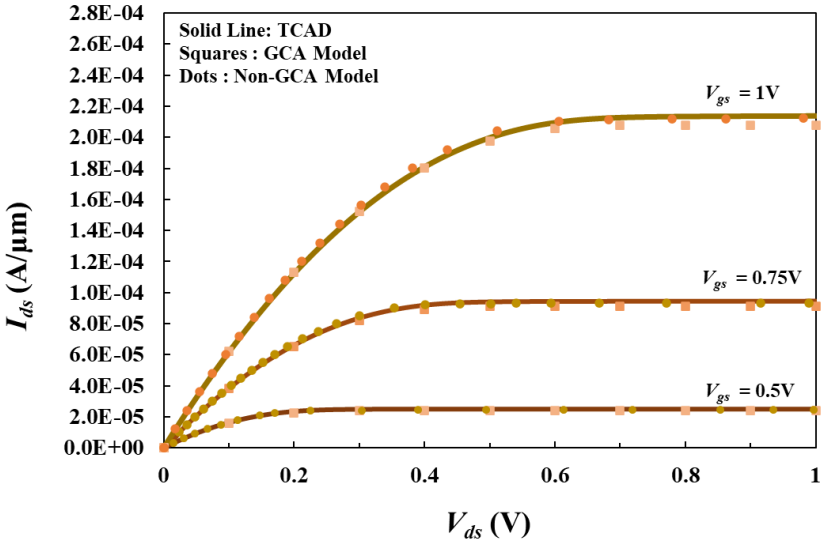


Figure 5.10 I_{ds}-V_{ds} characteristics generated by the model compared to TCAD. The squares are from the GCA model discussed in this section. The dots are from the non-GCA model discussed in the next section.

5.2.2 Non-GCA Model with $n = 1$ and $n = 2$ Velocity Saturation

Below 1 μm channel length, MOSFET currents are limited by velocity saturation. We consider two velocity saturation models here: n = 1 and n = 2. For n = 1,

$$I_{ds} = \frac{\mu_0 W Q_i}{1+(\mu_0/v_{sat})(dV/dy)} \frac{dV}{dy}. \tag{5.29}$$

For n = 2,

$$I_{ds} = \frac{\mu_0 W Q_i}{\sqrt{1+(\mu_0/v_{sat})^2 (dV/dy)^2}} \frac{dV}{dy}. \tag{5.30}$$

In the above, μ_0 is the low-field mobility, v_{sat} is the saturation velocity. Application of the GCA model from the previous section to Eqs. (5.29) or (5.30) yields unphysical I_{ds}-V_{ds} results: either a negative slope or no solution beyond the I_{ds} peak, because dV/dy diverges when v_{sat} is reached in the channel.

To deal with the problem, we apply a non-GCA model by adding a lateral field term, $Q_{iL} = \varepsilon_{si} d_{si} d^2V/dy^2$, to the mobile charge density, where d_{si} is a depth parameter ≤ x_s, the undoped region depth. Thus, for n = 1,

$$I_{ds} = \frac{\mu_0 W}{1+(\mu_0/v_{sat})(dV/dy)} \left(Q_i' + \varepsilon_{si} d_{si} \frac{d^2V}{dy^2} \right) \frac{dV}{dy} \tag{5.31}$$

The gate induced mobile charge density Q_i in the GCA model is given by Eq. (5.27) and plotted in Fig. 5.8 (solid curves). It never goes negative because Q_i is the total mobile charge density in the GCA model. Physically, however, the gate induced charge density, proportional to the oxide field in the gate direction, does go negative over the channel portion where the channel potential V becomes higher than the gate potential. In the non-GCA model, Eq. (5.31), Q_i' can go negative

while the total mobile charge density, $Q_i{}' + Q_{iL}$, stays positive. In this work, we make use of the well-known expression for $Q_i{}'$:

$$\mathrm{Q_i{}' = C_{inv}(V_{gs}\text{-}V_t\text{-}mV)}, \tag{5.32}$$

with *m* the body effect factor given by $1 + (\varepsilon_{si}/x_s)/(\varepsilon_{ox}/t_{ox})$. The other parameters are extracted from the GCA model as follows: $C_{inv}(V_{gs} - V_t)$ is given by $Q_i(V = 0)$, i.e., the initial Q_i value in Fig. 5.8. The slope, mC_{inv}, is set so that the maximum positive area under Eq. (5.32), $\int_0^{\frac{V_{gs}-V_t}{m}} Q_i{}'(V)dV = \frac{C_{inv}(V_{gs}-V_t)^2}{2m}$, equals the integrated area of Eq. (5.27) in Fig. 5.8, $\int_0^{\infty} Q_i(V)dV$, for each V_{gs}. The so-obtained $Q_i{}'(V)$ are shown as dashed lines in Fig. 5.8. $Q_i{}'$ changes sign when the potential in the channel exceeds $\frac{V_{gs}-V_t}{m}$ and the field in the oxide reverses. The extracted C_{inv} is somewhat lower than C_{ox} because of the finite inversion layer capacitance. It can be seen in Fig. 5.8 that C_{inv}, proportional to the slope $-dQ_i/dV$, decreases towards lower gate voltages. In this case, C_{inv} = 0.9 C_{ox}, 0.85 C_{ox}, and 0.75 C_{ox}, respectively for the three V_{gs} shown.

By multiplying the denominator to the LHS, Eq. (5.31) can be integrated once:

$$\frac{\mathrm{I_{ds}}}{\mu_0 \mathrm{W}}\mathrm{y} + \frac{\mathrm{I_{ds}}}{\mathrm{v_{sat}W}}\mathrm{V} = \mathrm{C_{inv}}\left[(\mathrm{V_{gs}}\text{-}\mathrm{V_t})\mathrm{V}\text{-}\frac{\mathrm{m}}{2}\mathrm{V}^2\right] + \frac{\varepsilon_{si} d_{si}}{2}\left[\left(\frac{\mathrm{dV}}{\mathrm{dy}}\right)^2 \text{-}\mathrm{E_1^2}\right] \tag{5.33}$$

where $E_1 = (dV/dy)|_{y=0}$ at the source. Since d^2V/dy^2 is negligible at the source, setting $V = 0$ in Eq. (5.31) gives

$$\mathrm{E_1} = \frac{\mathrm{I_{ds}}}{\mu_0 \mathrm{W Q_i{}'(V=0)}\text{-}(\mu_0/\mathrm{v_{sat}})\mathrm{I_{ds}}}. \tag{5.34}$$

For a given I_{ds}, Eq. (5.33) is a 1st order ordinary differential equation that can be solved numerically for $V(y)$. V_{ds} is then the value of $V(y)$ when y reaches L. The n = 1 I_{ds}-V_{ds} characteristics generated by the model are shown in Fig. 5.11. They are in close agreement with TCAD simulations.

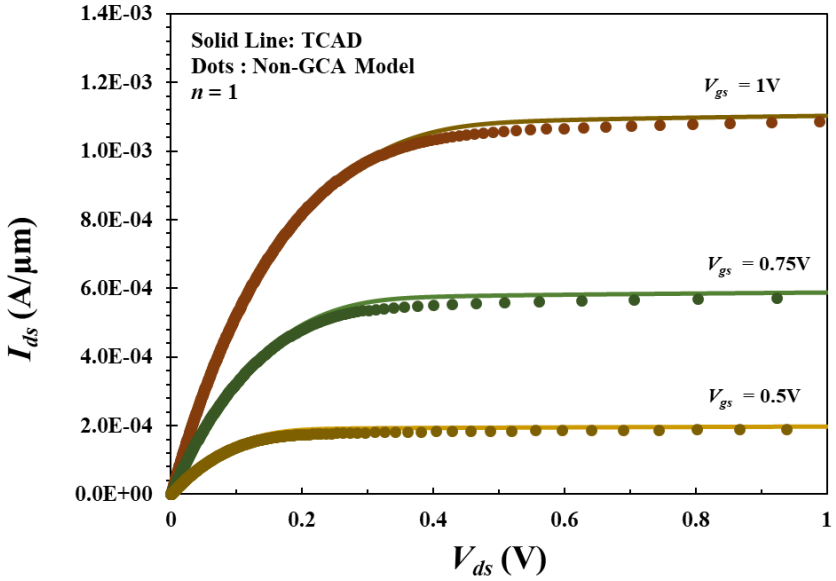


Figure 5.11 I_{ds}-V_{ds} characteristics generated by the n = 1 non-GCA model compared to TCAD. The parameters are L = 100 nm, v_{sat} = 10^7 cm/s, d_{si} = 5 nm. The rest of parameters are the same as those in the caption to Fig. 5.9.

Similarly, the n = 2 non-GCA model looks like

$$I_{ds} = \frac{\mu_0 W}{\sqrt{1+(\mu_0/v_{sat})^2(dV/dy)^2}}\left(Q_i' + \varepsilon_{si} d_{si}\frac{d^2V}{dy^2}\right)\frac{dV}{dy}. \tag{5.35}$$

To numerically solve this equation, a different procedure from that of n = 1 is followed. Equation (5.35) is first converted to a first-order differential equation in $g(V) \equiv (dV/dy)^2$ [3]:

$$I_{ds}^{\ 2}\left[1+\left(\frac{\mu_0}{v_{sat}}\right)^2 g\right]=W^2\mu_0^{\ 2}g\left[Q_i'(V)+\frac{\varepsilon_{si}d_{si}}{2}\frac{dg}{dV}\right]^2. \tag{5.36}$$

The initial condition g(V = 0) is obtained by neglecting d^2V/dy^2 in Eq. (5.35) and solving for $(dV/dy)^2$:

$$g(V=0)=\frac{(I_{ds}/\mu_0 W)^2}{\left[Q_i'(V=0)\right]^2-(I_{ds}/Wv_{sat})^2} \tag{5.37}$$

After $g(V)$ is solved numerically, $g^{-1/2} = dy/dV$ is readily integrated from V = 0 to V_{ds} where $y = L$ is reached. The n = 2 I_{ds}-V_{ds} characteristics solved by the model are shown in Fig. 5.12 to be consistent with TCAD results.

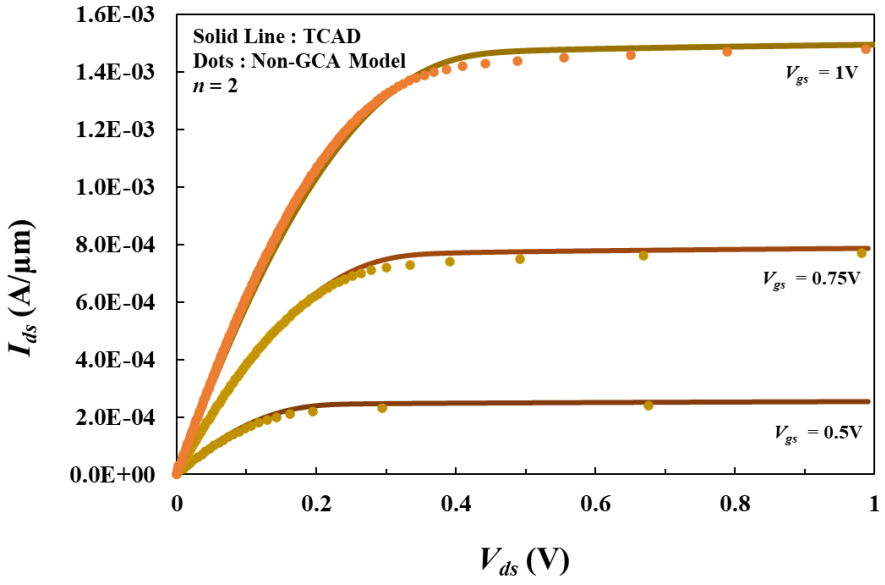


Figure 5.12 I_{ds}-V_{ds} characteristics generated by the n = 2 non-GCA model compared to TCAD. The parameters are the same as those in Fig. 5.11.

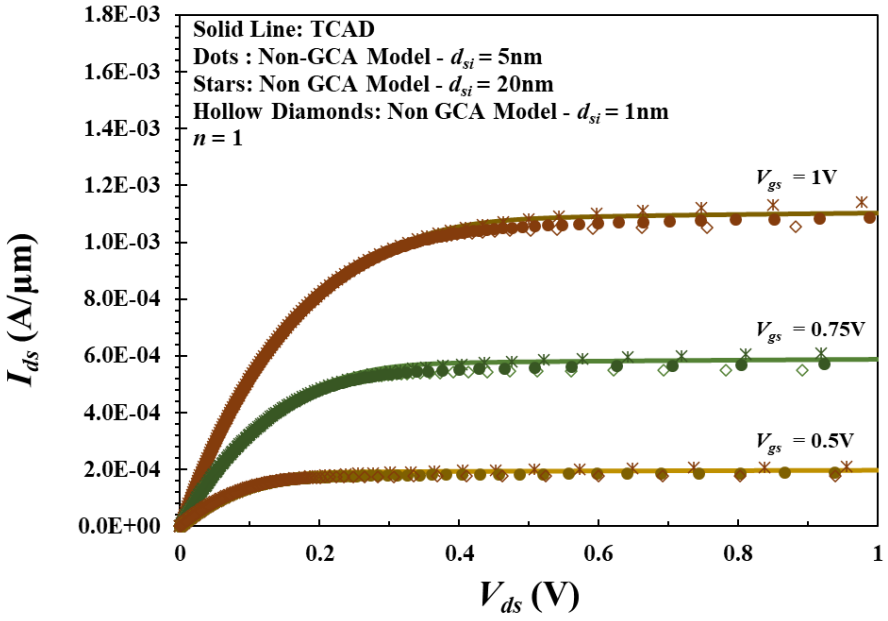


Figure 5.13 I_{ds}-V_{ds} characteristics generated by the n = 1 non-GCA model compared to TCAD with different d_{si}. The rest of parameters are the same as those in the caption to Figs. 5.9 and 5.11.

5.2.3 Comparison with Hardware Data by Adding R_{sd}

Model with Parasitic Source and Drain Resistance

In reality, there are source and drain series resistances in a MOSFET device that can adversely affect the drain current. An example is shown in Fig. 5.14(a) where the I_{ds}-V_{ds} characteristics generated by the intrinsic model of Section III are plotted alongside with the published data of 20 nm bulk MOSFETs. It is relatively straightforward to add source and drain series resistance to the non-GCA model since it computes V_{ds} for a given I_{ds}. Models that compute I_{ds} for a given V_{ds} would require multiple iterations.

With a source resistance R_s, the applied V_{gs} is reduced by the *IR* drop such that the gate voltage experienced by the intrinsic device is

$$V_{gs}' = V_{gs} - R_s I_{ds}. \tag{5.38}$$

For each given set of V_{gs}' and I_{ds} as the input, the $n = 1$ velocity saturation model in the previous section is called upon to calculate V_{ds}' of the intrinsic device. Then the external source-to-drain voltage is given by

$$V_{ds} = V_{ds}' + (R_s + R_d) I_{ds}. \tag{5.39}$$

By repeating the procedure for a series of I_{ds} values with the same V_{gs}, an I_{ds}-V_{ds} characteristic is generated for that V_{gs}. Note that V_{gs}' takes on different values as I_{ds} is varied under the same V_{gs}.

Figure 5.14(b) shows that, with proper gate-voltage dependent series resistance added to the $n = 1$ velocity saturation model, the I_{ds}-V_{ds} characteristics generated closely match the 20 nm MOSFET data. Here, R_s and R_d vary from 150 Ω-μm to 200 Ω-μm to 600 Ω-μm for V_{gs} = 0.9, 0.7, and 0.5 V, respectively. The other device parameters in the model are listed in the figure captions.

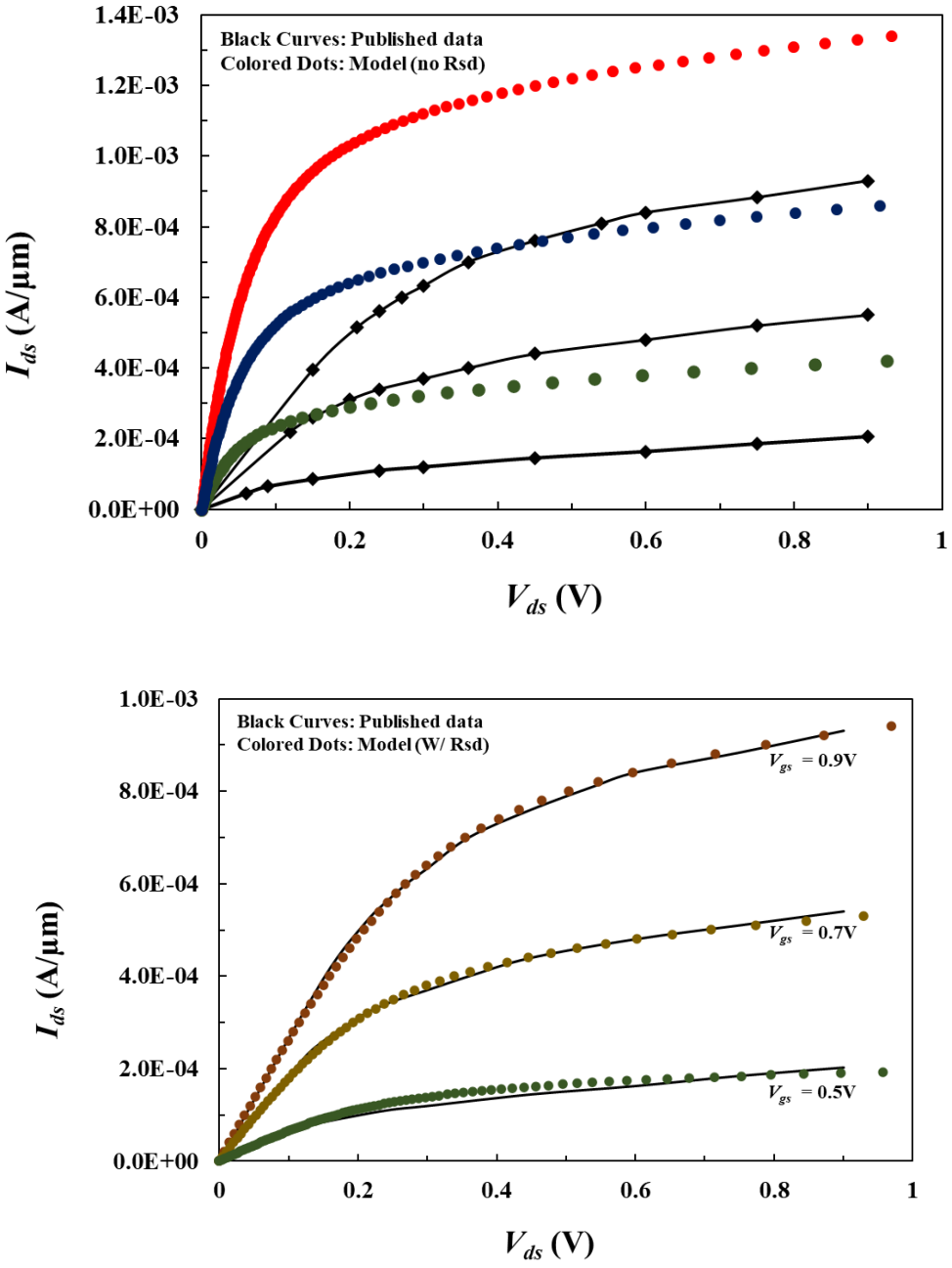


Figure 5.14 I_{ds}-V_{ds} characteristics generated by the n = 1 velocity saturation model compared to the published data of 20 nm MOSFETs. (a) No source and drain resistance. (b) With source and drain resistance (values given in the main text) added to the model. In both (a) and (b), solid curves are the published hardware data, dots are the model results. Parameters used in the model are: L = 20 nm, EOT = 1.2 nm, m = 1.3, V_t = 0.25 V, d_{si} = 15 nm. The mobility and saturation velocity assumed are the same as in the earlier figures. Similar C_{inv}, namely, C_{inv} = 0.87 C_{ox}, 0.71 C_{ox}, and 0.65 C_{ox}, for V_{gs} = 0.9, 0.7, and 0.5 V respectively are used.

References:

[1] M.-H. Su, C. Hong, and Y. Taur, "A Non-GCA Model for Ground-Plane MOSFETs", *Solid-State Electronics*, vol. 209, p. 108754, Nov. 2023

[2] Yuan Taur, Tak H Ning, "Fundamentals of modern VLSI devices", *Cambridge university press*, December 2021

Acknowledgments

Chapter 5, in full, is a reprint of the material as it appears in M.-H. Su, C. Hong, and Y. Taur, "A Non-GCA Model for Ground-Plane MOSFETs", *Solid-State Electronics*, vol. 209, p. 108754, Nov. 2023. The dissertation author was the primary investigator and author of this paper.

CHAPTER 6 SCE OF ET-SOI MOSFETS

SOI CMOS involves building more or less conventional MOSFETs on a thin layer of crystalline silicon, as illustrated in Fig. 6.1. The thin layer of silicon is separated from the substrate by a thick layer (typically 25 nm or more) of buried SiO_2 film, thus electrically isolating the devices from the underlying silicon substrate and from each other. An SOI CMOS process can be readily developed due to the compatibility with established bulk processing technology.

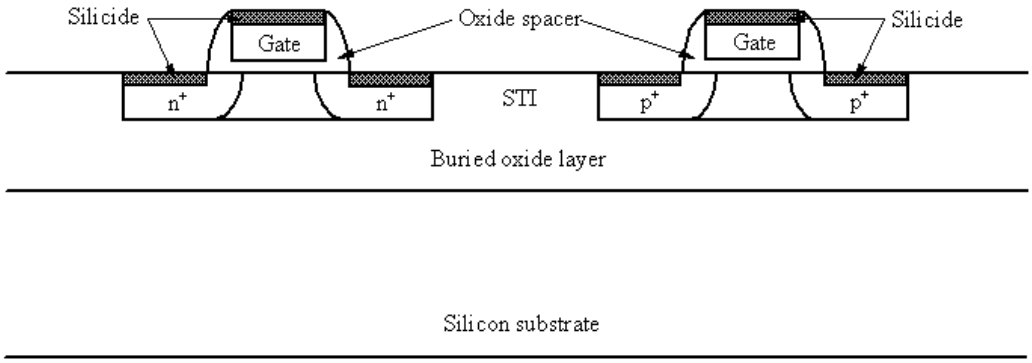


Figure 6.1. A schematic cross-section of SOI CMOS, with shallow trench isolation, dual polysilicon gates, and self-aligned silicide.

6.1 Short-Channel SOI MOSFETs

Short-Channel SOI MOSFETs

It has long been reported in the literature that fully-depleted SOI MOSFETs are more susceptible to short-channel effects (SCE) for lack of a conducting plane not too far below the device region (Su *et al.*, 1994; Wong *et al.*, 1994). The 2-D scale length model, however, does not apply to SOI MOSFETs because no closed rectangular region can be defined with known

potential values on its boundary. TCAD has become a necessary tool for investigating SCE in SOI MOSFETs (Xie *et al*., 2013).

2-D Fields in the Buried Oxide

Figure 6.2 compares the constant potential contours of a bulk ground-plane MOSFET with an SOI MOSFET side by side. In the bulk case, the 2-D fields are confined to the depletion (undoped) region bounded below by the conducting substrate. In the SOI case, on the other hand, the 2-D fields from the source and drain penetrate into the thick BOX region. Conceptually, since the scale length is given by the effective vertical distance between the gate and the bottom conductor, deeper field penetration would worsen the SCE. The mitigating factor is that the depth of field penetration is channel length dependent. Only for very long channel devices is the vertical distance given by the entire BOX thickness. For short channel devices where it matters, the effective depth of field penetration is much less than the BOX thickness.

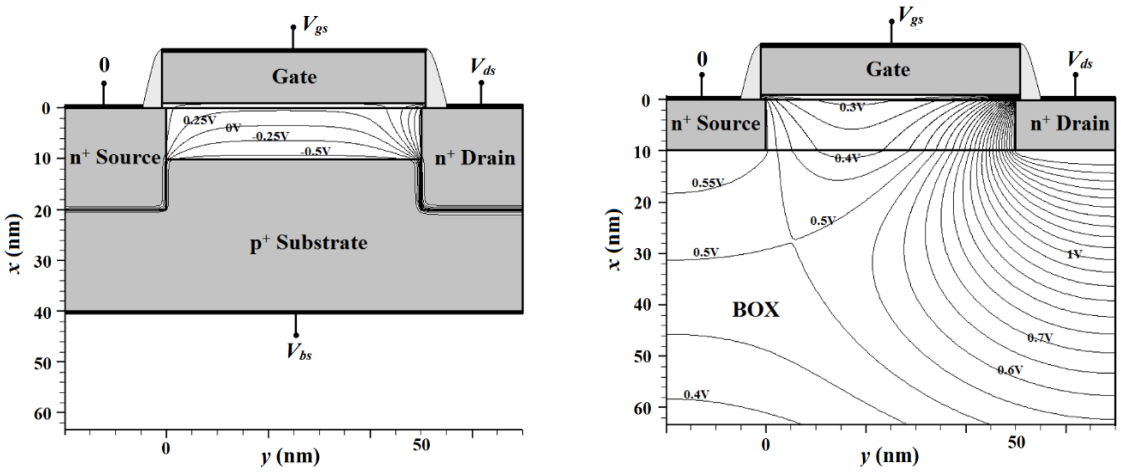


Figure 6.2. 2-D constant potential contours of (a) bulk and (b) SOI MOSFETs. For both devices, t_{ox} = 1 nm, L = 50 nm, V_{ds} = 1.0 V, V_{bs} = V_{bg} = 0. For the bulk, the depletion region (undoped) depth is 10 nm. For the SOI, the silicon thickness is 10 nm, and the BOX thickness is 200 nm. The labels refer to the potential as that defined $\frac{d^2\psi}{dx^2} = \frac{q}{\varepsilon_{si}} n_i e^{q(\psi - V)/kT}$, i.e., $\psi(x, y) \equiv -[E_i(x, y) - E_{fs}]/q$. The value of V_{gs} is such that the minimum surface potential between the source and drain, $\psi_{s,min}$, is 0.29 V (after Xie *et al*., 2013).

Figure 6.3 compares the V_t roll-off curves of the bulk and SOI MOSFETs in Fig. 6.2. By defining an $L_{\min}$ where the V_t roll-off is ΔV_t = –50 mV, we obtain $L_{\min}$ = 29 nm for the bulk MOSFET[1] and $L_{\min}$ = 58 nm for the SOI MOSFET. To gain further insight, $\Delta\psi_{s,\min}$, the minimum surface potential between the source and drain of a short channel device with respect to that of the long channel device, is plotted as a function of L in Fig. 6.4. For the bulk MOSFET, $\Delta\psi_{s,\min}$ versus L is largely proportional to $\exp[-\pi L/(2\lambda)]$, as expected from the scale length model with a λ of 12.6 nm given by $\frac{1}{\varepsilon_{ox}} tan\left(\frac{\pi t_{ox}}{\lambda}\right) + \frac{1}{\varepsilon_{si}} tan\left(\frac{\pi W_{dm}}{\lambda}\right) = 0$ for W_{dm} = 10 nm and t_{ox} = 1 nm. For the SOI MOSFET, first, the exponential slope is far less steep compared to that of the bulk

device, indicating longer λ and worse SCE. Second, the exponential slope is not constant, but increases towards shorter L, i.e., the effective λ decreases with decreasing L. This is attributed to the decrease of the depth of 2-D field penetration in Fig. 6.2(b) as L is shortened.

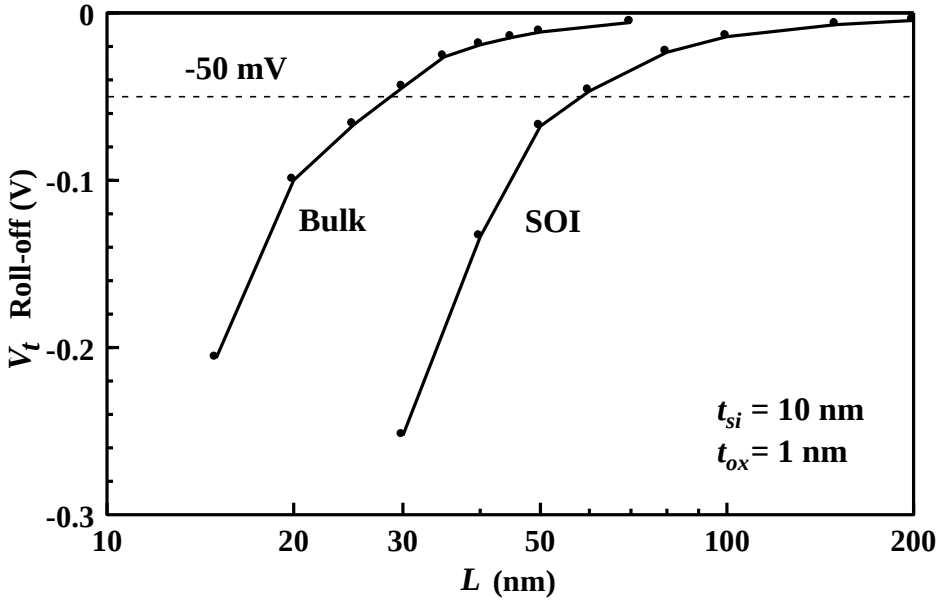


Figure 6.3. Short-channel V_t roll-off of the bulk and SOI MOSFETs in Fig. 6.2. Here, V_t is defined as the V_{gs} value where $I_{ds} = 10^{-8}$ A ($W/L = 1$), and V_t roll-off is defined as $\Delta V_t = V_t(L) - V_t$(Long channel). The -50 mV intercepts are L = 29 nm for bulk and 58 nm for SOI.

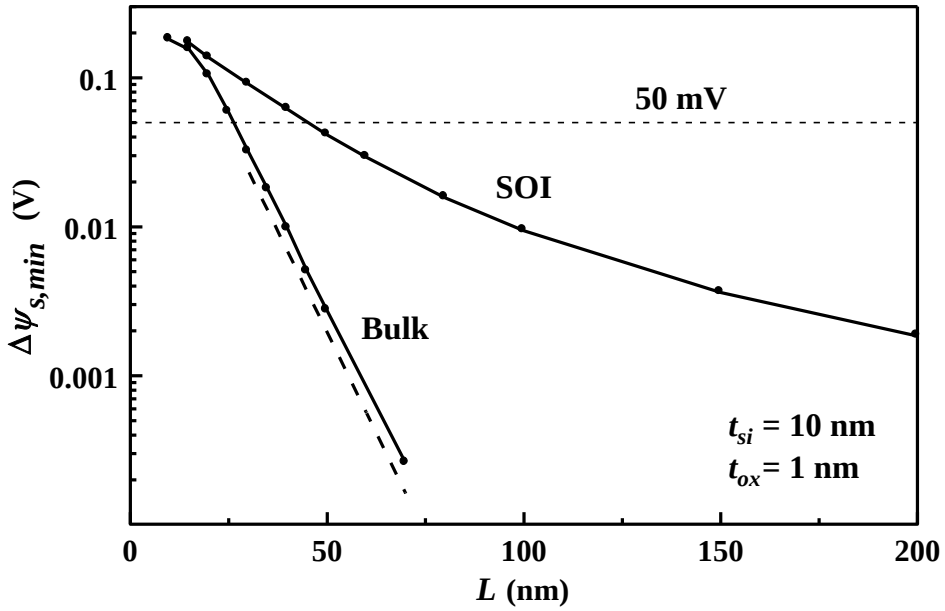


Figure 6.4. $\Delta\psi_{s,min}$, the minimum surface potential between the source and drain of a short channel device with respect to that of the long channel device for the SOI and bulk MOSFETs in Fig. 6.5. The dashed line is $\exp[-\pi L/(2\lambda)]$ with λ = 12.6 nm. The 50 mV intercepts are L = 26 nm for bulk and 45 nm for SOI.

6.2 Effects of BOX Thickness, Silicon Thickness, and Backgate Bias on SCE

Tremendous progress has recently been made on ET-SOI (extremely thin silicon-on-insulator) material and technology. Silicon film as thin as 5 nm and BOX (buried oxide) layer as thin as 10 nm are currently available. They are expected to allow scaling of MOSFET channel lengths to a regime competitive with FinFETs.

An analytic scale length model has been developed that works well for predicting the SCE (short-channel effect) of bulk and DG (double-gate) MOSFETs. However, no such analytic model is available for SOI MOSFETs. By using TCAD simulations, an empirical expression of minimum

channel length has been worked out for SOI devices with BOX layers 200 nm thick. But it is not clear how it may improve with thinner BOX layers.

In this work, we extend the investigation of SCE in ET-SOI MOSFETs in terms of the minimum channel length as a function of the BOX thickness and silicon thickness. Another factor is the effect of backgate bias on SCE. To realize a threshold voltage target in the range of 0.3-0.4 V (nMOS), either a midgap gate with positively biased backgate or an n^+ silicon work function gate with negatively biased backgate can be used. They make a significant difference on SCE because in subthreshold the mobile charge density peaks at the back surface in the former case and peaks at the front surface in the latter case.

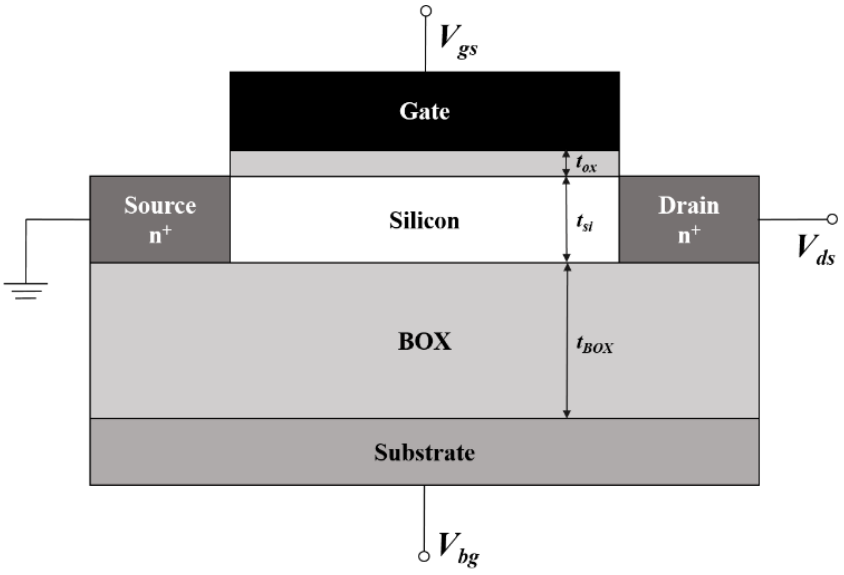


Figure 6.5. Cross-section of ET-SOI MOSFET investigated in this work. The range of BOX thickness is 10-200 nm. The range of silicon thickness is 2-10 nm. An EOT of 1 nm is assumed. V_{ds} = 1.0 V. Different type and concentration of substrate doping have been studied.

Effect of BOX and Silicon Thickness on SCE

Fig. 6.5 shows a schematic cross-section of the ET-SOI n-channel MOSFET studied in this work. The silicon body is undoped. The substrate is either lightly doped (10^{15} cm^{-3}, n^- or p^-) or a ground plane (5×10^{18} cm^{-3}, n^+ or p^+ GP). A gate oxide thickness of 1 nm is assumed. Both midgap and n^+ silicon gate work functions are considered.

Fig. 6.6 shows the threshold roll-off curves for different BOX thickness obtained from TCAD simulations. V_t-roll-offs are extracted from high-drain bias (V_{ds} = 1.0 V) subthreshold I_{ds}-V_{gs} characteristics at a constant current level of 10^{-8} A (W/L = 1) for different channel lengths. The minimum channel length (L_{min}) is defined as the channel length with a V_t-roll-off of 100 mV.

The improvement of L_{min} with BOX thickness is rather moderate—about 20% from a BOX thickness of 200 nm to 10 nm. This is because the depth of 2-D field penetration into BOX is channel length dependent, ~ $0.2\times L$ empirically [4]. Therefore, thick BOX does not pay a penalty, in terms of the field penetration hence SCE, as much as the physical BOX thickness.

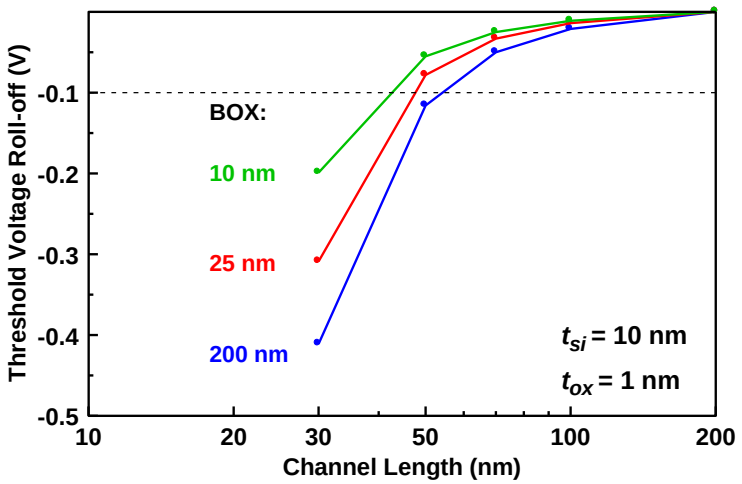


Figure 6.6. Short-channel V_t roll-off versus BOX thickness. A p^+ GP connected to the n^+ source is assumed in all three cases. L_{min} are 52, 47, 42 nm for BOX thickness of 200, 25, 10 nm, respectively.

Fig. 6.7 plots the V_t-roll-off of ET-SOI MOSFETs with different silicon thickness for the same BOX thickness of 10 nm. In this case, L_{min} is very sensitive to the silicon thickness, improving by over 2× when t_{si} is reduced from 10 nm to 2 nm. The latter is close to the quantum limit below which the threshold voltage becomes highly sensitive to the silicon thickness. An empirical expression for L_{min} is

$$L_{\min} \approx 3.3 \times (t_{si} + l_0)$$

where $l_0 \sim 3$ nm for $t_{ox} = 1$ nm and $t_{BOX} = 10$ nm. The experimental result of 20 nm MOSFETs with 3.5 nm silicon film lends support to the above expression.

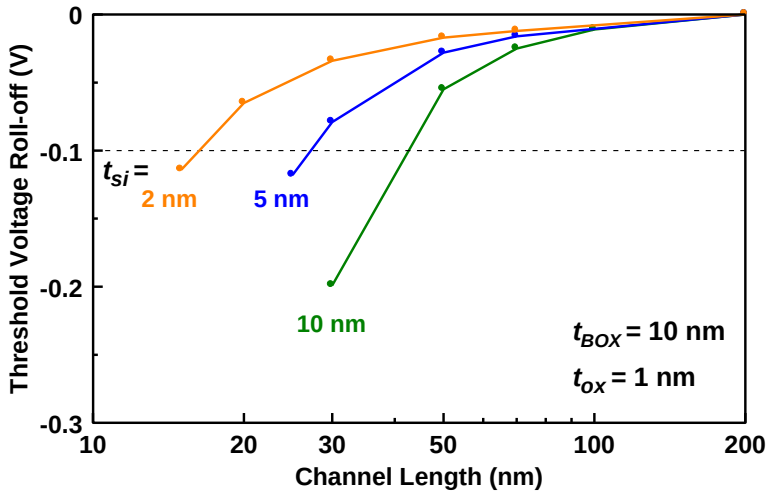


Figure 6.7. Short-channel V_t roll-off versus silicon thickness. A p^+ GP connected to the n^+ source is assumed in all three cases. L_{min} are 42, 27, 16 nm for silicon thickness of 10, 5, 2 nm, respectively.

Effects of Substrate Doping, Gate Work Function, and Back Gate Bias on SCE

For thin BOX MOSFETs, the substrate doping type and concentration have an effect on SCE. Figure 6.8 compares the V_t-roll-off with a p^- (10^{15} cm^{-3}) substrate, a p^+ ground plane (5×10^{18} cm^{-3}), an n^- (10^{15} cm^{-3}) substrate, and an n^+ ground plane (5×10^{18} cm^{-3}). The p^+ ground plane helps SCE slightly compared to that of a p^- substrate. But with an n^+ ground plane or n^- substrate, the SCE is significantly worse. The case of nMOS on n^+ ground plane is relevant because its SCE is equivalent to that of a pMOS on p^+ ground plane if the p^+ layer is formed uniformly over the entire substrate. From the SCE point of view, it is most desirable to have a p^+ ground plane under nMOS and an n^+ ground plane under pMOS, much like the p-well and n-well configuration in a bulk CMOS technology.

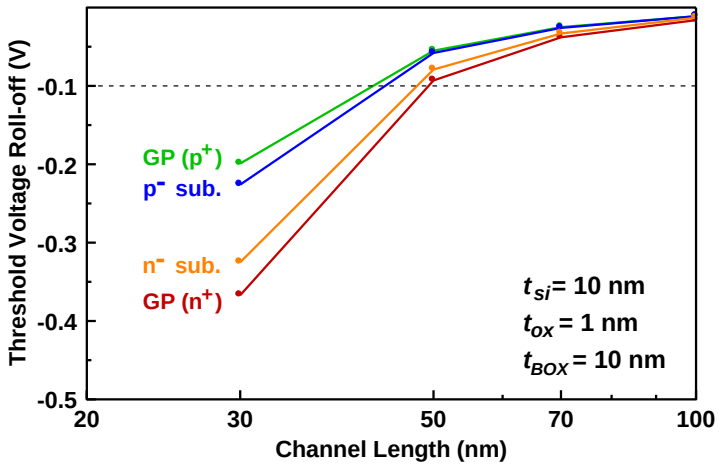


Figure 6.8. Comparison of V_t roll-off of nMOS with respect to substrate doping type and concentration. The substrate is grounded to the source in all cases.

To realize a desirable threshold voltage in the range of 0.3-0.4 V, the choice of gate work function plays a major role. Midgap work function gives too high a value and n^+ silicon work function too low. Since it is difficult to fine tune the gate work function to the precise value between midgap and n^+, the common practice is to adjust V_t by a substrate bias, V_{bg}. For midgap gates, a positive V_{bg} is applied to lower V_t while for n^+ silicon gates, a negative V_{bg} is applied to raise V_t. They have opposite effects on the SCE.

Figure 6.9 compares the V_t-roll-off of three devices, all with long channel threshold within 0.3-0.4 V. The first device has midgap work function on the front gate with V_{bg} = 3 V to lower the threshold. The second device has V_{bg} = 0 and relies on the work function of front gate to adjust V_t to the desired value. In this case, the SCE is independent of the V_t value. The third device has n^+ work function on the front gate with V_{bg} = –3 V to raise the threshold. The figure shows that the

first device with $V_{bg} > 0$ has the worst SCE while the third device with $V_{bg} < 0$ has the best SCE, with an L_{min} about 30% shorter than the one with $V_{bg} = 3$ V.

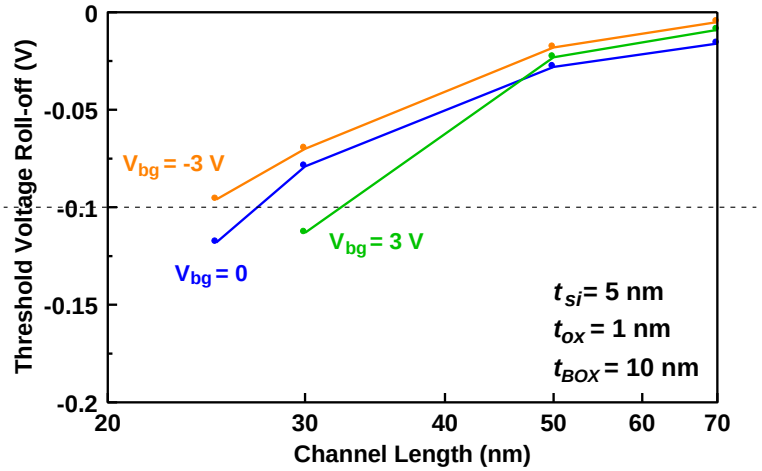


Figure 6.9. Comparison of SCE for different gate work function and backgate bias. With midgap work function, $V_{bg} = 3$ V is applied to lower V_t and with n^+ silicon work function, $V_{bg} = -3$ V is applied to raise V_t. The middle case relies on the gate work function with no V_{bg} to tune V_t to the right range. L_{min} are 32, 27, 24.5 nm respectively for the three cases.

The underlying reason is made clear in Fig. 6.9 where the potential versus depth is compared between the devices with positive and with negative V_{bg}. The device with $V_{bg} = -3$ V has a field in silicon such that the potential is higher at the front surface. Its subthreshold swing is ~ 65 mV/decade or

$$SS \approx \frac{(\varepsilon_{si}/\varepsilon_{ox})t_{ox} + t_{si} + (\varepsilon_{si}/\varepsilon_{ox})t_{BOX}}{t_{si} + (\varepsilon_{si}/\varepsilon_{ox})t_{BOX}} \times 60 \text{ mV/decade}$$

as expected. In the above, the numerator is the dielectric distance between the gate and the substrate, and the denominator is the dielectric distance from the front channel to the substrate. The device with $V_{bg} = 3$ V, however, has a field in the opposite direction such that the potential is highest at the back surface. The subthreshold swing is ~ 77 mV/decade or

$$SS \approx \frac{(\varepsilon_{si}/\varepsilon_{ox})t_{ox} + t_{si} + (\varepsilon_{si}/\varepsilon_{ox})t_{BOX}}{(\varepsilon_{si}/\varepsilon_{ox})t_{BOX}} \times 60 \text{ mV/decade}$$

Here, the denominator is reduced to the dielectric distance between the back channel and the substrate. This means that in addition to the worse V_t roll-off, the midgap device with V_{bg} = 3 V has degraded subthreshold swing such that its off current level is orders of magnitude higher than the device with V_{bg} = –3 V.

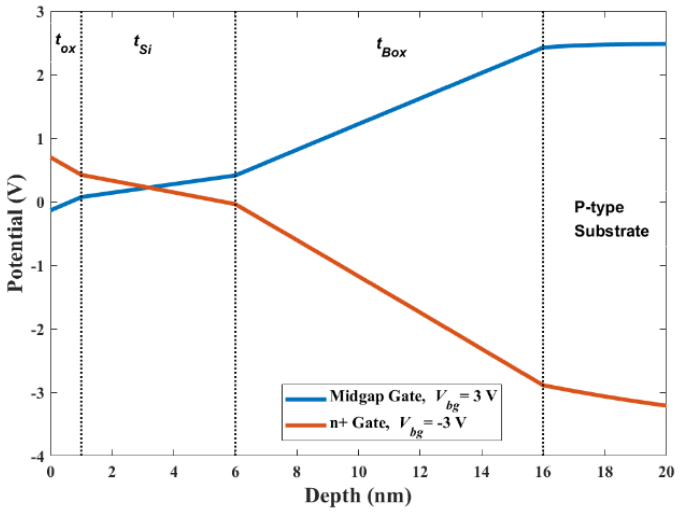


Figure 6.10. Potential versus depth for the cases of V_{bg} = 3 V and V_{bg} = -3 V in Fig. 6.8. The gate biases are in subthreshold such that $I_{ds} = 10^{-8}$ A (W/L = 1) in both devices.

References:

[1] Yuan Taur, Tak H Ning, "Fundamentals of modern VLSI devices", *Cambridge university press*, December 2021

[2] M.-H. Su, C. Hong, S. Cristoloveanu and Y. Taur, "Effects of BOX Thickness, Silicon Thickness, and Backgate Bias on SCE of ET-SOI MOSFETs," Microelectronic Engineering, 238, 111506, Jan. 2021.

Acknowledgments

Chapter 6, in full, is a reprint of the material as it appears in M.-H. Su, C. Hong, S. Cristoloveanu and Y. Taur, “Effects of BOX Thickness, Silicon Thickness, and Backgate Bias on SCE of ET-SOI MOSFETs,” Microelectronic Engineering, 238, 111506, Jan. 2021. The dissertation author was the primary investigator and author of this paper.

CHAPTER 7 CONCLUSION

In this dissertation, a continuous MOSFET model has been developed that takes the effect of lateral field gradient on carrier density into account. It goes beyond the GCA model and produces finite output conductance in the saturation region, without invoking CLM. It also explains why the carrier density is not pinched off even though the oxide field is zero or negative beyond the saturation point. Model generated I_{ds}-V_{ds} and g_{dc}-V_{ds} curves are consistent with TCAD simulations.

By capturing the essential physics, namely, the effect of lateral field gradient on carrier density, the model reduces the 2D potential problem to a first-order ordinary differential equation that can be solved readily on a spread sheet or with a standard mathematical tool. With regional approximations, the differential equation is solved analytically for $V(y)$ in the velocity saturation region. When coupled with modified CLM relations between current and GCA length, closed-form expressions are derived for the output conductance under both $n = 1$ and $n = 2$ models. The analytic solution derived for the velocity saturation region can be used to construct a compact model by connecting it to the conventional GCA solution for the triode region.

A physical model for ground-plane MOSFETs near the limit of bulk CMOS scaling is also developed. It starts with a GCA model for long channel devices by analytically solving 1-D Poisson's equation, taking into account depletion in the ground-plane. A non-GCA model continuous into the velocity saturation region is then formulated with the addition of a lateral-field driven mobile charge density in the current continuity equation. By incorporating series source

and drain resistance to the model, it produces I_{ds}-V_{ds} characteristics similar to the published 20 nm MOSFET data.