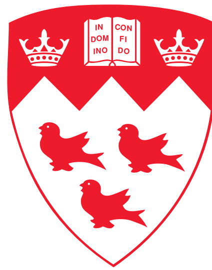
Theory and Simulation of Novel Low-Power Nanotransistors



Raphaël Joachim Prentki

Department of Physics, McGill University

Montréal, Québec, Canada

November 2021

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of *Doctor of Philosophy.*

© Raphaël Prentki, 2021

This thesis is dedicated to my grandmother, Marysia Prentki.

Abstract

Moore's law predicts an exponential growth of the number of transistors on integrated circuits (ICs). Transistors are now being downscaled to nanometric dimensions, making it increasingly difficult to maintain their power consumption at an acceptable level. Indeed, thermodynamics and electrostatics set a lower bound for the subthreshold swing (STS) of the industry-standard silicon fin field-effect transistor (FinFET) and, in turn, the power supply voltage of FinFET-based ICs. To resolve this power dissipation problem, the semiconductor industry will need to adopt transistors with novel channel materials, gate geometries, and/or charge transport mechanisms. Low-dimensional materials, such as silicon nanowires (NWs), are required for gate-all-around (GAA) field-effect transistors (FETs) and improved device electrostatics. The tunnel field-effect transistor (TFET) harnesses band-to-band tunnelling to achieve low STS. In this thesis, through various analytical and numerical tools of electrostatics, solid-state physics, as well as quantum and statistical mechanics, I investigate the nanoscale device physics of these novel transistors and propose potential solutions to the power dissipation problem.

Low-dimensional semiconductors exhibit weak screening, which is detrimental to the performance and scalability of nanotransistors. Typically, screening in semiconductors is strengthened by chemical doping. However, semiconductor doping is limited by such practical concerns as bandgap narrowing and solid solubility limits of dopants. To resolve this issue, I introduce bound-charge engineering (BCE), a novel and relatively simple scheme where a surface bound charge is engineered on the interface between a semiconductor and a neighbouring oxide to strengthen screening. I establish BCE by basic electrostatics; BCE is thus widely applicable to emerging materials and novel devices, in principle. For FET applications, several oxides should be used in conjunction: a low-permittivity spacer oxide for strong screening and a high-permittivity gate oxide for high gate control. I substantiate

the BCE scheme by atomistic quantum transport simulations based on the nonequilibrium Green's function (NEGF) formalism and the tight-binding (TB) model. In silicon NW TFETs, I demonstrate that BCE increases the on-state current by orders of magnitude, and the combination of oxides yields minimal STS. This enables the practical application of TFETs at higher clock frequency and lower power supply voltage, paving a way toward improved low-power transistors. To expand our understanding of BCE qualitatively and quantitatively, I derive an analytical surface potential model for cylindrical GAA BCE-assisted silicon NW FETs with arbitrary and possibly distinct spacer and gate oxides. This model is based on scaling theory and verified against NEGF–TB simulations; it provides an intuitive formalism for developing and modelling devices with BCE. Finally, I apply BCE to reduce direct source-to-drain tunnelling (DSDT) leakage in very short-channel FETs; DSDT is generally understood to set the ultimate scaling limit of FETs. Supported by NEGF–TB simulations and the surface potential model, I demonstrate that BCE can reduce DSDT down to acceptable levels in FETs with channel lengths as small as 1.5 nm, thereby paving a way toward ultra-scaled FETs.

Abrégé

La loi de Moore prédit une croissance exponentielle du nombre de transistors (T) sur les circuits intégrés (CI) et une miniaturisation des T. De nos jours, la taille des T approche le nanomètre, ce qui rend difficile le maintien de leur consommation d'énergie à un niveau acceptable. En effet, la thermodynamique et l'électrostatique (ES) fixent une limite supérieure pour la pente de sous-saturation (PSS) du T à effet de champ à ailettes en silicium (FinFET), le standard de l'industrie, et par conséquent, pour la tension d'alimentation des CI composés de FinFET. Pour résoudre ce problème de dissipation d'énergie, l'industrie des semi-conducteurs (SC) devra adopter des T avec de nouveaux matériaux de canal, de nouvelles géométries de grille et/ou de nouveaux mécanismes de transport de charge. Les matériaux à une ou deux dimensions, tels que les nanofils (NF) de silicium (Si), sont nécessaires pour les T à effet de champ (FET) à grille enrobante (GAA) et donc pour l'ES optimisée de ces dispositifs. Le T à effet tunnel (TFET) exploite l'effet tunnel de bande à bande pour obtenir une PSS élevée. Dans cette thèse, grâce à divers outils analytiques et numériques de l'ES, de la physique du solide, de la mécanique quantique et de la mécanique statistique, j'étudie la physique de ces dispositifs nanométriques et je propose des solutions possibles au problème de dissipation d'énergie.

Les SC à une ou deux dimensions sont caractérisés par un faible écrantage (EC) des charges électriques, ce qui nuit à la performance et à la miniaturisation des T. Dans les SC, l'EC est typiquement renforcé par le dopage chimique. Cependant, le dopage des SC est limité par des problèmes pratiques tels que le rétrécissement de la bande interdite et les limites de solubilité solide des dopants. Pour résoudre ce problème, je développe l'ingénierie des charges liées (en anglais, « bound-charge engineering », ou « BCE »), un nouveau processus relativement simple dans lequel on forme une charge de surface liée à l'interface entre un SC et un oxyde (OX) avoisinant afin de renforcer l'EC. J'établis le BCE par des concepts

élémentaires de l'ES ; en principe, le BCE est donc applicable aux matériaux émergents et aux nouveaux dispositifs. Pour les applications aux FET, plusieurs OX doivent être utilisés conjointement : un OX espaceur à faible permittivité pour un EC puissant et un OX de grille à permittivité élevée pour un contrôle de grille élevé. Je justifie la méthode du BCE par des simulations de transport quantique atomistique basées sur le formalisme de la fonction de Green hors-équilibre (NEGF) et un modèle de liaisons fortes (TB). Dans les TFET composés de NF de Si, je démontre que le BCE augmente le courant électrique de l'état passant de plusieurs ordres de grandeur. De plus, la combinaison d'OX maximise la PSS. Cela permet l'application pratique des TFET à une fréquence d'horloge plus élevée et à une tension d'alimentation plus faible, ouvrant la voie à des T prometteurs pour l'électronique de faible puissance. Afin de mieux comprendre le BCE sur les plans qualitatif et quantitatif, je calcule un modèle analytique du potentiel de surface des FET GAA cylindriques de NF de Si assistés par le BCE et composés d'OX espaceurs et de grille arbitraires et éventuellement distincts l'un de l'autre. Ce modèle est basé sur la théorie de l'échelle et vérifié à la lumière de simulations NEGF–TB ; il fournit un formalisme intuitif pour développer et modéliser des dispositifs exploitant le BCE. Enfin, j'applique le BCE pour réduire les courants de fuite par effet tunnel de la source au drain (ETSD) dans les FET à canal très court ; l'ETSD est généralement considéré comme le phénomène qui définit la taille minimale possible des FET. Grâce aux simulations NEGF–TB et au modèle de potentiel de surface, je démontre que le BCE peut réduire l'ETSD à des niveaux acceptables dans des FET ayant longueur de canal aussi petite que 1,5 nm, ouvrant ainsi la voie à des FET de taille minime.

Table of Contents

Abstract (English), Abrégé (Français) iii

Table of Contents, Lists of Figures and Tables vii

Acknowledgements xviii

Statement of Originality xxii

Mathematical Notation, Units, Physical Constants xxvi

Lists of Symbols and Abbreviations xxviii

Chapter 1 Introduction **1**

Chapter 2 The Power Dissipation Problem **7**

2.1 A Fundamental Switching Limit 8

2.1.1 Fundamentals of Field-Effect Transistors 8

2.1.2 Charge Transport in Metal–Oxide–Semiconductor FETs 11

2.1.3 The Subthreshold Swing 20

2.2 Power Consumption of Processors 23

2.2.1 Dynamic and Leakage Power 24

2.2.2 Minimizing Power Consumption with MOSFETs 27

2.2.3 Minimizing Power Consumption with Novel Devices 31

2.3 Summary 34

Chapter 3 Steep-Slope Field-Effect Transistors **37**

3.1 Large Transport Factor and the Tunnel FET 40

3.2 Large Body Factor and Negative Capacitance 45

3.3 Discussion and Comparisons 51

3.4 Summary 57

Chapter 4 Simulations of Semiconductor Devices **59**

4.1 From Many-Body Physics to TCAD 60

4.2 The Tight-Binding Model 66

4.2.1 The Slater–Koster Hamiltonian 66

4.2.2 Electronic Properties of Silicon Nanowires 74

4.3 The Nonequilibrium Green's Function Formalism 80

4.3.1 The Landauer–Büttiker Formalism 83

4.3.2 Formal Theory 87

4.4 Atomistic Quantum Transport 89

4.4.1 Iterative Procedure 89

4.4.2 Initial Guess for the Potential 93

4.5 Summary 96

Chapter 5 Bound-Charge Engineering **98**

5.1 Harnessing Surface Bound Charges 101

5.1.1 Basic Electrostatics 101
5.1.2 Application to FETs 103
5.2 Atomistic Quantum Transport Simulations 105
5.2.1 Depletion Length 106
5.2.2 Electric Field 109
5.2.3 Free and Bound Charges 109
5.3 High-Performance Tunnel FETs 114
5.4 Conclusion 117
Chapter 6 The Surface Potential of BCE-Assisted FETs **119**
6.1 The Scaling Theory of FETs 121
6.2 Potential Model and Depletion Length 123
6.3 The Physics of Bound-Charge Engineering 127
6.4 Comparison to NEGF–TB Simulations 128
6.5 Conclusion 132
Chapter 7 Reducing Tunnelling Leakage with BCE **133**
7.1 Tunnelling and Electrostatics 135
7.2 Atomistic Quantum Transport Simulations 137
7.3 High-Performance, Short-Channel MOSFETs 140
7.4 Conclusion 146
Chapter 8 Conclusion and Outlook **148**

Appendix A More Steep-Slope Field-Effect Transistors I

A.1 FETs with Large Transport Factors I

A.1.1 The Dirac-Source FET I

A.1.2 The Cold-Source FET V

A.1.3 The Superlattice FET and Bandstructure Engineering IX

A.1.4 The Impact–Ionization FET X

A.1.5 The Feedback FET and Band-Modulation FETs XV

A.2 FETs with Large Body Factors XIX

A.2.1 The Nanoelectromechanical FET XIX

A.2.2 The Piezoelectric FET XXI

A.3 FETs with Large Transport and Body Factors XXIII

Appendix B Validation of Simulations XXVI

B.1 Density of Mesh Points XXVI

B.2 Smoothness of the Permittivity XXVIII

Bibliography XXXI

List of Figures

2.1 Schematic of a typical field-effect transistor and its basic components. 9

2.2 Band diagrams of a typical metal–oxide–semiconductor field-effect transistor in the off state and on state. 13

2.3 Current–voltage characteristics (transfer and output) of a typical metal–oxide–semiconductor field-effect transistor. 15

2.4 Off-state band diagrams, transfer characteristics, and output characteristics of a typical metal–oxide–semiconductor field-effect transistor that exhibits drain-induced barrier lowering. 18

2.5 Off-state band diagrams illustrating direct source-to-drain tunnelling in typical long-channel and short-channel metal–oxide–semiconductor field-effect transistors. 19

2.6 Transfer characteristics of typical metal–oxide–semiconductor field-effect transistors with low and high subthreshold swing. 20

2.7 Circuit diagram of the various capacitive couplings to the channel of a field-effect transistor. 21

2.8 Ratio of leakage power to total processor power consumption as a function of power supply voltage. 29

2.9 Transfer characteristics of a typical metal–oxide–semiconductor field-effect transistor in cases of low and high gate metal workfunction. . . 31

2.10 Total processor power consumption per transistor as a function of power supply voltage for various values of transistor subthreshold swing. . . 32

3.1 Transfer characteristics of a typical metal–oxide–semiconductor field-effect transistor and an ideal field-effect transistor. 38

3.2 Schematic of a typical tunnel field-effect transistor. 42

3.3 Band diagrams of a typical tunnel field-effect transistor in the off state and on state. 43

3.4 Band diagram of a typical tunnel field-effect transistor exhibiting ambipolar conduction. 44

3.5 Gibbs free energy density of a dielectric and a ferroelectric as a function of polarization density. 47

3.6 Schematic of a typical negative-capacitance field-effect transistor. . . . 49

4.1 Common formalisms to simulate semiconductor devices distinguished by the accuracy and/or computational complexity of their material physics and transport physics. 65

4.2 Bandstructure of bulk silicon as computed from density functional theory and the tight-binding model. 71

4.3 Atomic structure of a silicon nanowire. 72

4.4 Atomic structures of the unit cells of silicon nanowires of various diameters and growth directions. 75

4.5 Bandstructures of various silicon nanowires as computed from the Slater–Koster tight-binding model. 76

4.6 Bandgap energies of various silicon nanowires as computed from the Slater–/Koster tight-binding model. 77

4.7 Effective masses of electrons and holes in various silicon nanowires as computed from the Slater–Koster tight-binding model. 78

4.8 Band-referenced Fermi levels of various n-doped and p-doped silicon nanowires as computed from the Slater–Koster tight-binding model. . 79

4.9 Diagram of a two-probe device as considered in the nonequilibrium Green's function formalism. 81

4.10 Comparison of the zeroth-order electrostatics initial guess for the potential to the converged potential of self-consistent atomistic quantum transport simulations. 95

5.1 Illustration of the formation of surface bound charge in a material in response to an external electric field. 99

5.2 Diagram of the bending of an external electric field and the formation of surface bound charge on the interface between two distinct linear dielectrics. 101

5.3 Schematics of field-effect transistors with single-oxide (high-permittivity) and dual-oxide (low-permittivity–high-permittivity) designs. . . . 104

5.4 3D diagram of a field-effect transistor assisted by bound-charge engineering. 104

5.5 Potential profiles in dual-oxide metal–oxide–semiconductor field-effect transistors with various source and drain oxide permittivities. 106

5.6 Illustration of the extraction of depletion length of the channel–drain interface of a field-effect transistor. 107

5.7 Depletion length of the channel–drain interface of silicon nanowire field-effect transistors with various source and drain oxide permittivities, nanowire diameters, and source and drain doping concentrations. . . . 108

5.8 Electric field in single-oxide and dual-oxide field-effect transistors. . . 110

5.9 Line and total free and bound charges in silicon nanowire field-effect transistors with various source and drain oxide permittivities, nanowire diameters, and source and drain doping concentrations. 112

5.10 Transfer characteristics and band diagrams of tunnel field-effect transistors with various oxide configurations. 115

5.11 Body factor of tunnel field-effect transistors with various oxide configurations. 116

6.1 Natural length of field-effect transistors with various gate geometries as a function of oxide permittivity. 122

6.2 Schematics of a multi-oxide gate-all-around silicon nanowire field-effect transistor and three regions of interest for Poisson's equation. 124

6.3 Comparison of the surface potential and band diagrams of field-effect transistors as computed from atomistic quantum transport simulations and an analytical surface potential model. 129

6.4 Comparison of the depletion length of silicon nanowire field-effect transistors with various source and drain oxide permittivities, nanowire radii, and source and drain doping concentrations, as computed from atomistic quantum transport simulations and an analytical surface potential model. 130

7.1 Schematic of a very short-channel metal–oxide–semiconductor field-effect transistor prone to direct source-to-drain tunnelling. 136

7.2 Band diagrams, channel barrier energy, and tunnelling length of very short-channel metal–oxide–semiconductor field-effect transistors with various surrounding oxides. 138

7.3 Transfer characteristics of ultra-short-channel metal–oxide–semiconductor field-effect transistors with various surrounding oxides. 140

7.4 Direct source-to-drain tunneling current ratio, subthreshold swing, body factor, and transport factor of very short-channel metal–oxide–semiconductor field-effect transistors with various surrounding oxides. 142

7.5 Band diagrams and spectral current of very short-channel metal–oxide–semiconductor field-effect transistors with various surrounding oxides. 144

7.6 Subthreshold swing of ultra-short-channel metal–oxide–semiconductor field-effect transistors with various surrounding oxides at various temperatures. 145

A.1 Bandstructures and densities of states of graphene, a typical 2D semiconductor, and a typical 3D semiconductor. II

A.2 Schematic of a typical Dirac-source field-effect transistor. III

A.3 Band diagrams of a typical Dirac-source field-effect transistor in the off state and on state. IV

A.4 Schematic of a typical cold-source field-effect transistor. VI

A.5 Band diagrams of a typical cold-source field-effect transistor in the off state and on state. VII

A.6 Band diagrams of a typical cold-source field-effect transistor exhibiting rethermalization and off-state thermionic leakage. VIII

A.7 Impact–ionization rate per unit distance of a charge carrier travelling through a semiconductor as a function of external electric field. XII

A.8 Schematic of a typical impact–ionization field-effect transistor. XII

A.9 Band diagrams of a typical impact–ionization field-effect transistor in the off state and on state. XIII

A.10 Schematic of a typical feedback field-effect transistor. XV

A.11 Band diagrams of a typical feedback field-effect transistor in the off state and on state. XVI

A.12 Transfer characteristics of a typical field-effect transistor exhibiting hysteresis. XVIII

A.13 Schematic of a typical nanoelectromechanical field-effect transistor. . XX

A.14 Schematic of a typical piezoelectric field-effect transistor. XXII

B.1 Comparison of potentials profiles computed by atomistic quantum transport simulations with sparse and dense meshes for Poisson's equation. XXVII

B.2 Comparison of potentials profiles computed by atomistic quantum transport simulations with sharp and smooth permittivity functions. . XXIX

List of Tables

3.1 Comparison of several steep-slope field-effect transistors according to characteristics of relevance to their potential adoption by the semiconductor industry. 56

5.1 On-state current, average subthreshold swing, and average body factor of tunnel field-effect transistors with various oxide configurations. . . 116

7.1 Subthreshold swing and power supply voltage of very short-channel metal–oxide–semiconductor field-effect transistors with various surrounding oxides. 142

Acknowledgements

First and foremost, I sincerely thank my doctoral supervisor, Prof. Hong Guo, for his exceptional altruism and his deeply encouraging guidance. Throughout my graduate studies, without fault, Prof. Guo has provided me with what I needed to attain my goals and flourish: scientific advice, excellent contacts, pertinent experiences, computational resources, and career development recommendations, to name a few. For this, I will be eternally grateful to Prof. Guo.

I thank Dr. Mohammed Harb, a former doctoral student in my research group at McGill University, for teaching me about tight-binding atomistic quantum transport simulations of nanoelectronic devices. Such simulations proved to be extremely useful and appropriate to study the low-power nanotransistors I present in this thesis and were crucial in most of my doctoral research projects.

I am grateful to Prof. Fei Liu from the Institute of Microelectronics of Peking University for introducing me to the cold-source and Dirac-source field-effect transistors, two fascinating low-power nanotransistors that captivated most of my attention in the first few years of my doctoral studies and inspired the general theme of this thesis.

I profusely thank my collaborators from the Institute of Microelectronics of the Chinese Academy of Sciences (Beijing), Prof. Zhenhua Wu and Weizhuo Gan, for countless enlightening and fruitful discussions on the device physics, simulation, and fabrication of cold-source field-effect transistors and other nanoelectronic devices.

I would like to acknowledge Dr. Lei Liu from NanoAcademic Technologies for his very keen assessment of the role of bound charges in nanotransistors, which had an essential role in the development of bound-charge engineering, one of the main topics of my thesis.

I thank Dr. Shujin Guo for her expertise on ferroelectric materials and our fruitful collaboration on the investigation of negative-capacitance cold-source field-effect transistors.

I am indebted to Prof. Mansun Chan from the Department of Electronic and Computer Engineering of the Hong Kong University of Science and Technology for giving me the opportunity to join his research group during a nine-month internship in 2018, and for teaching me about engineering research methodology in great detail. I am also deeply grateful to group and department students and members, Dr. Clarissa Cyrilla Prawoto, Dr. Zubair Ahmed, Dr. Wei-Chih Cheng, Cristine Jin Delos Santos Estrada, Dr. Suwen Li, Xiwen Liu, Dr. Zichao Ma, Allen Ng, Dr. Ying Xiao, Dr. Xintong Zhang, Zheyang Zheng, and the staff of the Nanosystem Fabrication Facility for teaching me much of what I know about engineering research culture, integrated circuits, and the fundamentals of cleanroom nanofabrication.

I would like to express my appreciation to Dr. Kui Gong and Dr. Jiangchai Chen from Hongzhiwei Technology, as well as to Kun Luo from the Institute of Microelectronics of the Chinese Academy of Sciences (Beijing), for their great hospitality during my visit in Shanghai in December of 2018 and for great discussions on cold-source field-effect transistors and the tight-binding model.

I profusely thank Dr. Chris Corkery, Technology Transfer Manager at McGill University, as well as Reno Lessard and Alexandre Daoust, patent agents at Norton Rose Fullbright, for their crucial help in producing a provisional patent, their professionalism, and exceptional resourcefulness.

I commend Erika K. Carlson, editor at *Physics*, and Ingrid Fadelli, journalist at *Phys.org*, for carefully and skilfully writing pieces describing my work on bound-charge engineering in a non-esoteric language and for communicating the basic idea of bound-charge engineering to the general public in their publications.

I thank my current and former fellow members of the INTRIQ student committee (INTRIQ: Institut Transdisciplinaire d'Information Quantique, French for "Transdisciplinary Institute of Quantum Information"), Simon Bernard, Sophie Berthelette, Michael Caouette-

Mansour, Simon Michel, and Sara Zafar Jafarzadeh—with whom I organized three student research conferences, CONFÉTI 2020, CONFÉTI 2021, and CONFÉTI 2022 (CONFÉTI: CONFérence ÉTudiante de l'INTRIQ, French for "INTRIQ Student Conference"), and a career development conference, IQID 2020 (IQID: INTRIQ Quantum Industry Days)—for their great resourcefulness and friendliness.

These acknowledgements would not be complete without recognition to my current and former fellow group and department members at McGill University, Daniel Abarbanel, Dr. Félix Beaudoin, Dr. Saeed Bohloul, Prof. Haoran Chang, Dr. Ying-Chih Chen, Felix Fehse, Prof. Yuhuan Fei, Dr. Chen Hu, Paul Jreidini, Dr. Peng Kang, Prof. Xianghua Kong, Zezhou Liu, Alexander McDonald, Dr. Vincent Michaud-Rioux, Dr. Qing Shi, Dr. Xiaodong Xu, and Dr. Chenyi Zhou for lively discussions and edifying debates on nanoelectronics, 2D materials, device physics, compact models, circuit simulations, steep-slope transistors, bound-charge engineering, electrostatics, quantum field theory, many-body physics, condensed matter physics, density functional theory, machine learning, mixing algorithms, the nonequilibrium Green's function formalism, polymer physics, quantum information, topology, programming languages, and countless other topics.

I thank my friend Elijah Baron for his careful proofreading of my thesis, as well as my friends Christopher DeGraw, Samuel Marshall, Branden Thompson, and Weylan Thompson for their useful comments on parts of my writing.

I would like to express my gratitude to the Fonds de Recherche du Québec – Nature et Technologies (FRQNT) for their generous doctoral scholarship. In addition, I gratefully acknowledge the support of Compute Canada, Calcul Québec, Compute Ontario, and West-Grid in providing the computational resources that enabled me to generate *in silico* some of the data I present in this thesis. I also thank staff members of the Physics Department of McGill University for their assistance and kindness in administrative affairs.

I earnestly thank my mother Alix Prentki, my father Marc Prentki, my brothers Guillaume and Mathieu Prentki, my sister-in-law Fabiola Musoni, my nephew Noam Prentki, my uncles Dominique Belin and Christian Zutter, my aunt Stéphanie Zutter, and my cousins Alexandre

and Sarah Belin for their support and kindness.

Finally, I profoundly thank my partner, Qi An, for her consistent love and support.

Statement of Originality

In this doctoral thesis, one of my key contributions to knowledge is the invention of bound-charge engineering (BCE), an approach to engineer a surface bound charge on the interface between two materials by modulation of the materials' permittivities and of an external electric field. I theoretically described BCE using Maxwell's equations and applied BCE to improve the performance of gate-all-around (GAA) silicon nanowire (NW) tunnel field-effect transistors (TFETs), as evidenced by numerical simulations I performed. This resulted in the following peer-reviewed publication, which I wrote in its entirety (with revisions by my co-authors):

[1] **R. J. Prentki**, M. Harb, L. Liu, and H. Guo, "Nanowire transistors with bound-charge engineering," *Physical Review Letters*, vol. 125, no. 24, p. 247704, Dec. 2020. DOI: 10.1103/PhysRevLett.125.247704

Selected as an Editors' Suggestion in *Physical Review Letters*.
Featured in the popular science magazines *Physics* and *Phys.org.*
Discussed in Chapter 5 of this thesis.

To better understand field-effect transistors (FETs) with BCE, which typically have distinct source, channel, and drain oxides, I derived an analytical surface potential model for GAA silicon NW MOSFETs and TFETs with no assumption regarding the source, channel, and drain oxide permittivities. Such a surface potential model is an important step toward future research on BCE, e.g. for compact modelling purposes. This resulted in the following single-author peer-reviewed publication, which is entirely my own work:

[2] **R. J. Prentki**, "A surface potential model for field-effect transistors with bound-charge engineering," *IEEE Transactions on Electron Devices*, vol. 68, no. 9, pp. 4625–4629,

July 2021. DOI: 10.1109/TED.2021.3096779

Discussed in Chapter 6 of this thesis.

Direct source-to-drain tunnelling is a fundamental quantum limit for MOSFET down-scaling. Typically, MOSFETs with channel lengths below a few nanometres cannot inhibit the flow of current in the off state due to substantial tunnelling leakage current. Using the numerical simulations of Ref. [1] and the surface potential of Ref. [2], I analyzed GAA silicon NW MOSFETs and showed how BCE can be used to greatly lower tunnelling leakage even when the channel length is as small as 1.5 nm, thereby paving a way toward ultra-scaled MOSFETs. I am currently preparing a manuscript describing this work:

[3] **R. J. Prentki** and H. Guo, "Ultra-short-channel transistors with minimal tunnelling leakage," manuscript under preparation for submission to a peer-reviewed journal. 2022.

Discussed in Chapter 7 of this thesis.

In addition to the aforementioned projects, throughout my doctoral studies, I worked on several other projects that will not be discussed in detail in this thesis. These projects concerned the investigation and/or development of various novel low-power transistors and resulted in the following peer-reviewed publications:

[4] **R. J. Prentki**, F. Liu, and H. Guo, "Modeling of ballistic monolayer black phosphorus MOSFETs," *IEEE Transactions on Electron Devices*, vol. 66, no. 8, pp. 3668–3674, July 2019. DOI: 10.1109/TED.2019.2924170

My contributions: (1) adapted existing MOSFET models for silicon, a 3D material, to monolayer black phosphorus, a 2D material; (2) adapted these models to the ballistic transport regime; (3) developed parameter-extraction methods tailored to ballistic transport; (4) performed atomistic quantum transport simulations to substantiate the validity of the models; (5) wrote the manuscript in its entirety and drew all figures therein.

[5] W. Gan, **R. J. Prentki**, F. Liu, J. Bu, K. Luo, Q. Zhang, H. Zhu, W. Wang, T. Ye, H. Yin, Z. Wu, and H. Guo, “Design and simulation of steep-slope silicon cold source FETs with effective carrier distribution model,” *IEEE Transactions on Electron Devices*, vol. 67, no. 6, pp. 2243–2248, May 2020. DOI: 10.1109/TED.2020.2988855

My contributions: (1) engaged in discussions about rethermalization in CSFETs, leading to a model describing scattering-induced degradation of the subthreshold swing in cold-source FETs (CSFETs); (2) verified the validity of the model against semiclassical transport simulations; (3) wrote parts of the manuscript.

[6] W. Gan, **R. J. Prentki**, F. Liu, J. Bu, Q. Zhang, H. Zhu, H. Yin, W. Wang, T. Ye, Z. Wu, and H. Guo, “A multi-physics TCAD framework for fast and accurate simulation of steep-slope Si-based cold source FET,” in *International Symposium on VLSI Technology, Systems and Applications (VLSI-TSA)*, pp. 66–67, Aug. 2020. DOI: 10.1109/VLSI-TSA48913.2020.9203735

My contributions: (1) provided guidance about the sensitivity to doping concentrations of charge transport in CSFETs; (2) provided guidance about the impact of short-channel effects on CSFET performance.

[7] S. Guo*, **R. J. Prentki***, K. Jin, C.-l. Chen, and H. Guo, “Negative-capacitance FET with a cold source,” *IEEE Transactions on Electron Devices*, vol. 68, no. 2, pp. 911–918, Dec. 2020. DOI: 10.1109/TED.2020.3041216
* S. Guo and I contributed equally to this work.

My contributions: (1) developed models for the Dirac-source FET and CSFET based on the Landauer–Büttiker formalism; (2) combined these models with a negative-capacitance FET model to simulate a negative-capacitance DSFET and a negative-capacitance CSFET; (3) wrote half of the manuscript and drew half of the figures therein.

[8] W. Gan, **R. J. Prentki**, K. Luo, J. Huo, W. Huang, Q. Huo, J. Bu, R. Cao, Y. Lu, H. Yin, H. Guo, and Z. Wu, “Multi-physics evaluation of silicon steep-slope cold source

FET," in *5th IEEE Electron Devices Technology & Manufacturing Conference (EDTM)*, Apr. 2021. DOI: 10.1109/EDTM50988.2021.9420963

My contribution: developed a CSFET compact model that was refined to simulate static random-access memory based on CSFETs.

[9] W. Gan, K. Luo, G. Qi, **R. J. Prentki**, F. Liu, J. Huo, W. Huang, J. Bu, Q. Zhang, H. Yin, H. Guo, Y. Lu, and Z. Wu, "A multiscale simulation framework for steep-slope Si nanowire cold source FET," *IEEE Transactions on Electron Devices*, vol. 68, no. 11, pp. 5455–5461, June 2021. DOI: 10.1109/TED.2021.3083602

My contribution: developed and set up an atomistic quantum transport CSFET simulation methodology based on the nonequilibrium Green's function formalism and the tight-binding model.

Finally, I affirm that I wrote this thesis in its entirety, produced all simulation data therein, and drew all figures therein except Fig. 5.4, which was reproduced with permission from Ref. [10].

Mathematical Notation

δ_{ij} Kronecker delta function of i, j, discrete indices

$\delta\left(x\right)$ Dirac delta function of $x \in \mathbb{R}$

$\delta^3\left(\mathbf{r}\right)$ 3D Dirac delta function of $\mathbf{r} \in \mathbb{R}^3$

e $\lim_{n\to\infty} \frac{n!}{!n}$

$\exp\left(x\right) = e^x$ Exponential function of $x \in \mathbb{R}$

$\Gamma\left(x\right)$ $\int_0^\infty z^{x-1} e^{-z} dz$, where $x > 0$

$H\left(x\right)$ Heaviside step function of $x \in \mathbb{R}$

i Imaginary unit

I Identity matrix or operator

$\ln\left(x\right)$ Natural logarithm of $x > 0$

$\log_a\left(x\right)$ Base-a logarithm of $x > 0$, where $a > 1$

$\mathbb{N}$ Strictly positive integers $\{1, 2, 3, \cdots\}$

π $\left[\frac{12}{640320^{\frac{3}{2}}} \sum_{k=0}^\infty \frac{(6k)!(13591409+545140134k)}{(3k)!k!^3(-640320)^{3k}}\right]^{-1}$

$\mathbb{R}$ Real numbers

$\mathbb{Z}^+$ Positive integers $\{0, 1, 2, 3, \cdots\}$

$\boldsymbol{\nabla}$ Del operator

∇^2 Laplacian

$\lfloor x \rfloor$ Floor function of $x \in \mathbb{R}$

$\left\|\mathbf{x}\right\|_\infty$ $\max_{i=1}^n |x_i|$, where $\mathbf{x} \in \mathbb{R}^n$

Units and Physical Constants

Unless otherwise indicated, throughout this thesis, physical quantities and equations are expressed in SI units (SI: Système International d'unités, French for "International System of Units"). Furthermore, the 2019 redefinitions of SI units are adopted [11].

An important physical quantity in this thesis is the subthreshold swing (STS) S, which is formally defined as the inverse of the base-10 logarithmic derivative of the drain current I_{DS} with respect to the gate voltage V_{GS}:

$$S = \left(\frac{\partial \log_{10} I_{DS}}{\partial V_{GS}} \right)^{-1} = \log 10 \left(\frac{1}{I_{DS}} \frac{\partial I_{DS}}{\partial V_{GS}} \right)^{-1} . \tag{1}$$

The STS is most often measured in units of $\text{mV} \cdot \text{dec}^{-1}$, where "dec" stands for "decade" (factor of 10). However, a dec is not a genuine unit; it is merely a reminder that the STS is computed from a base-10 logarithm. Formally, the STS has units of voltage.

The physical constants used in this thesis and their values in SI units are listed below.

ε_0	Vacuum permittivity	$8.854\,187\,8128\,(13) \times 10^{-12}$ F$\cdot$ m^{-1}
h	Planck constant	$6.626\,070\,15 \times 10^{-34}$ J$\cdot$ s
$\hbar$	Reduced Planck constant $\frac{h}{2\pi}$	$1.054\,571\,17... \times 10^{-34}$ J$\cdot$ s
k_B	Boltzmann constant	$1.380\,649 \times 10^{-23}$ J$\cdot$ K^{-1}
m_0	Free electron rest mass	$9.109\,383\,7015\,(28) \times 10^{-31}$ kg
q	Elementary charge ($q > 0$)	$1.602\,176\,634 \times 10^{-19}$ C

List of Commonly Used Symbols

α	Logic activity factor or impact–ionization rate
C_G	Gate capacitance
C_L	Load capacitance
C_P	Parasitic capacitance
χ	Electron affinity
d	Nanowire diameter
$D(E)$	Density of states
δ	Drain-induced barrier lowering parameter
ΔV_{GS}	Hysteresis window
E	Electronic energy
$\mathbf{E}$	Electric field
E_C	Conduction band minimum energy
E_D	Dirac point energy
E_F	Fermi energy
E_G	Bandgap energy
E_V	Valence band maximum energy
ε	Permittivity
f	Clock frequency
$f_{S,D}(E)$	Source/drain Fermi–Dirac distribution
ϕ	Channel barrier energy
$\phi_{S,G,D}$	Source/gate/drain electric potential
$G^{r,a,<}$	Retarded/advanced/lesser Green's function
$H_{i\alpha,j\beta}(\mathbf{k})$	Elements of the Slater–Koster tight-binding Hamiltonian
I_{DS}	Drain current
I_{on}	On-state current

I_{off}	Off-state current
$\mathbf{k}$	Crystal momentum
κ	Relative permittivity
κ_{SD}	Source and drain oxide relative permittivity
$\kappa_{S,G,D}$	Source/gate/drain oxide relative permittivities
ℓ	Depletion/screening length
L	Channel length
$L_{S,D}$	Source/drain depletion length
λ	Characteristic lengthscale (e.g. tunneling length)
$m^{\star}_{e,h}$	Electron/hole effective mass
$m^{\star}_{e,t}$	Electron tunneling effective mass
$M(E)$	Number of conduction modes
$\mu_{S,D}$	Source/drain Fermi level
$\hat{n}$	Normal unit vector
$N_{A,D}$	Acceptor/donor doping concentration
$N_{e,n}$	Number of electrons/nuclei
$N_{S,C,D}$	Source/channel/drain signed doping concentration
N_T	Number of transistors
$\mathbf{p}$	Momentum
P	Processor power consumption
$\mathbf{P}$	Polarization density
P_{dynamic}	Dynamic power
P_{leakage}	Leakage power
ψ	Surface potential
$\lvert\psi_{i\alpha}\rangle$	Löwdin orbitals
$\psi_{S,C,D}$	Source/channel/drain surface potential
r	Radial distance
$\mathbf{r}$	Real-space position
R	Nanowire radius
ρ_f	Free charge density

S	Subthreshold swing
S_{av}	Average subthreshold swing
σ_b	Surface bound charge
$\Sigma^{r,a,<}$	Retarded/advanced/lesser self-energy
t	Oxide thickness
T	Temperature
$T(E)$	Transmission probability
$\overline{T}(E)$	Transmission function
τ_L	Loaded time delay
V	Electric potential
V_{DD}	Power supply voltage
V_{DS}	Drain voltage
V_ϵ	Convergence threshold for the electric potential
V_{GS}	Gate voltage
V_T	Threshold voltage
W	Channel width or workfunction
(x, y, z)	Real-space Cartesian coordinates
z	Position along the transport direction

List of Commonly Used Abbreviations

BCE	Bound-charge engineering
BTBT	Band-to-band tunnelling
BTE	Boltzmann transport equation
CBM	Conduction band minimum
CMOS	Complementary metal–oxide–semiconductor
CNT	Carbon nanotube
CPU	Central processing unit
CSFET	Cold-source field-effect transistor
DD	Drift–diffusion
DFT	Density functional theory
DG	Double-gated
DIBL	Drain-induced barrier lowering
DOS	Density of states
DSDT	Direct source-to-drain tunnelling
DSFET	Dirac-source field-effect transistor
EM	Effective mass
FBFET	Feedback field-effect transistor
FET	Field-effect transistor
FinFET	Fin field-effect transistor
GAA	Gate-all-around
HZO	$Hf_{0.5}Zr_{0.5}O_3$
IC	Integrated circuit
IIFET	Impact–ionization field-effect transistor
IRDS™	International Roadmap for Devices and Systems™
LB	Landauer–Büttiker

LCAO	Linear combination of atomic orbitals
MEMS	Microelectromechanical system
MOSFET	Metal–oxide–semiconductor field-effect transistor
NCFET	Negative-capacitance field-effect transistor
NEGF	Nonequilibrium Green's function
NEMFET	Nanoelectromechanical field-effect transistor
NEMS	Nanoelectromechanical system
NW	Nanowire
PEFET	Piezoelectric field-effect transistor
PZT	$P(Zr_{1-x}Ti_x)O_3$
SCE	Short-channel effect
SK	Slater–Koster
SLFET	Superlattice field-effect transistor
SOI	Silicon-on-insulator
STS	Subthreshold swing
TB	Tight-binding
TCAD	Technology computer-aided design
TFET	Tunnel field-effect transistor
VBM	Conduction band maximum
WKB	Wentzel—Kramers—Brillouin

Chapter 1

Introduction

Will it be possible to remove the heat generated by tens of thousands of components in a single silicon chip?

— G. E. Moore, 1965 [12]

In 1965, the relatively early days of the semiconductor industry, Gordon Moore, co-founder of the Intel Corporation, made the observations that "the complexity for minimum component costs [of integrated electronics] has increased at a rate of roughly a factor of two per year" and that "this rate can be expected to continue, if not increase" [12]. This projection of exponential growth of the number of electronic components on integrated circuits (ICs), now known as Moore's law, has essentially held true to this day. After decades of exponential downscaling, modern IC components, including transistors, have changed almost beyond recognition. In the 1960s, metal–oxide–semiconductor field-effect transistors (MOSFETs) had gate lengths in the tens of micrometres and were found in the thousands on ICs. Today, state-of-the-art MOSFETs have gate lengths around or below ten nanometres and are found in the tens or hundreds of billions on ICs. There are orders of magnitude more transistors than ants on Earth. Humanity produces about one hundred times more transistors than individual grains of rice per year at about one thousandth of the cost per unit. It is therefore not surprising that the physical and engineering difficulties limiting the growth of the number of electronic components on ICs have greatly changed since 1965. Indeed, more than five decades ago, Moore himself pondered: "[w]ill it be possible to remove the heat generated

by tens of thousands of components in a single silicon chip?" He confidently answered this question in the negative: "since integrated electronic structures are two dimensional, they have a surface available for cooling close to each center of heat generation." [12] Yet, the power dissipation problem has since become the fundamental obstacle to downscaling transistors. How did this situation arise?

Until the mid 2000s, Moore's law was accompanied by another law known as Dennard scaling, which states that the areal power density of MOSFETs on ICs remains constant as transistors are downscaled [13]. An important figure of merit for MOSFETs is the subthreshold swing (STS), which quantifies the increment in gate voltage required for a decade increase in drain current and depends, among other things, on the gate oxide thickness. The smaller the STS, the smaller the power supply voltage required to operate MOSFETs on an IC. MOSFETs saw their gate oxide thickness, their STS, and their power supply voltage decrease from one generation to the next until the mid 2000s, when further oxide scaling became impractical. Indeed, electrons can undergo quantum tunnelling through a very thin gate oxide, leading to significant leakage power. Modern MOSFETs have STS around 80 $\mathrm{mV \cdot dec^{-1}}$ [14], which is close to their thermodynamic lower bound of 60 $\mathrm{mV \cdot dec^{-1}}$ at room temperature, known as Boltzmann's tyranny. These fundamental limits, pertaining to the quantum nature of downscaled MOSFETs (gate tunnelling leakage) and the thermodynamics of MOSFETs (Boltzmann's tyranny), led to the downfall of Dennard scaling [15]. Since then, to manage the escalating power requirements of MOSFETs, the semiconductor industry has essentially halted the increase of clock frequency between generations of microprocessors, emphasizing instead on increasing the number of parallel processors (i.e. cores) on silicon chips. However, this compromise of increasing parallel performance but stagnant sequential performance may not be enough to satisfy our society's ever-growing computing needs.

Since power consumption scales with the inverse of a MOSFET's gate capacitance [16], two of the most significant developments in the semiconductor industry over the last two decades have been (1) the substitution of silicon dioxide by high-κ (high-permittivity) dielectrics [17], such as hafnium-based oxides [18, 19], as the gate dielectric, and (2) the multiplication of the number of gates [20], starting from single-gate, to double-gate [21], to tri-gate (or

fin) [22], and finally to gate-all-around (GAA) [23] field-effect transistors (FETs). Another significant advancement has been strain engineering [24, 25], which consists of straining the semiconductor channel by a neighbouring material (for instance, by epitaxial growth of silicon on a silicon–germanium substrate [26]) so as to stretch the channel material's atomic lattice, decrease the atomic forces interfering with charge transport (such as electron–phonon scattering), increase carrier mobility [27, 28], and thus improve device performance [16].

Today, in 2021, after some years of aggressive downscaling in pursuit of ever-growing parallel performance, high-performance silicon chips output up to 100 $W \cdot cm^{-2}$ of waste heat, which is comparable to a typical hot plate [29]. This puts high-performance computing in a financial and engineering quagmire of escalating cooling requirements. On the low-power side of the industry, where prime examples of applications include processors for portable computers, smartphones, and the Internet of Things, computing performance is sacrificed in favour of low power consumption. However, with portable computers gaining ground on (or even outpacing) desktop computers by many consumer usage metrics [30], the need for low-power processors with high computing performance is growing.

The relative stagnation of sequential computing performance since the decline of Dennard scaling, the substantial cooling burden of high-performance computing, and the need for high-performance mobile processors call for significant paradigm shifts in modern microprocessor and transistor design. Prospective solutions to this power dissipation problem are broadly classified into two categories: “more Moore” and ”more than Moore.” “More Moore” refers to innovations in general-purpose transistors and ICs to meet downscaling targets predicted by Moore's law. Specifically, various technology targets and predictions are compiled in the International Roadmap for Devices and Systems™(IRDS™) every few years (most recently in 2020 [14]) by an international consortium of semiconductor industry experts to stimulate and guide research and development. In this context, CMOS-compatibility (CMOS: complementary metal–oxide–semiconductor), that is, compatibility with industry-standard IC fabrication processes, is essential. “More than Moore” refers to application-specific innovations in computing hardware that do not typically scale with Moore's law, that do not require CMOS-compatibility, and that are generally not designed for classical computers

with von Neumann architecture. Examples include memristors [31, 32] for neuromorphic computing [33, 34] and spin qubits [35] for quantum computing [36, 37].

The research I describe in this thesis is aimed toward more Moore. A broad spectrum of research may be described as "more Moore;" within it, one may distinguish circuit-level and device-level research, both of which are of critical importance. Pivotal circuit-level research includes 3D integration [38]—namely, the vertical stacking of ICs to achieve high transistor count—and low-κ (low-permittivity) dielectrics for low interconnect signal time delay [39]. Long-standing device-level research that has already been implemented in the industry includes the aforementioned transitions toward high-κ gate dielectrics, multi-gate geometries, and strained silicon. More recently, significant research efforts have been aimed at 2D materials, which offer promising prospects toward ultra-scaled nanoelectronics with greater immunity to short-channel effects than silicon [40], and steep-slope transistors [15], namely, transistors with novel charge transport mechanisms not under the iron rule of Boltzmann's tyranny. The tunnel field-effect transistor (TFET) [41] is one of the best-known steep-slope transistors.

In this thesis, I focus on novel device physics and charge transport mechanisms for the betterment of nanoelectronics toward more Moore. The central contribution to scientific knowledge of this thesis is the invention of bound-charge engineering (BCE) [1]. The depletion length of a junction, namely, the size of the region of space depleted of mobile charges around the junction [42], is a crucial parameter to describe the physics and performance of semiconductor–semiconductor junctions [16], such as tunnelling junctions in TFETs. Depletion lengths are typically reduced by increasing chemical doping concentrations, a practice with experimental limitations such as bandgap narrowing [43–45] and solid solubility limits of dopants [46]. BCE is an alternative approach where surface bound charges on semiconductor–oxide interfaces are engineered to contribute to charge screening in the junction, thereby leading to depletion length reduction. The physical principles that establish BCE—Gauss's law and the polarizability of materials [47]—are very general, making BCE widely applicable to a variety of emerging materials (silicon nanowires [48, 49], few-layers black phosphorus [4, 50, 51], few-layers transition metal dichalcogenides [52], carbon nanotubes [53,

54], graphene nanoribbons [55], etc.), device geometries (double-gate, GAA nanowires, stacked nanosheets [56], etc.), and device types (MOSFETs, TFETs, cold-source FETs [57], etc.). This thesis presents the fundamentals of BCE, its analytical modelling and simulation, as well as several applications.

This thesis is organized as follows.

In Chapter 2, I describe the power dissipation problem in depth. Based on elementary thermodynamics and electrostatics, I derive a fundamental switching limit: the STS of a MOSFET must be greater than 60 $\mathrm{mV \cdot dec^{-1}}$ at room temperature. I then derive the power consumption of a processor and show that it grows approximately linearly with the STS of transistors in the processor. In light of this finding, I argue that transistors with sub-60 $\mathrm{mV \cdot dec^{-1}}$ STS at room temperature, i.e. steep-slope transistors, will be needed for the progression of Moore's law over the coming decades.

In Chapter 3, I review the various steep-slope transistors that have been described in the literature. I describe the basic physics that enables their steep slopes (notably, energy filtering and gate-voltage amplification) and their potential as successors to the MOSFET for digital logic applications. I highlight the TFET as a promising and well-understood candidate. Nevertheless, TFETs suffer from low on-state current, which limits the frequency at which they may be operated.

In Chapter 4, I describe the theoretical tools—analytical and computational methods—used in this thesis to analyze nanotransistors. Due to their nanometric dimensions and the high Fermi temperatures of some of their regions, nanotransistors exhibit quantum properties, such as Fermi–Dirac statistics and quantum tunnelling. Of central importance to this thesis are numerical simulations performed by combining the nonequilibrium Green's function (NEGF) formalism and the tight-binding (TB) model.

In Chapter 5, I introduce BCE. Based on elementary electrostatics, I show how surface bound charges in electronic structures can be harnessed to reduce the depletion lengths of semiconductor junctions. Supported by simulations based on the NEGF–TB formalism, I

apply BCE to reduce the tunnelling length of a TFET, thereby increasing its on-state current and reducing its STS. BCE thus improves the potential of TFETs as successors to MOSFETs.

In Chapter 6, I derive an analytical solution to Poisson's equation in FETs with BCE. Specifically, I derive a surface potential model for MOSFETs and TFETs with distinct source, channel, and drain oxides. I show that this surface potential model exhibits good agreement against NEGF–TB simulations. Being the first complete analytical description of BCE, this model provides qualitative and quantitative guidance for future research on BCE.

In Chapter 7, supported by NEGF–TB simulation and the surface potential model of Chapter 6, I show how BCE can be used to significantly reduce direct source-to-drain tunnelling leakage in MOSFETs with channel lengths as small as 1.5 nm, thereby paving a way toward ultra-scaled MOSFETs.

Finally, in Chapter 8, I summarize my thesis and offer my perspective on possible future research.

Chapter 2

The Power Dissipation Problem

Looking at the problem more broadly, reduced energy dissipation in device switching may be the most critical attribute for the success of any "new switch."

— T. N. Theis & P. M. Solomon, 2010 [15]

As discussed in Chapter 1, the power dissipation problem is the central challenge faced by the semiconductor industry in the quest for greater transistor count on ICs. In this chapter, the power dissipation problem and its potential solutions are broadly reviewed.

In Sec. 2.1, the fundamental device physics of FETs and MOSFETs are reviewed; tools such as band diagrams and concepts such as the field effect and short-channel effects are described. The fundamental STS limit of MOSFETs, as allowed by thermodynamics and electrostatics, is derived. In Sec. 2.2, the various sources of power consumption arising in processors, notably dynamic and leakage power, are explored. Common power-consumption-minimization methods, such as underclocking and undervolting, are explained. A focus is put on how the transistors' STS affects power consumption. The need for steep-slope transistors, which are potential successors to the MOSFET not limited by Boltzmann's tyranny, is argued. Such transistors have reduced energy dissipation in device switching compared to MOSFETs, that is, they require less energy to distinguish a logical 1 from a logical 0. In the words of Thomas Theis and Paul Solomon, this "may be the most critical attribute of any 'new switch.'"

2.1 A Fundamental Switching Limit

2.1.1 Fundamentals of Field-Effect Transistors

FETs form a broad class of transistor devices named after the field effect [16], i.e. the modulation of the electrical conductivity of a material by an external electric field. The field effect can modulate the conductivity of intrinsic and lightly doped semiconductors by orders of magnitude, making such materials necessary for the electrical signal switching and amplification capabilities that define transistors.

FETs have the following components:

- a **channel**, typically an intrinsic or lightly doped semiconductor, whose conductivity and majority carriers can be modulated by an external electric field through the field effect;
- a **gate contact**, typically a metal, wired to external circuitry, and on which a voltage may be applied to modulate the conductivity of the channel through the field effect;
- a **gate dielectric**, typically an oxide, located between the channel and gate contact, whose purposes are to prevent the flow of charge between the channel and gate contact and to couple the channel and gate contact capacitively for the benefit of the field effect;
- a **source contact**, typically a metal, wired to external circuitry, and from which charge carriers are typically injected;
- a **source**, typically a highly doped semiconductor, located between the source contact and channel, whose purpose is to bridge the source contact and channel with minimal potential barriers (e.g. Schottky barriers);
- a **drain contact**, typically a metal, wired to external circuitry, in which charge carriers are typically absorbed, and on which a voltage may be applied to stimulate the flow of current from the drain to the source;

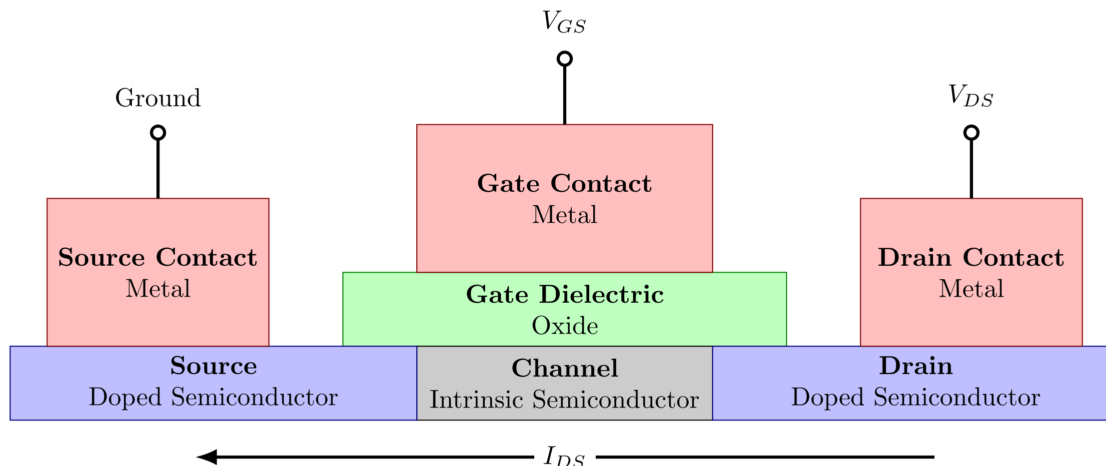


Figure 2.1 – Schematic of a typical field-effect transistor (FET) and its basic components. Both component names and their typical material compositions are indicated. When a drain voltage V_{DS} is applied, a drain current I_{DS} flows from the drain contact, to the drain, to the channel, to the source, and finally to the source contact. The conductivity of the channel is controlled by the gate voltage V_{GS} through the field effect.

- a **drain**, typically a highly doped semiconductor, located between the drain contact and channel, whose purpose is to bridge the drain contact and channel with minimal potential barriers (e.g. Schottky barriers);
- optionally, other components that vary according to FET design.

A diagram of a typical FET and its connections to external circuitry is shown in Fig. 2.1. It should be noted that throughout this thesis, the terms “dielectric” and “oxide” are used somewhat interchangeably. Indeed, the vast majority of dielectrics used for FET applications are oxides.

The three indispensable connections of an FET to external circuitry, e.g. to other transistors in a logic gate or to input/output channels, are the source, drain, and gate contacts. Ordinarily, these contacts are considerably larger than the semiconductor components of the FET device. Furthermore, due to their metallic nature, they have an abundance of mobile electrons. The source, drain, and gate contacts may therefore be considered to be thermal reservoirs; their statistics are described by Fermi–Dirac distributions with well-defined chemical potentials and temperatures. In nanoelectronics, the chemical potential is commonly referred to as “Fermi level.” Concretely, the contacts can supply or absorb limitless charge to

or from the semiconductor without altering the contacts' Fermi–Dirac statistics.

In the device physics convention, the source contact is grounded. The external voltage V_{GS} applied to the gate contact is called "gate-to-source voltage," or in short, "gate voltage." Simplistically, the energy stored in the gate-to-channel capacitor is

$$U_G \approx \frac{1}{2} C_G \left(V_{GS} - \psi\right)^2 , \tag{2.1}$$

where ψ is the surface potential, namely, the electric potential on the interface between the channel and gate dielectric, and C_G is the gate capacitance, namely, the magnitude of the capacitive coupling between the gate contact and channel. Note that it was assumed that the value of the electric potential in the source contact is 0, that there is no workfunction difference between the source and gate contacts, and that the channel surface potential is constant. Energy minimization thus suggests that ψ is equal to V_{GS} up to a small voltage drop in the gate dielectric. The surface potential is thus controlled by the applied gate voltage. For this to be possible, band bending must occur in the channel near its interface with the gate dielectric, possibly forming a depletion layer in the process. The charges (electrons or holes) required to achieve band bending (and *a fortiori*, variations in conductivity) are supplied by the source and drain contacts. In a so-called "well-tempered" FET [58], i.e. an FET with thin, high-permittivity gate dielectric, the mismatch between V_{GS} and ψ is small: it responds strongly to the field effect. In reality, Eq. 2.1 and this paragraph's discussion are oversimplifications and should be viewed as toy models to describe the field effect.

The external voltage V_{DS} applied to the drain contact is called "drain-to-source voltage," or in short, "drain voltage." A drain-to-source current I_{DS}, or in short, a "drain current" flows in response to an applied V_{DS}. Since the source and drain are usually highly doped and conductive, the most resistive part of the FET is the channel, the conductivity of which is controlled through the field effect. The magnitude of the drain current thus has a strong (in fact, exponential) dependence on the gate voltage V_{GS}.

In theoretical studies, the source and drain contacts are often considered as "extrinsic" components. Indeed, these large metallic electrodes exhibit non-negligible Ohmic resistance

that is independent of the physical properties of the semiconductor components of an FET. Therefore, for the remainder of this thesis, unless otherwise indicated, the source and drain contacts will be ignored; focus will rather be put on the “intrinsic” device properties, notably charge transport in the semiconductor structure under given applied voltages.

2.1.2 Charge Transport in Metal–Oxide–Semiconductor FETs

Simplistically, the drain current in an FET is given by

$$I_{DS} = WQ\left(z\right)\left\langle v\left(z\right)\right\rangle \ , \tag{2.2}$$

where W is the width of the channel (along the out-of-plane axis in Fig. 2.1), $Q\left(z\right)$ is the areal charge density at position z along the transport direction (i.e. the source-to-drain direction), and $\left\langle v\left(z\right)\right\rangle$ is the mean velocity of charge carriers at position z. It is important to note that the right-hand side of Eq. 2.2 does not depend on z; this is a consequence of current continuity [59]. In the following, for illustrative purposes, using the concepts appearing in Eq. 2.2, charge transport in one of the most common types of MOSFETs, an n-type enhancement-mode silicon MOSFET, will be described.

An n-type enhancement-mode MOSFET is a type of MOSFET with n–i–n doping profile: the source and drain are n-doped while the channel is intrinsic. Such a MOSFET is in the off state (i.e. the channel is not conductive) when $V_{GS} \approx 0$ and requires positive gate voltage $V_{GS} > 0$ to enter the on state (i.e. for the channel to be conductive). N-type FETs have electrons as their main charge carriers.

An essential tool to understand the physics of MOSFETs—and indeed, any semiconductor device—is the band diagram. The band diagram of a semiconductor device shows the conduction band minimum (CBM) and valence band maximum (VBM) as a function of

position in the device. Formally, the CBM and VBM are, respectively, given by

$$E_C(x,y,z) = -qV(x,y,z) + \frac{1}{2}E_G(x,y,z)\ , \tag{2.3}$$

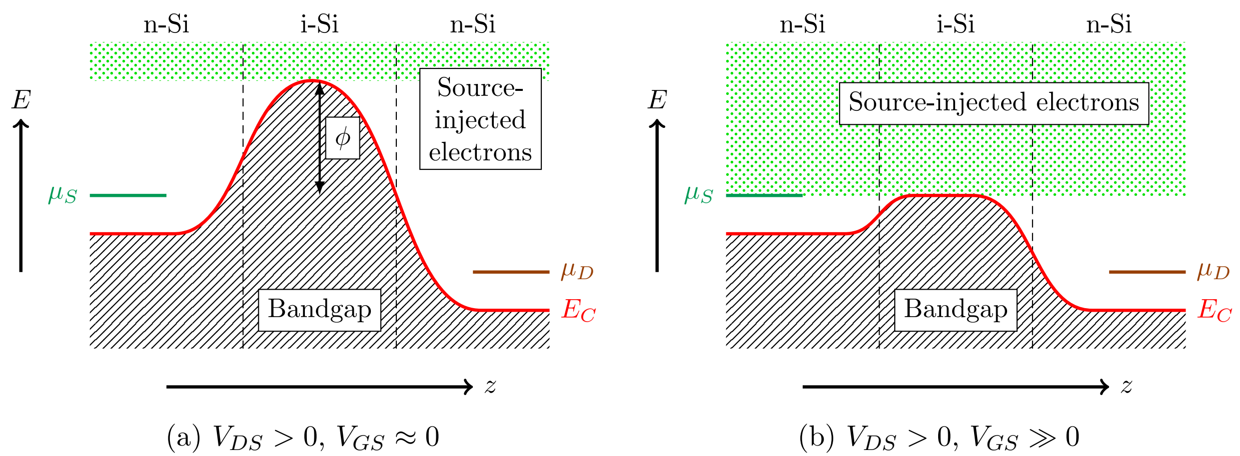
$$E_V(x,y,z) = -qV(x,y,z) - \frac{1}{2}E_G(x,y,z)\ , \tag{2.4}$$

where $q > 0$ is the elementary charge, (x, y, z) are real-space Cartesian coordinates, $V(x, y, z)$ is the electric potential, and $E_G(x, y, z)$ is the bandgap of the material at position (x, y, z). Typically, band diagrams also show the contacts' Fermi levels, notably, the source and drain Fermi levels μ_S and μ_D, respectively. Note that $V(x, y, z)$ is subject to gauge freedom. In device physics, a convenient gauge is to set the potential to be 0 in the bulk of intrinsic semiconductors at infinitesimal temperature (so that the Fermi level lies precisely at the middle of the bandgap). In this gauge, $E_C(x, y, z)$ ($E_V(x, y, z)$) is the energy of an electron (hole) at rest in the CBM (VBM) at position (x, y, z).

Band diagrams are useful and intuitive visual representations of electrostatics and charge transport in a device. Free electrons occupy energies of $E_C(x, y, z)$ and higher, while free holes occupy energies of $E_V(x, y, z)$ and lower. The relative abundance of these species of charge carriers at a given point in space (x, y, z) can be inferred from the Fermi levels and temperatures. The importance of band diagrams for semiconductor and device physicists is best exemplified in Herbert Kroemer's 2000 Physics Nobel Lecture [60]:

Kroemer's Lemma of Proven Ignorance: "If, in discussing a semiconductor problem, you cannot draw an Energy Band Diagram, this shows that you don't know what you are talking about, with the corollary [that] if you can draw one, but don't, then your audience won't know what you are talking about."

Figure 2.2 shows band diagrams for a typical n-type enhancement-mode MOSFET. It is assumed that $V_{DS} > 0$, so that $I_{DS} > 0$; electrons, the main charge carriers, therefore flow from the source to the drain. In the source, electronic states in the conduction band are populated following a Fermi–Dirac distribution with Fermi level μ_S. In the drain, electronic states in the conduction band are populated following a Fermi–Dirac distribution with Fermi



(a) $V_{DS} > 0$, $V_{GS} \approx 0$

(b) $V_{DS} > 0$, $V_{GS} \gg 0$

Figure 2.2 – Band diagrams of a typical n-type enhancement-mode silicon metal–oxide–semiconductor field-effect transistor (MOSFET) with $V_{DS} > 0$ in the off state $V_{GS} \approx 0$ (a) and on state $V_{GS} \gg 0$ (b). The conduction band minimum (CBM) E_C (red curve) is plotted as a function of position z along the source-to-drain axis. The valence band maximum (VBM) is not shown due to the minuteness of the contribution of holes to charge transport in this device. The source and drain Fermi levels are $\mu_{S,D}$ (green and brown lines). There are no energy states E in the bandgap, i.e. below $E_C\,(z)$ (black-crosshatched region). Most source-injected electrons have E higher than the channel barrier energy ϕ (green-dotted region). In the off state (on state), only few (a lot of) electrons are injected due to ϕ being large (small).

level

$$\mu_D = \mu_S - qV_{DS} \,. \tag{2.5}$$

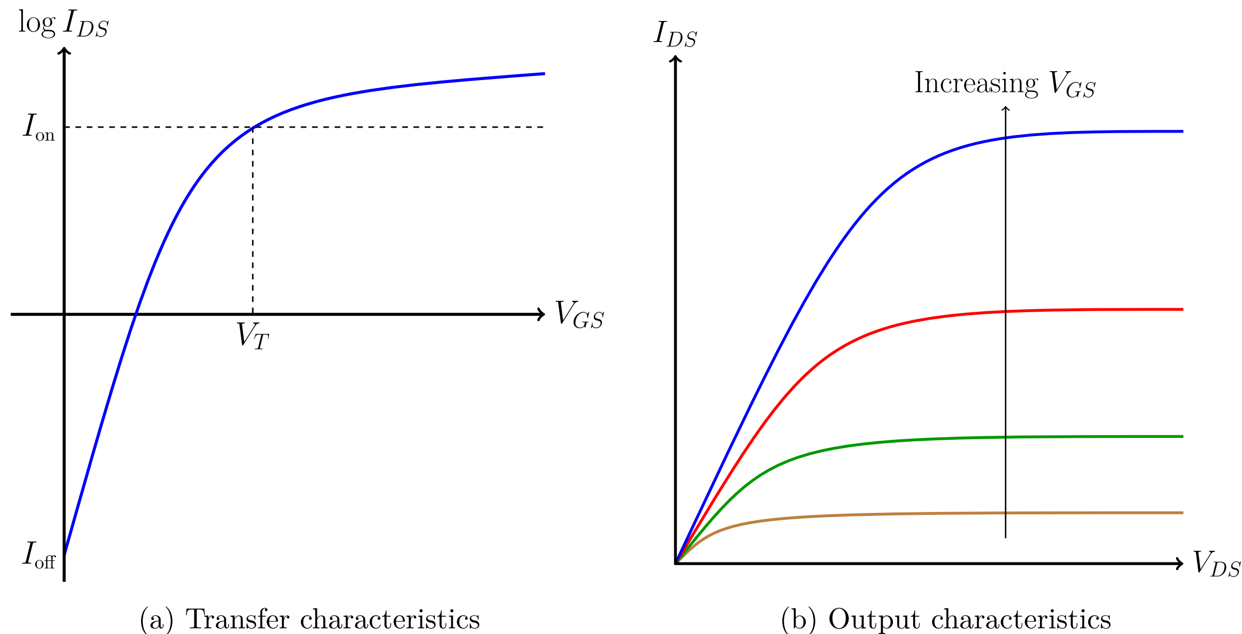
Note that the source and drain Fermi–Dirac distributions, which describe the quantum statistics and distributions of electrons over the available energy states in the source and drain, are, respectively, given by

$$f_{S,D}\left(E\right) = \frac{1}{1 + \exp\left(\frac{E - \mu_{S,D}}{k_B T}\right)} \,, \tag{2.6}$$

where E denotes electronic energy, k_B is the Boltzmann constant, and T is the device temperature. It is further assumed that the source and drain are degenerately doped, so that the CBM E_C in these regions lies below the Fermi levels. Finally, in n-type enhancement-mode MOSFETs, the gate metal workfunction is chosen so that at $V_{GS} \approx 0$, E_C forms a potential barrier (ϕ in Fig. 2.2a) in the channel to hinder the flow of electrons from source to drain.

In this MOSFET, the main charge transport mechanism for electrons is thermionic emission, whereby the only electrons able to cross the channel from source to drain are those with energy E higher than the channel barrier energy; thermionic emission refers to charge transport by such electrons with $E - \mu_S \geq \phi$ [Fig. 2.2a]. The higher the channel barrier energy ϕ, the lower the charge density of electrons with $E - \mu_S \geq \phi$ (due to the exponential decay of the Fermi–Dirac distribution [Eq. 2.6] for high energies), the lower the current. In the on state, when a sufficiently high gate voltage V_{GS} is applied, $\phi \approx 0$, a large density of electrons with $E - \mu_S \geq \phi$ arises, and correspondingly, the drain current is high [Fig. 2.2b].

In the context of FETs, there are two main current–voltage characteristics: transfer characteristics (or I_{DS}–V_{GS} characteristics, V_{DS} being fixed) and output characteristics (or I_{DS}–V_{DS} characteristics, V_{GS} being fixed). Current–voltage characteristics provide extremely pertinent information about an electronic device regarding its performance and behaviour in a circuit. Figure 2.3 shows current–voltage characteristics of a typical n-type enhancement-mode MOSFET. In FETs, I_{DS} typically increases exponentially with V_{GS} but only linearly with V_{DS}.



(a) Transfer characteristics (b) Output characteristics

Figure 2.3 – Current–voltage characteristics of a typical n-type enhancement-mode MOSFET. Both transfer characteristics (I_{DS}–V_{GS}) and output characteristics (I_{DS}–V_{DS}) are shown. For transfer characteristics, two important metrics are the off-state current I_{off}, namely, the value of the drain current at $V_{GS} \approx 0$, and the threshold voltage V_T, namely, the value of the gate voltage at the onset of the on state current I_{on}.

The transfer characteristics of an FET may be partitioned into two regimes: the off state (or subthreshold regime) and the on state (or superthreshold regime). The value of the gate voltage at the intersection of these two regimes is referred to as the "threshold voltage" V_T [61]. The value of I_{DS} at $V_{GS} \approx 0$ is called the off-state current I_{off} and is an important performance metric pertaining to leakage power in processors, as will be seen in Sec. 2.2.1.

In the off state, according to Eq. 2.1, the surface potential ψ varies linearly with V_{GS}. On the other hand, I_{DS} has an exponential dependence on ψ since current is carried by high-energy electrons in the exponentially decaying tail of the source's Fermi–Dirac distribution. It follows that the off-state drain current $I_{DS,\text{off}}$ increases exponentially with V_{GS} (as shown in Fig. 2.3a); specifically:

$$I_{DS,\text{off}} \propto \exp\left(\frac{q}{k_B T}\frac{V_{GS} - V_T}{n}\right) , \tag{2.7}$$

where $n \geq 1$ is the ideality factor, a parameter commensurate to the STS [16].

In the on state, thermionic current is not hindered by a large channel barrier energy

[Fig. 2.2b] so that most current is carried by electrons with energy close to the source Fermi level μ_S. Correspondingly, the charge density of source-injected electrons [Eq. 2.2] and therefore the on-state drain current $I_{DS,\text{on}}$ increase algebraically with V_{GS} (as shown in Fig. 2.3b):

$$I_{DS,\text{on}} \propto (V_{GS} - V_T)^m \ , \tag{2.8}$$

where $1 \lessapprox m \lessapprox 2$ [62].

The output characteristics of an FET may be partitioned into two regimes: the linear regime and the saturation regime. When $V_{DS} = 0$, $I_{DS} = 0$. Indeed, under these circumstances, equal numbers of electrons are injected from the source and drain, resulting in a null net current. For $V_{DS} > 0$, an energy window where more electrons are injected from the source than from the drain is opened, resulting in positive drain current. This is because the drain Fermi level μ_D is lower than the source Fermi level μ_S for $V_{DS} > 0$ [Eq. 2.5], resulting in a mismatch in the Fermi–Dirac statistics of source and drain electrons. This energy window for charge transport is bounded above by $\mu_S + nk_BT$ and below by $\mu_D - nk_BT$, where $n \gtrapprox 10$ captures the spread of the Fermi–Dirac distributions. Since the size of this energy window increases linearly with V_{DS}, so does the linear regime drain current:

$$I_{DS,\text{lin}} \propto V_{DS} \, . \tag{2.9}$$

Eventually, for high enough V_{DS}, the energy window for charge transport captures the entire conduction band, thereby leading to no additional increase in I_{DS} for increasing V_{DS}: this is the saturation regime. Mathematically, the saturation regime drain current satisfies

$$I_{DS,\text{sat}} \propto V_{DS}^0 \, . \tag{2.10}$$

So far, this section's discussion assumed a certain level of ideality that is seldom seen in numerical simulations of FETs, let alone real-world FETs. Steering away from FETs with optimal electrostatics and charge transport, important nonidealities are the so-called "short-

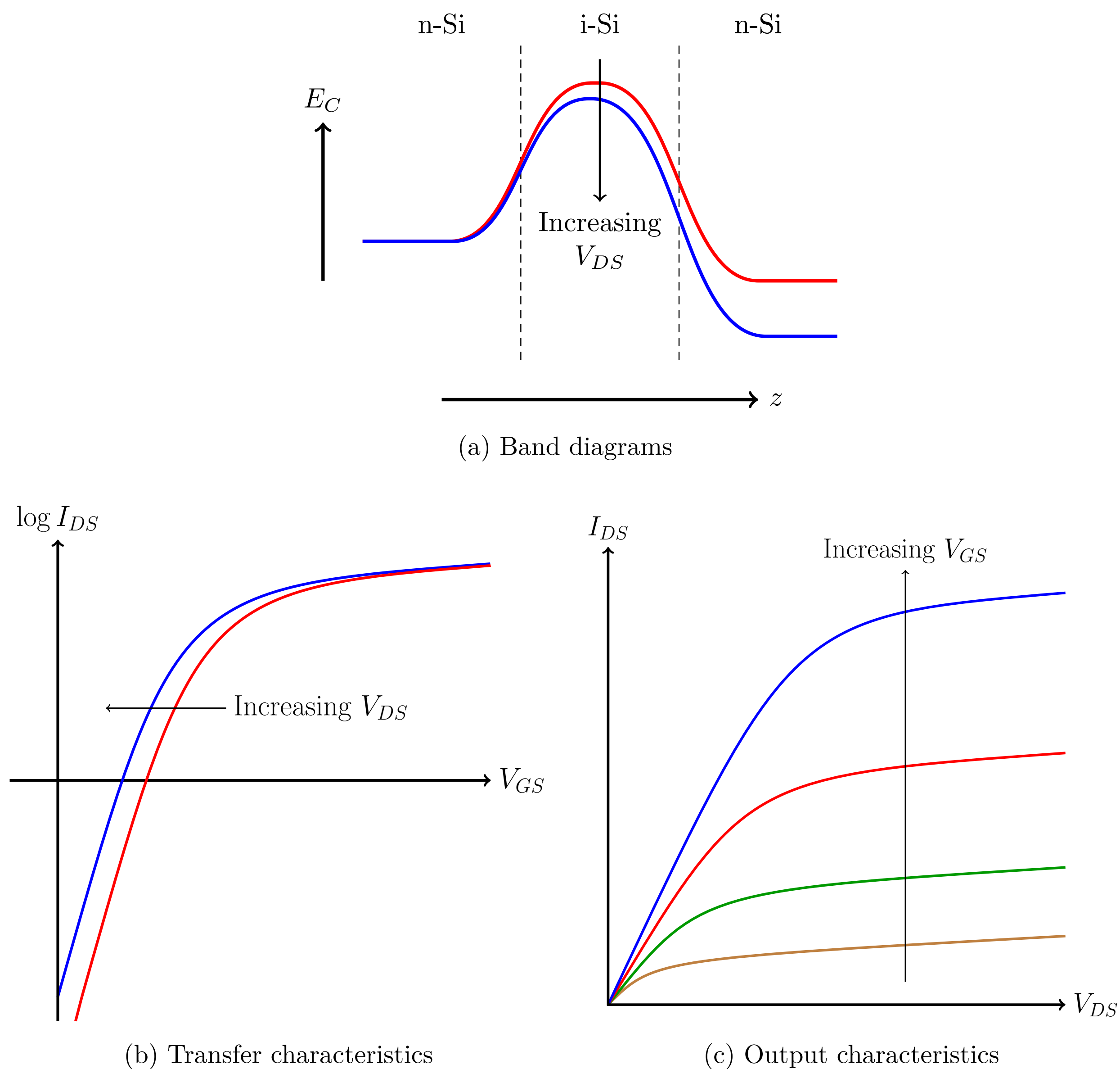
channel effects" (SCEs)—FET physics that arises when the channel length is downscaled to about 150 nm or below. A very wide variety of SCEs are described in the literature [63, 64]; there is some level of ambiguity with regards to what is and what isn't an SCE. Two SCEs are of particular relevance to this thesis: drain-induced barrier lowering (DIBL) and direct source-to-drain tunnelling.

Originally called *the* short-channel effect [65], DIBL is a reduction of the channel barrier energy ϕ that arises as V_{DS} is increased. In an FET with optimal electrostatics, the channel barrier energy is controlled by and only by V_{GS}. Contrarily, in short-channel FETs, due to the short distance between the channel and drain, the channel-to-drain capacitance is sufficiently high to significantly couple the channel and drain. As a result, increasing V_{DS} leads to lower ϕ. Band diagrams of an n-type enhancement-mode silicon MOSFET with DIBL are illustrated in Fig. 2.4a. Phenomenologically, DIBL results in (1) a reduction in threshold voltage V_T, which is seen as a "leftward movement" of transfer characteristics [Fig. 2.4b], and (2) a positive slope of output characteristics in the saturation regime [Fig. 2.4c]. The magnitude of the DIBL effect is quantified through the DIBL parameter δ, defined through

$$\delta = -\frac{V_{T,1} - V_{T,2}}{V_{DS,1} - V_{DS,2}}\,, \tag{2.11}$$

where $V_{DS,i}$ are two values of drain voltage and $V_{T,i}$ is the value of the threshold voltage at $V_{DS} = V_{DS,i}$ for $i = 1, 2$. The parameter δ is typically measured in units of $\text{mV} \cdot \text{V}^{-1}$. Note that $V_{DS,1}$ and $V_{DS,2}$ must both be in the saturation regime to ensure that the leftward movement of transfer characteristics captured in Eq. 2.11 is caused by DIBL and not by current saturation [Eq. 2.9].

Direct source-to-drain tunnelling is a charge transport mechanism that consists of the quantum tunnelling of charge carriers from the source to the drain, under the channel potential barrier. All FETs, including long-channel FETs, exhibit some amount of tunnelling. However, the strength of this effect decays exponentially with channel length L. Indeed, in a typical MOSFET, under the Wentzel—Kramers—Brillouin (WKB) approximation [66] and assuming a square channel potential [7], the probability for direct-source-to-drain tunnelling for an

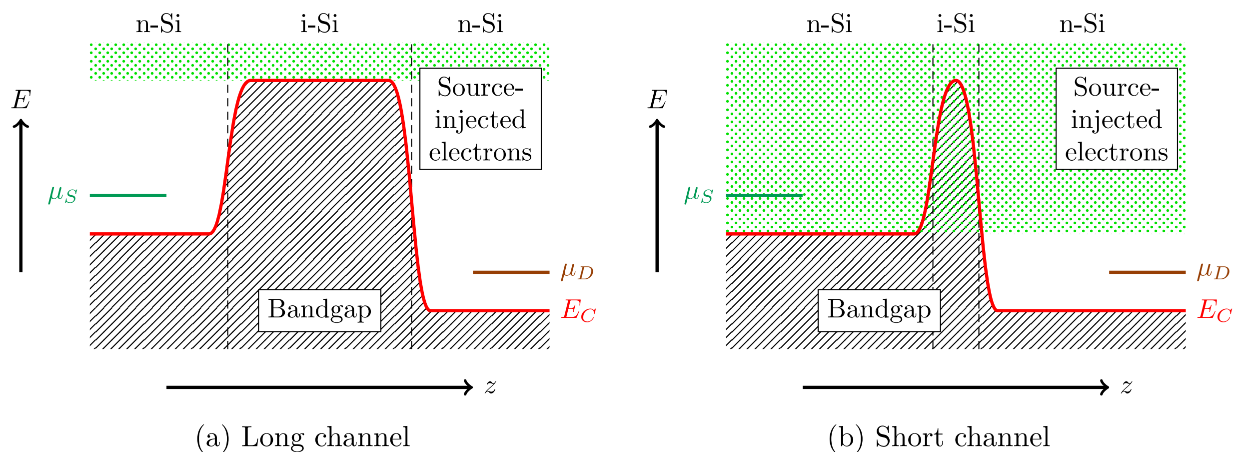


(a) Band diagrams

(b) Transfer characteristics

(c) Output characteristics

Figure 2.4 – Off-state band diagrams, transfer characteristics, and output characteristics of a typical n-type enhancement-mode silicon MOSFET that exhibits drain-induced barrier lowering (DIBL). In FETs with DIBL, increasing V_{DS} leads to a reduction of the channel barrier energy (a). Correspondingly, this results in a reduction of the threshold voltage (b). Similarly, this leads to a positive slope of the output characteristics in the saturation regime (c).



(a) Long channel

(b) Short channel

Figure 2.5 – Direct source-to-drain tunnelling in a typical n-type enhancement-mode silicon MOSFET. Off-state band diagrams in the long-channel limit and short-channel limit are shown. Long-channel FETs only exhibit thermionic emission, while short-channel FETs exhibit both thermionic emission and direct source-to-drain tunnelling (green-dotted region).

electron with energy $E = \mu_S$ is given by

$$T_{\text{tunn}} = \exp\left(-\frac{L}{\hbar}\sqrt{8m^{\star}_{e,t}\phi}\right) , \tag{2.12}$$

where $\hbar$ is the reduced Planck constant and $m^{\star}_{e,t}$ is the electron tunnelling effective mass (EM). Thus, assuming that the channel length is long enough (typically, greater than a few nanometres [67]), tunnelling does not play a significant role in charge transport.

Tunnelling is a charge transport mechanism for low-energy charge carriers that would otherwise not be able to carry current from source to drain, as illustrated in Fig. 2.5. In extreme cases [Fig. 2.5b], this can result in very high off-state current, poor transistor switching, and large power dissipations. In fact, direct source-to-drain tunnelling is generally considered to set the ultimate scaling limit of MOSFETs [67]. Short-channel MOSFETs will be thoroughly investigated in Chapter 7, where it will be shown how BCE can greatly reduce tunnelling leakage.

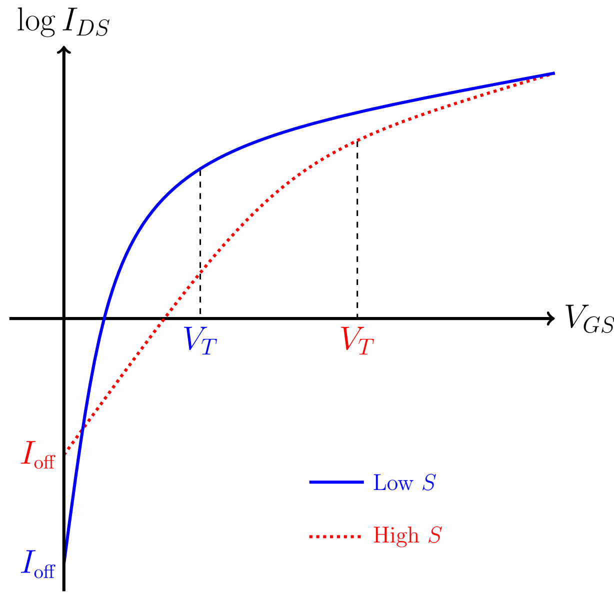


Figure 2.6 – Transfer characteristics of typical MOSFETs with low and high subthreshold swing (STS) S. Keeping the on-state current fixed, an FET with low V_T and low I_{off} has lower S than an FET with high V_T and high I_{off}.

2.1.3 The Subthreshold Swing

A crucial figure of merit for FETs is the subthreshold swing (STS). Formally, it is defined as

$$S = \left(\frac{\partial \log_{10} I_{DS}}{\partial V_{GS}} \right)^{-1} \tag{2.13}$$

and is typically measured in units of mV$\cdot$ dec^{-1}. Graphically, the STS is the inverse of the slope of the base-10 logarithm of the drain current $\log_{10} I_{DS}$ as a function of gate voltage V_{GS} [Fig. 2.6]. The STS quantifies how much of an increment in V_{GS} is required to obtain a decade (factor of ten) increase in I_{DS}. While S is generally V_{GS}-dependent, in MOSFETs, it is roughly constant in the subthreshold regime [16]. FETs with low S may be characterized by low threshold voltage V_T, low off-state current I_{off}, or both [Fig. 2.6].

To compute the STS of a MOSFET, consider that, using the chain rule, it may be

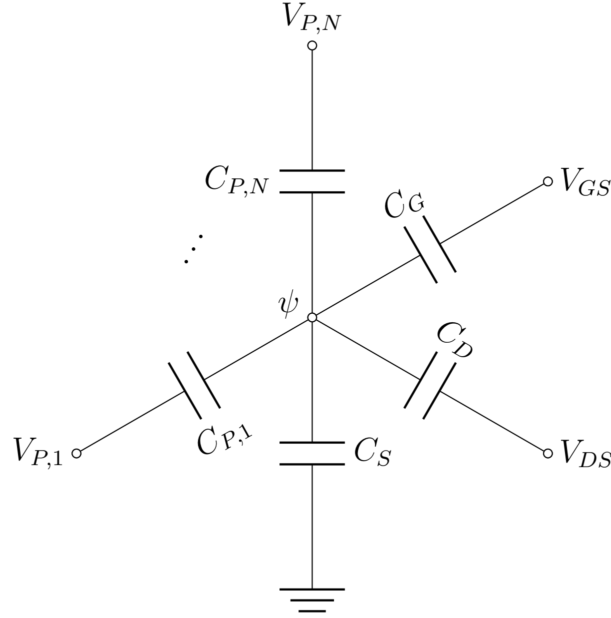


Figure 2.7 – Circuit diagram of the various capacitive couplings of an FET channel with surface potential ψ to neighbouring contacts. These include gate, source, drain, and an arbitrary number N of parasitic contacts. Electric potentials at contacts and capacitive couplings to contacts are, respectively, denoted by V and C.

expressed as the product of two terms:

$$S = \left(\frac{\partial \log_{10} I_{DS}}{\partial \phi} \frac{\partial \phi}{\partial V_{GS}} \right)^{-1} , \tag{2.14}$$

where ϕ is the channel barrier energy illustrated in Fig. 2.2a.

The first term $\frac{\partial \log_{10} I_{DS}}{\partial \phi}$ is known as the thermodynamics or transport factor. In the off state, $\phi \gg k_B T$, so that the Fermi–Dirac distribution describing the statistics of high-energy electrons, which carry current through thermionic emission, can be approximated as a Boltzmann distribution. Given the roughly constant density of states (DOS) of source-injected electrons, which are typically confined to a plane [68]:

$$I_{DS} \propto \exp\left(-\frac{\phi}{k_B T} \right) \tag{2.15}$$

$$\implies \frac{\partial \log_{10} I_{DS}}{\partial \phi} = -\frac{1}{\ln 10} \frac{1}{k_B T} . \tag{2.16}$$

The second term $\frac{\partial \phi}{\partial V_{GS}}$ is known as the body factor. The FET channel surface potential ψ can be thought of as being set through capacitive coupling not only to the gate contact but also to other neighbouring contacts. This includes the source and drain contacts with capacitive couplings denoted by C_S and C_D, respectively. Parasitic capacitive couplings to the channel may also exist; for example, to the contacts of neighbouring transistors or other electronic components on an IC. The potentials at these contacts are denoted by $V_{P,i}$ and the corresponding capacitive couplings are denoted by $C_{P,i}$ with $i = 1, 2, \cdots, N$. These various couplings to the channel charge are illustrated in Fig. 2.7. Thus, extending on the idea established in Eq. 2.1, the capacitive charging energy associated with a surface potential ψ is

$$U_\psi = \frac{1}{2} C_G \left(V_{GS} - \psi\right)^2 + \frac{1}{2} C_S \psi^2 + \frac{1}{2} C_D \left(V_{DS} - \psi\right)^2 + \frac{1}{2} \sum_{i=1}^{N} C_{P,i} \left(V_{P,i} - \psi\right)^2 . \tag{2.17}$$

By energy minimization, it can be inferred that

$$\frac{dU_\psi}{d\psi} = 0 \implies \psi = \frac{C_G V_{GS} + C_D V_{DS} + \sum_{i=1}^{N} C_{P,i} V_{P,i}}{C_G + C_S + C_D + \sum_{i=1}^{N} C_{P,i}} . \tag{2.18}$$

Since ψ is equal to $-\frac{1}{q}\phi$ up to an additive constant [Eq. 2.3], it follows that the body factor is

$$\frac{\partial \phi}{\partial V_{GS}} = -q \frac{C_G}{C_G + C_S + C_D + \sum_{i=1}^{N} C_{P,i}} \stackrel{\text{def}}{=} -q\alpha_G , \tag{2.19}$$

where $0 \leq \alpha_G \leq 1$ is the gate control parameter [62]. Note that Eq. 2.17 can be formalized into an accurate ballistic MOSFET drain current model known as the capacitor model [69], including a nanoscale monolayer black phosphorus MOSFET model I published during my doctoral studies [4].

Combining Eqs. 2.16 and 2.19, it follows that the STS of a MOSFET in the off state is

$$S = \ln 10 \frac{k_B T}{q} \frac{1}{\alpha_G} \geq \ln 10 \frac{k_B T}{q} . \tag{2.20}$$

Equation 2.20 sets a lower bound on the STS of a MOSFET due to physical limits in the thermodynamics of MOSFET charge transport (thermionic emission) [Eq. 2.16] and the electrostatics of MOSFET charge inversion (the field effect) [Eq. 2.19]. In particular, at room

temperature ($T = 300$ K), one has

$$S \geq 60 \text{ mV} \cdot \text{dec}^{-1} \,. \tag{2.21}$$

The limit set in Eq. 2.20 is sometimes described as “Boltzmann’s tyranny” since it arises as a result of Boltzmann statistics and has significant repercussions on power consumption in MOSFET-based ICs, as will be shown in the next section. Parenthetically, the phrase “Boltzmann’s tyranny” appears to have first been coined in 2008 by Zhirnov and Cavin [70].

2.2 Power Consumption of Processors

Modern central processing units (CPUs) in general-purpose personal computers consume between 10 W and 100 W of power, whereas mobile CPUs, such as those found in smartphones, may consume as little as 1 W of power. The power consumption P in a processor comes in many flavours:

$$P = P_{\text{interconnect}} + P_{\text{short}} + P_{\text{dynamic}} + P_{\text{leakage}} + \cdots \tag{2.22}$$

Before defining the various terms in this equation, it should be noted that all sources of power consumption are proportional to V_{DD}^2 (or in some situations, higher powers of V_{DD}), where V_{DD} is the power supply voltage. Specifically, on a processor, V_{DD} is the voltage applied to the gate and drain contacts (V_{GS} and V_{DS}) of a transistor to switch it on. In modern and optimized processors, V_{DD} is typically related to the transistors’ threshold voltage V_T as follows [14, 71]:

$$V_{DD} \approx 3V_T \,. \tag{2.23}$$

The term $P_{\text{interconnect}}$ represents the Joule heating of the various structures and wiring connecting different components of an IC electrically, namely, interconnects. Given a power

supply voltage V_{DD}, the voltage drop across an interconnect in a serial circuit containing interconnects and other electronic components (e.g. transistors) is proportional to $\rho_{\text{interconnect}} V_{DD}$, where $\rho_{\text{interconnect}}$ is the resistivity of the interconnect material. Thus, from Ohm's law,

$$P_{\text{interconnect}} \propto \rho_{\text{interconnect}} V_{DD}^2 \,. \tag{2.24}$$

The interconnects' Joule heating can thus be significantly reduced by reducing V_{DD}. Minimizing $\rho_{\text{interconnect}}$ is also important; low-resistivity materials such as copper are thus sought for interconnects. A large body of research about low-resistivity interconnects can be found in the literature, for example, research pertaining to the minimization of grain-boundary resistance in copper interconnects [72, 73]. It should be noted that in optimized, modern processors, $P_{\text{interconnect}}$ is not a significant part of total power consumption. Alternatively, $P_{\text{interconnect}}$ can be thought of as being implicitly captured in P_{dynamic}, as will be explained in Sec. 2.2.1.

The term P_{short} is the short-circuit power loss, namely, the power consumed as a result of transient short circuits arising during logic gate toggles. Since transistors in a logic gate require a very small but finite amount of time to change state, transistors in a toggling logic gate may be simultaneously conducting, leading to an anomalous short circuit and associated Joule heating. These circuit-level power losses can be significant and are the object of a substantial body of research [74, 75], but will not be further discussed in this thesis.

The last two terms in Eq. 2.22 arise at the transistor level and account for around 90% of processor power consumption [76]. The dynamic power P_{dynamic} and the leakage power P_{leakage} are described in detail in the next section.

2.2.1 Dynamic and Leakage Power

On ICs, logic and memory modules take the forms of groups of interconnected transistors, for instance, static random-access memory cells and logic gates. These modules store and manipulate bits of information—0s and 1s. During such a bit manipulation, the state of

one or more transistors in the module may switch from the off state to the on state or vice versa. Switching the state of a transistor from the off state to the on state involves a transfer of charge to the transistor's gate contact. This requires energy; specifically, to charge the channel-to-gate capacitor as well as the associated interconnects, sometimes referred to as the "load." Since interconnects have a significantly larger physical size than the transistor channel, the load capacitance C_L dominates the gate-to-channel capacitance C_G. Thus, given a power supply voltage V_{DD}, the energy required to charge a transistor is

$$E_L = C_L V_{DD}^2 \,, \tag{2.25}$$

half of which is stored in the load capacitor and half of which is dissipated through Joule heating in interconnects; note that this assumes the charging to be non-adiabatic [15]. Furthermore, the time required to charge a transistor and the load capacitor, called loaded time delay τ_L, is set by the transistor's on-state current I_{on}:

$$\tau_L \approx \frac{C_L V_{DD}}{I_{\text{on}}} \,. \tag{2.26}$$

This is because the final charge being stored is $C_L V_{DD}$ and the rate of charging is on the order of I_{on}. This sets the minimal time required for one cycle of logic operations in an IC. Correspondingly, the clock frequency f at which an IC can be run must satisfy

$$f \leq \frac{1}{\tau_L} \,. \tag{2.27}$$

In modern processors, according to the IRDS™ [14], $C_L \approx 7$ fF, $V_{DD} \approx 0.7$ V, and $I_{\text{on}} \approx 50$ μA, so that $\tau_L \approx 100$ ps and $\frac{1}{\tau_L} \approx 10$ GHz. On the other hand, typically, $f \approx 3$ GHz. Thus, as discussed in Chapter 1, modern processors operate far from their maximum possible clock frequency.

The dynamic power P_{dynamic} represents the power required to charge transistors. Let the number of transistors on an IC be N_T and suppose a fraction α of them gets charged per clock cycle; α is known as the logic activity factor and is typically around 0.01 [41]. The

dynamic power is then given by

$$P_{\text{dynamic}} = \alpha N_T E_L f = \alpha N_T C_L V_{DD}^2 f = \alpha N_T I_{\text{on}} V_{DD} \tau_L f \,. \tag{2.28}$$

The leakage power P_{leakage} represents the off-state power consumption of transistors due to various current leakage mechanisms, such as thermionic emission, direct source-to-drain tunnelling, and gate leakage [77]. It is simply given by

$$P_{\text{leakage}} = N_T I_{\text{off}} V_{DD} \,. \tag{2.29}$$

Given an average STS S_{av} and conventions under which the on-state current I_{on} is reached for gate voltage $V_{GS} = V_T = \frac{1}{3}V_{DD}$ [Eq. 2.23] and the off-state current I_{off} is reached for $V_{GS} = 0$, the off-state current can be expressed in terms of S_{av} [Eq. 2.14]:

$$S_{\text{av}} = \frac{\frac{1}{3}V_{DD}}{\log_{10}(I_{\text{on}}) - \log_{10}(I_{\text{off}})} \tag{2.30}$$

$$\implies I_{\text{off}} = I_{\text{on}} 10^{-\frac{V_{DD}}{3S_{\text{av}}}} \,. \tag{2.31}$$

Inserting Eq. 2.31 into Eq. 2.29, one has

$$P_{\text{leakage}} = N_T V_{DD} I_{\text{on}} 10^{-\frac{V_{DD}}{3S_{\text{av}}}} \,. \tag{2.32}$$

Thus, adding Eq. 2.28 to Eq. 2.32, ignoring short-circuit power loss, the power consumption of a processor is given by

$$P = N_T I_{\text{on}} V_{DD} \left(\alpha \tau_L f + 10^{-\frac{V_{DD}}{3S_{\text{av}}}} \right) \,. \tag{2.33}$$

In modern processors, the ratio of leakage to dynamic power is around 0.4 [78]. Equation 2.33 should be interpreted as a relatively rough approximation for power consumption since:

- it does not account for the algebraic dependence of the drain current on V_{DD} in the on state [Eq. 2.8];

- it ignores the resistance of interconnects and contacts, which may result in applied gate and drain voltages lower than V_{DD};
- it only considers thermionic leakage and not gate leakage, which may account for as much as half of total leakage [77].

Nevertheless, Eq. 2.33 offers great qualitative and acceptable quantitative pictures of power consumption in processors and will be used in the following sections to further describe the power dissipation problem.

The goal of this thesis and the broad field of research in which it fits is to minimize the power consumption in Eq. 2.33, all the while maximizing processor performance:

$$\text{Processor performance} \propto N_T f \,. \tag{2.34}$$

In the following, various avenues to achieve this goal are reviewed. They broadly fall into two categories: (1) high-level methods that alter IC architecture or how voltage is supplied to IC components but do not alter the fundamental IC building blocks, i.e. MOSFETs, and (2) low-level methods where MOSFETs are replaced by other types of transistors or electronic switches.

2.2.2 Minimizing Power Consumption with MOSFETs

Interconnect Capacitance Reduction

The root of dynamic power lies in the charging of the load capacitor C_L [Eq. 2.25], namely, the capacitor formed by two neighbouring interconnects or contacts. Processor power consumption can thus be reduced by reducing C_L. Simplistically, using the parallel-plate formula [47],

$$C_L \propto A_{\text{interconnect}} \varepsilon_{\text{interconnect}} \,, \tag{2.35}$$

where $A_{\text{interconnect}}$ is the interconnect surface area and $\varepsilon_{\text{interconnect}}$ is the permittivity of the dielectric between interconnects. Thus, the reduction of C_L can be achieved in two ways.

First, C_L can be reduced by reducing the physical dimensions of interconnects and contacts. Just like transistors, interconnects have undergone exponential downscaling over the past decades. Among other things, this was made possible by technological advances in photolithography, most recently with the development of extreme ultraviolet lithography [79, 80]. Furthermore, ruthenium has been proposed as a potential successor to copper for highly scaled interconnect metallization [81, 82] since ruthenium has smaller grains than copper and thus lower grain-boundary resistance; ruthenium wires with cross-sectional areas as small as 33 nm^2 have been realized experimentally [83].

Second, C_L can be reduced by reducing the permittivity of the dielectric between interconnects [39, 84]. In the early days of the semiconductor industry, silicon dioxide (SiO_2), which has a relative permittivity of 3.9 [16], was the dielectric of choice both for interconnects isolation and transistor gates due to its ease of fabrication. There has since been an "oxide divergence;" the industry now seeks low-κ oxides for interconnects and high-κ oxides for transistors. Low-κ oxides include fluorine-doped [85] and porous [86] SiO_2, as well as porous oxides in general [87]. The lowest possible relative permittivity for conventional materials is 1, that is, the permittivity of air. This calls for air-gapped interconnects, namely, interconnects separated by pockets of air. While this comes with issues related to IC mechanical stability, air gaps have been successfully realized in ICs in recent years [88, 89].

Underclocking and Undervolting

One of the most direct and common methods to reduce processor power consumption, dynamic power specifically, is underclocking, namely, the reduction of the clock frequency f. Indeed, the power required to charge load capacitors is trivially lowered when they are charged less often. Underclocking comes with a cost to processor computing performance that can be partially offset by techniques such as dynamic frequency scaling (also known as throttling), whereby f is automatically adjusted in response to the temperature of the

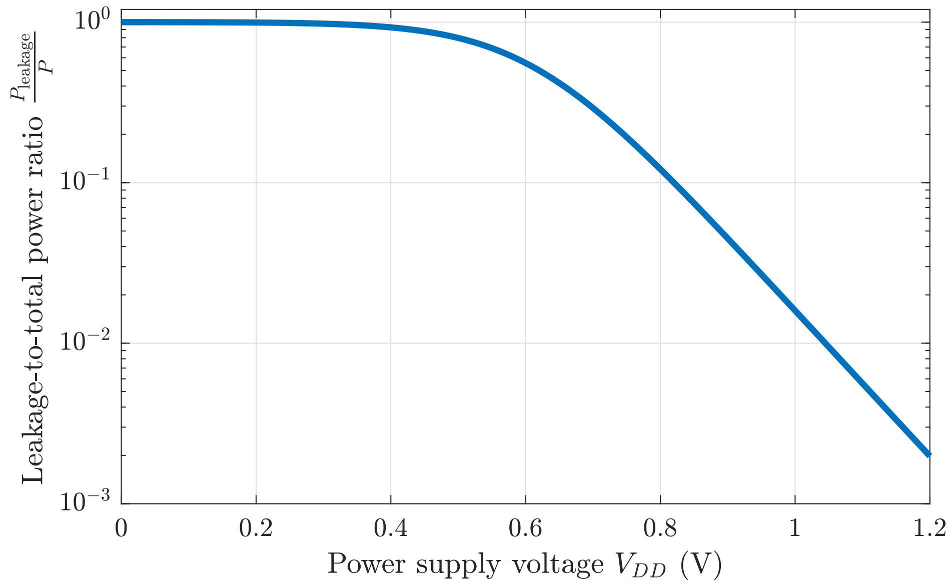


Figure 2.8 – Ratio $\frac{P_{\text{leakage}}}{P}$ of leakage power P_{leakage} [Eq. 2.32] to total processor power consumption P [Eqs. 2.33 and 2.26] as a function of power supply voltage V_{DD}. The average STS is set to $S_{\text{av}} = 80$ mV·dec^{-1}, the on-state current is set to $I_{\text{on}} = 50$ μA, the load capacitance is set to $C_L = 7$ fF, the logic activity factor is set to $\alpha = 0.01$, and the clock frequency is set to $f = 3$ GHz.

processor and workload conditions. Due to this lower performance, underclocking is most often used in mobile processors, for which power management is essential.

Undervolting, namely, the reduction of the power supply voltage V_{DD}, is often used in conjunction with underclocking. Typically, this reduction in V_{DD} implies a reduction in I_{on} due to the exponential dependence of a MOSFET's drain current on V_{DD} in the subthreshold regime [Eq. 2.7]. Since most sources of processor power consumption are proportional to V_{DD}^2 and/or I_{on}, undervolting is particularly effective. It should be noted that aggressive underclocking and/or undervolting may lead to significant leakage power P_{leakage} [Eq. 2.32]. For example, as shown in Fig. 2.8 in which modern processor operation specifications are assumed, for $V_{DD} = 0.9$ V, 4.5% of power consumption is leakage power, while for $V_{DD} = 0.5$ V, this percentage rises to 80%.

Parallelism

Parallel computing is a form of computing where several computations are performed simultaneously—as opposed to sequential computing, where computations are performed one after the other. Parallel processors can thus be run at lower clock frequencies than sequential processors while maintaining the processor performance constant. Parallel computing can sometimes provide very substantial benefits to processor performance and sometimes virtually no benefit depending on the computing task. In terms of processor architecture, parallelism typically requires a multi-core processor, in which multiple, separate processing units (cores) are combined on a single IC and can execute computations independently from each other. Multi-cores processors became ubiquitous starting from the mid 2000s, when Dennard scaling [13] started to fail [Chapter 1]. A consequence of Dennard scaling was that clock frequency underwent an exponential increase from one generation to the next up to the mid 2000s; clock frequencies have since then plateaued. Parallelism can thus be thought of as a compromise to the "forced underclocking" that processors have undergone since the downfall of Dennard scaling.

Other Methods

There is a cornucopia of other MOSFET-based methods to minimize power consumption in processors. One method worth mentioning consists of inserting layers of materials with high heat conductivity on ICs in an effort to cool down processors as quickly as possible. Since the STS of a MOSFET and thus processor power consumption increase with increasing temperature [Eq. 2.20], keeping the processor temperature low is an effective way to minimize power.

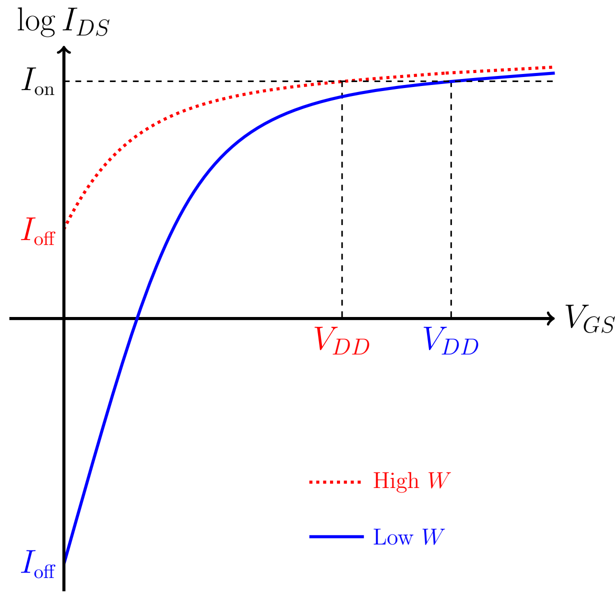


Figure 2.9 – Transfer characteristics of a typical MOSFET in cases of low and high gate metal workfunction W. Increasing W effectively shifts transfer characteristics to the left. Compared to a low-W MOSFET, a high-W MOSFET requires lower V_{DD} to reach I_{on} but has higher I_{off}.

2.2.3 Minimizing Power Consumption with Novel Devices

Steep-Slope FETs

An ideal method to reduce power consumption [Eq. 2.33] while maximizing processor performance [Eq. 2.34] would require:

- reduction in power supply voltage V_{DD};
- no reduction in clock frequency f;
- no reduction in on-state current I_{on} ($f \propto I_{\text{on}}$ [Eq. 2.27]).

Such a method differs from undervolting in the second and third requirements.

In FETs in the off state, the channel barrier energy ϕ [Fig. 2.2a] satisfies [16]

$$\phi \propto -qV_{GS} + W - \chi\,, \tag{2.36}$$

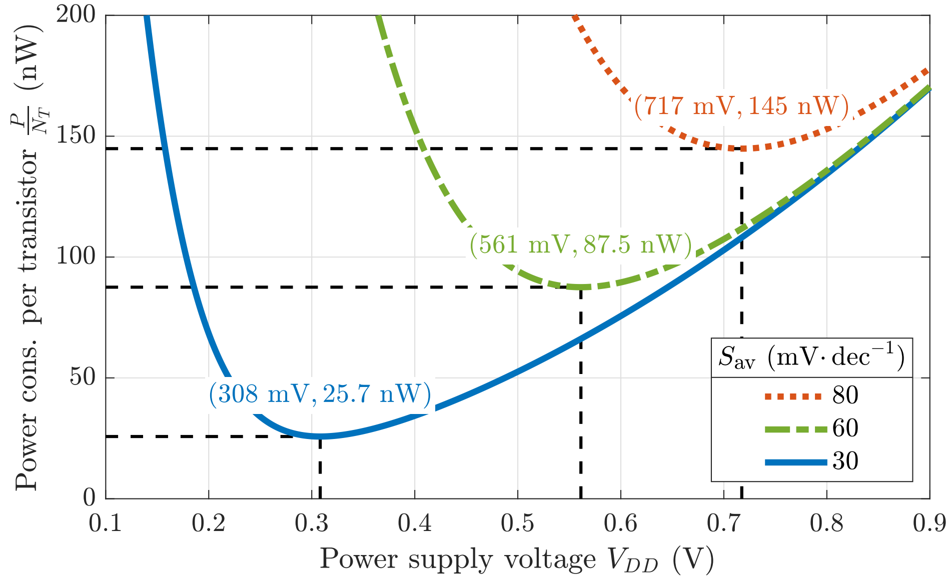


Figure 2.10 – Total processor power consumption per transistor $\frac{P}{N_T}$ [Eqs. 2.33 and 2.26] as a function of power supply voltage V_{DD} for an average STS S_{av} of 80 mV· dec^{-1} (typical FinFET at room temperature), 60 mV· dec^{-1} (ideal MOSFET at room temperature), and 30 mV· dec^{-1} (hypothetical steep-slope FET). It is assumed that the on-state current is $I_{on} = 50$ µA, the load capacitance is $C_L = 7$ fF, the logic activity factor is $\alpha = 0.01$, and the clock frequency is $f = 3$ GHz.

where W is the gate metal workfunction and χ is the semiconductor's electron affinity. Correspondingly, V_{DD}, which can be interpreted as the value of V_{GS} for which $\phi \approx 0$, satisfies

$$V_{DD} \propto \chi - W\,. \tag{2.37}$$

It can thus be seen that V_{DD} can be reduced by choice of a gate metal with high workfunction. This could also be achieved by choice of a semiconductor with low electron affinity, although this is less practical due to a multitude of significant repercussions (notably, changes in carrier mobility and bandgap). Graphically, transfer characteristics can be shifted to the left by increasing W, as shown in Fig. 2.9. This results in lower V_{DD} and exponentially higher I_{off}, which may result in significant leakage power. For MOSFETs, this "leakage power bottleneck" can be attributed to Boltzmann's tyranny, i.e. to the $\frac{k_B T \ln 10}{q}$ STS limit. To lower V_{DD} while keeping leakage power at an acceptable level, a novel FET with sub-$\frac{k_B T \ln 10}{q}$ STS would be needed.

To be precise, dynamic power increases quadratically with V_{DD} [Eq. 2.28] while leakage power increases exponentially as V_{DD} approaches 0 [Eq. 2.32]. This trade-off between dynamic and leakage power sets the optimal value of V_{DD} that minimizes the total power P [Eqs. 2.33 and 2.26], as illustrated in Fig. 2.10. Quantitatively:

$$\frac{\partial P}{\partial V_{DD}} = 0 \implies N_T \left[2\alpha f C_L V_{DD} + I_{\text{on}} \left(1 - \frac{V_{DD} \ln 10}{3S_{\text{av}}} \right) 10^{-\frac{V_{DD}}{3S_{\text{av}}}} \right] = 0 \tag{2.38}$$

$$\implies V_{DD} = \frac{3}{\ln 10} S_{\text{av}} \ln \left(\frac{I_{\text{on}} \ln 10}{6\alpha f C_L} \frac{1}{S_{\text{av}}} \right), \tag{2.39}$$

where it has been assumed that $\frac{V_{DD} \ln 10}{3S_{\text{av}}} \gg 1$. Thus, the optimal value of V_{DD} as well as the corresponding total power consumption increase linearly (approximately) with the average STS S_{av}. Under typical industry requirements, a processor with hypothetical steep-slope FETs with $S_{\text{av}} = 30$ mV$\cdot$dec^{-1} dissipates about 25.7 nW of power per transistor, which is about 29% that of a processor with ideal MOSFETs ($S_{\text{av}} = 60$ mV$\cdot$dec^{-1}), and only about 18% that of a processor with industry-standard FETs, i.e. FinFETs ($S_{\text{av}} = 80$ mV$\cdot$dec^{-1}).

The quest for steep-slope FETs, which do not fall under the same thermodynamics and electrostatics constraints as MOSFETs, is thus essential to minimize processor power and enable future generations of processors toward more Moore. Candidates are broadly reviewed and compared against each other in Chapter 3.

Adiabatic Computing

Loosely speaking, the path toward reduced power consumption that has been discussed in the previous section consists of reducing the energy stored in the load capacitor to distinguish a logical 0 from a logical 1. An alternative and emerging approach is to conserve this energy from one logical operation to the next; this approach is known as adiabatic or reversible computing [15,90–92]. Adiabatic computing can be implemented using conventional MOSFETs [93], steep-slope FETs [15], as well as ferromagnetic logic [94].

In 1961, using the second law of thermodynamics, Rolf Landauer showed that the smallest

amount of energy required to erase one bit of information is [95]

$$E_B = k_B T \ln 2 \,. \tag{2.40}$$

This equation, known as Landauer's principle, has since been verified experimentally [96–98] and is generally accepted to be a fundamental limit of irreversible computing. At room temperature (T = 300 K), Eq. 2.40 implies that 17.9 meV of energy is required to erase a bit of information. The energy stored in load capacitors in conventional CMOS circuits [Eq. 2.25] to distinguish bits of information is currently around 8 keV and will approach the Landauer limit around the year 2050. This can be estimated from Koomey's law [99], which states that the number of computations per unit of energy dissipated in state-of-the-art processors doubles every year and a half. Thus, beyond that point, some amount of "energy recycling" between logical operations will be required to further increase the energy efficiency of computing.

Not unlike conventional computing, adiabatic computing suffers from a trade-off between computation speed and power consumption. As a result, adiabatic computing typically requires very low clock frequencies, making it unsuitable for practical applications. It is, however, an important research direction for future classical computing. Nevertheless, adiabatic computing will not be further discussed in this thesis.

2.3 Summary

The field effect refers to the modulation of the conductivity of materials by an external electric field or voltage. Certain materials, such as lightly doped or intrinsic semiconductors, respond very strongly to the field effect. This makes such materials suitable to be used in switches such as field-effect transistors (FETs). FETs are electronic devices with three terminals: the source, the drain, and the gate. The source is typically grounded. The drain voltage V_{DS} sets the energy window for charge transport; the drain current I_{DS} typically

scales linearly with V_{DS}. The gate voltage V_{GS} modulates the conductivity of the device's channel through the field effect; I_{DS} typically scales exponentially with V_{GS}.

In a metal–oxide–semiconductor field-effect transistor (MOSFET), the most common transistor, the main charge transport mechanism is thermionic emission: the only charge carriers able to carry current are those with energy higher than the channel barrier energy. The number of injected carriers is set by the exponentially decaying tail of the contacts' Boltzmann distributions. Band diagrams are particularly useful tools to visualize this process as well as charge transport in any semiconductor device.

The subthreshold swing (STS) is an FET performance metric that is defined as

$$S = \left(\frac{\partial \log_{10} I_{DS}}{\partial V_{GS}} \right)^{-1} . \tag{2.41}$$

For MOSFETs, it can be derived from thermodynamics and electrostatics that

$$S \geq \ln 10 \frac{k_B T}{q} , \tag{2.42}$$

where T is the device temperature; at room temperature, $S \geq 60$ mV$\cdot$dec^{-1}.

There are many sources of power consumption in processors; the two most important of which are dynamic and leakage power. Dynamic power is the power required to switch transistors from the off state to the on state. Leakage power is the power leaked by transistors in the off state. By adding them, the power consumption of a processor can be approximated as

$$P = N_T V_{DD} \left(\alpha f C_L V_{DD} + I_{\mathrm{on}} 10^{-\frac{V_{DD}}{3 S_{\mathrm{av}}}} V_{DD} \right) , \tag{2.43}$$

where N_T is the number of transistors on the processor, V_{DD} is the power supply voltage, α is the logic activity factor, f is the clock frequency, C_L is the load capacitance, and S_{av} is the average STS of transistors. There are many methods to reduce a processor's power consumption. The most common methods are underclocking and undervolting, but such methods also reduce processor performance.

At fixed N_T, α, f, C_L, and S_{av}, the value of V_{DD} which minimizes P roughly scales linearly with S_{av}; the corresponding minimized P also roughly scales linearly with S_{av}, a parameter with a lower bound in MOSFETs. The power dissipation problem refers to the inability to reduce P below what is allowed by the fundamental MOSFET STS thermodynamics limit of $\frac{k_B T \ln 10}{q}$; for this reason, this limit is called "Boltzmann's tyranny."

To further reduce power consumption while maximizing processor performance, novel FETs with sub-$\frac{k_B T \ln 10}{q}$ STS are required. In the next chapter, such novel FETs, steep-slope FETs, are investigated.

Chapter 3

Steep-Slope Field-Effect Transistors

In a nutshell, smaller subthreshold slopes correspond to faster devices and lower switching energies.

— V. V. Zhirnov & R. K. Cavin, 2008 [70]

As seen in the previous chapter, in principle, a reduction in processor power consumption without compromise to performance can be achieved using steep-slope FETs, i.e. FETs with sub-$\frac{k_B T \ln 10}{q}$ STS. Specifically, given an on-state current I_{on} set by technological requirements (notably, clock frequency requirements) [14], the following FET attributes would be required:

- a low off-state current I_{off} for low leakage power;
- a low threshold voltage V_T to enable operation at low power supply voltage $V_{DD} \approx 3V_T$ for low dynamic and leakage powers;
- a drain current greater than or equal to I_{on} when the applied voltages satisfy $V_{GS} = V_{DS} = V_{DD}$.

The transfer characteristics of such an ideal FET are illustrated in Fig. 3.1.

It is important to note that a low STS is a necessary but not sufficient condition to satisfy these three requirements. Thus, even ignoring practical concerns such as ease of fabrication, scalability, and cost, not all steep-slope FETs are suitable successors to the MOSFET to resolve the power dissipation problem. Furthermore, the suitability of a novel steep-slope FET as a MOSFET replacement is very sensitive to its intended application (e.g. low-power

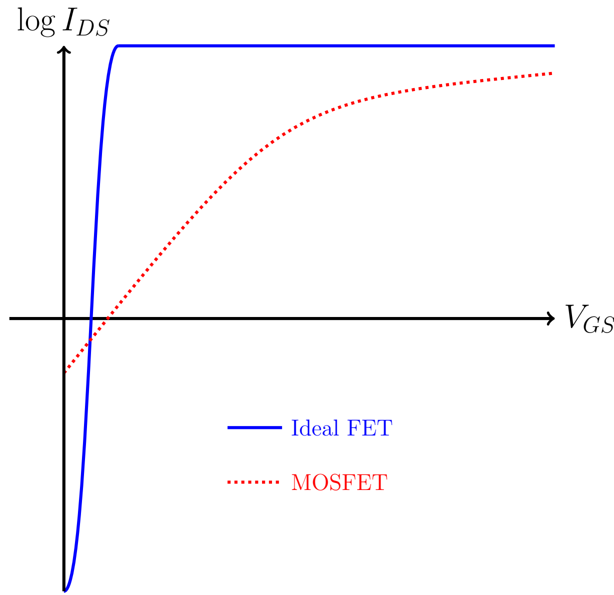


Figure 3.1 – Transfer characteristics of a typical MOSFET and ideal FET. The very sharp switching of the ideal FET—within a small range of gate voltage V_{GS} and over many decades of drain current I_{DS}—enables its operation at low power supply voltage with minimal leakage and dynamic power. For the MOSFET, which cannot switch faster than 60 mV$\cdot$ dec^{-1} at room temperature, these performance metrics are limited by Boltzmann's tyranny.

or high-performance computing). Nevertheless, the STS remains an excellent metric to gauge the worth of a steep-slope FET. This is because more tangible metrics such as the threshold voltage, the off-state current, and the on-state current reflect not only the intrinsic device performance but also extrinsic properties. For example, the threshold voltage can be shifted arbitrarily by changing the gate metal workfunction [Sec. 2.2.3]. Such metrics are thus strongly tied to circuit architecture and application. On the other hand, the STS only depends on intrinsic device properties, such as charge transport mechanisms, gate geometry, gate electrostatics, and material composition of the transistor channel. In other words, the STS reflects, with little to no external noise, the extent to which fundamental device physics may enable lower power consumption.

The most obvious sub-60 mV$\cdot$ dec^{-1}-STS FET is a MOSFET operated at cryogenic (or low) temperature T; indeed, the STS of a MOSFET is proportional to T [Eq. 2.20]. In some applications, such as quantum computing [100] and gravitational wave detection [101], MOSFETs are naturally operated at cryogenic temperatures, leading to low power consumption.

However, for most other applications, this is impractical due to the significant engineering, power, and cost burdens of cryogenic cooling.

Since the subthreshold slope (the reciprocal of the STS) of an FET is given by

$$\frac{1}{S} = \underbrace{\left(\frac{\partial \log_{10} I_{DS}}{\partial \phi}\right)}_{\text{Transport factor}} \overbrace{\left(\frac{\partial \phi}{\partial V_{GS}}\right)}^{\text{Body factor}}, \tag{3.1}$$

there are three ways to reduce the STS:

- increase the transport factor $\frac{\partial \log_{10} I_{DS}}{\partial \phi}$, that is, the sensitivity of the drain current I_{DS} on the channel barrier energy ϕ; this requires alternative charge transport mechanisms; large-transport-factor FETs are reviewed in Sec. 3.1 and Appendix A.1;
- increase the body factor $\frac{\partial \phi}{\partial V_{GS}}$, that is, the sensitivity of the channel barrier energy ϕ on the applied gate voltage V_{GS}; this requires alternative gate stack electrostatics; large-body-factor FETs are reviewed in Sec. 3.2 and Appendix A.2;
- simultaneously increase the transport and body factors, as is discussed in Appendix A.3.

This chapter describes the tunnel FET (TFET) and the negative capacitance FET (NCFET) in depth; these two devices are generally considered to be the most promising potential surrogates to the MOSFET. Interested readers may consult Appendix A for in-depth reviews of several other large-transport-factor FETs—the Dirac source FET (DSFET), the cold-source FET (CSFET), the superlattice FET (SLFET), the impact–ionization FET (IIFET), and the feedback FET (FBFET)—as well as several other large-body-factor FETs—the nanoelectromechanical FET (NEMFET) and the piezoelectric FET (PEFET). In Sec. 3.3, all these steep-slope FETs are compared against each other; their potential as MOSFET replacements to solve the power dissipation problem is discussed.

3.1 Large Transport Factor and the Tunnel FET

A general estimate for the off-state thermionic leakage current of an FET is given by

$$I_{DS}(\phi) \approx M_S(\phi) \exp\left(-\frac{\phi}{k_B T}\right) . \quad (3.2)$$

Here, ϕ is the channel barrier energy. The term $M_S(\phi)$ is the number of conduction modes for source-injected charge carriers at energy ϕ; this term is related to the DOS of source-injected carriers. The term $\exp\left(-\frac{\phi}{k_B T}\right)$ arises as a result of the Boltzmann statistics of source-injected carriers in the off state. For a MOSFET, $M_S(\phi)$ is slowly varying [102], resulting in a transport factor of $\frac{\partial \log_{10} I_{DS}}{\partial \phi} = -\frac{1}{\ln 10} \frac{1}{k_B T}$ [Eq. 2.16].

Now, consider an FET engineered in such a way that $M_S(\phi)$ has a minimum at an arbitrary value ϕ_0 of the channel barrier energy in the off state. For example, one could consider a gapped case

$$M_S(\phi) \propto H\left(|\phi - \phi_0| - \frac{E_G}{2}\right) , \quad (3.3)$$

where H is the Heaviside step function and $E_G > 0$ is a bandgap energy, resulting in a transport factor of

$$\begin{aligned} \frac{\partial \log_{10} I_{DS}}{\partial \phi} = &-\frac{1}{\ln 10} \frac{1}{k_B T} H\left(|\phi - \phi_0| - \frac{E_G}{2}\right) \\ &+ \log_{10}\left[I_{DS}\left(\phi_0 + \frac{E_G}{2}\right)\right] \delta\left(\phi - \phi_0 - \frac{E_G}{2}\right) \\ &+ \log_{10}\left[I_{DS}\left(\phi_0 - \frac{E_G}{2}\right)\right] \delta\left(\phi - \phi_0 + \frac{E_G}{2}\right) . \end{aligned} \quad (3.4)$$

In this model, the transport factor (and subthreshold slope) are infinite at $\phi = \phi_0 \pm \frac{E_G}{2}$. Note that this only considers thermionic leakage; a more realistic description would include the effects of other forms of leakage, which would result in finite, albeit high, transport factor.

One may also consider an FET with

$$M_S(\phi) \propto |\phi - \phi_0|^n \ , \tag{3.5}$$

where $n > 0$, resulting in a transport factor of

$$\frac{\partial \log_{10} I_{DS}}{\partial \phi} = -\frac{1}{\ln 10}\left(\frac{1}{k_B T} + \frac{n}{\phi_0 - \phi}\right) . \tag{3.6}$$

The transport factor is infinite at $\phi = \phi_0$ and high around $\phi = \phi_0$.

Intuitively, the reason for the large transport factors in Eqs. 3.4 and 3.6 is the “filtering” of high-energy charge carriers around $\phi = \phi_0$ that, if present, would result in much greater thermionic leakage and STS. For this reason, FETs that have a limited number of conduction modes for source-injected carriers in the off state are known as energy-filtering FETs. Most large-transport-factor FETs are energy-filtering FETs. This includes the TFET, the CSFET [Appendix A.1.2], and the SLFET [Appendix A.1.3], which are described by Eq. 3.3, and the DSFET [Appendix A.1.1], which is described by Eq. 3.5. Nonetheless, not all large-transport-factor FETs are energy-filtering FETs. Instead, some FETs harness semiconductor physics characterized by positive feedback, such as the IIFET [Appendix A.1.4] and the FBFET [Appendix A.1.5].

As indicated in its name, the main charge transport mechanism in a tunnel FET is tunnelling, specifically, band-to-band tunnelling (BTBT). Devices exhibiting BTBT have been investigated at least since the 1930s, including Zener's diode in 1934 [103], Stuetzer's junction fieldistor in 1952 [104], Esaki's tunnel diode in 1957 [105], and Hofstein and Warfield's insulated gate tunnel junction triode in 1965 [106]. In its modern form, the TFET was first proposed in 1978 by Quinn, Kawamoto, and McCombe [107] and modelled in 1987 [108]. Sub-$\frac{k_B T \ln 10}{q}$ STS was first demonstrated experimentally in a carbon nanotube TFET in 2004 by Appenzeller, Lin, Knoch, and Avouris [54], making the TFET the first steep-slope FET.

In its most basic form, the structure of a TFET is identical to that of a MOSFET, with the exception that the source and drain are doped with opposite polarities. For example,

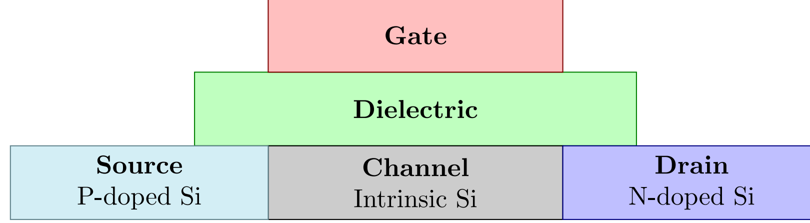


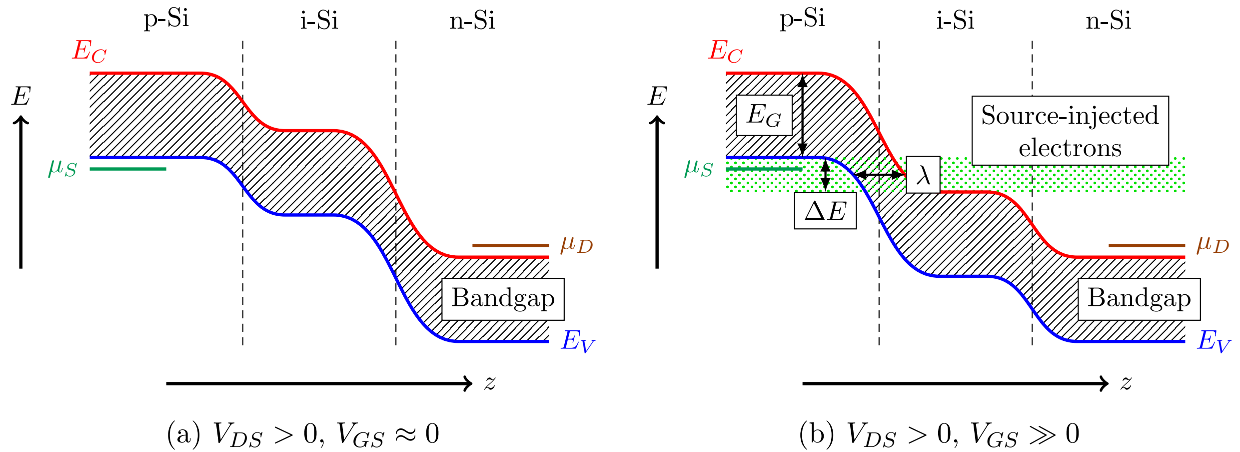
Figure 3.2 – Schematic of a typical n-type enhancement-mode silicon tunnel field-effect transistor (TFET). The source is p-doped, the channel is intrinsic, and the drain is n-doped. The main charge transport mechanism is valence-band-to-conduction-band tunnelling at the source–channel junction.

as illustrated in Fig. 3.2, an n-type enhancement-mode silicon TFET has a p-doped source and an n-doped drain. The gate metal is chosen so that in the off state, the source–channel junction forms a staggered gap band alignment, as illustrated in Fig. 3.3a. Off-state leakage in TFETs is thus dominated by Shockley–Read–Hall generation in the source [109], direct source-to-drain tunnelling, and thermionic emission of holes, making the off-state current extremely small. In the on state, a positive gate voltage V_{GS} pushes the channel CBM E_C below the source VBM E_V, thereby forming a broken gap band alignment at the source–channel junction, as shown in Fig. 3.3b. This band alignment creates a narrow energy window ΔE over which electrons in the source valence band can tunnel to the channel conduction band over a tunnelling length λ. This energy window is bounded above by the source VBM, thereby effectively suppressing the thermionic leakage of high-energy electrons observed in MOSFETs and providing the TFET with a low STS.

The BTBT current of a TFET is proportional to the BTBT probability, which can be estimated using the WKB approximation as [110]

$$T_{\mathrm{BTBT}} \approx \exp\left(-\frac{4}{3q\hbar}\frac{\sqrt{2m_t^\star E_G^3}}{E_G+\Delta E}\lambda\right) , \tag{3.7}$$

where $m_t^\star$ is the BTBT EM and E_G is the bandgap energy. Note that this formula assumes that the source–channel junction is a homojunction and ignores phonon-assisted tunnelling, which can be dominant in indirect-bandgap semiconductors [111]. Importantly, both ΔE and λ are V_{GS}-dependent, making T_{BTBT} a super-exponential function of V_{GS} and the STS



(a) $V_{DS} > 0$, $V_{GS} \approx 0$

(b) $V_{DS} > 0$, $V_{GS} \gg 0$

Figure 3.3 – Band diagrams of a typical n-type enhancement-mode silicon TFET in the off state (a) and on state (b). The CBM E_C (red curve) and VBM E_V (blue curve) are plotted as a function of position z along the source-to-drain axis. The bandgap energy is E_G. The source and drain Fermi levels are $\mu_{S,D}$ (green and brown lines). There are no energy states in the bandgap, between $E_V(z)$ and $E_C(z)$ (black-crosshatched region). In the off state, essentially no electrons are injected from the source. In the on state, electrons are injected (green-dotted region) through quantum tunnelling from the source valence band to the channel conduction band in an energy window ΔE over a tunnelling length λ.

V_{GS}-dependent. Due to the exponentially decaying form of the BTBT probability, the on-state current of TFETs tends to be significantly lower than that of MOSFETs. Correspondingly, a significant part of TFET research and development has been geared toward realizing high BTBT current. There are several ways to achieve a high on-state BTBT current in TFETs:

- lower the tunnelling EM $m_t^\star$; this can be achieved through choice of semiconductors with low EM and through strain engineering, as has been proposed and demonstrated for strained silicon [112], strained germanium [113], strained silicon–germanium [114], and strained III–V semiconductors [115, 116]; for example, in all-silicon TFETs, strain can boost the on-state current of a TFET by one order of magnitude [112];
- lower the bandgap energy E_G; however, reducing E_G tends to increase the STS since the energy-filtering property of TFETs deteriorates in the narrow-bandgap limit [Fig. 3.3a]; in the limiting case of no bandgap, graphene TFETs have been theoretically shown to have on-state current within one order of magnitude of that of silicon MOSFETs but do not break the $\frac{k_B T \ln 10}{q}$ STS limit [117];
- increase the BTBT energy window ΔE; this can be achieved by increasing the source

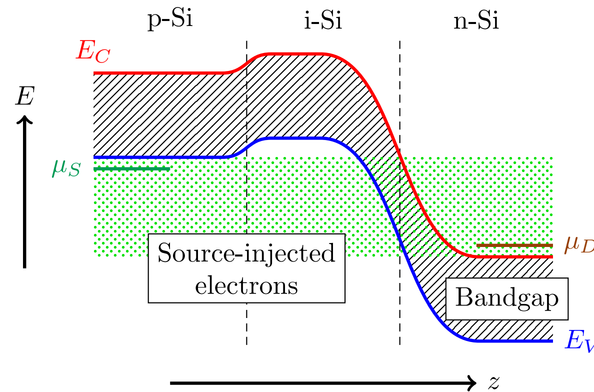


Figure 3.4 – Band diagram of a typical n-type enhancement-mode silicon TFET with $V_{DS} > 0$ and $V_{GS} \ll 0$. Electrons are injected from the drain through quantum tunnelling from the drain conduction band to the channel valence band over a wide energy window. In this state, the TFET exhibits ambipolar conduction, an undesired off-state conduction mechanism.

doping concentration; degenerate doping is typically required; alternatively, this can be achieved by having distinct source and channel semiconductors that, through a mismatch in their electron affinities, naturally form broken gap or staggered gap heterojunctions, as was proposed and demonstrated for various III–V semiconductor heterojunctions [118–120];

- decrease the tunnelling length λ; this can be achieved, as previously mentioned, with increased doping concentrations and III–V semiconductor heterojunctions; alternatively, λ can be reduced through bound-charge engineering, as I will show in Chapter 5.

The techniques used to increase the on-state current of a TFET may inadvertently increase leakage current in the deep off state, when $V_{DS} > 0$ and $V_{GS} \ll 0$ (for an n-type enhancement-mode device). Indeed, as illustrated in Fig. 3.4, such conditions lead to a short tunnelling length at the channel–drain junction, leading to a very significant BTBT current over an energy window ranging from the source VBM to the drain CBM. This is known as ambipolarity, a phenomenon that generally refers to high and comparable conduction for both positive and negative gate voltages while the drain voltage is positive. Ambipolar conduction prevents a TFET from being used in certain logic circuits such as inverters, which is a severe barrier for the implementation of TFETs in CMOS logic [121]. A large number of device-level engineering techniques to alleviate ambipolarity have been proposed [122], including pocket

doping [123, 124], whereby a "pocket" of very highly doped silicon is placed at the junction with the channel in a relatively lightly doped silicon source so as to reduce the tunnelling length λ for high on-state current and reduce the energy window for ambipolar conduction.

Thanks to its relatively simple design that is very similar to that of a MOSFET, silicon TFETs could be produced on a large scale using modern, CMOS-compatible process flows [125], although silicon may not be the best material for TFETs. In 2016, Memisevic, Svensson, Hellenbrand, Lind, and Wernersson experimentally demonstrated TFET based on an InAs–GaAsSb–GaSb heterojunction with on-state current $I_{\text{on}} = 10.6\ \text{A}\cdot\text{m}^{-1}$, minimal STS of 48 $\text{mV}\cdot\text{dec}^{-1}$, and sub-60 $\text{mV}\cdot\text{dec}^{-1}$ STS sustained over 2 decades of drain current (at room temperature); this device could outperform silicon MOSFETs at a power supply voltage $V_{DD} \lessapprox 0.3$ V [126]. More recently, in 2020, Tomioka, Gamo, Motohisa, and Fukui demonstrated a GAA InGaAs–Si heterojunction TFET with $I_{\text{on}} = 2.4\ \text{A}\cdot\text{m}^{-1}$, minimal STS of 21 $\text{mV}\cdot\text{dec}^{-1}$, and sub-60 $\text{mV}\cdot\text{dec}^{-1}$ STS sustained over 4 decades of drain current at drain voltage $V_{DS} = 0.35$ V [127]. Despite these promising reports, TFETs have not been applied commercially to date. Regardless, conceptually, the TFET is extremely important as it was the first energy-filtering FET and, furthermore, the first steep-slope FET.

3.2 Large Body Factor and Negative Capacitance

The body factor of a MOSFET is given by [Eq. 2.19]

$$\frac{\partial \phi}{\partial V_{GS}} = -q \frac{C_G}{C_G + C_P}\,, \tag{3.8}$$

where ϕ is the channel barrier energy, V_{GS} is the gate voltage, C_G is the gate capacitance, and C_P is the parasitic capacitance. Since $C_G, C_P > 0$,

$$\left| \frac{\partial \phi}{\partial V_{GS}} \right| \leq q\,. \tag{3.9}$$

In other words, ϕ can be decreased no more than 1 meV for every increment of 1 mV in V_{GS}. This is a fundamental limit on the effect that an external gate voltage can have on internal device electrostatics in MOSFETs.

It is ultimately Boltzmann's distribution that sets bounds on the off-state thermionic leakage and transport factor in a MOSFET [Eq. 2.16]; these bounds are thus referred to as "Boltzmann's tyranny." Similarly, it is ultimately Gauss's law that sets the bound described by Eq. 3.9 since the parameters in this equation (i.e. the capacitances) stem from Gauss's law [47]. This bound could thus be called "Gauss's tyranny," thereby making Boltzmann and Gauss surprising accomplices in impeding the switching of MOSFETs. Parenthetically, as one of the reviewers of this thesis pointed out, it may unfair to call Boltzmann and Gauss tyrants: "It is perhaps remarkable that the thermionic limit to STS is as low as 60 mV$\cdot$ dec^{-1}. Nature need not have been so kind."

To surpass Gauss's tyranny, some FETs, known as gate-voltage-amplifying FETs, employ novel gate stack electrostatics going beyond that of a MOSFET's metal–oxide–semiconductor capacitor. Such novel gate electrostatics broadly fall into three categories: ferroelectricity for the NCFET, electromechanical gates for the NEMFET [Appendix A.2.1], and piezoelectricity for the PEFET [Appendix A.2.2].

Materials polarize in response to external electric fields: negative and positive charges within the material (atomic nuclei and electrons) slightly shift away from each other, thereby forming atomic-scale dipoles. For most materials, known as linear dielectrics, this response is linear. Their Gibbs free energy density is given by

$$G_D = \frac{1}{2\left(\varepsilon - \varepsilon_0\right)} P^2 - \mathbf{E} \cdot \mathbf{P}\,, \tag{3.10}$$

where ε is the material's permittivity, ε_0 is the vacuum permittivity, $\mathbf{P}$ is the polarization density, and $\mathbf{E}$ is the external electric field. Thermodynamic equilibrium thus requires $dG_D = 0 \implies \mathbf{P} = \left(\varepsilon - \varepsilon_0\right)\mathbf{E}$, as expected for linear dielectrics [47]. In particular, for $E = 0$, G_D is minimized for $P = 0$, as illustrated in Fig. 3.5.

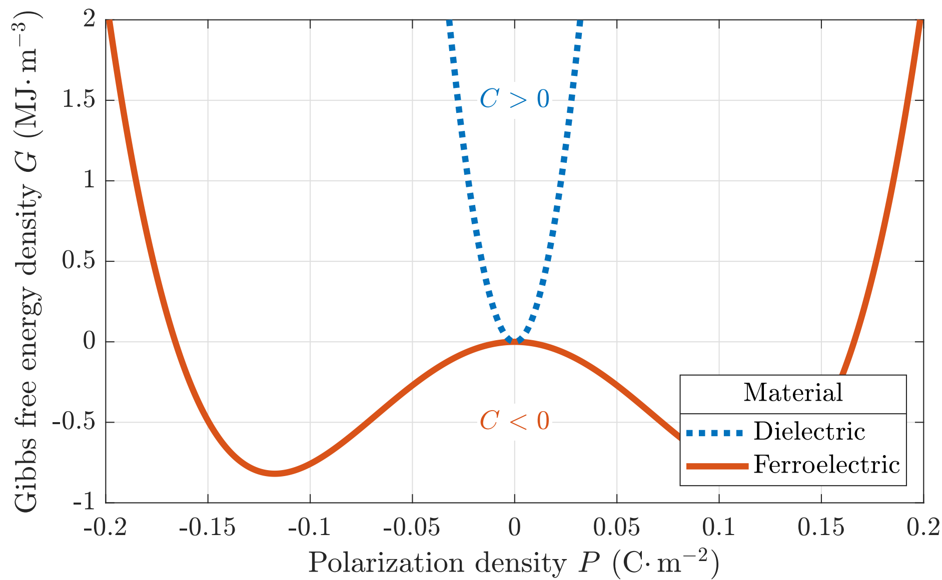


Figure 3.5 – Gibbs free energy density G of a dielectric and a ferroelectric as a function of polarization density P in the absence of an external electric field. The dielectric's Gibbs free energy is computed from Eq. 3.10; the dielectric is assumed to be hafnium dioxide (HfO_2) with a permittivity of $30\varepsilon_0$ [17]. The ferroelectric's Gibbs free energy is computed from Eq. 3.11; the ferroelectric is assumed to be $Hf_{0.5}Zr_{0.5}O_3$ with anisotropy constants given by $\alpha = -1.19 \times 10^8$ m$\cdot$ F^{-1}, $\beta = 4.32 \times 10^9$ m$^5\cdot$ F$^{-1}\cdot$ C^{-2}, and $\gamma = 0$ [128]. Because of the negative curvature of the ferroelectric's Gibbs free energy at $P = 0$, ferroelectric capacitors can exhibit negative capacitance $C < 0$.

In contrast, some materials, known as ferroelectrics, exhibit spontaneous polarization, even in the absence of an external electric field. Their Gibbs free energy density is given in Landau–Khalatnikov theory [129] by

$$G_F = \alpha P^2 + \beta P^4 + \gamma P^6 - \mathbf{E} \cdot \mathbf{P}\,, \tag{3.11}$$

where α, β, and γ are known as the anisotropy constants. Typically $\alpha < 0$ and $\gamma > 0$; if $\gamma = 0$, then $\beta > 0$. This ensures that for $E = 0$, G_F has two minima, one with $P < 0$ and one with $P > 0$, as illustrated in Fig. 3.5. The spontaneous polarization of a ferroelectric exhibits hysteresis: it depends on the electric fields that were previously applied to it.

Capacitors can be constructed with dielectrics and ferroelectrics alike; the differential capacitance of a capacitor is given by

$$C = \left(\frac{\partial^2 U}{\partial Q^2}\right)^{-1}, \tag{3.12}$$

where U is the energy stored in the capacitor and Q is the charge stored on its plates. Since $U \propto G_{D,F}$ and $Q \propto P$, dielectric capacitors always have positive capacitance, while ferroelectric capacitors can exhibit transient, negative capacitance, as can be seen from the negative curvature of G_F at $P = 0$ in Fig. 3.5. Negative capacitance has been experimentally demonstrated in ferroelectric capacitors from direct measurements of polarization density as a function voltage applied across the capacitor [130]. Furthermore, negative capacitors have the unusual property that, when connected in series with a dielectric (positive) capacitor, the resulting equivalent capacitance is larger than that of the dielectric capacitor; this was experimentally observed for dielectric–ferroelectric nanoscale heterostructures [131] and superlattices [132].

Plugging in a negative gate capacitance $C_G < 0$ such that $|C_G| > C_P$ in Eq. 3.8 results in a body factor $\left|\frac{\partial \phi}{\partial V_{GS}}\right| > q$, in contrast to Gauss's tyranny [Eq. 3.9]. This inspired Salahuddin and Datta to introduce the negative-capacitance FET in 2007 [133, 134]. A typical NCFET is shown in Fig. 3.6. Its structure is identical to that of a MOSFET, with the exception

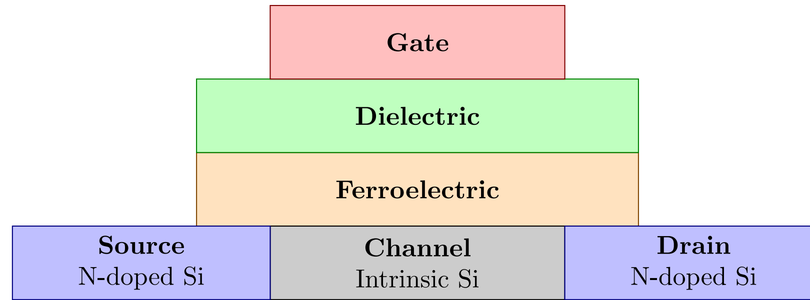


Figure 3.6 – Schematic of a typical n-type enhancement-mode silicon negative-capacitance field-effect transistor (NCFET). The source is n-doped, the channel is intrinsic, and the drain is n-doped. A ferroelectric layer is inserted in the gate stack between the channel and the dielectric, which enables amplification of the gate voltage. The main charge transport mechanism is thermionic emission.

that a ferroelectric layer is inserted between the silicon channel and the dielectric layer. The ferroelectric layer provides gate-voltage amplification, while the dielectric layer stabilizes the ferroelectric layer's transient, negative-capacitance state, and reduces hysteresis, to an extent. Popular ferroelectric materials for NCFET applications include perovskites such as $P(Zr_{1-x}Ti_x)O_3$ (PZT) [135] and doped hafnium dioxides such as $Hf_{0.5}Zr_{0.5}O_3$ (HZO) [136,137]. Hafnium dioxide has been part of semiconductor-processing technology for decades, thereby favoring HZO for CMOS compatibility. Furthermore, the ferroelectric layer thickness in HZO-based NCFETs can be scaled below 10 nm while maintaining significant gate-voltage amplification [138], which is typically not possible for NCFET based on other ferroelectric materials. This makes HZO a prime ferroelectric candidate for NCFETs.

The insertion of the ferroelectric layer in the gate stack does not affect the physics of charge transport in the semiconductor channel. Consequently, the main charge transport mechanism in NCFETs is thermionic emission; NCFETs can thus, in principle, reach MOSFET-like on-state current at low power supply voltage. One of the main drawbacks of NCFETs is the significant hysteresis [Fig. A.12] that they typically exhibit for clock frequencies $f \gtrapprox 100$ kHz, which makes NCFETs unsuitable for operation at GHz frequencies. NCFETs generally suffer from a trade-off between hysteresis and STS: a thicker (thinner) ferroelectric layer leads to higher (lower) hysteresis and lower (higher) STS. On one end of this spectrum, in 2017, Ko, Lee, and Shin experimentally realized a PZT-based tri-gate silicon NCFET with average

STS below 20 mV$\cdot$ dec^{-1} at room temperature and hysteresis greater than $\Delta V_{GS} = 0.48$ V at drain voltage $V_{DS} = 0.1$ V (and unspecified voltage sweep frequency) [139]. On the other end of this spectrum, in 2019, Kwon *et al.* experimentally realized a PZT-based single-gated silicon NCFET with essentially no hysteresis at $f = 1$ MHz and average STS of 63 mV$\cdot$ dec^{-1} at room temperature, which is not lower than the 60 mV$\cdot$ dec^{-1} limit of MOSFETs, but was shown to be 4 mV$\cdot$ dec^{-1} lower than a MOSFET with comparable material composition and geometry; the drain voltage was set to $V_{DS} = 0.1$ V [140]. Both of these NCFETs were shown to have very high on-state current $I_{\text{on}} \gtrapprox 10^3$ A$\cdot$ m^{-1}.

The NCFET is arguably the second most studied steep-slope FET, after the TFET. Much like for the TFET, the progress of research toward low-power electronics based on NCFETs has been slow; indeed, the NCFET has not had commercial applications yet [141]. Furthermore, serious doubts have been cast on the NCFET concept since its inception. Some reports question the validity of key assumptions that are required to establish the idea that a ferroelectric's negative-capacitance state can be stabilized by a dielectric capacitor [142–144]. Other reports question whether negative capacitance could be applied or could provide any benefit to the ultra-scaled transistors of tomorrow [145, 146]. Furthermore, an internal metal gate is most often inserted between the dielectric and ferroelectric layers of experimentally realized NCFETs; this gate is used to measure the gate stack's internal voltage in research reports. However, NCFETs do not require such an internal gate, in principle. It has been suggested that the steep slope observed in such NCFETs could be artefacts of the internal metal gate, the gate voltage scan rate, and the measurement setup in general; indeed, transistors with a metal–dielectric–metal–dielectric–semiconductor gate stack (without any ferroelectric layer) have been shown to have sub-$\frac{k_B T \ln 10}{q}$ STS [147]. There thus appears to be a gap in our theoretical understanding of NCFETs that needs to be filled for the field to develop.

3.3 Discussion and Comparisons

In this section, I share some of my thoughts and opinions about steep-slope FETs. Numerous steep-slope FETs are investigated in this chapter and in Appendix A. They do not form an exhaustive list of steep-slope FETs; rather, they are some of the most prominent candidates. A type of steep-slope FET that is not discussed is the phase-change FET, which contains a material or structure that can sharply transition between low-resistance and high-resistance phases, as can be achieved through the formation of current filaments in metal–insulator–metal structures [148, 149]. In another recently-proposed steep-slope FET, the channel bandgap is modulated via an electric-field-induced topological phase transition, as can be achieved in quantum spin Hall materials with honeycomb lattices [150].

Broadly speaking, a steep slope can be achieved in an FET using one (or more) of the three following methods.

Energy filtering: Unlike the MOSFET, the TFET, DSFET, CSFET, and SLFET have fewer conduction modes in the off state than in the on state; this is a form of bandstructure engineering. This heavily limits these devices' off-state current and enables them to surpass Boltzmann's tyranny. The principle of energy filtering is powerful: in principle, a material or structure with a Fermi level lying close (within a few hundreds of meV or less) to a band edge can be used as the injection source of an FET to achieve sub-$\frac{k_B T \ln 10}{q}$ STS. More generally, a material or structure with a Fermi level lying close to a point where the DOS sharply decreases (or increases) can be used to achieve a steep slope; the sharper the decrease (or increase) in DOS, the lower the STS. To be practically useful, such a material or structure would further need to form an Ohmic contact with the source contact metal and a suitable band alignment with the channel semiconductor.

Positive feedback: The IIFET and FBFET achieve low STS through positive feedback: the voltages applied to these FETs tune the onset of a chain reaction, i.e. a self-amplifying chain of events. In an IIFET, impact–ionized charge carriers impact–ionize other carriers; in

an FBFET, thermionically injected charge carriers lower the thermionic injection barrier for other carriers. In principle, other positive-feedback mechanisms could be harnessed to design steep-slope FETs. Ideally, such a mechanism could arise in semiconductor structures. Both the IIFET and FBFET require impractically large power supply voltage to achieve low STS. This is because the positive-feedback mechanisms in these devices couple the valence and conduction bands; voltages larger than the bandgap are therefore required. This issue could, perhaps, be resolved with "intra-band positive feedback," specifically, a positive-feedback mechanism that occurs in a narrow energy window, within a single band.

Dual-purpose gate: In a MOSFET, the gate has a singular purpose: to modulate the channel conductivity through the field effect. In the NCFET, NEMFET, and PEFET, the gate has an additional purpose. In an NCFET, it flips the polarization state of the ferroelectric layer, thereby enabling negative capacitance and gate-voltage amplification. In a NEMFET, it physically moves to modulate the gate capacitance. In a PEFET, it strains the piezoelectric layer and semiconductor channel, thereby modulating the semiconductor's bandgap and electron affinity. Electric fields and voltages have legions of effects on matter and structures. An effect that modulates the conductivity of a semiconductor could be used to design a dual-purpose-gate FET with a steep slope.

Steep-slope FETs form an extraordinary garden rich in ingenious ideas stemming from physics and engineering. It is likely to expand in the future following these three principles, perhaps most straightforwardly with the principle of energy filtering.

Modern MOSFETs on ICs have STS roughly between 72 and 82 $\text{mV}\cdot\text{dec}^{-1}$. The IRDS™predicts that, by 2034, this range will drop to 65 to 75 $\text{mV}\cdot\text{dec}^{-1}$ by replacing tri-gate devices (FinFETs) by GAA devices [14]. In the MOSFET paradigm, further scaling of the STS will be difficult beyond that point. It can therefore be expected that the semiconductor industry will be compelled to adopt a steep-slope FET starting in around 15 to 20 years. This steep-slope FET has perhaps already been invented. Table 3.1 compares the FETs reviewed in this chapter and in Appendix A according to several attributes that may enable their future adoption by the industry.

The *raison d'être* of a steep-slope FET is to reduce the power supply voltage V_{DD} of processors since their power consumption is proportional to V_{DD}^2 [Eq. 2.33]. All the steep-slope FETs reviewed in this thesis are compatible with low-V_{DD} operation, except the IIFET and FBFET. These devices both require $V_{DD} \gtrapprox \frac{E_G}{q}$, where E_G is the semiconductor bandgap energy. This is because their low STS is contingent on inter-band positive-feedback mechanisms. The adoption of these devices is therefore conditional on the replacement of silicon by a narrow-bandgap semiconductor or on novel positive-feedback physics (such as ballistic impact–ionization for the IIFET [151, 152]).

Ambipolarity, namely, high conductivity at negative gate voltage and positive drain voltage (for n-type enhancement-mode devices), hinders the use of FETs in certain circuits, such as inverters. Ambipolarity is a significant issue for TFETs [Fig. 3.4], but can be alleviated by orders of magnitude (in terms of ambipolar leakage current) using various techniques [122].

Like the MOSFET, the NCFET and PEFET both have roughly constant STS in the off state. This is a convenient property that enables smooth tuning of the off-state current (and therefore, leakage power [Eq. 2.32]) but is not strictly required to lower the power consumption of processors.

In principle, all steep-slope FETs reviewed in this thesis exhibit low STS both in the diffusive (long-channel) and ballistic (short-channel) limits, except the CSFET. Indeed, scattering-induced rethermalization may lead to significant thermionic leakage in a CS-FET [Fig. A.6]. In practice, CSFET-based processors may require long-mean-free-path semiconductors (e.g. gallium arsenide) and devices with extremely small physical features.

Usually, FETs are three-terminal devices: they have a gate contact, a source contact, and a drain contact. The DSFET requires an additional control gate to modulate electrostatic doping in its graphene source. The PEFET requires two gates in its gate stack to achieve low STS [153]. Finally, the NCFET may require an internal gate in its gate stack to achieve low STS, although this is the object of some controversy [147]. These additional gates may be frequently charged and discharged, leading to additional power dissipations, and present additional fabrication complexity.

The FBFET requires a preconditioning step, specifically, the trapping of electrons and holes in its gate dielectric, to achieve low STS. Very high voltages need to be applied to trap charges, which may result in a power consumption burden. Furthermore, it is not clear what is the retention time of these trapped charges.

The FBFET, NCFET, and NEMFET all exhibit significant hysteresis [Fig. A.12], which hinders the operation of logic gates, and in extreme cases, may limit the scaling of the power supply voltage. Hysteresis can be significantly lowered in NCFETs [140]. However, this has not been demonstrated at GHz frequencies. Therefore, these devices may not be suitable for high-performance computing.

The NEMFET and PEFET both have moving parts (the gate, for the NEMFET, and the piezoelectric layer, for the PEFET). This causes significant mechanical stability issues, especially for the NEMFET, which may strongly limit device longevity.

Except for the DSFET, all the reviewed devices can be built with silicon. This even includes the SLFET, which is usually based on a III–V-semiconductor superlattice; indeed, an SLFET based on a silicon geometric superlattice has been proposed [154]. On the other hand, the DSFET requires a graphene source to achieve low STS. Silicon is the material of choice of the semiconductor industry, whereas graphene may prove difficult to mass-produce.

Importantly, a low STS has been experimentally demonstrated for all reviewed devices except the CSFET, SLFET, and PEFET. Without an unambiguous, real-world proof, the physics, models, and simulations that were developed for these devices remain subject to inaccuracies and oversights. The NCFET has had numerous experimental demonstrations, but they also raised some doubts [142–144, 147].

Finally, a sufficiently high on-state current $I_{\mathrm{on}} \gtrapprox 10^3\ \mathrm{A \cdot m^{-1}}$ is required for short loaded time delay [Eq. 2.26] and high clock frequency [Eq. 2.27]; the IRDS™predicts that this stringent requirement on I_{on} will last at least until the year 2034 [14]. Such a high I_{on} has been experimentally demonstrated in the IIFET and NCFET only. On-state current is strongly limited by tunnelling in the TFET and CSFET, and by low source DOS in the DSFET;

it is therefore unlikely that these devices will ever attain $I_{\mathrm{on}} \gtrapprox 10^3\ \mathrm{A \cdot m^{-1}}$. The SLFET, FBFET, NEMFET, and PEFET could possibly attain such high I_{on} after optimization. The geometrical downscaling of load capacitors as well as the downscaling of V_{DD}, which are expected in future generations of processors, may put less stringent requirements on I_{on}. It is possible that $I_{\mathrm{on}} \gtrapprox 10^2\ \mathrm{A \cdot m^{-1}}$ may be sufficient for future generations of processors, especially for applications in low-power computing; the TFET and DSFET both have had experimental demonstrations of I_{on} within or close to this range.

It remains to determine the "best" steep-slope FET. In terms of number of checkmarks in Table 3.1, the NCFET "wins" with a score of 10 (if controversies are ignored), although its hysteresis at high frequencies may be a roadblock to adoption in high-performance computing. Then comes the IIFET with a score of 9, but its V_{DD} scaling issues make its low STS moot. The TFET and SLFET follow with a score of 8; however, the SLFET remains, for now, theoretical. The experimentally demonstrated DSFET follows with a score of 7, which is noteworthy, considering how recently it was introduced (in 2018). Further optimization may make the DSFET an even more promising candidate. In terms of interest garnered by the research community, the TFET and NCFET are the clear winners. The TFET has already been shown to outperform the MOSFET at $V_{DD} = 0.3$ V in experiments [126]. The main barrier to adoption of the TFET is its low on-state current, although exceptional progress has been made in recent years [127]. In Chapter 5, I will introduce bound-charge engineering, a novel method of manipulating surface bound charges that can be used to increase the on-state current of TFETs.

FET attribute	FETs with large transport factors						FETs with large body factors		
	TFET	**DSFET**	**CSFET**	**SLFET**	**IIFET**	**FBFET**	**NCFET**	**NEMFET**	**PEFET**
Compatible with low-V_{DD} operation?	✓	✓	✓	✓			✓	✓	✓
No ambipolarity?		✓	✓	✓	✓	✓	✓	✓	✓
Constant STS in the off state?							✓		✓
Compatible with non-ballistic operation?	✓	✓		✓	✓	✓	✓	✓	✓
Does not require control/internal gate?	✓		✓	✓	✓	✓	?	✓	
Does not require preconditioning?	✓	✓	✓	✓	✓		✓	✓	✓
No hysteresis at GHz frequency?	✓	✓	✓	✓	✓				✓
No moving parts?	✓	✓	✓	✓	✓	✓	✓		
Can be built with silicon?	✓		✓	✓	✓	✓	✓	✓	✓
Experimentally demonstrated?	✓	✓			✓	✓	?	✓	
$I_{\mathrm{on}} \gtrapprox 10^3\ \mathrm{A\cdot m^{-1}}$ in experiment?					✓		✓		

Table 3.1 – Comparison of several steep-slope FETs according to characteristics of relevance to their potential adoption by the semiconductor industry. A checkmark (✓) indicates a "yes," a blank space indicates a "no," and a question mark (?) indicates some level of controversy in the research community. Here, an "experimental demonstration" refers to the fabrication of a real-world device with sub-$\frac{k_B T \ln 10}{q}$ STS. The power supply voltage and the on-state current are denoted by V_{DD} and I_{on}, respectively.

3.4 Summary

The subthreshold slope, namely, the reciprocal of the subthreshold swing (STS), of a field-effect transistor (FET) is given by

$$\frac{1}{S} = \underbrace{\left(\frac{\partial \log_{10} I_{DS}}{\partial \phi}\right)}_{\text{Transport factor}} \overbrace{\left(\frac{\partial \phi}{\partial V_{GS}}\right)}^{\text{Body factor}}, \tag{3.13}$$

where I_{DS} is the drain current, ϕ is the channel barrier energy, and V_{GS} is the gate voltage. For a metal–oxide–semiconductor field-effect transistor (MOSFET), $S = \frac{k_B T \ln 10}{q}$, where T is the device temperature.

There are three types of steep-slope FETs, i.e. FETs with sub-$\frac{k_B T \ln 10}{q}$ STS.

The first is FETs with large transport factors. Among such FETs, one may distinguish between energy-filtering FETs and positive-feedback FETs. Energy-filtering FETs have fewer conduction modes for source-injected charge carriers in the off state than in the on state, which reduces off-state leakage current and STS. Energy filtering can be achieved using a bandgap in the source, as in the tunnel FET (TFET), the cold-source FET (CSFET), and the superlattice FET (SLFET). It can also be achieved using the low DOS around the Dirac point in graphene, as in a Dirac-source FET (DSFET). In positive-feedback FETs, the onset of a positive-feedback mechanism is tuned by the applied voltages. In the impact–ionization FET (IIFET), this mechanism is impact–ionization avalanche breakdown. In the feedback FET (FBFET), this mechanism is a charge–potential band modulation mechanism.

The second is FETs with large body factors. In such FETs, the gate modulates the semiconductor channel not only through the field effect but also through a secondary mechanism. In a negative-capacitance FET (NCFET), an applied gate voltage flips the polarization state of a ferroelectric layer in the gate stack, which enables a transient, negative-capacitance state, thereby amplifying the effect of the gate voltage on internal device electrostatics. In a

nanoelectromechanical FET (NEMFET), a suspended gate physically moves in response to an applied gate voltage, thereby modulating the gate capacitance from being small in the off state to high in the on state. In a piezoelectric FET (PEFET), an applied gate voltage strains a piezoelectric layer in the gate stack, which, in turn, strains the semiconductor channel, thereby narrowing the bandgap and reducing the electron affinity.

The third is “hybrid” FETs with large transport and body factors. Such FETs include the negative-capacitance tunnel FET and the piezoelectric tunnel FET.

The semiconductor industry may be compelled to adopt a steep-slope FET in around 15 to 20 years; indeed, around that time, further scaling of the STS and power supply voltage V_{DD} in MOSFET-based processors will be difficult. None of the steep-slope FETs investigated in this thesis are perfect candidates. Indeed, some of them suffer from V_{DD} scaling issues (IIFET and FBFET). One of them suffers from ambipolarity (TFET). Some of them require ballistic operation (CSFET), control/internal gates (DSFET, PEFET, and possibly NCFET), or preconditioning (FBFET). Some of them suffer from significant hysteresis at high clock frequency (FBFET, NCFET, and NEMFET). Some of them have moving parts or issues pertaining to mechanical stability (NEMFET and PEFET). One cannot be built with silicon (DSFET). Some were not experimentally demonstrated (CSFET, SLFET, PEFET, and possibly NCFET). Finally, among those that were experimentally demonstrated, some do not have on-state current $I_{\text{on}} \gtrapprox 10^3\ \text{A}\cdot\text{m}^{-1}$ (TFET, DSFET, FBFET, and NEMFET), which is required for a clock frequency in the GHz range.

Despite low I_{on} and ambipolarity, the TFET is a notable candidate due to (1) the extensive amount of research that has been conducted on it, (2) its experimentally demonstrated outperformance of MOSFETs at low V_{DD} [126], (3) recent progress to increase I_{on} [127], and (4) numerous advances to reduce ambipolarity [122].

This thesis introduces bound-charge engineering, which can be used to further increase the on-state current of TFETs. This is substantiated by quantum simulations of TFETs; the simulation scheme is introduced in the next chapter.

Chapter 4

Simulations of Semiconductor Devices

One shouldn't work on semiconductors, it's a mess; who knows whether semiconductors even exist.

— W. E. Pauli, 1931 [155]

Despite the ubiquity of semiconductor devices in the modern world, semiconductors were still poorly understood just a few generations ago. Experimental observations were sometimes not reproducible and their agreement with theoretical models was often poor. In a 1931 letter, Wolfgang Pauli, one of the founders of quantum mechanics, advised his student Rudolf Peierls against working on semiconductors, going as far as questioning their very existence. Retrospectively, these troubled beginnings may have been caused by the acute sensitivity to impurities of semiconductors. Indeed, some impurities can dope a semiconductor, increasing its conductivity by orders of magnitude for impurity concentrations as small as one part per billion. Other impurities form deep-level traps, rendering the semiconductor insulating. It goes without saying that our theoretical understanding of semiconductors has tremendously improved since the 1930s, starting with Felix Bloch's pioneering work on band theory [156]. Today, semiconductor device simulation and modelling is an essential part of the development of future generations of transistors and processors; several companies in the S&P 500 specialize in such tools for technology computer-aided design (TCAD). One such TCAD tool is the atomistic quantum transport package `Nanoskim 2.0` [157,158], which self-consistently combines the nonequilibrium Green's function (NEGF) formalism and the

tight-binding (TB) model.

This chapter describes the NEGF–TB simulation methodology used in this thesis, as implemented in `Nanoskim 2.0`. A broad overview of TCAD simulation paradigms is given in Sec. 4.1. The TB model is described in Sec. 4.2, the NEGF formalism is described in Sec. 4.3, and their self-consistent combination is described in Sec. 4.4. Complete and self-contained derivations and descriptions of the TB model and the NEGF formalism are beyond the scope of this thesis. Instead, in this chapter, the basic ideas of these theories and the fundamental assumptions required to derive them are reviewed.

4.1 From Many-Body Physics to TCAD

In its most general form, the many-body Hamiltonian of a system of interacting electrons and atomic nuclei is given by

$$\begin{aligned}\hat{\mathbb{H}} = &\sum_{i=1}^{N_e} \frac{\hat{\mathbf{p}}_i^2}{2m_0} + \frac{q^2}{4\pi\varepsilon_0} \sum_{i=1}^{N_e} \sum_{j>i}^{N_e} \frac{1}{|\hat{\mathbf{r}}_i - \hat{\mathbf{r}}_j|} \\ &+ \sum_{I=1}^{N_n} \frac{\hat{\mathbf{P}}_I^2}{2M_I} + \frac{q^2}{4\pi\varepsilon_0} \sum_{I=1}^{N_n} \sum_{J>I}^{N_n} \frac{Z_I Z_J}{\left|\hat{\mathbf{R}}_I - \hat{\mathbf{R}}_J\right|} \\ &- \frac{q^2}{4\pi\varepsilon_0} \sum_{i=1}^{N_e} \sum_{I=1}^{N_n} \frac{Z_I}{\left|\hat{\mathbf{r}}_i - \hat{\mathbf{R}}_I\right|}\,, \end{aligned} \tag{4.1}$$

where N_e and N_n are, respectively, the numbers of electrons and nuclei, m_0 is the free electron (rest) mass, M_I and Z_I are, respectively, the mass and atomic number of nucleus I, $\hat{\mathbf{r}}_i$ and $\hat{\mathbf{R}}_I$ are, respectively, the position operators of electron i and nucleus I, and where $\hat{\mathbf{p}}_i$ and $\hat{\mathbf{P}}_I$ are, respectively, the momentum operators of electron i and nucleus I. The first line in Eq. 4.1 represents the kinetic energy of electrons and their Coulombic repulsion. The second line represents the kinetic energy of nuclei and their Coulombic repulsion. The third line represents the Coulombic attraction between electrons and nuclei. To obtain the dynamics of

a system of electrons and nuclei, the many-body Schödinger equation must be solved:

$$i\hbar\frac{d}{dt}\left|\chi\left(t\right)\right\rangle = \hat{\mathbb{H}}\left|\chi\left(t\right)\right\rangle \ , \tag{4.2}$$

where t is time and $\left|\chi\left(t\right)\right\rangle$ is the many-body quantum state for electrons and nuclei; its real-space representation is $3\left(N_e + N_n\right)$-dimensional. In this context, the computational complexity of Schrödinger's equation is (at least) $\Theta\left(\left(N_e + N_n\right)^3\right)$. Due to its immense complexity, Eq. 4.2 can only be solved for extremely small systems, such as small atoms or simple molecules. Practical simulations of semiconductor devices require simplifications and/or approximations, some examples of which are outlined below.

Simulating a semiconductor device typically requires four ingredients.

The first ingredient is a basis (or bases) on which the device and the charge carriers therein are discretized. Although the device and its various properties are most often continuous, the equations that describe the state and evolution of a device can rarely be solved analytically. To be solved numerically, they need to be discretized, that is, device properties need to be expressed as linear combinations of basis vectors. A basis should be small enough to avoid significant computational burden but large enough to capture device properties with high accuracy; choosing an optimal basis is, to some extent, more of an art than a science. Common bases include real space, momentum space, and atomic orbitals [42].

The second ingredient is Poisson's equation,

$$\boldsymbol{\nabla}\cdot\left(\varepsilon\boldsymbol{\nabla}V\right) = -\rho_f \ , \tag{4.3}$$

which can be solved to obtain the electric potential V in a system as a function of free charge density ρ_f and permittivity ε [47]. It is arguably the simplest of the four ingredients. Indeed, Poisson's equation is valid and immutable over lengthscales ranging from light-years down to the Compton wavelength of the electron (about 10^{-12} m), in principle [159]. Thus, unlike other ingredients, Poisson's equation presents little burden of choice: it is the unique equation to relate potential and charge. Numerically, it is commonly solved using Gaussian

elimination [160], the method of image charges [47], and multipole expansions [161].

The third ingredient is a formalism that captures the electronic, mechanical, and/or optical properties of the materials within the device (for instance, their bandstructures and charge carrier mobilities). There is a broad range of material physics formalisms, which spans orders of magnitude of computational complexity.

Drift–diffusion model: In this ubiquitous transistor model, electrical current is expressed as a sum of drift and diffusion currents, both of which are parametrized in terms of charge carrier mobilities [16].

Effective mass approximation: In this approximation, the bandstructures of semiconductors are approximated as parabolic at band extrema, being parametrized solely in terms of effective masses (EMs) (which may be anisotropic) and possibly bandgap energies [162]. The EM approximation generally accurately captures low-energy excitations but not high-energy excitations, since it fails far from band extrema due to band nonparabolicity [163, 164].

$\mathbf{k} \cdot \mathbf{p}$ perturbation theory: In this model, the Schrödinger equation is solved perturbatively [165]; the perturbed Hamiltonian is proportional to $\mathbf{k} \cdot \mathbf{p}$, the scalar product of crystal momentum and real momentum. The elements of the perturbative expansion, known as "optical matrix elements," can be fitted from experimental measurements or more elaborate simulations [166].

Tight-binding model: In this model, the quantum states of electrons are approximated as linear combinations of atomic-like orbitals. The Hamiltonian is expressed in terms of two types of parameters: on-site integrals, which describe the energy costs associated with having orbitals filled, and hopping (or two-centre) integrals, which describe the probabilities of electrons tunnelling between nearby orbitals [42].

Density functional theory: In this model, thanks to the Hohenberg—Kohn theorems [167], the $3N$-dimensional problem of solving the many-body interacting-electrons Schödinger equation (where N is the number of electrons) is reduced, in principle without loss of accuracy, to a 3-dimensional problem consisting of minimizing an energy functional with

respect to electron density. This functional minimization can be further simplified to a set of single-electron Schrödinger equations, the Kohn–Sham equations [168, 169].

These formalisms fill different needs in semiconductor device modelling; they are often used in conjunction with each other. Density functional theory (DFT) is the only tool in this list capable of predicting material properties from first principles (*ab initio*). Indeed, DFT provides an atomistic description of matter and structures; in other words, all individual atoms, their quantum properties, and their interactions with each other are explicitly considered. Furthermore, the various parameters required to perform a DFT simulation (notably, the parameters of the exchange–correlation functional [170] and of the pseudopotentials [171]) are, for the most part, material-independent. As a result, barring experimental measurements, DFT is the only method to extract the TB Hamiltonian parameters, the EMs, and the mobilities needed in other listed formalisms. DFT is needed to simulate novel materials and structures but is computationally burdensome. For well-studied materials, for which parameters have been extracted, the TB model offers great accuracy, an atomistic description of matter, and captures some many-body effects. $\mathbf{k} \cdot \mathbf{p}$ perturbation theory captures some quantum effects but ignores atomistic effects. The EM approximation and the drift–diffusion (DD) model are the least computationally taxing but ignore the atomistic nature of matter and only roughly approximate bandstructures.

The final ingredient is a formalism that captures the statistics and transport properties of the device, such as its nonequilibrium response (notably, electrical current) to external stimuli (notably, voltages). Specifically, the previously described third ingredient provides the allowed energy states of a system; this final ingredient describes which of these states are occupied and the extent to which they contribute to nonequilibrium responses. Three main transport formalisms can be distinguished.

Local equilibrium statistics: At equilibrium, matter is described by Fermi–Dirac statistics (in the case of fermions, such as electrons) or Bose–Einstein statistics (in the case of bosons, such as phonons): energy states are populated according to well-defined functions of their energies, a unique temperature, and a unique chemical potential. Both the Fermi–Dirac and

Bose–Einstein statistics converge to Boltzmann statistics in the classical (high-temperature) limit [172]. Under nonequilibrium conditions, temperature and chemical potential are generally not well nor uniquely defined. A common approximation to nonequilibrium statistics consists of local equilibrium statistics, namely, equilibrium statistics with position-dependent temperature and chemical potential (sometimes known as quasi Fermi level). For example, the DD model assumes local Boltzmann statistics.

Boltzmann transport equation: Classically, the positions and velocities of particles in an ideal gas are described by the Maxwell–Boltzmann distribution [172]. External forces (notably, electromagnetic forces) and interactions between particles alter this distribution. The resulting nonequilibrium classical distribution may be computed from the Boltzmann transport equation (BTE), one of the most general approaches to classical transport [42].

Nonequilibrium Green's function formalism: Quantum systems distinguish themselves from classical systems in several ways, including by the operator nature of physical observables (and non-commutativity of many pairs of operators) and the Pauli exclusion principle (for fermions). This makes the BTE fundamentally inadequate to tackle quantum systems. The NEGF formalism is a general description of nonequilibrium quantum statistics. The main mathematical object of the NEGF formalism is the NEGF; physical observables, such as current, are expressed in terms of NEGFs [173, 174].

The NEGF formalism is the most computationally burdensome among the listed transport formalisms. This is because its numerical implementation typically involves the inversion of large matrices [173]. However, it is the only formalism that can accurately capture the various nonequilibrium quantum properties observed in nanotransistors, such as quantum tunnelling [Sec. 3.1] and quantum interferences [175].

Formalisms for material physics and transport physics can be paired in various ways to obtain self-consistent descriptions of semiconductor devices, as illustrated in Fig. 4.1. The material physics formalism provides the allowed states in the device as a function of electric potential. The transport physics formalism determines which of these states are occupied, resulting in expressions for charge density and electrical current. Finally, Poisson's equation

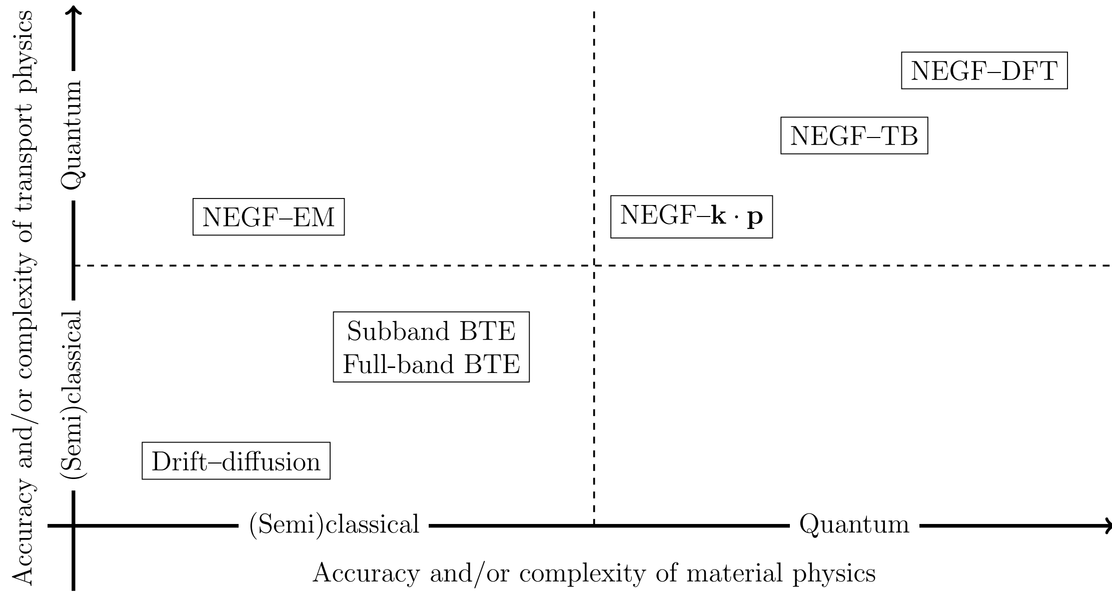


Figure 4.1 – Common formalisms to simulate semiconductor devices graphed in a "phase space" of accuracy and/or computational complexity of material physics and transport physics. They include the drift–diffusion model [16], the full-band Boltzmann transport equation (BTE) [176], the subband BTE [177], the nonequilibrium Green's function (NEGF)–effective mass (EM) approximation [162], the NEGF–$\mathbf{k} \cdot \mathbf{p}$ perturbation theory [178], the NEGF–tight-binding (TB) model [179], and the NEGF–density functional theory (DFT) [180–182].

provides the electric potential as a function of charge density. Self-consistency refers to compatibility between the charge density outputted by the transport formalism and the electric potential outputted by Poisson's equation.

The NEGF–DFT algorithm [180–182] is one of the most powerful, practical, and universal methods to predict the nonlinear and nonequilibrium quantum transport properties of nanoelectronic devices. However, it typically cannot simulate systems with more than a few hundreds of atoms, even on supercomputers [183]. On the other hand, the NEGF–TB algorithm [179] provides comparable accuracy (assuming TB parameters are known) at a fraction of the computational cost. Systems with hundreds of thousands of atoms can be simulated in NEGF–TB [158, 184]. The nanotransistors simulated in this thesis are composed of silicon, a material for which TB parameters are known [185]. In this context, the NEGF–TB algorithm offers an optimal balance between computational burden and accuracy: matter is described atomistically and transport is described quantum mechanically at a fraction of the

cost of NEGF–DFT simulations.

4.2 The Tight-Binding Model

4.2.1 The Slater–Koster Hamiltonian

The TB model finds its roots in the method of linear combination of atomic orbitals (LCAO), initially developed by Bloch in 1929 [156]. In its modern form, the TB model is based on a simplified LCAO Hamiltonian parametrization developed by Slater and Koster in 1954 [186]. The resulting Hamiltonian, the Slater–Koster (SK) Hamiltonian, has adjustable parameters fitted to interpolate the predictions of *ab initio* simulations (most often, DFT simulations). While the SK Hamiltonian is a single-electron Hamiltonian, it retains symmetries inherited from the crystal lattice as well as quantum mechanical rigour, making it an accurate and physically transparent atomistic descriptor of materials. How can the many-body Hamiltonian in Eq. 4.1, which describes the motion and interactions of electrons and nuclei, be simplified to the single-electron SK Hamiltonian, which is typically parametrized in terms of less than twenty constants?

The first step is the Born–Oppenheimer approximation, which consists of decoupling the electronic motion from the nuclear motion [187]. The mass of an electron m_0 is at least 10^3 (and more typically, about 10^4) times smaller than that of an atomic nucleus M_I. As a result, electrons move orders of magnitude faster than nuclei. Dynamically, electrons adiabatically follow the motion of the slow nuclei; the second line in Eq. 4.1, i.e. the nuclear Hamiltonian, is thus approximately constant over the timescales over which electronic states evolve. Meanwhile, the nuclei respond to a mean-field electron potential. The many-body quantum state $|\chi(t)\rangle$ in Eq. 4.2 can thus be approximated as the tensor product of an electronic state $|\chi_e(t)\rangle$ and a nuclear state $|\chi_n(t)\rangle$: $|\chi(t)\rangle = |\chi_e(t)\rangle \otimes |\chi_n(t)\rangle$. Thus, in the Born–Oppenheimer approximation, the electronic state evolves according to the following

many-electron Schrödinger equation:

$$i\hbar\frac{d}{dt}\left|\chi_e\left(t\right)\right\rangle = \hat{\mathbb{H}}_e\left|\chi_e\left(t\right)\right\rangle\ , \tag{4.4}$$

where

$$\hat{\mathbb{H}}_e = \sum_{i=1}^{N_e}\frac{\hat{\mathbf{p}}_i^2}{2m_0} + \frac{q^2}{4\pi\varepsilon_0}\sum_{i=1}^{N_e}\sum_{j>i}^{N_e}\frac{1}{|\hat{\mathbf{r}}_i - \hat{\mathbf{r}}_j|} - \frac{q^2}{4\pi\varepsilon_0}\sum_{i=1}^{N_e}\sum_{I=1}^{N_n}\frac{Z_I}{|\hat{\mathbf{r}}_i - \mathbf{R}_I|}\ , \tag{4.5}$$

where, importantly, the nuclear positions $\mathbf{R}_I$ were demoted from operators (as in Eq. 4.1) to constants.

The second step is the independent electron approximation, which consists of neglecting electron–electron interactions, as justified by Landau's Fermi liquid theory [59]. Suppose electron–electron interactions could be slowly turned on. The excitations of the non-interacting electron gas (the Fermi gas) would then adiabatically evolve to excitations of the interacting electron gas (the Fermi liquid). For excitations within about k_BT of the Fermi energy (where T is temperature), Lev Landau showed that this evolution satisfies adiabatic continuity, i.e. there is a one-to-one, momentum-conserving, spin-conserving, and charge-conserving correspondence between excitations of the Fermi gas and excitations of the Fermi liquid. An excitation of the Fermi liquid is a quasiparticle, specifically, a linear combination of a bare electron (with weight $\sqrt{Z}$, where $0 < Z < 1$ is known as the quasiparticle weight) and electron–hole excitations. These Landau quasiparticles are non-interacting and have finite but very long lifetimes (which corresponds to the electron–electron scattering rate) compared to typical scattering rates, e.g. electron–phonon scattering rates. Their mass is renormalized to $\widetilde{m} \approx \frac{m_0}{Z}$, but the Landau quasiparticles otherwise have physical properties very similar to those of electrons. For materials with Wigner–Seitz radius $r_S < 3$, $Z \gtrapprox 0.7$ [59]. Since many electronic properties of materials, including most transport properties, are predominantly determined by electrons within about k_BT of the Fermi energy, the many-electron state in Eq. 4.4 can be approximated as the tensor product of Landau quasiparticles states $\left|\zeta_j\left(t\right)\right\rangle$, with $j = 1, 2, \cdots, N_e$: $\left|\chi_e\left(t\right)\right\rangle = \bigotimes_{j=1}^{N_e}\left|\zeta_j\left(t\right)\right\rangle$. The states $\left|\zeta_j\left(t\right)\right\rangle$ evolve independently from

each other according to the following single-particle Schrödinger equation:

$$i\hbar\frac{d}{dt}\left|\zeta_j\left(t\right)\right\rangle = \hat{H}_j\left|\zeta_j\left(t\right)\right\rangle \ , \tag{4.6}$$

where

$$\hat{H}_j = \frac{\hat{\mathbf{p}}_j^2}{2\widetilde{m}} - \frac{q^2}{4\pi\varepsilon_0}\sum_{I=1}^{N_n}\frac{Z_I}{\left|\hat{\mathbf{r}}_j - \mathbf{R}_I\right|} \, . \tag{4.7}$$

Importantly, it should be noted that $\hat{\mathbb{H}}_e = \sum_{j=1}^{N_e}\hat{H}_j$. For simplicity, in the remainder of this chapter, the Landau quasiparticle index j will be dropped; the single-particle Hamiltonian then becomes

$$\hat{H} = \frac{\hat{\mathbf{p}}}{2m_0} - \frac{q^2}{4\pi\varepsilon_0}\sum_{I=1}^{N_n}\frac{Z_I}{\left|\hat{\mathbf{r}} - \mathbf{R}_I\right|} \, , \tag{4.8}$$

where $\hat{\mathbf{r}}$ and $\hat{\mathbf{p}}$ are, respectively, the quasiparticle position and momentum operators. Note that the quasiparticle mass $\widetilde{m}$ has been replaced by the free electron mass m_0, an assumption justified by the very small Wigner–Seitz radii of semiconductors, even in cases of degenerate doping [59]. Thus, for all intents and purposes, $\hat{H}$ can be viewed as a single-electron Hamiltonian. Its real-space representation is

$$\left\langle \mathbf{r}\,\middle|\,\hat{H}\,\middle|\,\mathbf{r}'\right\rangle = \left(-\frac{\hbar^2}{2m_0}\nabla^2 - \frac{q^2}{4\pi\varepsilon_0}\sum_{I=1}^{N_n}\frac{Z_I}{\left|\mathbf{r} - \mathbf{R}_I\right|}\right)\delta^3\left(\mathbf{r} - \mathbf{r}'\right) \, . \tag{4.9}$$

It remains to discretize $\hat{H}$. Consider an atomic lattice of N_c periodically repeating unit cells (i.e. a crystal) with lattice vectors $\mathbf{c}_n$, where $n = 1, 2, \cdots, N_c$. Within each unit cell are N_a atoms displaced by vectors $\mathbf{a}_i$ from the cell's lattice vector, where $i = 1, 2, \cdots, N_a$. The total number of nuclei in the lattice is thus $N_n = N_c N_a$, and their positions are $\mathbf{R}_I = \mathbf{c}_n + \mathbf{a}_i$, where $I = 1, 2, \cdots, N_n$. Under these assumptions, Eq. 4.9 can be expressed as

$$\left\langle \mathbf{r}\,\middle|\,\hat{H}\,\middle|\,\mathbf{r}'\right\rangle = \left(-\frac{\hbar^2}{2m_0}\nabla^2 - \frac{q^2}{4\pi\varepsilon_0}\sum_{n=1}^{N_c}\sum_{i=1}^{N_a}\frac{Z_i}{\left|\mathbf{r} - \mathbf{c}_n - \mathbf{a}_i\right|}\right)\delta^3\left(\mathbf{r} - \mathbf{r}'\right) \, , \tag{4.10}$$

where Z_i is the atomic number of the atom labelled by i in the crystal's unit cell. Associated with each atom i is a set of hydrogenic atomic orbitals $\left|\phi_{i\alpha}\right\rangle$, where α denotes the orbital's quantum numbers (the principal, azimuthal, and magnetic quantum numbers) [165].

The real-space representation of orbital α centred on atom i of unit cell n is denoted by $\langle \mathbf{r} - \mathbf{c}_n - \mathbf{a}_i \,|\, \phi_{i\alpha} \rangle$. In general, the orbitals centred on different atoms are not orthogonal, in which case Löwdin's method [188] is used to generate a set of Löwdin orbitals $|\psi_{i\alpha}\rangle$ such that

$$\int_{-\infty}^{+\infty} \langle \psi_{i\alpha} \,|\, \mathbf{r} - \mathbf{c}_n - \mathbf{a}_i \rangle \langle \mathbf{r} - \mathbf{c}_m - \mathbf{a}_j \,|\, \psi_{j\beta} \rangle \, d^3\mathbf{r} = \delta_{ij}\delta_{nm}\delta_{\alpha\beta} \,. \tag{4.11}$$

Importantly, the symmetry properties of the Löwdin orbital $\langle \mathbf{r} - \mathbf{c}_n - \mathbf{a}_i \,|\, \psi_{i\alpha} \rangle$ are similar to those of the corresponding atomic orbital $\langle \mathbf{r} - \mathbf{c}_n - \mathbf{a}_i \,|\, \phi_{i\alpha} \rangle$. Finally, for each atom i in the unit cell and for each orbital α, the following Bloch sum is defined:

$$\langle \mathbf{r} \,|\, \Psi_{\mathbf{k}i\alpha} \rangle = \frac{1}{\sqrt{N_c}} \sum_{n=1}^{N_c} e^{i\mathbf{k}\cdot\mathbf{c}_n} \langle \mathbf{r} - \mathbf{c}_n - \mathbf{a}_i \,|\, \psi_{i\alpha} \rangle \;, \tag{4.12}$$

where $\mathbf{k}$, the crystal momentum, is a quantum number by Bloch's theorem [156].

In the SK scheme, it is in the basis of the states $|\Psi_{\mathbf{k}i\alpha}\rangle$ that the Hamiltonian $\hat{H}$ of Eq. 4.10 is discretized. This basis block-diagonalizes $\hat{H}$, with each block corresponding to a single value of $\mathbf{k}$. The matrix elements $H_{i\alpha,j\beta}(\mathbf{k}) := \left\langle \Psi_{\mathbf{k}i\alpha} \middle| \hat{H} \middle| \Psi_{\mathbf{k}j\beta} \right\rangle$ of the block of $\hat{H}$ corresponding to momentum $\mathbf{k}$ are given by

$$\frac{1}{N_c} \sum_{n=1}^{N_c} \sum_{m=1}^{N_c} e^{i\mathbf{k}\cdot(\mathbf{c}_m - \mathbf{c}_n)} \iint_{\mathbb{R}^2} \langle \psi_{i\alpha} \,|\, \mathbf{r} - \mathbf{c}_n - \mathbf{a}_i \rangle \left\langle \mathbf{r} \middle| \hat{H} \middle| \mathbf{r}' \right\rangle \langle \mathbf{r}' - \mathbf{c}_m - \mathbf{a}_j \,|\, \psi_{j\beta} \rangle \, d^3\mathbf{r} d^3\mathbf{r}' \,. \tag{4.13}$$

Furthermore, by translational symmetry:

$$\begin{aligned} H_{i\alpha,j\beta}(\mathbf{k}) = \sum_{n=1}^{N_c} e^{i\mathbf{k}\cdot\mathbf{c}_n} \int_{-\infty}^{+\infty} \langle \psi_{i\alpha} \,|\, \mathbf{r} - \mathbf{c}_n - \mathbf{a}_i \rangle \left(-\frac{\hbar^2}{2m_0} \nabla^2 \right) \langle \mathbf{r} - \mathbf{a}_j \,|\, \psi_{j\beta} \rangle \, d^3\mathbf{r} \\ + \sum_{n=1}^{N_c} e^{i\mathbf{k}\cdot\mathbf{c}_n} \int_{-\infty}^{+\infty} \langle \psi_{i\alpha} \,|\, \mathbf{r} - \mathbf{c}_n - \mathbf{a}_i \rangle \left(-\frac{q^2}{4\pi\varepsilon_0} \sum_{n'=1}^{N_c} \sum_{i'=1}^{N_a} \frac{Z_{i'}}{|\mathbf{r} - \mathbf{c}_{n'} - \mathbf{a}_{i'}|} \right) \langle \mathbf{r} - \mathbf{a}_j \,|\, \psi_{j\beta} \rangle \, d^3\mathbf{r} \,. \end{aligned} \tag{4.14}$$

Since the Löwdin orbitals are known, the kinetic term in $H_{i\alpha,j\beta}(\mathbf{k})$ (the term in the first line of Eq. 4.14) can be readily computed. The potential term (the second line of Eq. 4.14) is a sum of integrals, the integrands of which have contributions from (1) the Löwdin orbital

$|\psi_{i\alpha}\rangle$ centred on atom i, (2) the Löwdin orbital $|\psi_{j\beta}\rangle$ centred on atom j, and (3) the Coulomb potential centred on atom i'. There are thus four categories of integrals in the expansion of the potential term:

- **on-site integrals**, corresponding to the case $i = j = i'$;
- **two-centre integrals**, corresponding to the case $i = i'$ or $j = i'$, with $i \neq j$;
- **three-centre integrals**, corresponding to the case $i \neq j \neq i' \neq i$;
- **intra-atomic integrals**, corresponding to the case $i = j \neq i'$.

To simplify the calculation of the SK Hamiltonian matrix elements $H_{i\alpha,j\beta}$, Slater and Koster proposed three approximations.

First, although each atom i has, in principle, infinitely many orbitals $|\phi_{i\alpha}\rangle$, only those energetically close to the Fermi energy, namely valence electron orbitals, contribute to interatomic orbital hybridization. In the case of silicon, early SK TB models considered one s and three p orbitals per atom [186]; the model would then be referred to as an "sp^3 model." Unfortunately, sp^3 models incorrectly predict that silicon has a direct bandgap. To resolve this issue, Vogl, Hjalmarson, and Dow proposed an $sp^3s^\star$ model [189]; $s^\star$ refers to an excited s orbital. At present, ten-orbital $sp^3d^5s^\star$ models are most commonly used [185, 190, 191].

Second, in three-centre integrals, there is only a small overlap between orbitals and the Coulomb potential; consequently, these integrals are significantly smaller than on-site and two-centre integrals. The SK TB model thus only considers on-site and two-centre integrals; this simplification, the two-centre approximation, is certainly not perfect, but generally leads to accurate predictions [192]. The SK model can be refined beyond the two-centre approximation to include three-centre integrals [193] and intra-atomic integrals [194], but these models will not be further considered in this thesis.

Third, the orbitals $\langle \mathbf{r} \,|\, \phi_{i\alpha}\rangle$ of each atom i decay exponentially with the distance r from atom i. It is thus natural to neglect two-centre integrals for which the two centres, the atoms i and j, are separated by more than a certain cutoff distance. Most typically, the cutoff distance is set to be the interatomic distance, in which case the SK TB model is referred

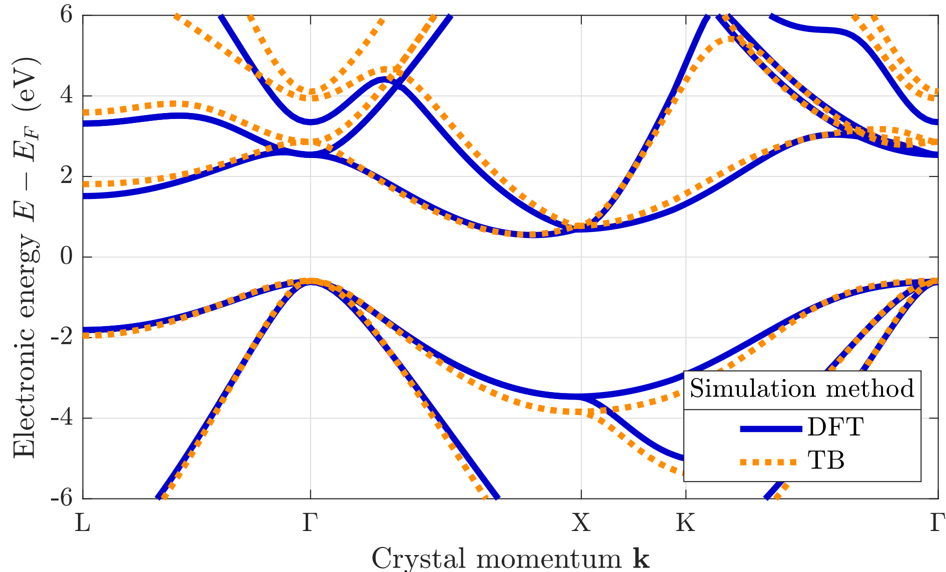


Figure 4.2 – Bandstructure of intrinsic bulk silicon as computed from DFT and the TB model. Electronic energies E (in reference to the Fermi energy E_F) are plotted as a function of crystal momentum $\mathbf{k}$ along high-symmetry lines of the Brillouin zone. The DFT simulation is performed in the Vienna *Ab initio* Simulation Package (VASP) [197–200]. The TB simulation is performed in the Slater–Koster (SK) scheme [Sec. 4.2.1] in a two-centre, nearest neighbour, ten-orbital $sp^3d^5s^\star$ model [185].

to as a "nearest neighbour model." Some SK TB models consider second nearest neighbour integrals [195] and even third nearest neighbour integrals [196], but these models will not be further considered in this thesis.

These approximations greatly simplify the sum in Eq. 4.14. In fact, for zincblende structures, under the two-centre and nearest neighbour approximations, the elements of the ten-orbital $sp^3d^5s^\star$ SK TB Hamiltonian $H_{i\alpha,j\beta}(\mathbf{k})$ can be parametrized in terms of fourteen constants only. Furthermore, for a monoatomic zincblende structure (such as silicon), four pairs of these constants are related by symmetry, resulting in ten independent SK parameters. The derivation and expression of $H_{i\alpha,j\beta}(\mathbf{k})$ in terms of these parameters is algebraically involved and will not be presented in this thesis. Interested readers may consult Refs. [189] and [191].

While the SK Hamiltonian parameters can, in principle, be computed analytically [190], they are not usually extracted in this way; this would lead to relatively poor predictions of

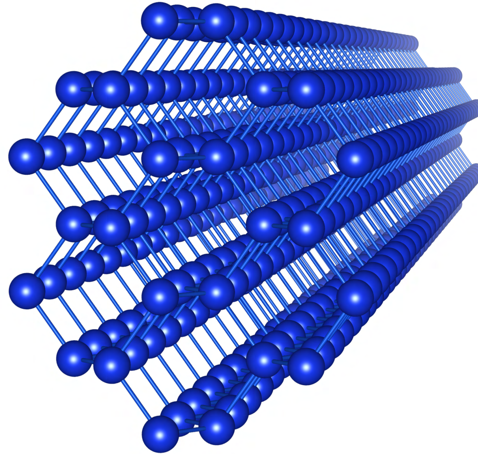


Figure 4.3 – Atomic structure of a silicon nanowire (NW) in the ball-and-stick model. Each sphere represents a silicon atom, while rods represent interatomic bonds. The NW is grown in [110] and has a diameter of 1.16 nm; the interatomic distance is 2.35 Å.

the electronic properties of materials. Rather, the SK parameters are typically extracted by means of a least squares fitting of the SK TB bandstructure to an appropriate DFT bandstructure over a large number of $\mathbf{k}$ points [192]. In Fig. 4.2, bandstructures of silicon, as computed from DFT and from the SK TB model, are compared. The simulations predict a bandgap energy of 1.14 eV, in agreement with experimental data. The SK TB bandstructure matches the DFT bandstructure well 6 eV around the Fermi energy E_F, and almost perfectly 1 eV around E_F. In addition, it correctly predicts that the VBM lies at the Γ point and that the CBM lies along the ΓX line. The SK TB model is thus seen to be a good predictor of the electronic properties of silicon. Thanks to its relatively low computational complexity, it is scalable to systems much larger than what could be simulated with DFT.

It remains to describe the simulation of confined systems, namely, systems that are periodically repeating in less than three dimensions. Examples of great relevance to this thesis are silicon nanowires (NWs), which are periodic along one direction (called the “growth direction” or the “transport direction”) but finite along other directions. Silicon atoms are tetravalent; however, the silicon atoms on the surface of an NW are bonded to less than four other silicon atoms; unpassivated surface silicon atoms would have dangling bonds. Typically, dangling bonds result in surface states within the bandgap, which greatly affect the electronic and transport properties of silicon NWs. In a real-world NW transistor, dangling bonds on

the surface of the NW would typically be passivated by bonds to atoms of neighbouring oxides. This passivation is random and difficult to model, in part due to the typically amorphous nature of oxides [201, 202]. Fortunately, the electronic properties of NWs do not strongly depend on the exact details of the passivation. In TB simulations, dangling bonds are typically dealt with using one of the two following methods:

- by hydrogen passivation, that is, by explicitly adding hydrogen atoms on the surface of the NW;
- by raising the dangling bond energy (which is related to the surface atoms' on-site integrals) until no surface state remains in the bandgap.

The first method suffers from three major drawbacks. First, it requires the calculation of additional parameters, specifically, on-site and two-centre integrals involving hydrogen atoms [203]. Second, it requires a larger Hamiltonian matrix since hydrogen atoms and their orbitals are explicitly considered. Third, it was shown to overestimate transistor performance [202]. Thus, in this thesis, dangling bond energies are raised, following the scheme of Ref. [204]; an energy shift ≥ 5 eV is used.

To summarize, the following approximations were needed to simplify the many-electron, many-nucleus Hamiltonian of Eq. 4.1 to the discrete, single-electron SK TB Hamiltonian used in this thesis:

- the Born–Oppenheimer approximation, to decouple electrons from nuclei;
- the independent electron approximation, to decouple electrons from each other;
- discretization in the basis of relatively few Löwdin orbitals energetically close to the Fermi energy (a ten-orbital $sp^3d^5s^\star$ model is used to model silicon), to simplify the Schödinger equation to a low-dimensional matrix equation;
- the two-centre approximation, to reduce the number of elements of the SK TB Hamiltonian matrix;
- the nearest neighbour approximation, also to reduce the number of elements of the SK TB Hamiltonian matrix;
- least squares fitting of the bandstructure to DFT simulations, to obtain the SK TB

Hamiltonian parameters;

- surface passivation by raising dangling bond energies, to model confined systems.

Having established the SK TB model used in this thesis, it is applied in the next section to analyze some electronic properties of silicon NWs.

4.2.2 Electronic Properties of Silicon Nanowires

The main vehicles for the projects described in this thesis are silicon NWs. Low-dimensional silicon-based materials such as silicon NWs [205] and silicon nanosheets [56] are very likely to be used for beyond-FinFET scaling in future technology nodes [206]. Indeed, these materials are needed for silicon GAA FETs, which have greater immunity against SCEs, higher body factor, and lower STS compared to FinFETs [Chapter 2]. Experimentally, silicon NWs can be realized in various ways; a common method is the vapour–liquid–solid method, where the growth of NWs on a substrate from chemical vapour deposition is stimulated by catalytic liquid droplets [207–209]. The diameter and length of an NW is controllable; the former typically ranges from 1 nm to 100 nm; the latter is typically no greater than a few micrometres. The crystallographic growth direction is also controllable; for silicon NWs, it is typically limited to [100], [110], and [111]. In addition, silicon NWs can be doped *in situ*, while they are grown [210]. Thus, a wealth of FETs and other semiconductor devices can be experimentally realized with silicon NWs.

In Fig. 4.4, the atomic structures of the unit cells of the various silicon NWs investigated in this thesis are shown. NWs with a diameter of 1 nm have as few as 21 atoms per unit cell (for the NW grown in [100]), while 7 nm-wide NWs have as many as 1 826 atoms per unit cell (for the NW grown in [111]). For simplicity, NWs with circular cross sections are considered. However, real-world NWs are prone to surface reconstruction and core distortion [211]. Furthermore, pristine atomic structures are assumed; real-world NWs are prone to defects such as vacancies [212]. These effects could be readily captured by the SK TB model, which can be used to simulate silicon NWs with various cross sections and atomic

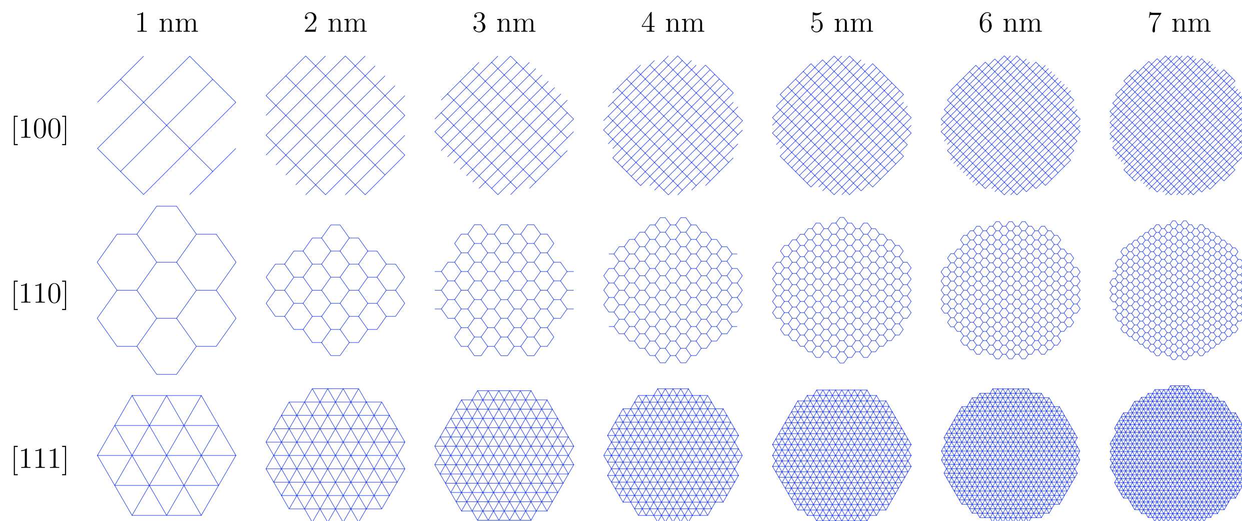


Figure 4.4 – Atomic structures of the unit cells of several silicon NWs in the wireframe model. Lines represent interatomic bonds. The NWs have diameters ranging from 1 nm to 7 nm (the structures are not to scale); their growth directions are [100], [110], and [111]; they all have circular cross sections.

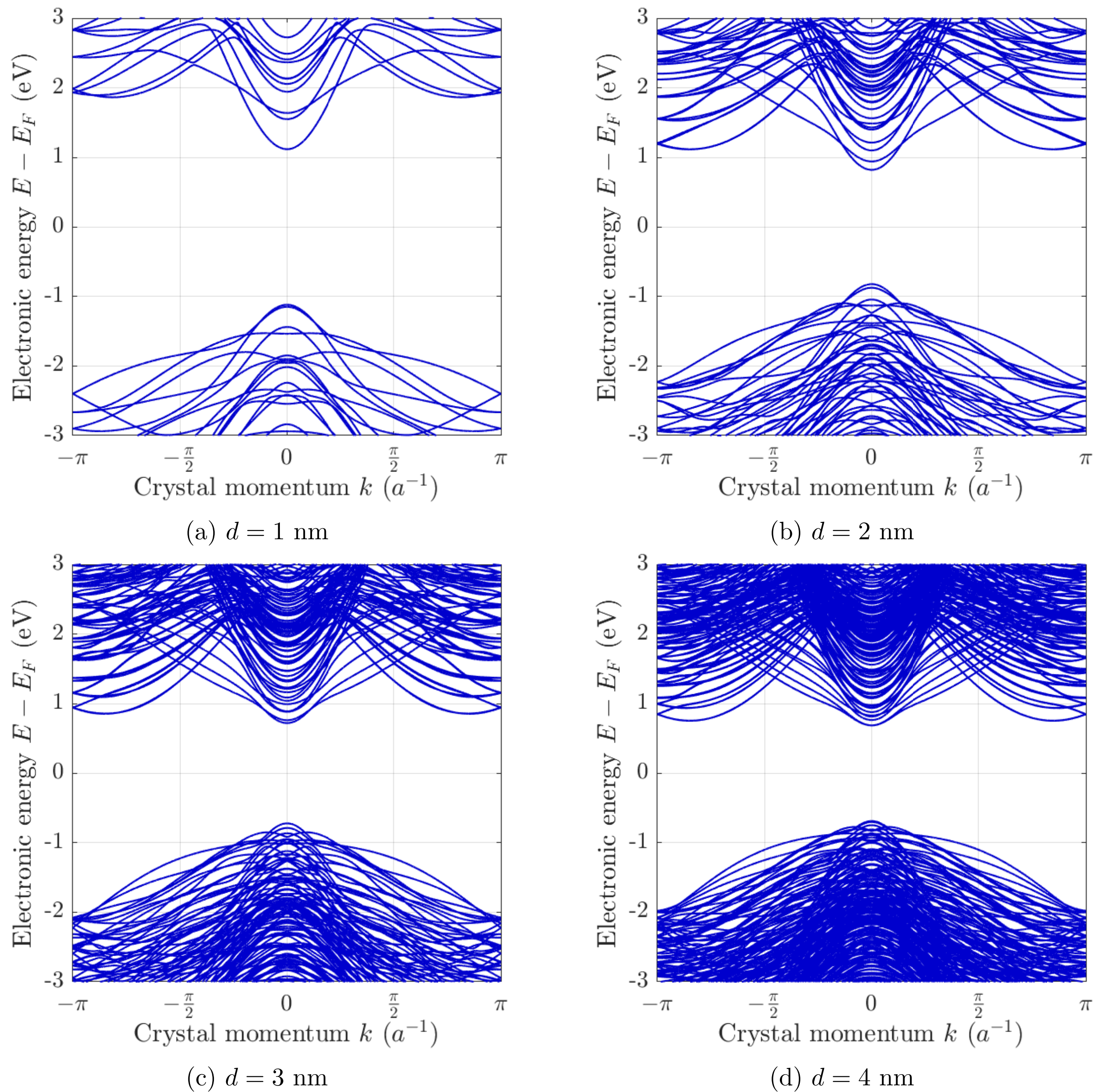
structures.

Figure 4.5 shows the bandstructures of four NWs grown in [110] with diameters d ranging from 1 nm to 4 nm. Simulations are performed in the SK TB scheme described in the previous section. Two important observations should be made:

- the bandgap energies E_G of the NWs are all greater than that of bulk silicon, for which $E_G \approx 1.14$ eV; furthermore, the thinner the NW, the wider its bandgap;
- the density of bands (also known as subbands) is greater for wide NWs than for narrow NWs.

These are consequences of quantum confinement. Electrons and holes in a silicon NW are confined along two dimensions of space, leading to zero-point energy and energy quantization, both of which are more significant for thinner NWs [48]. The zero-point energy translates to a wider bandgap (compared to bulk silicon), while energy quantization translates to spread-out subbands. Many important electronic properties of silicon NWs can be extracted from such bandstructures; some of these properties are reviewed below.

The bandgap energies of various silicon NWs are shown in Fig. 4.5. Overall, these data



(a) $d = 1$ nm

(b) $d = 2$ nm

(c) $d = 3$ nm

(d) $d = 4$ nm

Figure 4.5 – Bandstructures of various silicon NWs with circular cross sections, various diameters d, and growth direction [110], as computed from the $sp^3d^5s^\star$ SK TB model. Electronic energies E (in reference to the Fermi energy E_F) are plotted as a function of crystal momentum k (relative to the inverse of the lattice constant a).

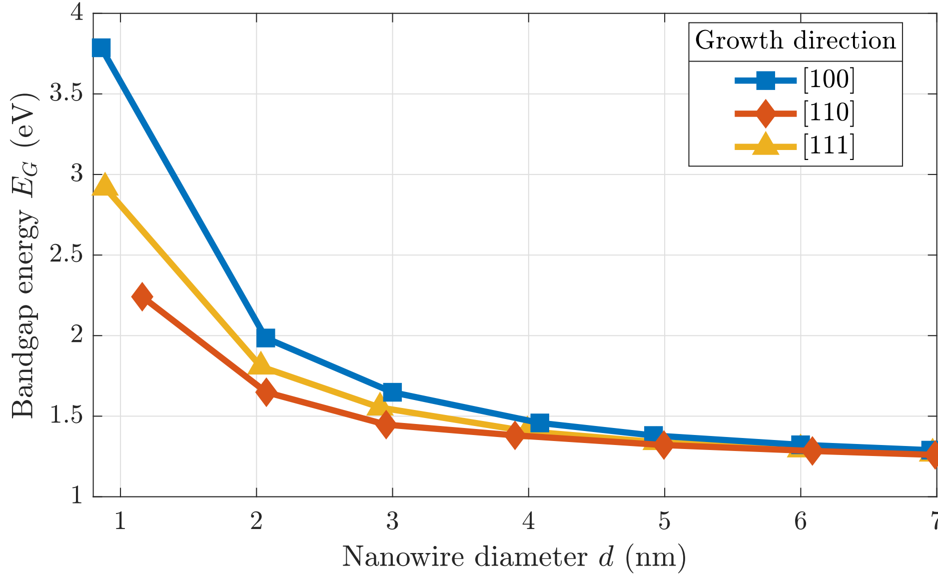


Figure 4.6 – Bandgap energies E_G of silicon NWs with circular cross sections, various diameters d, and various growth directions, as computed from the $sp^3d^5s^\star$ SK TB model.

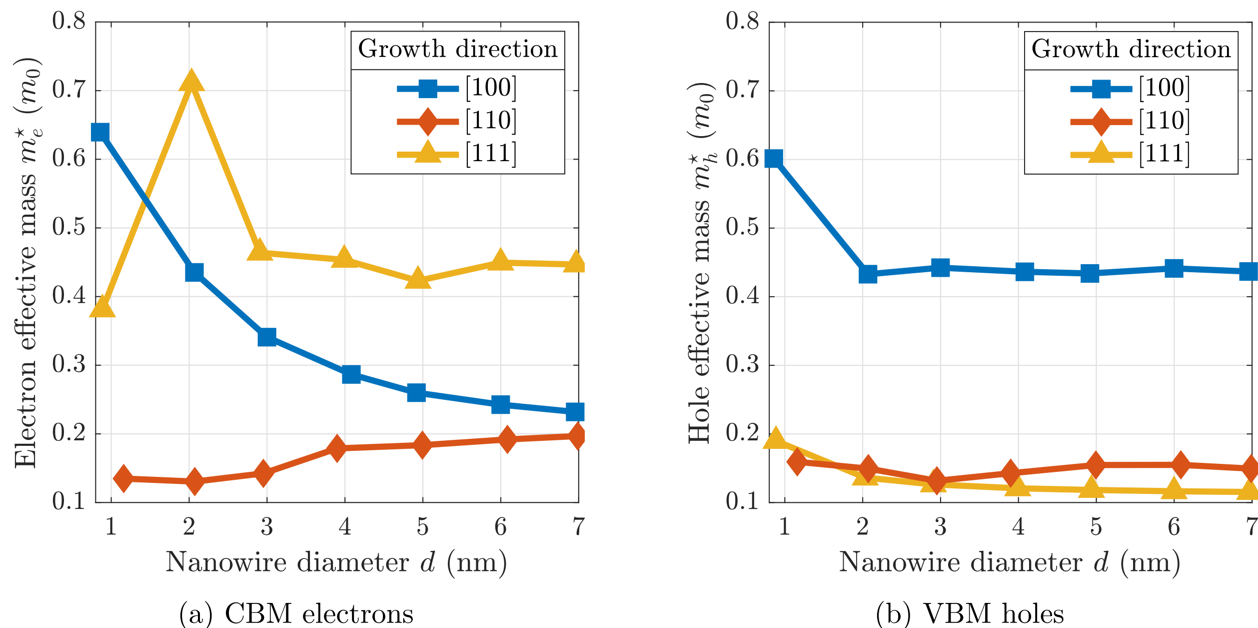
exhibit good agreement with previous experimental and theoretical reports [48, 213, 214]. Regardless of the growth direction, E_G approaches the bulk silicon bandgap energy for large d. However, NWs grown in [100] have larger E_G and NWs grown in [110] have smaller E_G. For FET applications, a bandgap energy not much larger than 1.5 eV is preferable; this criterion favours NWs grown in [110]. Parenthetically, all investigated NWs have a direct bandgap, except the NWs grown in [111] with $d \geq 2$ nm. This can be explained by geometrical arguments pertaining to the "folding" of the 3D Brillouin zone of bulk silicon onto the 1D Brillouin zone of the NW, which depends on growth direction [215, 216].

The electron and holes EMs, $m_e^\star$ and $m_h^\star$, of the silicon NWs are shown in Fig. 4.7. They are defined as

$$(m_e^\star)^{-1} = \frac{1}{\hbar^2} \left. \frac{d^2E}{dk^2} \right|_{\text{CBM}} , \tag{4.15}$$

$$(m_h^\star)^{-1} = -\frac{1}{\hbar^2} \left. \frac{d^2E}{dk^2} \right|_{\text{VBM}} , \tag{4.16}$$

where $E(k)$ is the bandstructure [Fig. 4.5] and where the derivatives are calculated at the CBM and VBM, respectively. The EM tensor of bulk silicon is anisotropic. The EMs of a



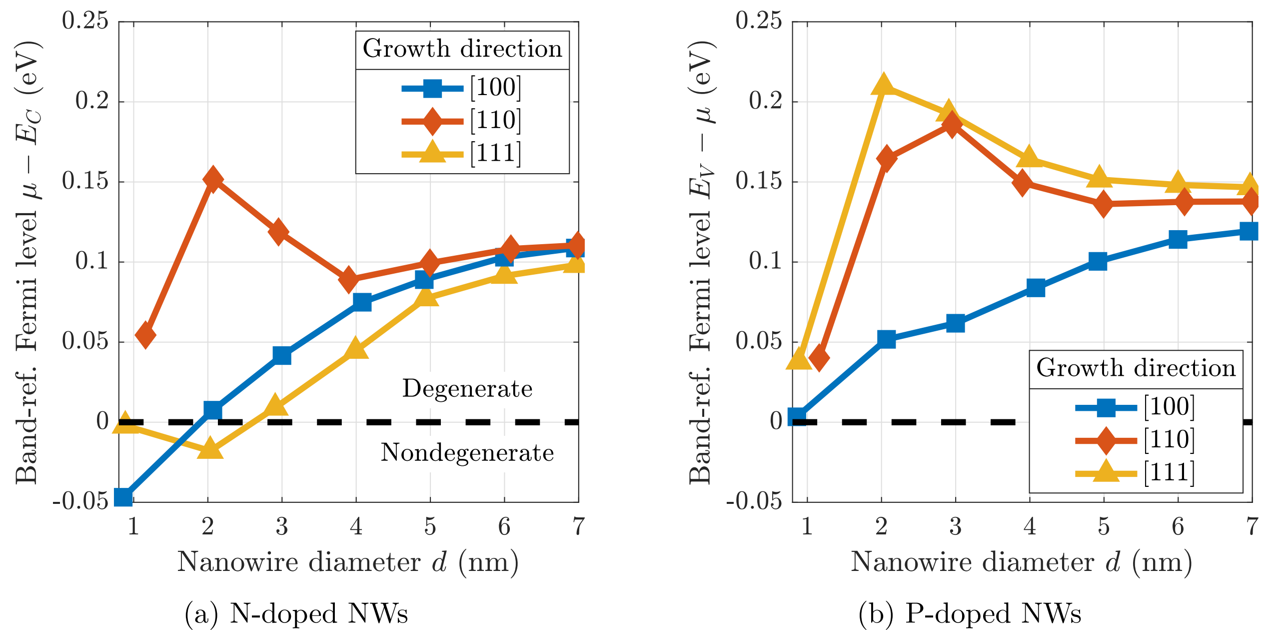
(a) CBM electrons

(b) VBM holes

Figure 4.7 – Effective masses (EMs) of electrons $m_e^\star$ and holes $m_h^\star$ in various silicon NWs with circular cross sections, various diameters d, and various growth directions, as computed from the $sp^3d^5s^\star$ SK TB model and Eqs. 4.15–4.16.

silicon NW grown in a given crystallographic direction is thus, to an extent, set by the EMs of bulk silicon along that same crystallographic direction; quantum confinement also has a significant effect for thin NWs. The NWs grown in [110] have low EMs both for electrons and holes. For most FET applications, a lower EM is preferable as it lead to higher mobility (even in the ballistic limit) and high on-state current [16, 217]. For TFET applications, a low EM is required for high BTBT probability and high on-state current [Eq. 3.7]. Thus, these criteria favour NWs grown in [110]. Parenthetically, the "jump" in the electron EM for the NWs grown in [111] at $d = 2$ nm can be attributed to the direct-to-indirect bandgap transition.

Figure 4.8 shows the band-referenced Fermi levels of highly n-doped and p-doped silicon NWs. The doping concentrations are close to the limit of what is experimentally achievable: a donor doping concentration of $N_D = 2 \times 10^{20}$ cm^{-3} for n-doped NWs and an acceptor doping concentration of $N_A = 3 \times 10^{20}$ cm^{-3} for p-doped NWs. In almost all surveyed NWs, such doping is degenerate, that is, the Fermi level lies within the conduction band (for n-doped NWs) or valence band (for p-doped NWs). NWs grown in [110] exhibit highly degenerate



(a) N-doped NWs

(b) P-doped NWs

Figure 4.8 – Band-referenced Fermi levels of n-doped and p-doped silicon NWs with circular cross sections, various diameters d, and various growth directions, as computed from the $sp^3d^5s^\star$ SK TB model at a temperature of $T = 300$ K. For n-doped (p-doped) NWs, the Fermi level μ is referenced to the CBM (VBM) energy E_C (E_V). The donor doping concentration of n-doped NWs is $N_D = 2 \times 10^{20}$ cm^{-3}; the acceptor doping concentration of p-doped NWs is $N_A = 3 \times 10^{20}$ cm^{-3}. Band-referenced Fermi levels below (above) the black-dashed lines correspond to nondegenerate (degenerate) doping.

doping both for n-doping and p-doping. To achieve a high BTBT probability in a TFET, a broken gap band alignment is needed [Fig. 3.3b]. Typically, this requires degenerate doping in the source of the TFET. Again, this criterion favours NWs grown in [110]. Parenthetically, the solid solubility limits of dopants in bulk silicon and silicon NWs may be very different; recent reports suggest that this limit is higher in silicon NWs [218, 219].

Overall, silicon NWs grown in [110] offer several advantages over other silicon NWs. These include their various attractive electronic properties listed in this section, as well as many others discussed in the literature. For example, these NWs offer greater immunity against interface roughness [220]. The investigations of this thesis will thus focus on silicon NWs grown in [110].

4.3 The Nonequilibrium Green's Function Formalism

Loosely speaking, a Green's function is the inverse of an operator. It is a useful mathematical tool in the theory of linear (ordinary and partial) differential equations, where it is used to solve inhomogeneous equations. For example, the Green's function of the Laplacian can be used to solve Poisson's equation for an arbitrary charge distribution [221]. In the context of many-body physics, the Green's function specifically refers to the inverse of a linear operator related to the Hamiltonian. For example, given a single-electron, time-independent Hamiltonian H of a system in thermal equilibrium, the retarded (causal) Green's function is defined as

$$G^r(E) = \lim_{\substack{\eta \to 0 \\ \eta > 0}} \left[(E + i\eta) I - H \right]^{-1} , \tag{4.17}$$

where E is the electronic energy and I is the identity operator. In essence, the Green's function is a solution to the Schrödinger equation: along with simple algebraic manipulations, it can be used to obtain the dynamics of the quantum system described by H as well as its DOS [59].

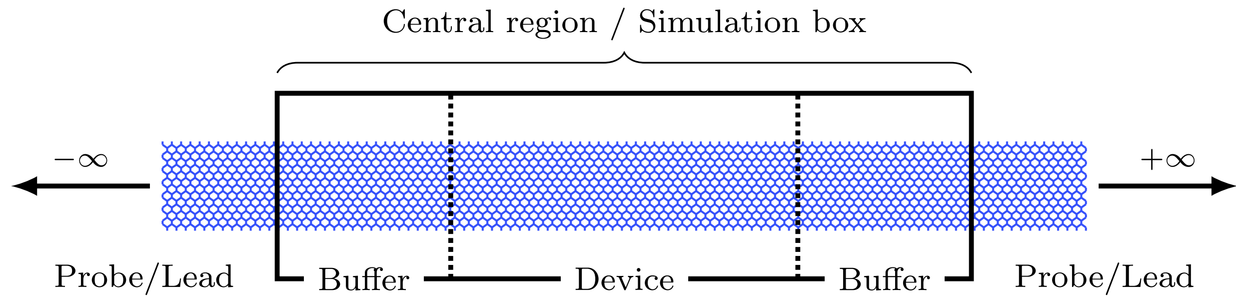


Figure 4.9 – Diagram of a two-probe device as considered in the nonequilibrium Green's function (NEGF) formalism. Two semi-infinite thermal reservoirs, the probes (or leads), are connected to a finite structure, the central region. An external voltage may be applied to the probes, which are assumed to exhibit equilibrium statistics. Charges are injected from the probes into the central region, whose nonequilibrium statistics depends on the applied voltage and scattering in the central region. In numerical implementations, the central region is called "simulation box" and includes the device and buffer regions, parts of the leads neighbouring the device that smoothly connect the device to the leads to preserve the equilibrium character of the leads. In this diagram, the device is composed of a silicon NW (blue lattice).

Equation 4.17 can be generalized to describe nonequilibrium systems, resulting in an NEGF from which nonequilibrium properties like electrical current can be computed; the NEGF formalism was first developed in the 1960s by Keldysh [222] and independently by Kadanoff and Baym [223]. Since the NEGF formalism is based on Schrödinger's equation, it can describe both classical (e.g. thermionic emission above a potential barrier) and quantum (e.g. tunnelling under a potential barrier) transport phenomena. Furthermore, it is indifferent to the underlying physics of the Hamiltonian; the NEGF formalism is thus compatible with a broad range of material physics formalisms [Fig. 4.1].

By assumption, the systems studied in the NEGF formalism comprise:

- a **central region**, namely, an arbitrary structure finite along one dimension of space, the transport direction, and confined or periodically repeating along the other dimension(s);
- connected to the central region, at least two **probes**, namely, semi-infinite thermal baths that maintain thermal equilibrium (possibly through strong inelastic scattering) and may supply and absorb an unlimited amount of charge and energy to and from the central region with no impact on the probes' physical properties; external voltages may be applied to probes.

A priori, it may appear that modelling such a system would be impossible. Indeed, it is an infinite system of interacting charges without any obvious periodicity nor symmetry. Furthermore, this system is characterized by complex statistics that arises as a result of a tangled interplay between what happens in the probes and what happens in the central region. What happens in the probes manifestly affects what happens in the central region; after all, an electrical current needs to flow in the central region in response to voltages applied to the probes. Likewise, what happens in the central region should also affect what happens in the probes. A fundamental assumption of the NEGF formalism is to ignore the central region's effect on the probes. No matter what happens in the central region, probes always follow equilibrium statistics with fixed chemical potentials and temperatures. To justify this assumption, one should note that systems modelled in the NEGF formalism are typically nanoscale electronic devices connected to microscale contacts; probes are thus orders of magnitude larger than central regions. This implies that the central region should have no effect on the bulk of the probes. On the other hand, the central region does affect neighbouring regions of the probes, within a few screening lengths. A proper implementation of the NEGF formalism thus requires these neighbouring regions, called "buffers," to be included in the central region. This ensures that what happens in the central region does not affect the probes, as well as a smooth matching of the charge density and electrical current at interfaces between the central region and probes [180]

The most commonly studied systems in the NEGF formalism are two-probe devices; such a device is illustrated in Fig. 4.9. In this thesis, FETs are modelled as two-probe devices. The bulk of the source and drain are the two probes. The channel, the parts of the source and drain within a few screening lengths from the channel, and the neighbouring oxides constitute the central region. For simplicity, the gate is not explicitly considered as a probe; rather, its effect is modelled through an appropriate boundary condition for Poisson's equation. Throughout this thesis, confinement along at least one of the two directions normal to the transport direction will be assumed. NW FETs are confined along both of these directions.

Voltages, i.e. differences in chemical potential [Eq. 2.5], may be applied across probes. In response, electrical currents flow between probes. A natural question is thus: what exactly

makes electrons flow? Supriyo Datta offers a particularly intuitive answer to this question: "Each contact [probe] seeks to bring the channel [central region] into equilibrium with itself. The source keeps pumping electrons into it, hoping to establish equilibrium. But equilibrium is never achieved as the drain keeps pulling electrons out in its bid to establish equilibrium with itself. The channel is thus forced in a balancing act between two reservoirs with different agendas and this sends it into a nonequilibrium state intermediate between what the source would like to see and what the drain would like to see." [224]

An important question ensues from this basic picture of transport in the NEGF formalism. How can one quantify the charge density in the central region and the electrical current between probes? This question is answered in two ways. First, Sec. 4.3.1 introduces the Landauer–Büttiker (LB) formalism, a physically transparent simplification of the NEGF formalism. The LB formalism is mathematically simple and useful in developing an intuitive understanding of charge transport but is not generally applicable to arbitrary devices. Second, Sec. 4.3.2 presents a brief overview of the more formal treatment of the generally applicable NEGF formalism used for the NEGF–TB numerical simulations performed in `Nanoskim 2.0`.

4.3.1 The Landauer–Büttiker Formalism

The LB formalism was initially developed by Rolf Landauer and Markus Büttiker [225–227]. It was later harnessed by Datta to develop a qualitatively clear and quantitatively accurate description of charge transport in many FETs [173, 224]. It is Datta's form of the LB formalism that is presented here. The key assumptions of the LB formalism are as follows:

- electrons respond to a mean-field potential due to other electrons and voltages [59]; this assumption is usually inadequate for strongly correlated transport, such as in single-electron transistors [228];
- transport is coherent, namely, electrons retain their quantum phase as they traverse the central region [173]; this implies that transport is elastic, namely, electrons flow between probes in independent energy modes.

It is important to note that neither of these two assumptions is required for the NEGF

formalism in its most general form. Parenthetically, coherent transport is sometimes referred to as "ballistic" in the field of statistical physics. In the field of electronic engineering, "ballistic transport" has a more restrictive definition: it refers to transport with no scattering whatsoever, including elastic scattering. It is the latter nomenclature that is used in this thesis.

Based on these assumptions, it can be shown that the total free charge in the central region is [102]

$$Q_f = -q \int_{-\infty}^{+\infty} D\left(E\right) \frac{1}{2} \left[f_S\left(E\right) + f_D\left(E\right)\right] dE \,, \tag{4.18}$$

where $D\left(E\right)$ is the electronic DOS of the central region and $f_{S,D}\left(E\right)$ are the source and drain Fermi–Dirac distributions [Eq. 2.6]. The central region is connected to each probe in only one of its two boundaries, thereby intuitively explaining the factor of $\frac{1}{2}$ in Eq. 4.18. This makes explicit the notion that the central region's statistics is intermediate between those of the source and drain.

Next, in the LB formalism, the drain current of an FET is expressed as [102, 173, 224]

$$I_{DS} = \frac{q}{h} \int_{-\infty}^{+\infty} M\left(E\right) T\left(E\right) \left[f_S\left(E\right) - f_D\left(E\right)\right] dE \,, \tag{4.19}$$

where $M\left(E\right) \in \mathbb{Z}^{+}$ is the number of conduction modes (including a possible spin degeneracy) at energy $E \in \mathbb{R}$ and $T\left(E\right) \in \left[0\,,1\right]$ is the transmission probability at E. This equation is intuitive and can be explained using an analogy to a sophisticated highway system for peculiar trucks that carry electrical current between source and drain. This highway system is divided into two parts: a highway for trucks driving from the source to the drain, and a highway for trucks driving from the drain to the source. Each highway has an infinite number of vertically stacked highway levels that are labelled by an energy index E with possible values of $0, \Delta E, -\Delta E, 2\Delta E, -2\Delta E, \cdots$; ΔE is a very small positive number. Electrons are trucks on this highway between probes. Each truck carries a load consisting of a charge of $-q$ and requires a time of $\frac{h}{\Delta E}$ to complete its journey between probes. Effectively, each truck carries a drain-to-source current of $\pm \frac{q}{h} \Delta E$; the positive (negative) sign is taken for trucks travelling

from source to drain (from drain to source). Each highway level has a certain number of lanes $M(E)$ (for each direction). Driving through this highway may be perilous: a truck on level E only has a probability $T(E)$ of completing its journey between probes. Traffic officers limit access to the highway. The probability that a truck on a lane labelled by E is admitted into the highway from source to drain (from drain to source) is $f_S(E)$ ($f_D(E)$). Finally, this highway system functions at peak efficiency within the previously described constraints: there are no traffic accidents nor congestion. To obtain the electrical current I_{DS} carried on the highway system, one should add up the current contributions of $\pm\frac{q}{h}\Delta E$ from all trucks. The analogy is perfected by taking the limit $\Delta E \to 0$. Datta considers Eq. 4.19 so clear and simple that he opined: "NEGF is generally regarded as an esoteric tool for specialists, but we believe it should be a part of the standard training of science and engineering students." [229]

How can the number of conduction modes $M(E)$ be computed? By assumption, the system modelled in the LB formalism is infinite along one dimension of space, the transport direction (including the central region and the two probes). The solution to Schrödinger's equation thus takes the form of a plane wave along the transport direction. A plane wave is characterized by its wavenumber k. The system is confined along at least one of the other dimensions of space, leading to discrete energy levels E_n indexed by a quantum number n. Simplistically, the energy modes of the system can thus be expressed as $E_n(k) = E_n + \frac{(\hbar k)^2}{2m^\star}$, where $m^\star$ is the EM along the transport direction. For this reason, the indices n can be thought of as labelling "subbands" or "transverse modes." Such a two-probe system can thus be thought of as a waveguide for electrons. Since the kinetic energy $\frac{(\hbar k)^2}{2m^\star}$ is positive and may be taken to be as large as necessary, the number of conduction modes $M(E)$ at energy E is given by the total number of subbands with energy lower than E [173]:

$$M(E) = \sum_n H(E - E_n) \,. \tag{4.20}$$

This is a general expression for the number of conduction modes. For a 2D central region of width W, assuming that the confining potential is an infinite square well, the subband energies are given by $E_n = \frac{\hbar^2}{2\widetilde{m}}\left(\frac{n\pi}{W}\right)^2$, where $\widetilde{m}$ is an appropriate EM and $n \in \mathbb{N}$ [165]. The Fermi energy E_F may be expressed in terms of the Fermi wavelength λ_F: $E_F = \frac{2\pi^2\hbar^2}{\widetilde{m}\lambda_F^2}$. The

number of conduction modes at $E = E_F$, that is, the number of subbands with energy lower than E_F, is given by [173]

$$M\left(E_F\right) = \left\lfloor \frac{2W}{\lambda_F} \right\rfloor . \tag{4.21}$$

Thus, the number of conduction modes can be understood as the number of half Fermi wavelengths that fit in the width of the central region. For a central region with large width W, the spacing between subband energies is very small. It then becomes more practical to view the subband energies as a continuum described by a bandstructure. This makes Eq. 4.20 inappropriate. Reference [230] describes how to compute $M\left(E\right)$ for an arbitrary bandstructure. In the case of a parabolic conduction band in 2D, one has

$$M\left(E\right) = Wg\frac{\sqrt{2m^{\star}\left(E - E_C\right)}}{\pi\hbar}H\left(E - E_C\right) , \tag{4.22}$$

where g is the valley degeneracy, $m^{\star}$ is the density-of-states EM (assumed to be equal to the EM along the transport direction), and E_C is the CBM energy [102]. Inserting Eq. 4.22 into Eq. 4.19 results in a well-known MOSFET drain current model [4, 62, 69, 231]:

$$I_{DS} = W\frac{q}{\hbar^2}\left(\frac{k_BT}{\pi}\right)^{\frac{3}{2}} g\sqrt{\frac{m^{\star}}{2}}T\left(E\right)\left[\mathscr{F}_{\frac{1}{2}}\left(\eta_S\right) - \mathscr{F}_{\frac{1}{2}}\left(\eta_D\right)\right] , \tag{4.23}$$

where $T\left(E\right)$ was assumed to be constant, $\eta_{S,D} = \frac{\mu_{S,D} - E_C}{k_BT}$ ($\mu_{S,D}$ are the source and drain Fermi levels), and $\mathscr{F}_{\frac{1}{2}}$ is the complete Fermi–Dirac integral of order $\frac{1}{2}$, defined as

$$\mathscr{F}_{\frac{1}{2}}\left(\eta\right) = \frac{1}{\Gamma\left(1 + \frac{1}{2}\right)}\int\limits_0^{\infty} \frac{x^{\frac{1}{2}}}{1 + \exp\left(x - \eta\right)}dx . \tag{4.24}$$

It remains to specify an expression for the transmission probability $T\left(E\right)$. The transmission probability reflects the extent to which electrons are able to traverse the central region without being "impeded" in their journey between source and drain, that is, without scattering. Thus, $T\left(E\right)$ should depend on the channel length L as well as the mean free path λ. In general, λ is E-dependent; here, for simplicity, it is assumed to be constant. In the limit $\frac{L}{\lambda} \to 0$, known as the ballistic limit, electrons are unlikely to undergo scattering, so that

$T(E) \to 1$. In the limit $\frac{L}{\lambda} \to \infty$, known as the diffusive limit, $T(E)$ must approach $\frac{\lambda}{L}$. This is because the predictions of the DD model and the Einstein relation only match those of Eq. 4.23 when $T(E) = \frac{\lambda}{L}$ [102]. To smoothly interpolate between these two limits, one may express the transmission probability as

$$T(E) = \frac{\lambda}{L + \lambda}\,. \tag{4.25}$$

Thanks to its correct limiting behaviour, Eq. 4.25 and the LB formalism are considered to bridge the gap between the DD transport of microscale FETs and the ballistic transport of nanoscale FETs [217]. In particular, the LB formalism can be adapted to model FETs operating in the quasiballistic regime [232].

Together, Eqs. 4.19, 4.20, and 4.25 describe an intuitive model for coherent transport in FETs. This model is used for the analytical investigations presented in this thesis. Computational investigations will, instead, make use of the formal NEGF theory presented in the next section.

4.3.2 Formal Theory

The LB formalism has limited applicability for two reasons. First, the expression for charge with the LB formalism [Eq. 4.18] is not spatially resolved. This makes it difficult to use the LB formalism in conjunction with Poisson's equation to model systems with complex electrostatics. Second, there are no general analytical expressions for the number of conduction modes $M(E)$ and transmission probability $T(E)$. Furthermore, where analytical expressions exist, they are usually only approximate. In this section, generally applicable formulas for charge density and electrical current within the NEGF formalism are provided. These formulas are compatible with numerical methods. A complete and formal treatment of the NEGF formalism is beyond the scope of this thesis. Interested readers may consult Refs. [173] or [174] for detailed derivations.

In the NEGF formalism, the free charge density per unit volume is computed through [173,

174, 180, 181]

$$\rho_f(\mathbf{r}) = -\frac{i}{2\pi} q \int_{-\infty}^{+\infty} G^<(\mathbf{r}, \mathbf{r}' = \mathbf{r}, E)\, dE\,, \tag{4.26}$$

where $G^<(\mathbf{r}, \mathbf{r}', E)$ is the lesser Green's function obtained by the Keldysh equation:

$$G^< = G^r \Sigma^< G^a\,. \tag{4.27}$$

The retarded and advanced Green's functions are calculated by

$$G^r = \lim_{\substack{\eta \to 0 \\ \eta > 0}} \left[(E + i\eta) I - H - \Sigma_S^r - \Sigma_D^r\right]^{-1}\,, \tag{4.28}$$

$$G^a = [G^r]^\dagger\,, \tag{4.29}$$

where H is the Hamiltonian of the device material. Compared to the equilibrium retarded Green's function [Eq. 4.17], the nonequilibrium retarded Green's function [Eq. 4.28] includes retarded self-energies $\Sigma_{S,D}^r$, which incorporate the effect of the source and drain on the central region's statistics, respectively. The term $\Sigma^<$ is the lesser self-energy, which describes the injection of charge from the source and drain. It is defined in terms of the source and drain broadening functions $\Gamma_{S,D}$ as

$$\Sigma^< = i\Gamma_S f_S(E) + i\Gamma_D f_D(E)\,, \tag{4.30}$$

$$\Gamma_{S,D} = i\left(\Sigma_{S,D}^r - \Sigma_{S,D}^a\right)\,, \tag{4.31}$$

$$\Sigma_{S,D}^a = [\Sigma_{S,D}^r]^\dagger\,. \tag{4.32}$$

The retarded self-energies $\Sigma_{S,D}^r$ are computed from an iterative scheme described in Ref. [233].

In the case of coherent transport, the drain current is computed through the Landauer formula

$$I_{DS} = \frac{q}{h} \int_{-\infty}^{+\infty} \overline{T}(E) \left[f_S(E) - f_D(E)\right] dE\,, \tag{4.33}$$

where $\overline{T}(E)$ is the transmission function, defined as

$$\overline{T}(E) = \mathrm{Tr}\left(\Gamma_D G^r \Gamma_S G^a\right) . \tag{4.34}$$

Here, spin degeneracy is included in $\overline{T}(E)$. Comparing Eq. 4.33 to Eq. 4.19, it can be seen that in the regime of applicability of the LB formalism, the transmission function is given by

$$\overline{T}(E) = M(E)\, T(E) . \tag{4.35}$$

Equation 4.34 can be generalized to describe non-coherent transport, notably electron transport with inelastic electron–phonon scattering [234–236].

4.4 Atomistic Quantum Transport

Previous sections introduced the SK TB model, with which bandstructure and other material properties are computed, as well as the NEGF formalism, with which nonequilibrium and transport properties are computed. This section provides a brief description of how the SK TB model and the NEGF formalism can be combined to provide a self-consistent description of two-probe systems like FETs, as implemented in `Nanoskim 2.0`; full details of the implementation are described in Refs. [157] and [158].

4.4.1 Iterative Procedure

The TB model provides a method to compute Hamiltonian H_0 of an isolated atomistic system, which describes the interactions between electrons and nuclei under zero external field. The external voltage applied to a two-probe system leads to a redistribution of charge compared to the equilibrium case; this is sometimes called "charge transfer." The charge

density of a two-probe system can thus be expressed as

$$\rho_f(\mathbf{r}) = \rho_{f,0}(\mathbf{r}) + \Delta\rho_f(\mathbf{r}) \ , \tag{4.36}$$

where $\rho_{f,0}(\mathbf{r})$ is the charge density at equilibrium. In this section, for simplicity, ρ_f and other physical quantities are expressed as functions of real-space position $\mathbf{r}$. However, in the numerical implementation of `Nanoskim 2.0`, these physical quantities are vectors or matrices in a basis of Bloch sums of Löwdin orbitals [Eq. 4.12]. Importantly, the potential due to $\rho_{f,0}(\mathbf{r})$, $V_0(\mathbf{r})$, is included in H_0. The term $\Delta\rho_f(\mathbf{r})$ describes the redistribution of charge. Similarly, the electric potential can be expressed as

$$V(\mathbf{r}) = V_0(\mathbf{r}) + \Delta V(\mathbf{r}) \ . \tag{4.37}$$

The Poisson equation for the potential $\Delta V(\mathbf{r})$ due to charge redistribution is given by

$$\boldsymbol{\nabla} \cdot [\varepsilon(\mathbf{r}) \boldsymbol{\nabla} \Delta V(\mathbf{r})] = -\Delta\rho_f(\mathbf{r}) \ , \tag{4.38}$$

where $\varepsilon(\mathbf{r})$ is the permittivity. To be complete, the boundary conditions for Poisson's equation need to be specified. In `Nanoskim 2.0`, the simulation box [Fig. 4.9] is taken to be a right rectangular prism. Position along the transport direction is labelled by z. Two opposite faces of the simulation box, at $z = z_S$ and $z = z_D$, correspond to boundaries with the source and drain, respectively. Portions of the four other faces correspond to the boundary with the gate; this interface of the simulation box in contact with the gate is labelled by Ω_G. Thus, the following Dirichlet boundary conditions are imposed:

$$V(\mathbf{r})|_{z=z_{S,D}} = \phi_{S,D} \, , \tag{4.39}$$

$$V(\mathbf{r})|_{\Omega_G} = \phi_G \, , \tag{4.40}$$

where $\phi_{S,G,D}$ are the values of the electric potential at the source, gate, and drain contacts, as determined by the gate voltage, the drain voltage, and relevant workfunctions [Sec. 2.2.3]. At every other surface of the simulation box, Neumann boundary conditions are imposed:

the normal derivative of $V(\mathbf{r})$ on these surfaces is set to 0. Numerically, Poisson's equation is solved on a uniform, cubic grid with a mesh point density of 0.1944 Å^{-3}; Appendix B.1 shows that this mesh point density is sufficient to capture the electrostatics of the systems investigated in this thesis.

Once the electric potential is known, the Hamiltonian of the two-probe system is expressed as

$$H = H_0 - q\Delta V\,. \tag{4.41}$$

At this stage, it can be seen that a set of coupled equations must be solved to obtain the Hamiltonian H. Indeed, H depends on ΔV [Eq. 4.41], an unspecified potential. To obtain $\Delta V(\mathbf{r})$, the Poisson equation needs to be solved [Eqs. 4.38–4.40]. The source term of the Poisson equation is $\Delta\rho_f(\mathbf{r}) = \rho_f(\mathbf{r}) - \rho_{f,0}(\mathbf{r})$ [Eq. 4.36]. In the NEGF formalism, $\rho_f(\mathbf{r})$ can be obtained by Eq. 4.26, which involves the computation of a Green's function, which is itself a function of H [Eqs. 4.27–4.32]. To solve this problem, a self-consistent iterative procedure is employed, as described in Algorithm 1. The algorithm starts by constructing an "initial guess" for the potential, $V^{(0)}(\mathbf{r})$. The algorithm then iterates over Eq. 4.41, Eqs. 4.27–4.32, and Eqs. 4.38–4.40; each iteration i results in an updated potential $V^{(i)}(\mathbf{r})$. The iteration stops when convergence is reached, namely when

$$\left\| V^{(i)}(\mathbf{r}) - V^{(i+1)}(\mathbf{r}) \right\|_\infty \leq V_\epsilon\,, \tag{4.42}$$

where V_ϵ is the convergence threshold. The smaller V_ϵ, the greater the number of iterations and the greater the accuracy. For the simulations presented in this thesis, a convergence threshold of $V_\epsilon = 5$ mV is used. After convergence, observables may be computed, notably the electrical current [Eqs. 4.33 and 4.34].

The NEGF–TB self-consistent iterative procedure described in Algorithm 1 is not mathematically guaranteed to converge, let alone converge in relatively few iterations. This problem also plagues other common calculations of condensed matter physics, notably those based on DFT. A crucial task is thus to optimize the algorithm so as to minimize the number

Algorithm 1: Self-consistent NEGF–TB atomistic quantum transport algorithm

1. Compute the equilibrium SK TB Hamiltonian H_0 [Sec. 4.2.1]
2. Compute the equilibrium charge density $\rho_{f,0}(\mathbf{r})$ and electric potential $V_0(\mathbf{r})$ [158]
3. Compute the retarded self-energies $\Sigma^r_{S,D}$ [233]
4. Construct an initial guess for the electric potential $V^{(0)}(\mathbf{r})$ [Sec. 4.4.2]
5. Set $i = 0$
6. **while** $\left\| V^{(i)}(\mathbf{r}) - V^{(i+1)}(\mathbf{r}) \right\|_\infty > V_\epsilon$ **do**
7. Compute the electric potential due to charge transfer $\Delta V^{(i)}(\mathbf{r}) = V^{(i)}(\mathbf{r}) - V_0(\mathbf{r})$ [Eq. 4.37]
8. Compute the nonequilibrium Hamiltonian $H^{(i)} = H_0 - q\Delta V^{(i)}$ [Eq. 4.41]
9. Compute the retarded Green's function $G^{r,(i)} = \lim_{\substack{\eta\to 0\\ \eta>0}} \left[(E + i\eta) I - H^{(i)} - \Sigma^r_S - \Sigma^r_D\right]^{-1}$ [Eq. 4.28]
10. Compute the lesser Green's function $G^{<,(i)}(\mathbf{r}, \mathbf{r}', E)$ [Eqs. 4.27–4.32]
11. Compute the free charge density $\rho^{(i)}_f(\mathbf{r}) = -\frac{i}{2\pi} q \int_{-\infty}^{+\infty} G^{<,(i)}(\mathbf{r}, \mathbf{r}' = \mathbf{r}, E)\, dE$ [Eq. 4.26]
12. Compute the free charge density due to charge transfer $\Delta\rho^{(i)}_f(\mathbf{r}) = \rho^{(i)}_f(\mathbf{r}) - \rho_{f,0}(\mathbf{r})$ [Eq. 4.36]
13. Solve Poisson's equation $\boldsymbol{\nabla} \cdot \left[\varepsilon(\mathbf{r}) \boldsymbol{\nabla} \Delta V^{(i+1)}(\mathbf{r})\right] = -\Delta\rho^{(i)}_f(\mathbf{r})$ [Eq. 4.38] with appropriate boundary conditions at the source, drain, and gate contacts [Eqs. 4.39 and 4.40]
14. Compute the electric potential $V^{(i+1)}(\mathbf{r}) = V_0(\mathbf{r}) + \Delta V^{(i+1)}(\mathbf{r})$ [Eq. 4.37]
15. Apply periodic Pulay mixing with Kerker preconditioner to $V^{(i+1)}(\mathbf{r})$ [Refs. [237] and [238]]
16. Increment i by 1
17. Compute the transmission function $\overline{T}(E) = \mathrm{Tr}\left(\Gamma_D G^r \Gamma_S G^a\right)$ [Eq. 4.34]
18. Compute the drain current from the Landauer formula $I_{DS} = \frac{q}{h} \int_{-\infty}^{+\infty} \overline{T}(E) \left[f_S(E) - f_D(E)\right] dE$ [Eq. 4.33]

of iterations required to reach convergence. Although I have not written the core of `Nanoskim 2.0`, I have implemented two methods to improve convergence in this software. The first is a periodic Pulay mixing scheme [237] with Kerker preconditioner [238]. Periodic Pulay mixing has been shown to be generally applicable to self-consistent iterations and to be one of the most stable and efficient mixing methods [237]. Kerker preconditioners have been shown to improve convergence significantly for systems with high DOS around the Fermi level, such as degenerately doped semiconductors [238]. The second is a high-quality initial guess, which is described in the next section.

4.4.2 Initial Guess for the Potential

In a two-probe system like an FET, the electric potential is fixed by the drain voltage at interfaces with the source and drain contacts [Eq. 4.39]. The simplest potential that satisfies these two boundary conditions is the following linear interpolation:

$$V^{(0)}_{\text{linear}}(\mathbf{r}) = \phi_S + \frac{\phi_D - \phi_S}{z_D - z_S}(z - z_S) \; . \tag{4.43}$$

While simple, this "linear drop" initial guess completely ignores internal device electrostatics, notably those arising as a result of the doping profile and the gate voltage.

A device can be partitioned into several regions according to its doping profile. Specifically, suppose there are N regions labelled by $i = 1, 2, \cdots, N$ with doping concentrations N_i located between $z = z_{i-1}$ and $z = z_i$; note that $z_0 = z_S$ and $z_N = z_D$. Ignoring external voltages and charge transfer at interfaces, the potential V_i in each region is fixed by the doping concentration N_i, following

$$\int_{-\infty}^{+\infty} \frac{D(E)}{1 + \exp\left(\frac{E - \mu_0 + qV_i}{k_B T}\right)} = N_0 \pm N_i \, , \tag{4.44}$$

where $D(E)$ is the DOS of the semiconductor material, μ_0 is the chemical potential in the intrinsic case, and N_0 is the concentration of electrons in the intrinsic case. The electric

potential V_i is measured in a gauge where it equals 0 in the intrinsic case. On the right-hand side of Eq. 4.44, $N_0 + N_i$ is taken for donor doping and $N_0 - N_i$ is taken for acceptor doping.

To improve on the linear drop of Eq. 4.43, I implemented an initial guess $V^{(0)}(\mathbf{r})$ based on the following procedure; the gate is assumed to lie between positions $z = z_{G,1}$ and $z = z_{G,2}$:

(1) for each region i, between $z = z_{i-1}$ and $z = z_i$, set the potential V_i to be consistent with the doping concentration N_i [Eq. 4.44];

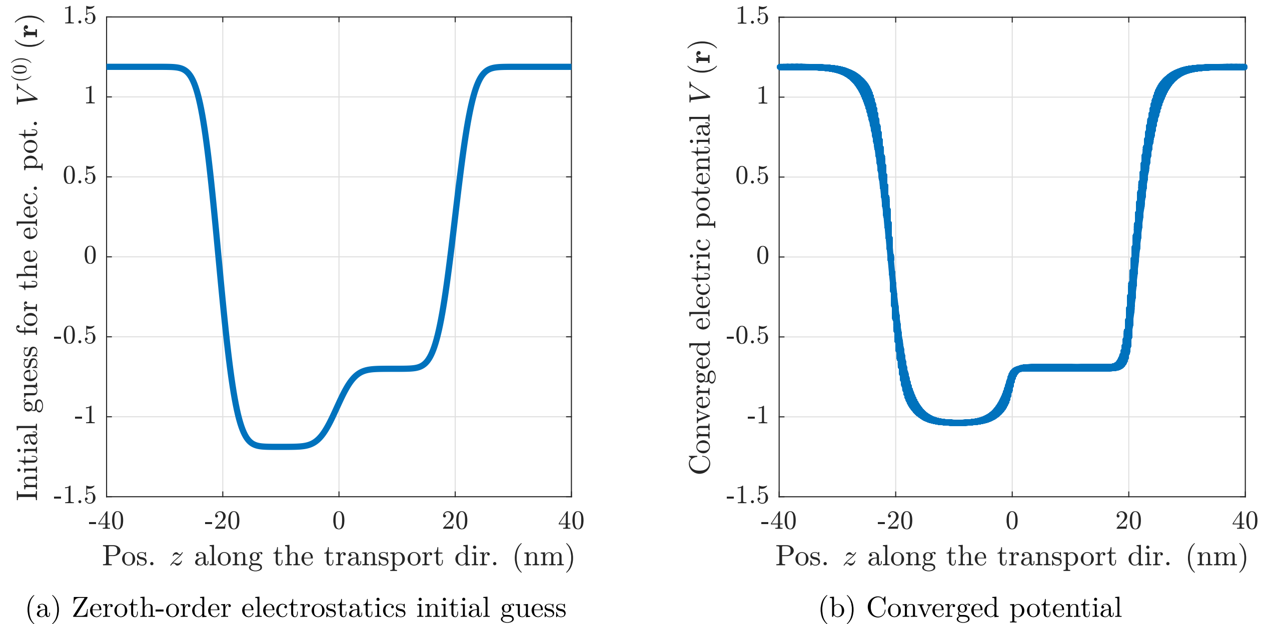
(2) between $z = z_{G,1}$ and $z = z_{G,2}$ (in the gated region), set the potential to be equal to ϕ_G, as determined by the applied gate voltage [Eq. 4.40];

(3) between $z = z_{N-1}$ and $z = z_D$ (in the drain), shift the potential by $-V_{DS}$, as determined by the applied drain voltage V_{DS} [Eq. 2.5];

(4) the resulting potential is a step function; to smooth it (in order to capture charge transfer at interfaces), take the convolution of the resulting potential with a unit-weight Gaussian function; the standard deviation of this Gaussian function sets the depletion length of junctions and should thus be on the order of a few nanometres.

Since this initial guess considers basic electrostatics due to the doping profile and applied voltages, it may be called the “zeroth-order electrostatics” initial guess.

Figure 4.10 compares, in the case of a gated n–p–i–n silicon NW, the zeroth-order electrostatics initial guess for the potential to the converged potential, as calculated in `Nanoskim 2.0`. It can be seen that this initial guess approximates the converged potential quite well. However, three important mismatches between these potentials can be noted. First, in the p-doped region (roughly between $z = -20$ nm and $z = 0$ nm), the converged potential is higher than the initial guess. This is most likely due to a bound state in the quantum well formed in this region. Naturally, this quantum effect is included in the converged potential of the NEGF–TB algorithm but not in the initial guess. Second, the initial guess depends on z but not on other spatial coordinates. In reality, the potential also has a weak dependence on other spatial coordinates. The NEGF–TB algorithm solves the 3D Poisson equation; the converged potential thereby depends on all three spatial coordinates, as evidenced by the “thickness” of the potential profile in Fig. 4.10b, which is a projection of the 3D potential on



(a) Zeroth-order electrostatics initial guess (b) Converged potential

Figure 4.10 – Comparison of the zeroth-order electrostatics initial guess for the potential $V^{(0)}(\mathbf{r})$ to the converged potential $V(\mathbf{r})$ in the NEGF–TB iterative procedure, in the case of a gated n–p–i–n silicon NW. The initial guess considers lowest order electrostatics due to the doping profile and applied voltages.

the z axis. Third, the width of depletion regions in the initial guess is set by an arbitrary constant (the standard deviation of the aforementioned Gaussian function). The initial guess therefore cannot accurately match the converged potential near interfaces.

It is difficult to precisely quantify the extent to which the zeroth-order electrostatics initial guess improves convergence relative to the linear drop initial guess. Indeed, this would depend on a dauntingly large number of variables: the doping profile, the applied voltages, etc. Anecdotally, based on numerical evidence, for relatively simple structures such as MOSFETs biased at low drain voltage, the zeroth-order electrostatics initial guess reduces the number of steps required to reach convergence by about one third. For complex structures, such as the one simulated in Fig. 4.10, convergence cannot be reached starting from the linear drop initial guess but can be reached starting from the zeroth-order electrostatics initial guess.

4.5 Summary

At a fundamental level, field-effect transistors (FETs) are quantum systems of interacting electrons and atomic nuclei. They are most accurately described by a many-body Schrödinger equation with dimensionality commensurate to the total number of particles. In practice, this equation is impossible to solve. To simulate FETs, four ingredients are needed. The first is a basis on which the system is discretized. The second is Poisson's equation, which describes the electric interactions between electrons and nuclei as well as the effect of the drain and gate voltages. The third is a material physics formalism that captures the electronic properties of the device, such as bandstructure and mobility; important examples include the drift–diffusion model, the tight-binding (TB) model, and density functional theory (DFT). The fourth is a transport physics formalism that captures the nonequilibrium response of the FET to applied voltages, notably current; important examples include the Boltzmann transport equation and the nonequilibrium Green's function (NEGF) formalism. The NEGF–TB algorithm offers an optimal balance between computational burden and accuracy. It describes matter atomistically and transport quantum mechanically; it can simulate systems with hundreds of thousands of atoms. The numerical simulations performed in this thesis are based on the NEGF–TB algorithm as implemented in the atomistic quantum transport package `Nanoskim 2.0`.

Several approximations are required to derive the TB Hamiltonian; these include the Born–Oppenheimer approximation, which decouples electrons from nuclei, and the independent electron approximation, which decouples electrons from each other, expressing the many-body Schrödinger equation as a set of single-electron Schrödinger equations. `Nanoskim 2.0` implements the TB model under the Slater–Koster scheme. The Hamiltonian is thereby discretized in a basis of Bloch sums of Löwdin orbitals and can be expressed in terms of about a dozen parameters, thanks to the two-centre and nearest neighbour approximations. These parameters are fitted from DFT simulations. Silicon nanowires (NWs), 1D allotropes of silicon offering promising prospects for future generations of gate-all-around FETs, can be accurately

simulated by the TB model. They are characterized by diameter and crystallographic growth direction. Quantum confinement plays an important role: thin NWs have wider bandgaps than thicker NWs. Overall, silicon NWs grown in [110] offer several advantages. They have low electron and hole EMs. Furthermore, with experimentally achievable doping concentrations, they can be degenerately n-doped and p-doped; this is necessary for devices with broken gap band alignment like tunnel field-effect transistors.

The systems studied in the NEGF formalism comprise a central region connected to at least two probes. By assumption, the probes are thermal baths that may supply or absorb limitless charge carriers to or from the central region; their properties do not depend on the state of the central region. When a voltage is applied across probes, they follow statistics with distinct Fermi levels. Each probe attempts to bring the central region in equilibrium with itself. The central region thereby follows nonequilibrium statistics intermediate between those of the probes, which leads to an electrical current flowing through the central region. The Landauer–Büttiker formalism provides a clear picture of transport in two-probe systems; electrical current is expressed in terms of the number of conduction modes and the transmission probability. The NEGF formalism expresses nonequilibrium charge density and electrical current in terms of Green's functions, which are inverses of operators related to the Hamiltonian.

The NEGF–TB algorithm is an iterative procedure in which the electric potential depends on charge density (through Poisson's equation), the charge density depends on the Hamiltonian (through the NEGF formalism), and the Hamiltonian depends on the electric potential. This algorithm may require a large number of iterations to converge. Convergence can be improved through periodic Pulay mixing with Kerker preconditioner and/or with a high-quality initial guess.

The projects described in the next three chapters of this thesis are substantiated by numerical simulations based on the NEGF–TB algorithm.

Chapter 5

Bound-Charge Engineering

Indeed, some authors give you the impression that bound charges are in some sense "fictitious"—mere bookkeeping devices used to facilitate the calculation of fields. Nothing could be further from the truth: [they are] perfectly genuine accumulations of charge.

— D. J. Griffiths, 2013 [47]

There are two types of charges in materials: free charges and bound charges. Free charges are loosely bound to atomic nuclei and free to move around in response to external fields. When the switch of an incandescent light bulb is flipped, it is free charges that zoom through the tungsten filament to make it glow. Likewise, it is free charges that carry current in transistors and any other electrical device. In contrast, bound charges are tightly bound to atomic nuclei; such a charge can only move within a few ångströms around the atom to which it is bound by a strong Coulombic attraction. Bound charges arise as a result of atomic polarizability. When an atom is subjected to an electric field, its positively charged nucleus and negatively charged electronic cloud slightly separate, resulting in an electric dipole moment. At the microscopic scale, these positive and negative charges typically cancel out with each other. A notable exception to this rule is illustrated in Fig. 5.1. Here, a material is subjected to a constant electric field, resulting in atomic polarization. In the bulk of the material, the positive ends of atomic dipoles cancel out with the negative ends of

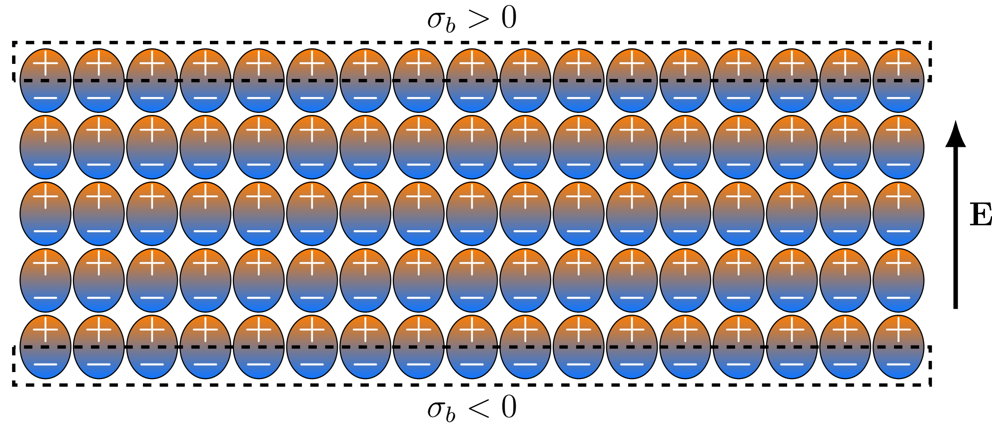


Figure 5.1 – Illustration of the formation of surface bound charges σ_b in a material. In the presence of a constant external electric field **E**, atoms (ellipses) polarize. The positively charged nuclei and negatively charged electrons (orange and blue halves of the ellipses, respectively) are subject to opposite Lorentz forces, leading to charge separation on the atomic level. These charges cancel out in the bulk of the material but not on its surfaces (dashed rectangles).

dipoles in the neighbouring row of atoms. On the surfaces, this cancellation is not possible, resulting in finite surface bound charges. As David Griffiths stressed in his textbook on electrodynamics, these are "perfectly genuine accumulations of charge."

Free charges respond more readily to external fields and voltages than bound charges. As a result, typically, it is by manipulating free charges that electronic devices work. In this chapter, I introduce bound-charge engineering (BCE), a general method to engineer bound charges to our advantage in materials and devices. BCE is especially useful for devices based on 1D and 2D materials. These low-dimensional materials have received tremendous attention from the scientific and technological research communities in recent years owing in part to their prospects toward future generations of nanoelectronic devices [Chapter 3]. Problematically, these emerging low-dimensional materials have a relatively small number of free charges compared to their 3D counterparts. Among other things, free charges play an important role in the screening effect: they tend to redistribute themselves to create sharp electric potential profiles in materials and devices. The greater the number of free charges, the stronger the screening effect, the sharper the electric potential profile [59].

Poor screening presents a major problem to the performance and scalability of any nan-

odevice requiring sharp potential interfaces, e.g. TFETs [Sec. 3.1], CSFETs [Appendix A.1.2], small-diameter short-channel GAA FETs, and novel memory devices containing many interfaces. Screening is typically strengthened by increasing the chemical doping concentration, which is in practice limited by the solid solubility limits of dopants [46] and bandgap narrowing [43–45]. In many industrial applications, the limit of how much semiconductors can be doped in nanoscale FETs without seriously disrupting material integrity and electronic properties is nearly reached [14]. In simulations, one often considers doping levels higher than what is feasible experimentally in order to generate sharp potential profiles across junctions. BCE partially solves the screening problem, making it a timely and meaningful discovery. My collaborators and I first reported BCE in Ref. [1].

In Sec. 5.1, BCE is established by fundamental principles of electrostatics: Gauss's law and dielectric polarization. It is shown how a surface bound charge can be engineered on the interface between a semiconductor like silicon and a low-κ oxide like silicon dioxide to strengthen screening in the semiconductor. For FET applications, to avoid compromising gate control, a high-κ oxide inside the gated region is combined with a low-κ oxide outside to achieve strong screening. In Sec. 5.2, this theoretical picture is confirmed by atomistic quantum transport simulations. BCE significantly improves charge screening in low-dimensional nanodevices without compromise to gate control. In Sec. 5.3, BCE is applied to silicon NW TFETs, thereby greatly increasing the overall device performance (in terms of on-state current and STS), thanks to the sharp band-to-band tunnelling junctions that BCE enables. It can be concluded that BCE paves a way toward improved low-power transistors.

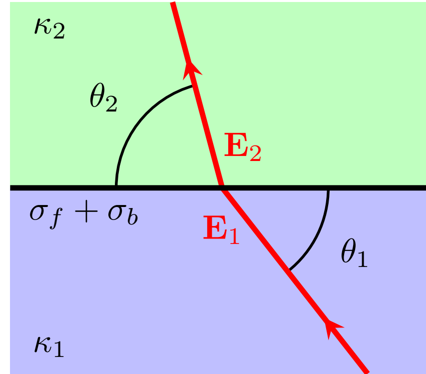


Figure 5.2 – Diagram of the bending of an external electric field and the formation of surface bound charge on the interface between two distinct linear dielectrics. When an electric field $\mathbf{E}_1$ (red arrow) is incident to the interface (black line) between two distinct linear dielectric media (green and blue regions) with relative permittivities $\kappa_{1,2}$, a surface bound charge σ_b forms on the interface.

5.1 Harnessing Surface Bound Charges

5.1.1 Basic Electrostatics

To explain BCE, consider the interface between a semiconductor with relative permittivity κ_1 and an oxide with relative permittivity κ_2, as shown in Fig. 5.2. The relative permittivity is a dimensionless material parameter that quantifies the extent to which a material polarizes in response to an external electric field; at the microscopic level, it is related to atomic polarizability. By Gauss's law,

$$(\kappa_2 \mathbf{E}_2 - \kappa_1 \mathbf{E}_1) \cdot \hat{n} = \frac{\sigma_f}{\varepsilon_0} \tag{5.1}$$

$$\implies \kappa_1 E_1 \sin\theta_1 + \frac{\sigma_f}{\varepsilon_0} = \kappa_2 E_2 \sin\theta_2 \,, \tag{5.2}$$

where $\hat{n}$ is the normal unit vector from the semiconductor to the oxide, σ_f is the surface free charge on the interface between the two media, $\mathbf{E}_{1,2}$ are the electric fields in the two media, and $\theta_{1,2}$ are the angles they make with the interface. Assuming the materials are

linear dielectrics, the polarization density in medium i near the interface is [47]

$$\mathbf{P}_i = (\kappa_i - 1)\,\varepsilon_0 \mathbf{E}_i\,. \tag{5.3}$$

Furthermore, the surface bound charge on the interface is

$$\sigma_b = \mathbf{P}_1 \cdot \hat{n} - \mathbf{P}_2 \cdot \hat{n}\,. \tag{5.4}$$

It follows from Eq. 5.2 that the total surface charge on the interface is

$$\sigma_f + \sigma_b = \varepsilon_0 E_1 \frac{\kappa_1 - \kappa_2}{\kappa_2} \sin\theta_1 + \frac{\sigma_f}{\kappa_2}\,. \tag{5.5}$$

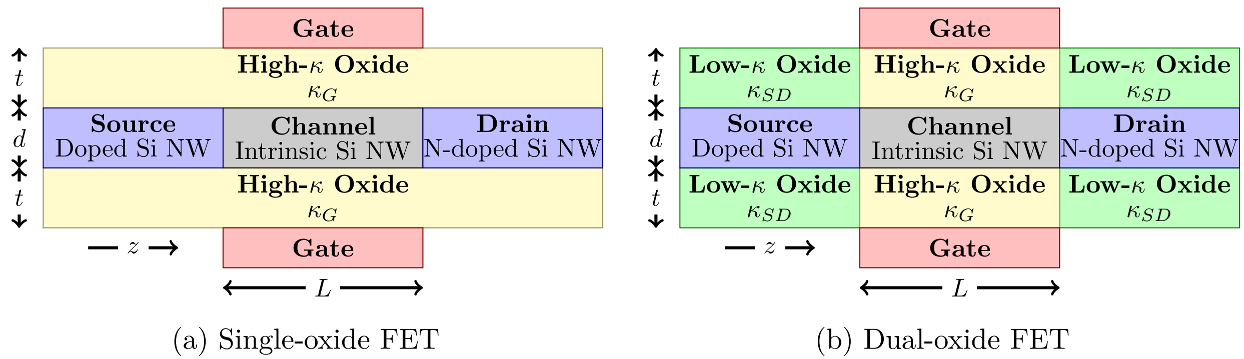
Before explaining how to apply BCE to FETs, a few observations can be made. First, Eq. 5.5 is established by some of the most basic principles of electrostatics like Gauss's law. It can thus be expected to be widely applicable to materials and devices. Although the rest of this thesis focuses on applications of BCE to FETs, the scope of its applications is potentially wider. Second, Eq. 5.5 assumes that the materials are *linear* dielectrics. However, it could be generalized to the nonlinear case. Third, Eq. 5.5 assumes that the permittivities are isotropic; its generalization to the anistropic case would need to consider the permittivities along the direction of the fields. Fourth, Eq. 5.5 assumes the permittivity to be a step function: it sharply transits from being equal to κ_1 in medium 1 to being equal to κ_2 in medium 2. Appendix B.2 shows that BCE is not affected by this assumption, both qualitatively and quantitatively. Fifth, Eq. 5.5 assumes that charges are *surface* charges. Formally, surface charges do not exist. Indeed, quantum mechanically, it is not possible to confine charge to a plane with *zero* volume. Instead, charge should be described by its volumetric density, as will be done in the 3D simulations described in the next sections. The volume bound charge could be computed through $\rho_b(\mathbf{r}) = -\boldsymbol{\nabla} \cdot \mathbf{P}(\mathbf{r})$, where $\mathbf{P}(\mathbf{r})$ is the polarization density. However, this would be more complex than the derivation of the surface bound charge σ_b presented in this section. Furthermore, in the system described by Fig. 5.2, the "surface" bound charge is confined to the region of space between the semiconductor and oxide, which has a thickness

on the order of the interatomic distance. For the intended applications of BCE, this distance is negligible. Thus, the concept of a surface charge is a useful theoretical tool to conceptualize BCE.

5.1.2 Application to FETs

Equation 5.5 suggests that under an external electric field, the surface bound charge σ_b on a semiconductor can be modulated by the permittivities $\kappa_{1,2}$. In an n-type FET biased such that the gate potential is lower than the source (drain) potential (i.e. in the off state), some electric field lines must flow from the source (drain) to the gate. Thus, positive charges accumulate over a depletion length ℓ in the source (drain) near its interface with the channel to screen the negative channel charge. If the oxide surrounding the source has a relative permittivity κ_{SD} smaller than that of the semiconductor source (for silicon, the relative permittivity is $\kappa_{\text{Si}} = 11.7$), then, by Eq. 5.5, the surface charge on the oxide–semiconductor interface is positive and maximized in the limit of small κ_{SD}, and thereby enhances the screening of the channel charge (i.e. ℓ is lower). This effect is expected to be stronger in FETs with larger surface-area-to-volume ratios, such as in thin cylindrical silicon NW FETs. Similar arguments can be applied for p-type FETs. In short, using a low-κ source and drain oxide enables the formation of a bound charge of the same sign as the depletion region free charge, thereby resulting in low depletion length.

High-κ oxides are needed and used in modern FETs for high gate control [Eq. 2.19], which may suggest the single-oxide FET design illustrated in Fig. 5.3a. Instead, the analysis of the last paragraph suggests combining two oxides: a high-κ one around the channel of the FET and a low-κ one around the source and drain for strong screening of the channel charge. This realizes an FET whose cross section and 3D structure are shown in Figs. 5.3b and 5.4, respectively. Thanks to surface bound charges in the source and drain near interfaces with the channel, which strengthen screening, the dual-oxide FET has a potential profile that is much sharper at the source–channel and drain–channel interfaces. Parenthetically, lowering the drain oxide permittivity also weakens DIBL, which scales with the channel-to-drain



(a) Single-oxide FET (b) Dual-oxide FET

Figure 5.3 – Schematics of the n-type cylindrical gate-all-around (GAA) silicon NW FETs investigated in this chapter. With its dual-oxide design, the FET in (b) is assisted by bound-charge engineering (BCE). Cross sections through the NWs' axis of rotational symmetry are shown.

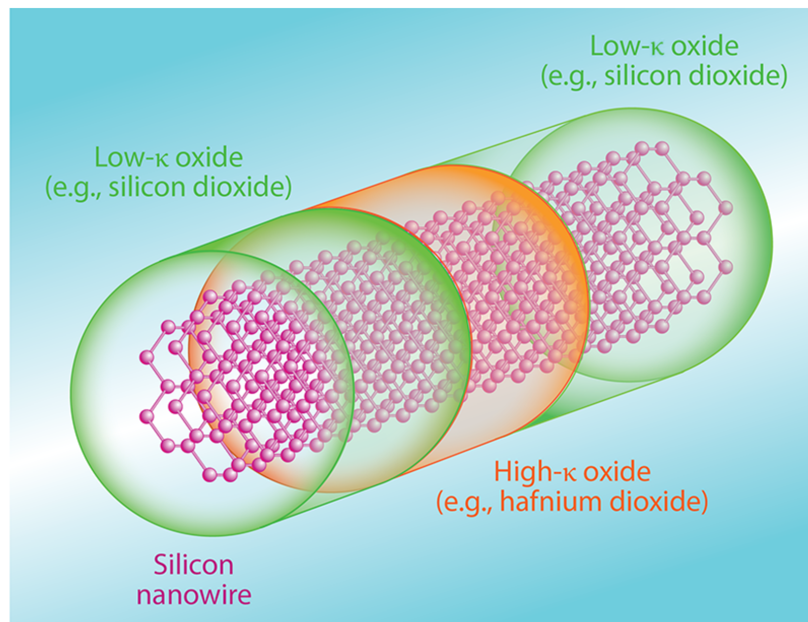


Figure 5.4 – 3D diagram of an FET assisted by BCE. The FET is composed of a silicon NW (purple lattice). The channel is surrounded by a high-κ oxide (orange-shaded region) while the source and drain are surrounded by a low-κ oxide (green-shaded regions). This figure is reproduced with permission from Ref. [10]; credit: APS/Carin Cain.

capacitance.

These arguments are general and make essentially no assumption about the nature of the device, thereby making BCE widely applicable to FETs. In principle, BCE is applicable to:

- a wide range of materials, e.g. silicon NWs [48, 49], silicon–germanium [239–241], III–V semiconductors [127, 242, 243], few-layers transition metal dichalcogenides [52, 244–247],

few-layers black phosphorus [4, 50, 51, 248, 249], graphene nanoribbons [55, 250], and carbon nanotubes [53, 251];

- novel device geometries, e.g. GAA stacked nanosheets [56];
- emerging devices, especially those that benefit from short depletion lengths, such as TFETs and CSFETs [Appendix A.1.2].

5.2 Atomistic Quantum Transport Simulations

Having established the general physics of BCE, its potential benefit to charge screening is concretely tested within the NEGF–TB atomistic quantum transport package `Nanoskim 2.0` described in Sec. 4.4. As a vehicle for this study, cylindrical GAA silicon NW FETs are considered [Figs. 5.3 and 5.4]. Semiconductor NWs, especially silicon NWs, are ideal CMOS-compatible materials for GAA transistors [205, 252]. When combined with high-κ gate dielectrics, GAA silicon NW MOSFETs display excellent gate control and performance [253, 254]. In this chapter, NWs grown in [110] and with diameters of $d = 1, 2, 3$ and 4 nm are considered. The atomic structures of their unit cells are shown in Fig. 4.4, and their electronic properties are investigated in Sec. 4.2.2. For all FETs investigated in this chapter, the surrounding oxide thickness is $t = 2$ nm and the channel length is $L = 9.98$ nm. The gate metal's workfunction is taken to be 0.5 eV greater than the silicon NW's electron affinity. The oxides that surround the NWs are treated in a continuum approximation, being described solely by their permittivities. This approximation is justified by the very large bandgaps of oxides and their resulting inertness in charge transport. Furthermore, experimental reports show that given proper device fabrication, the permittivities of nanometric thin-film oxides are close to the bulk permittivities [255, 256], indicating that such thin-film oxides can be treated in a continuum approximation.

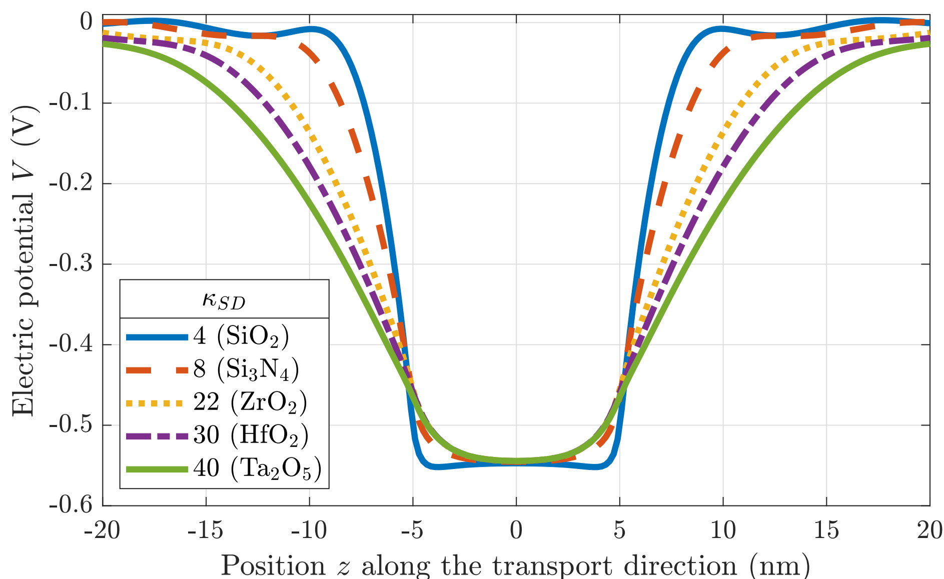


Figure 5.5 – Electric potential V as a function of position z along the transport direction from the channel centre in n-type $d = 1$ nm-wide cylindrical GAA silicon NW MOSFETs [Fig. 5.3b], as obtained from NEGF–TB simulations. The source and drain doping concentration is $N_D = 2 \times 10^{20}$ cm^{-3} and the gate oxide is HfO_2 ($\kappa_G = 30$). Several source and drain oxides (with different relative permittivities κ_{SD}) are compared. The drain and gate voltages are, respectively, set to $V_{DS} = 0$ V and $V_{GS} = 0.5$ V.

5.2.1 Depletion Length

Figure 5.5 shows the calculated electric potential profiles at equilibrium of n-type cylindrical GAA silicon NW MOSFETs with HfO_2 gate oxide ($\kappa_G = 30$) and various source and drain oxides [Fig. 5.3b]. A clear trend is seen: NWs with BCE, namely those with lower κ_{SD}, exhibit sharper potential profiles at the junctions that the channel makes with the source and drain. The sharpness of the potential profile is described by a lengthscale ℓ over which the potential decays to 0; it can be quantified using an exponential fit, as illustrated in Fig. 5.6. The potential decay lengthscale ℓ is commensurate to the depletion length (or screening length) of the channel–drain interface. Although ℓ is not formally equal to the depletion length, for simplicity, it will be referred to as such throughout this chapter. In Fig. 5.7, the extracted ℓ as a function of κ_{SD} is shown. The depletion length is seen to increase monotonically with the permittivity of the source and drain oxide, as expected from Eq. 5.5. Furthermore, $\ell \to 0$ as $\kappa_{SD} \to 0$; this is a consequence of the $\frac{1}{\kappa_2}$ singularity in

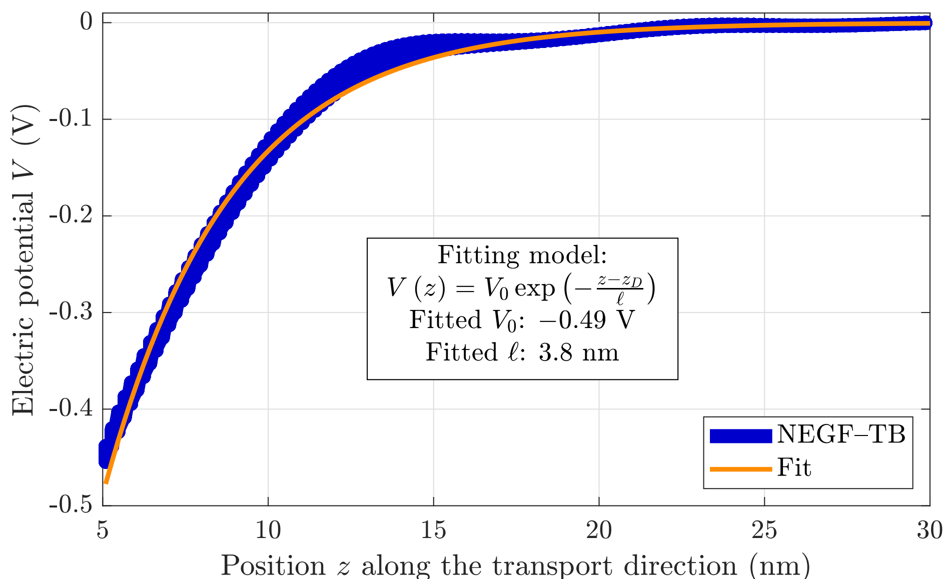
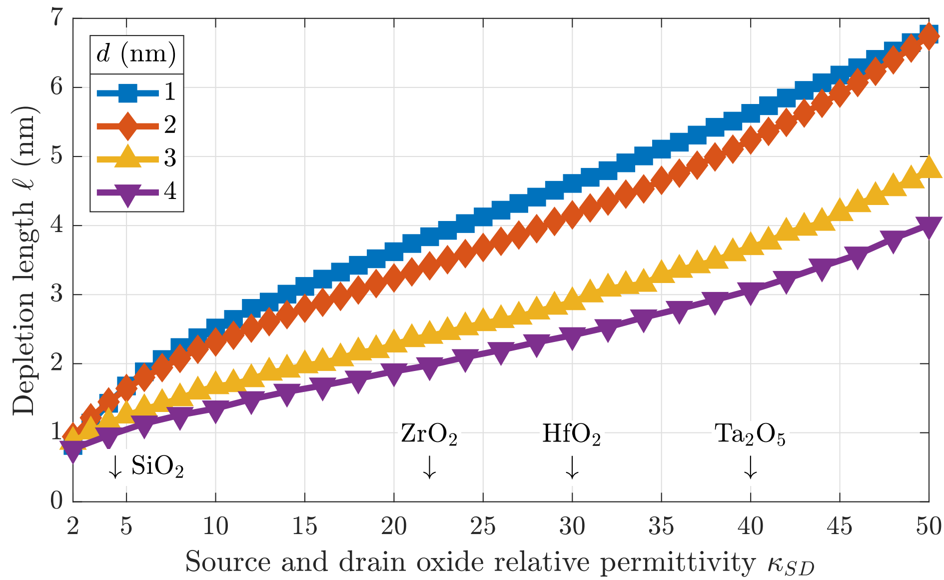


Figure 5.6 – Illustration of the extraction of the depletion length ℓ in the case of $\kappa_{SD} = 22$ from Fig. 5.5. The term $z_D = 4.99$ nm refers to the position of the channel–drain interface.

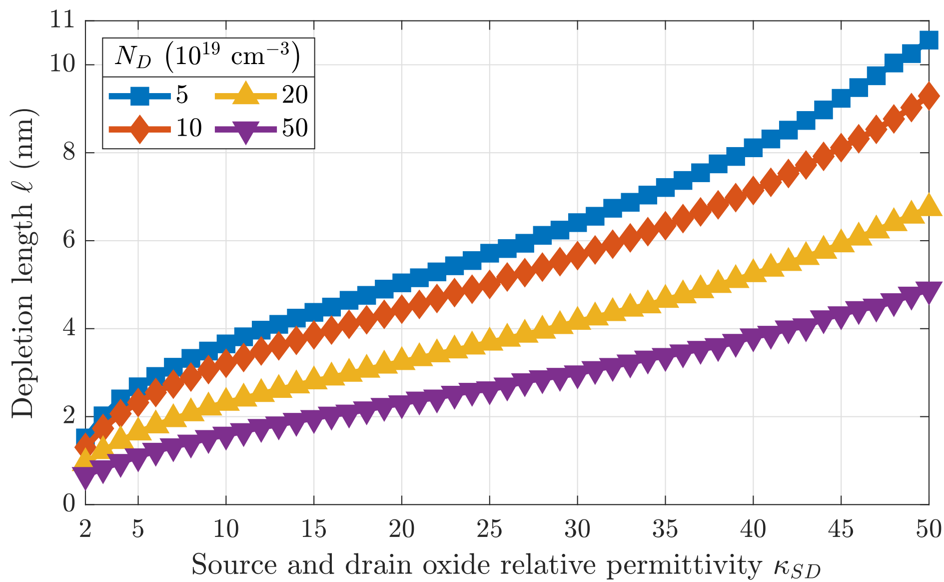
Eq. 5.5.

In Fig. 5.7a, the ℓ–κ_{SD} relationship is calculated for NW diameters d ranging from 1 nm to 4 nm. Broadly speaking, thicker NWs display stronger screening as expected from their larger densities of free charges. Furthermore, BCE provides greater improvement to screening in thinner NWs, for which the relative contribution of surface bound charges is higher, as expected from their larger surface-area-to-volume ratio. Indeed, consider a BCE-assisted device with SiO_2 as the source and drain oxide ($\kappa_{SD} \approx 4$) and a single-oxide device with HfO_2 as the source and drain oxide ($\kappa_{SD} = 30$). Comparing these devices in the case of $d = 4$ nm, BCE reduces the depletion length by a factor of $\frac{\ell(\kappa_{SD}=30)}{\ell(\kappa_{SD}=4)} = 2.51$. For a thinner NW with $d = 3$ nm ($d = 2$ nm, $d = 1$ nm), this ratio is 2.53 (2.87, 3.19).

In Fig. 5.7b, BCE is benchmarked for various source and drain doping concentrations N_D. Expectedly, devices with lower N_D have smaller depletion lengths. However, by similar arguments as before, BCE is found to be more effective at lower doping concentrations. BCE thereby provides an alternative to achieve stronger charge screening in applications where higher chemical doping is infeasible. Overall, the NEGF–TB simulation results quantitatively substantiate the BCE picture. Furthermore, BCE is found to be more effective in thinner



(a) Various d, fixed $N_D = 2 \times 10^{20}$ cm^{-3}



(b) Various N_D, fixed $d = 2$ nm

Figure 5.7 – Depletion length ℓ as a function of source and drain oxide permittivity κ_{SD} in n-type cylindrical GAA silicon NW MOSFETs [Fig. 5.3b], as obtained from the method of Fig. 5.6. The gate oxide is HfO_2 ($\kappa_G = 30$). The applied drain and gate voltages are set to $V_{DS} = 0$ V and $V_{GS} = 0.5$ V, respectively. In (a), the source and drain doping concentration is fixed to $N_D = 2 \times 10^{20}$ cm^{-3} and several NW diameters d are investigated. In (b), the NW diameter is fixed to $d = 2$ nm and several N_D are investigated.

NWs with lower doping concentrations. Therefore, BCE would be especially useful to future generations of silicon NW transistors, for which d is expected to shrink to a few nanometres.

5.2.2 Electric Field

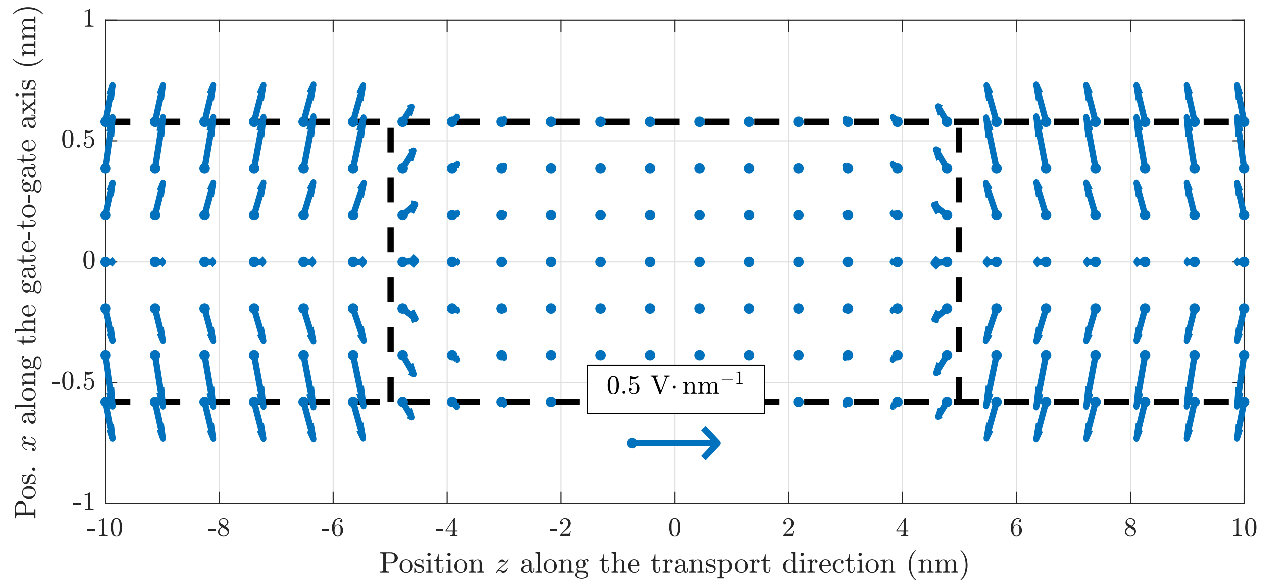
The enhanced screening that BCE provides to NW transistors relies on the formation of a surface bound charge on the interface between the semiconductor NW and the low-κ oxide surrounding the source and drain, as illustrated in Fig. 5.2. Ignoring surface free charge, Eq. 5.5 shows that the surface bound charge is non-zero when the electric field makes a non-zero angle with the semiconductor–oxide interface. Figure 5.8 shows the electric field inside some of the silicon NW MOSFETs investigated in this chapter. The electric field is obtained by numerical differentiation of the electric potential V obtained in the NEGF–TB simulations through

$$\mathbf{E}(\mathbf{r}) = -\boldsymbol{\nabla} V(\mathbf{r}) \ . \tag{5.6}$$

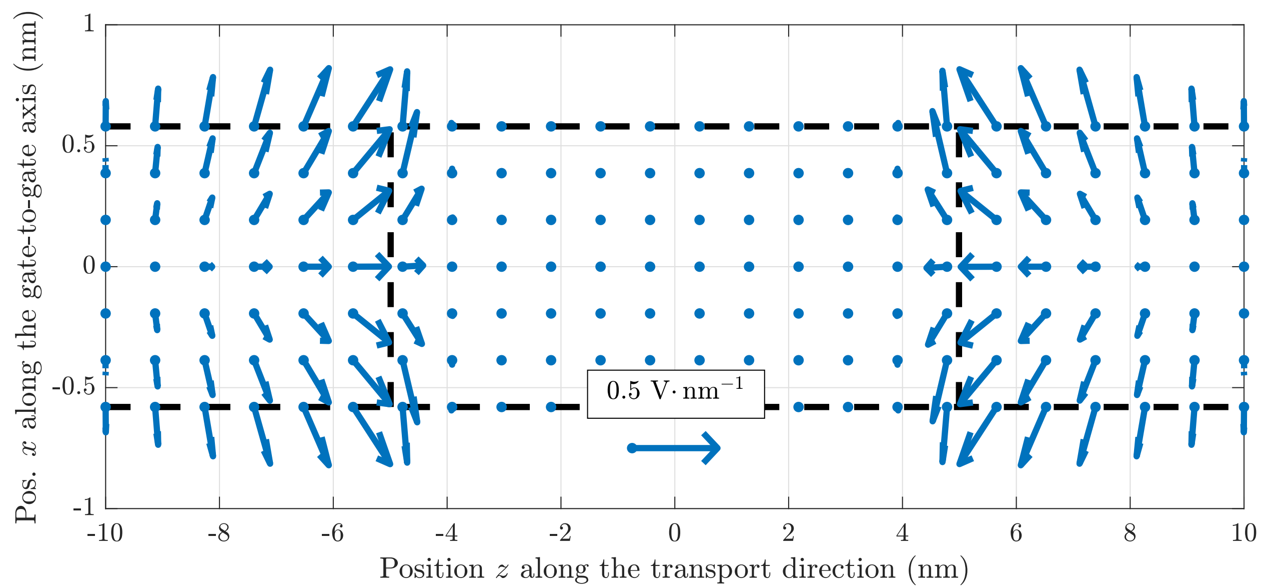
Importantly, Fig. 5.8 shows that both in cases of high κ_{SD} and low κ_{SD}, the electric field is non-parallel to the source-oxide–semiconductor and drain-oxide–semiconductor interfaces, thereby leading to surface bound charges. This confirms the relevance of Eq. 5.5 to NW transistors, and thus corroborates the proposed BCE. Parenthetically, the electric fields in the devices simulated in Fig. 5.8 are quite different, both in terms of the magnitude and direction. Thus, bound-charge engineering could perhaps alternatively be termed "electric-flux engineering."

5.2.3 Free and Bound Charges

Two types of charges are considered in the simulations of this chapter: a volume free charge ρ_f and a surface bound charge σ_b. In the NEGF–TB simulation scheme, ρ_f is explicitly considered and computed through Eq. 4.26. On the other hand, σ_b is implicitly captured in the simulations through the discontinuity of the electric field $\mathbf{E}$ at surfaces of discontinuity of



(a) $\kappa_{SD} = 30$ (HfO_2)



(b) $\kappa_{SD} = 4$ (SiO_2)

Figure 5.8 – Electric field **E** (blue arrows) in n-type $d = 1$ nm-wide cylindrical GAA silicon NW MOSFETs [Fig. 5.3b], as obtained from NEGF–TB simulations. Cross sections through the $y = 0$ plane are shown. The source and drain doping concentration is $N_D = 2 \times 10^{20}$ cm^{-3} and the gate oxide is HfO_2 ($\kappa_G = 30$). The applied drain and gate voltages are set to $V_{DS} = 0$ V and $V_{GS} = 0.5$ V, respectively. The horizontal (vertical) black-dashed lines indicate the boundaries of the NW (transistor channel).

the relative permittivity κ.

To compare the contributions of ρ_f and σ_b to the total charge involved in screening, they may be converted to a line free charge $\lambda_f(z)$ and a line bound charge $\lambda_b(z)$, respectively. Specifically, $\lambda_{f,b}(z)$ are the charges per unit length on a cross section of the NW normal to the source-to-drain axis z. Specifically, given that ρ_f is explicitly computed in the simulations, the line free charge is

$$\lambda_f(z) = \int_0^{2\pi}\int_0^{\frac{d}{2}} \rho_f(z,r,\theta)\, r dr d\theta\,, \tag{5.7}$$

where r is the distance from the z axis and θ is the angle around the z axis. Note that (z,r,θ) defines a cylindrical coordinates system. To compute the line bound charge, the electric field $\mathbf{E}$ is computed through the method described in Sec. 5.2.2. The line bound charge is then given by integration of Eq. 5.5, namely

$$\lambda_b(z) = \frac{d}{2}\int_0^{2\pi} \varepsilon_0 \frac{\kappa_{\mathrm{Si}} - \kappa_{SD}}{\kappa_{SD}} \mathbf{E}\left(z, r = \frac{d}{2}, \theta\right) \cdot \hat{\mathbf{n}}(\theta)\, d\theta\,, \tag{5.8}$$

where $\hat{\mathbf{n}}(\theta)$ is the unit vector normal to the NW surface; its Cartesian coordinates are $(\cos\theta, \sin\theta, 0)$. The free and bound line charges are compared in Figs. 5.9a and 5.9b for two of the MOSFETs investigated in this chapter. Far from the source–channel interface, $\lambda_b \approx 0$ since there, the electric field $\mathbf{E}$ is parallel to the NW–oxide interface. Near the source–channel interface, $\mathbf{E}$ has a non-zero component normal to the NW surface [Fig. 5.8]; correspondingly, $\lambda_b \neq 0$. In the case of HfO_2 (SiO_2) source and drain oxide ($\kappa_{SD} = 30$) ($\kappa_{SD} = 4$), the bound charge λ_b is negative (positive), as can be seen in Fig. 5.9a [Fig. 5.9b], and thereby weakens (strengthens) screening.

Furthermore, the total free and bound charges Q_f and Q_b involved in the screening of the channel charge can be obtained by integration of Eqs. 5.7 and 5.8, namely

$$Q_{f,b} = \int_{-\infty}^{-\frac{L}{2}} \lambda_{f,b}(z)\, dz\,, \tag{5.9}$$

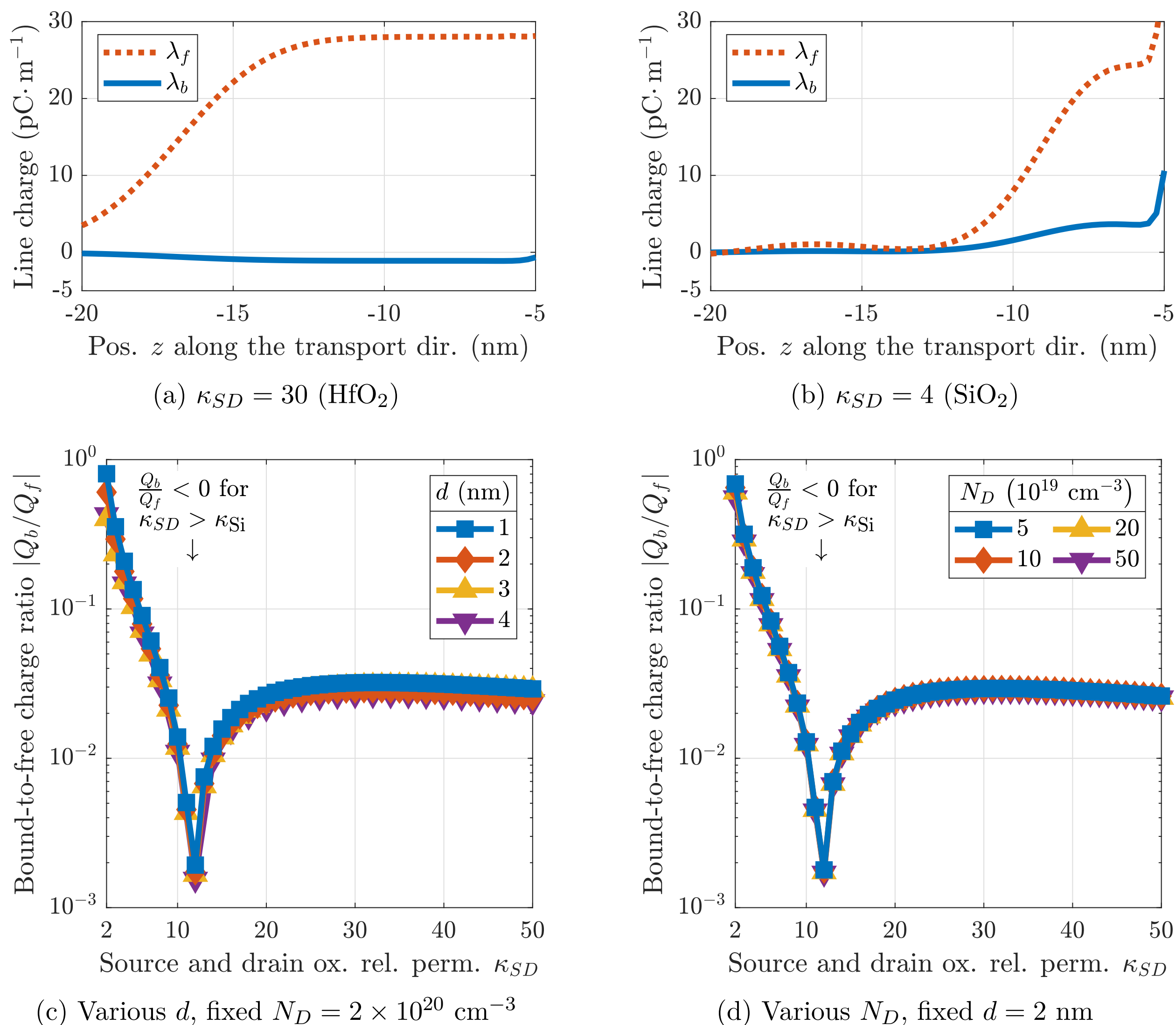


Figure 5.9 – (a)–(b) Free and bound line charges λ_f and λ_b in n-type $d = 1$ nm-wide cylindrical GAA silicon NW MOSFETs [Fig. 5.3b], as obtained from NEGF–TB simulations. The source and drain doping concentration is $N_D = 2 \times 10^{20}$ cm^{-3} and the gate oxide is HfO_2 ($\kappa_G = 30$). The applied drain and gate voltages are set to $V_{DS} = 0$ V and $V_{GS} = 0.5$ V, respectively. (c)–(d) Ratio of total bound charge to total free charge in the devices investigated in Fig. 5.7, namely for silicon NW MOSFETs with various source and drain oxide relative permittivities κ_{SD}, various NW diameters d, and various source and drain doping concentrations N_D.

The ratio of total bound charge to total free charge $\frac{Q_b}{Q_f}$ for all silicon NW MOSFETs investigated in Fig. 5.7 is shown in Fig. 5.9. In the limit of low κ_{SD}, bound charges play a major role in screening. Indeed, Q_b is comparable to Q_f for $\kappa_{SD} = 2$. Furthermore, at fixed κ_{SD}, the ratio $\frac{Q_b}{Q_f}$ is higher for thinner NWs [Fig. 5.9c] and for lower source and drain doping concentration [Fig. 5.9d], in accordance with the results of Sec. 5.2.1, showing BCE to be more effective for those types of NW devices.

It can be noted that the ratio $\frac{Q_b}{Q_f}$ approaches a constant in the limit of high κ_{SD}. Thus, in this limit, this ratio alone would predict the depletion length ℓ to be independent of κ_{SD}. This would contradict the results shown in Fig. 5.7, where it is seen that ℓ increases with κ_{SD} for all sampled values of κ_{SD}. To resolve this contradiction, two effects may play an important role. The first is the effect of the gate. The capacitive coupling between the gate and the regions of the source and drain that neighbour the channel is higher for higher κ_{SD}. When this capacitive coupling is higher, the electric potential in those regions is closer to the gate electric potential [Eq. 2.19] and to the channel potential. As a result, ℓ increases with κ_{SD}. For the second, consider a degenerate 3D electron gas. In such a system, the Thomas–Fermi screening length is given by [42]

$$\ell = \sqrt{\frac{\varepsilon_0 h^2}{4 m_g^\star q^2 k_F} \sqrt{\kappa_g}}\,, \tag{5.10}$$

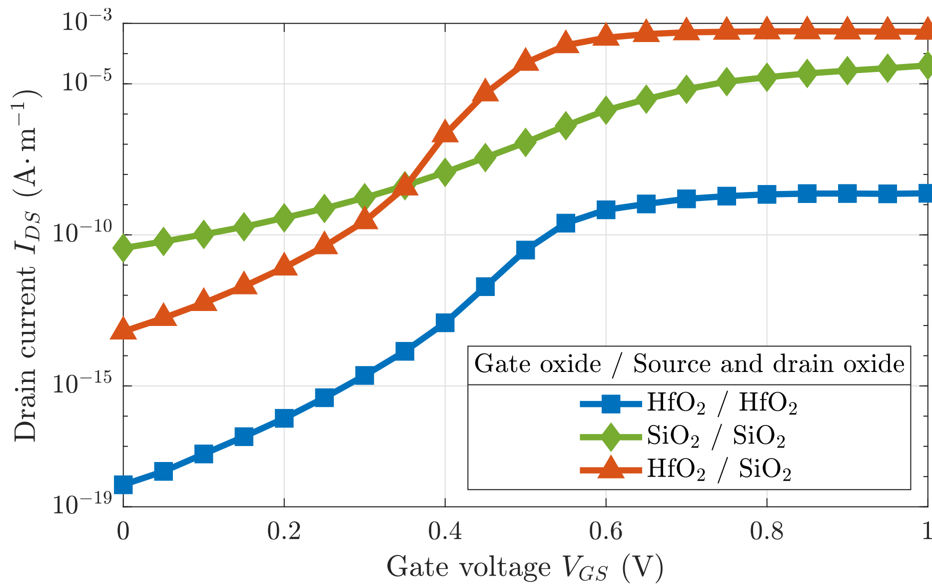
where $m_g^\star$, k_F, and κ_g are, respectively, the EM, Fermi wavenumber, and relative permittivity of the gas. Equation 5.10 accurately describes screening in degenerately doped, bulk silicon. On the other hand, in a silicon NW, the term κ_g should reflect not only the permittivity of silicon but also that of the oxide surrounding the NW. For thin NWs, κ_g should approach the relative permittivity of the surrounding oxide since most electric field lines involved in charge screening lie within the oxide. This argument would predict that $\ell \propto \sqrt{\kappa_{SD}}$ in the limit of large κ_{SD}.

5.3 High-Performance Tunnel FETs

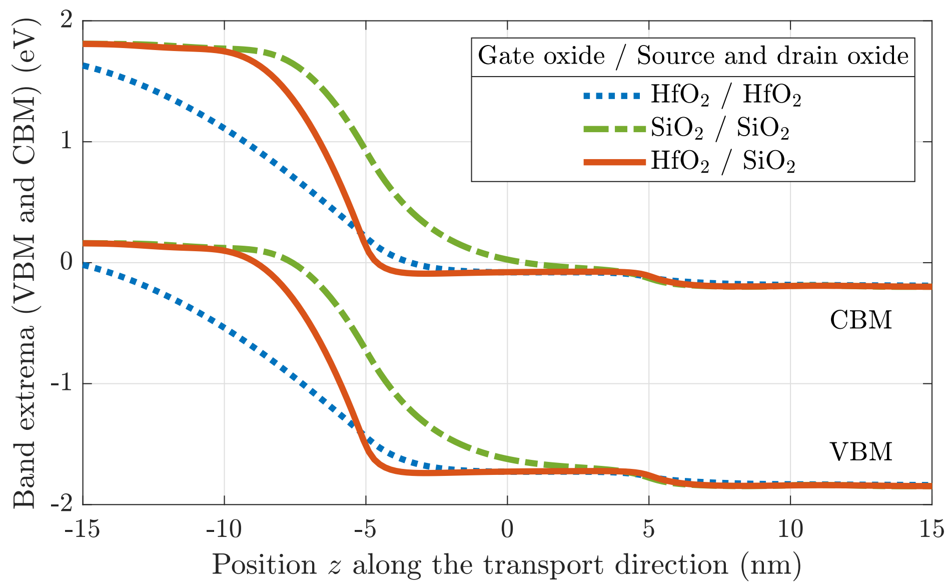
Having understood the effect of BCE on electrostatic screening, its benefits to device performance are investigated, using TFETs as examples. As discussed in Sec. 3.1, charge transport in TFETs involves band-to-band tunnelling from the source to the channel; the tunnelling probability decays exponentially with the tunnelling length [Eq. 3.7]. TFETs with sharper junctions thus exhibit higher on-state current, which enables their practical applications at higher clock frequencies [Eqs. 2.26 and 2.27].

Three different n-type $d = 2$ nm-wide cylindrical GAA silicon NW TFETs [Fig. 5.3b] are simulated in the NEGF–TB algorithm, as shown in Fig. 5.10. In the first TFET, the silicon NW is entirely surrounded by the high-κ oxide HfO_2 ($\kappa_G = \kappa_{SD} = 30$). In the second TFET, the silicon NW is entirely surrounded by the low-κ oxide SiO_2 ($\kappa_G = \kappa_{SD} = 3.8$). The third is a BCE-assisted TFET: the silicon NW is surrounded by HfO_2 and SiO_2 as in Fig. 5.3b ($\kappa_G = 30$ and $\kappa_{SD} = 3.8$). Parenthetically, the channel of the devices is 9.98 nm long and includes 1 664 silicon atoms. The entire simulation box (source and drain buffers and channel) is up to 117 nm long and includes up to 19 456 silicon atoms.

The calculated transfer characteristics at $V_{DS} = 50$ mV of these devices are shown in Fig. 5.10a. Several observations can be made. First, overall, the best performing device is the BCE-assisted TFET (red curve). Its low-κ source oxide significantly sharpens the potential profile at the source–channel interface [Fig. 5.10b], which reduces tunnelling length, thus increasing the on-state current by orders of magnitude compared to the HfO_2-only TFET (blue curve). Second, the SiO_2-only TFET (green curve) has an on-state current orders of magnitude higher than that of the HfO_2-only TFET (blue curve), although it is smaller than that of the BCE-assisted TFET, which has greater gate control due to HfO_2 surrounding the channel. Third, the calculated band diagram in Fig. 5.10b shows that the tunnelling length in the BCE-assisted TFET is 4.3 nm (red curve), to be compared to 11.9 nm for the HfO_2-only TFET (blue curve). This is the physical reason for the much higher on-state current in the BCE-assisted TFET. Fourth, devices employing HfO_2 as the gate oxide (red



(a) Transfer characteristics



(b) Band diagrams

Figure 5.10 – (a) Transfer characteristics of n-type $d = 2$ nm-wide cylindrical GAA silicon NW TFETs [Fig. 5.3b] at a temperature of $T = 300$ K and a drain voltage of $V_{DS} = 50$ mV, as obtained from NEGF–TB simulations. The donor doping concentration of the n-doped source is $N_D = 2 \times 10^{20}$ cm^{-3}; the acceptor doping concentration of the p-doped drain is $N_A = 3 \times 10^{20}$ cm^{-3}. The current is normalized by the NW perimeter. (b) Band diagrams of the TFETs at a gate voltage of $V_{GS} = 0.6$ V.

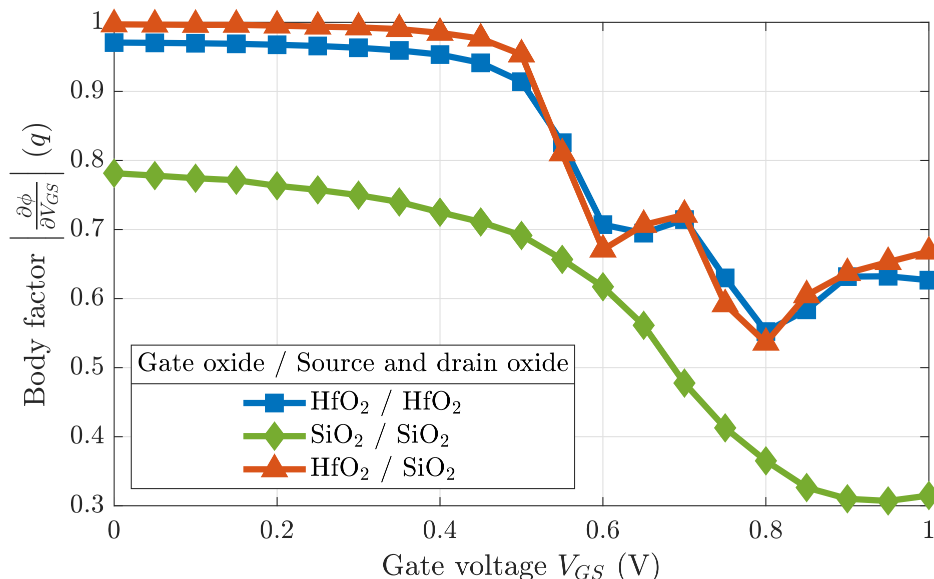


Figure 5.11 – Body factor of the TFETs of Fig. 5.10 at $T = 300$ K and $V_{DS} = 50$ mV as a function of gate voltage V_{GS}. The body factor is obtained by numerical differentiation of the NEGF–TB simulation data and expressed in units of elementary charge q.

Gate oxide	Source and drain oxide	I_{on} (A$\cdot$ m^{-1})	S_{av} (mV$\cdot$ dec^{-1})	$\left\|\frac{\partial\phi}{\partial V_{GS}}\right\|_{\mathrm{av}}$ (q)
HfO_2	HfO_2	6.8×10^{-10}	58.0	1.08
SiO_2	SiO_2	1.3×10^{-6}	112	1.40
HfO_2	SiO_2	3.4×10^{-4}	52.6	1.06

Table 5.1 – On-state current I_{on}, average STS S_{av}, and average body factor $\left|\frac{\partial\phi}{\partial V_{GS}}\right|_{\mathrm{av}}$ of the TFETs of Fig. 5.10. The on state (off state) is defined to be reached when the gate voltage is $V_{GS} = 0.6$ V ($V_{GS} = 0.2$ V) and the drain voltage is $V_{DS} = 50$ mV. The averages of the STS and the body factor are taken between the off state and the on state.

and green curves) have much lower STS than that which employs SiO_2 (blue curve). The much-improved STS is due to the lower body factor that a high-κ gate oxide confers, as shown in Fig. 5.11. The values of on-state current, average STS, and average body factor for the three devices are reported in Table 5.1. It can be noticed that the BCE-assisted TFET has an average STS 9.3% lower than the HfO_2-only TFET, but its average body factor is only 1.9% lower. Therefore, the benefits of BCE to STS do not merely come from an improved body factor but also from an improved transport factor. The BCE-assisted TFET has large transport factor thanks to its tunnelling length being more sensitive to the channel barrier energy ϕ than the HfO_2-only TFET [Eq. 3.7].

The benefits of BCE to TFETs are considerable. According to Eq. 2.26, keeping the load capacitance and power supply voltage fixed, the high on-state current of the BCE-assisted TFET enables its operation at a clock frequency 250 times higher than the SiO_2-only TFET and 500 000 times higher than the HfO_2-only TFET. Still, the on-state current of the BCE-assisted TFET investigated here lags that of a MOSFET with comparable geometry and doping by several orders of magnitude. This is due to the relatively large bandgap energy (1.65 eV) of the silicon NW [Fig. 4.6], which significantly reduces the BTBT probability [Eq. 3.7]. As investigated in Sec. 3.1, state-of-the-art TFETs achieve high on-state current by the use of (1) narrow-bandgap materials and (2) tunnelling heterojunctions. In this thesis, BCE was not directly applied to heterojunction-based TFETs, since TB model parameters do not exist for these systems. In principle, BCE is applicable to such TFETs and can be viewed as a third, separate means to achieve high on-state current in TFETs.

5.4 Conclusion

In this chapter, I introduced BCE, a novel and relatively simple scheme where oxides of different permittivities are combined to greatly sharpen the potential profiles of NW transistors at interfaces along the transport direction, with no compromise to gate control. As an example, when BCE is applied to silicon NW TFETs, it very effectively reduces tunnelling length while maintaining high gate control, leading to much-improved device characteristics regarding on-state current and STS. Since both the high-κ hafnium dioxide and low-κ silicon dioxide are well-known, industry-standard oxides (and there are many other material choices, as shown in Fig. 5.7), BCE-assisted FETs should be realizable experimentally. Recent reports of ultra-low-κ dielectrics, e.g. amorphous boron nitride with $\kappa \approx 1.5$ [257], offer especially promising prospects. BCE may thereby offer significant performance gains with relatively few changes to CMOS processes and assembly. The principle of BCE is general; it could also be applied to other low-dimensional materials, notably 2D semiconductors as well as semiconducting NWs, nanotubes, and nanosheets. BCE could also be applied to scale down

low-dimensional nanodevices, especially systems with a large number of junctions such as NAND memory and the CSFET. Finally, given the very general physics underlying BCE, it could find applications in areas of research beyond nanoelectronics in which strong charge screening might be desired, e.g. molecular electronics, electrochemistry, and material sciences.

Chapter 6

The Surface Potential of BCE-Assisted FETs

Physics-based compact models for electronic devices play two important and distinct roles. First, the kernel of the model serves as a compact mathematical description of our understanding of the device. This conceptual model helps us to interpret experiments and detailed simulations and guides our thinking in device research and development. Second, the complete model with extensions [...] enables circuit design.

— M. S. Lundstrom & D. A. Antoniadis, 2014 [217]

The three pillars of research in materials and device physics are experiment, computation, and theory. Experiments are the ultimate judges of scientific ideas. Only a well-controlled experiment can prove or disprove, beyond doubt, a hypothesis. For novel nanoscale transistors, experiments typically require costly cleanrooms, nanofabrication equipment, and human effort. As a result, the initial stages of research on modern innovations in nanoelectronics typically involve computation and theory. Even the largest corporations of the semiconductor industry heavily rely on these tools. First-principles computational tools, such as the NEGF–DFT algorithm [Sec. 4.1], lend themselves particularly well to the investigation of systems that are not deeply understood. To further our understanding of such systems, theoretical models need to be developed. Ideally, such a model provides analytical formulas that relate the system's internal properties to adjustable, external parameters. In the field of IC design,

such an analytical model may be referred to as a "compact model" and may be required to simulate and develop semiconductor circuits.

Chapter 5 is a good example of the relevance of computations to nanoelectronics research. There, the novel concept of BCE and the hypothesis that it benefits TFETs and other FETs is tested against computations. The natural next step in the research process is to develop an analytical model for BCE-assisted FETs. How do the depletion lengths and surface bound charges depend on the applied voltages and permittivities of oxides? To answer these questions, an important task is to devise a physics-based analytical surface potential model for FETs with distinct gate and spacer oxides; here, "spacer oxide" is synonymous to "source and/or drain oxide." The surface potential is simply the potential on the interface between the semiconductor and oxides; it is an essential FET modelling tool since charge transport occurs within a narrow inversion layer close to that interface in FETs. Existing MOSFET models such as those based on scaling theory [58, 258–261] ignore the source and drain depletion regions as well as spacer oxides. Several TFET models consider these depletion regions but only in the case of identical spacer and gate oxides [64, 262–264]. In this chapter, I aim to fill this void in the literature by reporting a surface potential model for cylindrical GAA FETs assisted by BCE without any assumption about the permittivities of the spacer and gate oxides. I first reported this model in Ref. [2].

In Sec. 6.1, the scaling theory of FETs is introduced. In scaling theory, the surface potential of an FET is expressed in term of a so-called "natural length," a lengthscale related to the FET's electrostatic properties. In Sec. 6.2, a BCE-assisted tri-oxide cylindrical GAA silicon NW FET is introduced. A surface potential model for this FET is derived. This model is based on scaling theory. In Sec. 6.3, the physics of BCE is elucidated: in much the same way that semiconductors can be doped by chemical dopants, they can effectively be doped by surface bound charges. Finally, in Sec. 6.4, the validity of the analytical surface potential model is verified against atomistic quantum transport simulations.

6.1 The Scaling Theory of FETs

How much can the gate length (or channel length) L of an FET be shortened before the gate loses its electrostatic grip on the channel? Since a high FET body factor is required for low processor power consumption [Chapter 2], answering this question has been a crucial theoretical endeavour toward denser nanoelectronics. In 1992, Yan, Ourmazd, and Lee introduced one of the first physics-based, qualitatively clear, and quantitatively accurate models to answer this question: the scaling theory of the silicon-on-insulator (SOI) MOSFET [58].

An SOI MOSFET is a single-gated, planar MOSFET; its structure is shown in Fig. 2.1. By examining the various symmetries of SOI MOSFETs, Yan *et al.* showed that in the channel, Poisson's equation can be expressed as

$$\frac{\partial^2 \psi(z)}{\partial z^2} - \frac{\psi(z)}{\lambda_{\text{SOI}}^2} = 0\,, \tag{6.1}$$

where $\psi(z)$ is the surface potential and where

$$\lambda_{\text{SOI}} = \sqrt{\frac{\kappa_{\text{Si}}}{\kappa_G} t_{\text{Si}} t} \tag{6.2}$$

is the natural length for SOI FETs. The parameters κ_{Si}, κ_G, t_{Si}, and t are, respectively, the silicon relative permittivity, the gate oxide relative permittivity, the silicon layer thickness, and the oxide layer thickness. This scaling theory was later generalized to double-gated (DG) [258] and cylindrical GAA [259] MOSFETs, for which the natural lengths were, respectively, shown to be

$$\lambda_{\text{DG}} = \sqrt{\frac{\kappa_{\text{Si}}}{2\kappa_G}\left(1 + \frac{\kappa_G t_{\text{Si}}}{4\kappa_{\text{Si}} t}\right) t_{\text{Si}} t}\,, \tag{6.3}$$

$$\lambda_{\text{GAA}} = \frac{t_{\text{Si}}}{4}\sqrt{1 + 2\frac{\kappa_{\text{Si}}}{\kappa_G}\ln\left(1 + \frac{2t}{t_{\text{Si}}}\right)}\,. \tag{6.4}$$

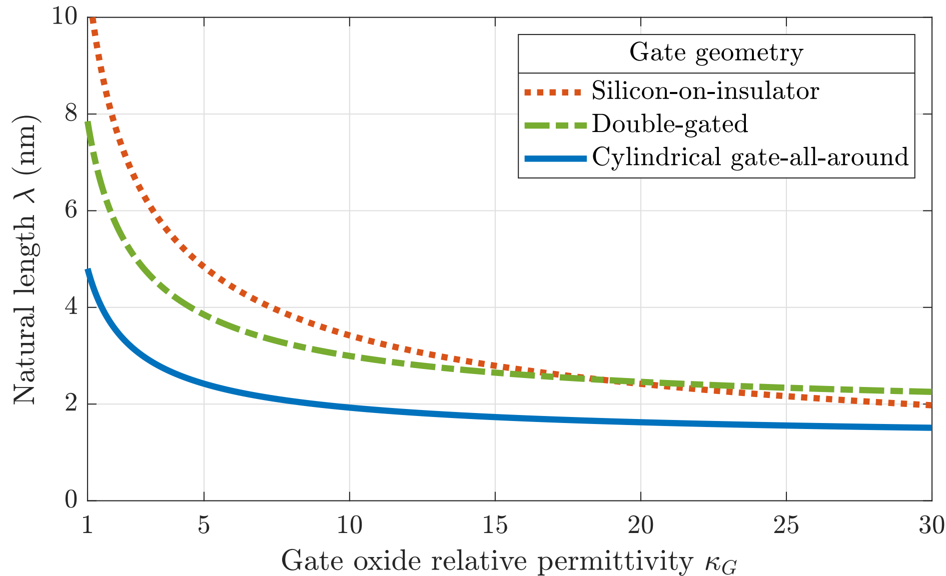


Figure 6.1 – Natural length λ of single-gated, double-gated, and GAA FETs as a function of gate oxide relative permittivity κ_G. The silicon relative permittivity is set to $\kappa_{\text{Si}} = 11.7$, the silicon layer thickness is set to $t_{\text{Si}} = 5$ nm, and the oxide layer thickness is set to $t = 2$ nm.

Note that in the cylindrical GAA geometry, in which the FET channel is composed of a silicon NW, $t_{\text{Si}} = 2R$, where R is the NW radius.

The natural lengths for these three gate geometries are graphed in Fig. 6.1. Suzuki *et al.* showed that the dimensionless parameter $\alpha = \frac{L}{\lambda}$ (where λ is chosen within Eqs. 6.2–6.4 according to gate geometry) uniquely sets the STS of a MOSFET, in principle [258]. Furthermore, the STS is a strictly decreasing function of α. These facts guided much of the innovations of the semiconductor industry in the past decades, especially since the onset of the power dissipation problem in the mid 2000s. Specifically, to avoid unmanageable processor power consumption, FETs on ICs have been downscaled in such a way that α has remained roughly constant over successive generations of FETs, thereby enabling high-performance FETs with nanometric L. To make this possible, the other key physical dimensions of FETs, t_{Si} and t, have also been downscaled. Furthermore, SiO_2, a low-κ oxide, was progressively replaced as the gate oxide by high-κ dielectrics [Chapter 1]. Finally, gate geometries changed: SOI FETs were replaced by FinFETs. The IRDS™predicts that FinFETs will be replaced by GAA FETs around the year 2025 [14]. Indeed, the GAA gate geometry offers smaller natural lengths than the SOI and DG geometries [Fig. 6.1].

In addition, Eq. 6.1 is a simple ordinary differential equation, which may be solved to obtain the surface potential. In the next section, scaling theory is generalized to FETs with multiple oxides (such as BCE-assisted FETs) to obtain a surface potential model in these devices.

6.2 Potential Model and Depletion Length

As a vehicle for this study, consider the cylindrical GAA silicon NW FET illustrated in Fig. 6.2a. The silicon NW (with relative permittivity κ_{Si} and radius R) is partitioned into source, channel (with length L), and drain with doping concentrations N_S, N_C, and N_D, respectively. By convention, throughout this chapter, positive (negative) doping concentrations denote n-doping (p-doping); for example, an n-type TFET would have $N_S < 0$, $N_C \approx 0$, and $N_D > 0$ (p–i–n). The source, channel, and drain are surrounded by oxides with thickness t and relative permittivities κ_S, κ_G, and κ_D, respectively. The source (drain) depletion length is L_S (L_D). Finally, the electric potential is set to ϕ_S, ϕ_G, and ϕ_D at the source, gate, and drain contacts, respectively; the values of $\phi_{S,G,D}$ are set by the doping concentrations $N_{S,D}$, the silicon NW electron affinity, the gate metal workfunction, and the applied drain and gate voltages [Sec. 2.2.3].

Depending on the values of permittivities, this FET is assisted by BCE. Since this device has cylindrical symmetry, the surface bound charge on the interface between the source depletion region and the source oxide (regions A and B in Fig. 6.2b) can be expressed as [Eq. 5.5]

$$\sigma_b = \varepsilon_0 \frac{\kappa_{\mathrm{Si}} - \kappa_S}{\kappa_S} \mathbf{E}_{\mathrm{Si}} \cdot \hat{r}\,, \tag{6.5}$$

where $\mathbf{E}_{\mathrm{Si}}$ is the external electric field inside the silicon NW and $\hat{r}$ is the unit vector pointing radially outward [Fig. 6.2a]. This surface bound charge arises as a result of the fact that the silicon and oxide polarize to different degrees under the field $\mathbf{E}_{\mathrm{Si}}$, thereby explaining the factor of $(\kappa_{\mathrm{Si}} - \kappa_S)$ in Eq. 6.5.

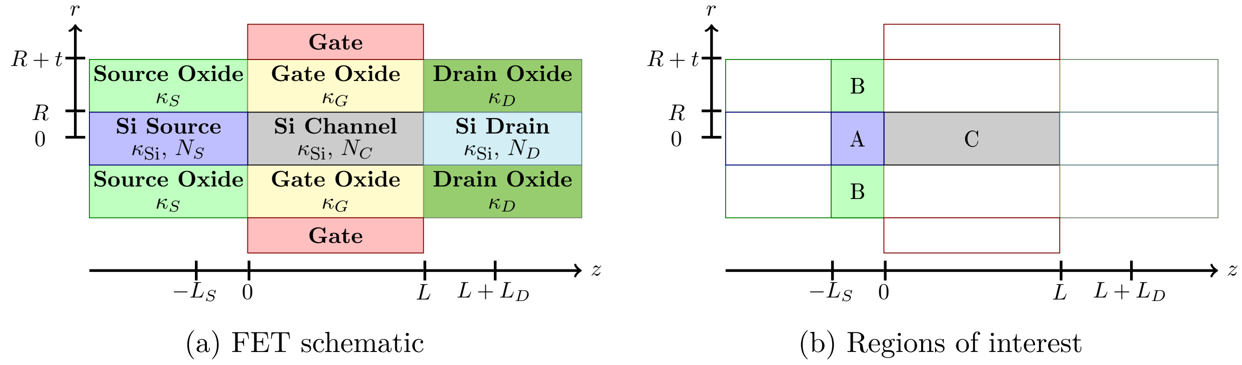


Figure 6.2 – (a) Schematic of the cylindrical GAA silicon NW FET investigated in this chapter. A cross section through the NW's axis of rotational symmetry is shown. Interface positions are defined in a cylindrical coordinates system (r, z). (b) Three regions of interest, A, B, and C, are distinguished. A is the source depletion region. B is the source oxide surrounding A. C is the channel.

In this section, the surface potential in the device of Fig. 6.2a is derived. The potential $C(r, z)$ in the channel is derived in Ref. [259] under the approximations of full depletion and potential quadratic in r [58, 259]:

$$C(r, z) = \left[C_S \sinh\left(\frac{L - z}{\lambda}\right) + C_D \sinh\left(\frac{z}{\lambda}\right) + \frac{q N_C \lambda^2}{\kappa_{\text{Si}} \varepsilon_0} \right] \left(1 - \frac{r^2}{4\lambda^2}\right) + \phi_G \,, \tag{6.6}$$

where $\lambda = \frac{R}{2}\sqrt{1 + 2\frac{\kappa_{\text{Si}}}{\kappa_G} \ln\left(1 + \frac{t}{R}\right)}$ is the natural length for cylindrical GAA FETs [Eq. 6.4] and where C_S and C_D are constants to be determined by the boundary conditions.

Similarly, one may approximate the potential in region A [Fig. 6.2b] to be quadratic in r:

$$A(r, z) = a_0(z) + a_1(z)\, r + a_2(z)\, r^2 \,. \tag{6.7}$$

In region B, one may approximate the potential as

$$B(r, z) = b_0(z) + b_n(z)\, r^n + b_l(z) \ln\left(\frac{R + t}{r}\right) , \tag{6.8}$$

where $n > 0$ is an adjustable parameter. Here, the r-independent term is required for potential continuity. Furthermore, the potential in an infinite cylindrical capacitor is proportional to $\ln r$ [47]; the third term in Eq. 6.8 thus captures electrostatics arising as a result of cylindrical

geometry. Finally, the term proportional to r^n provides a degree of freedom (i.e. n) to improve agreement of the model with numerical simulations [Sec. 6.4].

Several boundary conditions may be imposed on the potentials $A(r,z)$ and $B(r,z)$. First, to avoid an infinite line charge at $r=0$,

$$\left.\frac{\partial A}{\partial r}\right|_{r=0} = 0\,. \tag{6.9}$$

Second, to ensure that no electric field traverses the oxide through its exterior surface,

$$\left.\frac{\partial B}{\partial r}\right|_{r=R+t} = 0\,. \tag{6.10}$$

Third, to ensure that the potential and displacement field are continuous on the semiconductor–oxide interface,

$$A(R,z) = B(R,z)\,, \tag{6.11}$$

$$\kappa_{\mathrm{Si}}\left.\frac{\partial A}{\partial r}\right|_{r=R} = \kappa_S\left.\frac{\partial B}{\partial r}\right|_{r=R}\,. \tag{6.12}$$

These four boundary conditions constrain $A(r,z)$ and $B(r,z)$ to

$$A(r,z) = a_0(z) + a_2(z)\,r^2\,, \tag{6.13}$$

$$B(r,z) = a_0(z) + k_0 a_2(z) + k_n a_2(z)\,r^n + k_l a_2(z)\ln\left(\frac{R+t}{r}\right)\,, \tag{6.14}$$

where $k_l = 2\frac{\kappa_{\mathrm{Si}}}{\kappa_S}R^2\left[\left(\frac{R}{R+t}\right)^n - 1\right]^{-1}$, $k_n = \frac{k_l}{n(R+t)^n}$, and $k_0 = R^2 - k_n R^n - k_l\ln\left(1+\frac{t}{R}\right)$.

The potentials $A(r,z)$ and $B(r,z)$ satisfy Poisson's equation, which in the full depletion approximation reads

$$\nabla^2 A(r,z) = -\frac{qN_S}{\kappa_{\mathrm{Si}}\varepsilon_0}\,, \tag{6.15}$$

$$\nabla^2 B(r,z) = 0\,. \tag{6.16}$$

Inserting Eqs. 6.13 and 6.14 into Eqs. 6.15 and 6.16 and evaluating at $r = R$, it can be shown that $a_0(z)$ and $a_2(z)$ are quadratic and constant functions of z, respectively. Thus, the surface potential of the source depletion region is

$$\psi_S(z) = \alpha_S(z + L_S)^2 + \phi_S\,, \tag{6.17}$$

where $-L_S \leq z \leq 0$. Note that continuity of the potential and field at $z = -L_S$ was imposed, namely $\psi_S(-L_S) = \phi_S$ and $\psi'_S(-L_S) = 0$. Likewise, the surface potential of the drain depletion region is

$$\psi_D(z) = \alpha_D(z - L - L_D)^2 + \phi_D\,, \tag{6.18}$$

where $L \leq z \leq L + L_D$. Here,

$$\alpha_{S,D} = -\frac{qN_{S,D}}{2\kappa_{\text{Si}}\varepsilon_0}\left[1 + \frac{2}{n}\frac{\kappa_{S,D}}{\kappa_{\text{Si}}}\frac{(R+t)^n - R^n}{R^n}\right]^{-1}. \tag{6.19}$$

The channel surface potential is

$$\psi_C(z) \coloneqq C(R, z)\,, \tag{6.20}$$

where $0 \leq z \leq L$ [Eq. 6.6]. Finally, for $z \leq -L_S$ ($z \geq L + L_D$), the surface potential is identically equal to ϕ_S (ϕ_D).

Four unknowns remain to be determined: $L_{S,D}$ and $C_{S,D}$. These can be found by imposing continuity of the potential and field at $z = 0$ and $z = L$:

$$\begin{cases} \psi_S(0) = \psi_C(0) \\ \psi_C(L) = \psi_D(L) \\ \psi'_S(0) = \psi'_C(0) \\ \psi'_C(L) = \psi'_D(L) \end{cases}. \tag{6.21}$$

This is a set of coupled transcendental equations that can generally only be solved numerically.

However, for well-tempered FETs, i.e. for $\frac{L}{\lambda} \gg 1$, a closed-form solution exists:

$$L_{S,D} = \lambda \left(-1 + \sqrt{1 + \frac{\overline{\phi_G} - \phi_{S,D}}{\alpha_{S,D}\lambda^2}} \right) , \tag{6.22}$$

$$C_{S,D} = -2\lambda \exp\left(-\frac{L}{\lambda}\right) \alpha_{S,D} L_{S,D} , \tag{6.23}$$

where $\overline{\phi_G} = \phi_G + \frac{qN_C\lambda^2}{\kappa_{\text{Si}}\varepsilon_0} \left(1 - \frac{R^2}{4\lambda^2}\right)$.

Equations 6.6, 6.17–6.20, 6.22, and 6.23 define an analytical surface potential model for the BCE-assisted FET described in Fig. 6.2a. To summarize, the surface potential is:

$$\psi(z) = \begin{cases} \phi_S & \text{for } z \leq -L_S \\ \psi_S(z) & \text{for } -L_S \leq z \leq 0 \\ \psi_C(z) & \text{for } 0 \leq z \leq L \\ \psi_D(z) & \text{for } L \leq z \leq L + L_D \\ \phi_D & \text{for } L + L_D \leq z \end{cases} . \tag{6.24}$$

Note that the effect of radial quantum confinement [214, 265], i.e. the increase of the NW bandgap energy relative to its bulk value as $R \to 0$, is implicitly considered in the model. This is because the model parameters $\phi_{S,G,D}$ depend on the NW's electronic properties and, *a fortiori*, the NW bandgap energy [Sec. 4.2.2].

6.3 The Physics of Bound-Charge Engineering

The surface bound charge σ_b on the interface between regions A and B [Fig. 6.2b] can be computed by inserting $\mathbf{E}_{\text{Si}} = -\left.\boldsymbol{\nabla} A\right|_{r=R}$ [Eq. 6.13] into Eq. 6.5:

$$\sigma_b = qN_S \frac{\frac{R}{2}\left(\frac{1}{\kappa_S} - \frac{1}{\kappa_{\text{Si}}}\right)}{\left[1 + \frac{n}{2}\frac{\kappa_{\text{Si}}}{\kappa_S}\frac{R^n}{(R+t)^n - R^n}\right]} . \tag{6.25}$$

The volumetric concentration of the surface bound charge N_b may then be obtained by multiplying σ_b by the surface-area-to-volume ratio of region A: $N_b = \frac{2\sigma_b}{qR}$. Furthermore, for $\frac{L_S}{\lambda} \gg 1$, Eq. 6.22 simplifies to

$$L_S \approx \sqrt{\frac{2\kappa_{\mathrm{Si}}\varepsilon_0 V_{bi}}{qN}}, \tag{6.26}$$

where $V_{bi} = \phi_S - \overline{\phi_G}$ and $N = N_S + \left(\frac{1}{\kappa_S} - \frac{1}{\kappa_{\mathrm{Si}}}\right)^{-1} N_b$. The terms V_{bi} and N can be thought of as the effective built-in potential between the source and channel and the effective total charge in region A (including both free and bound charges contributions), respectively.

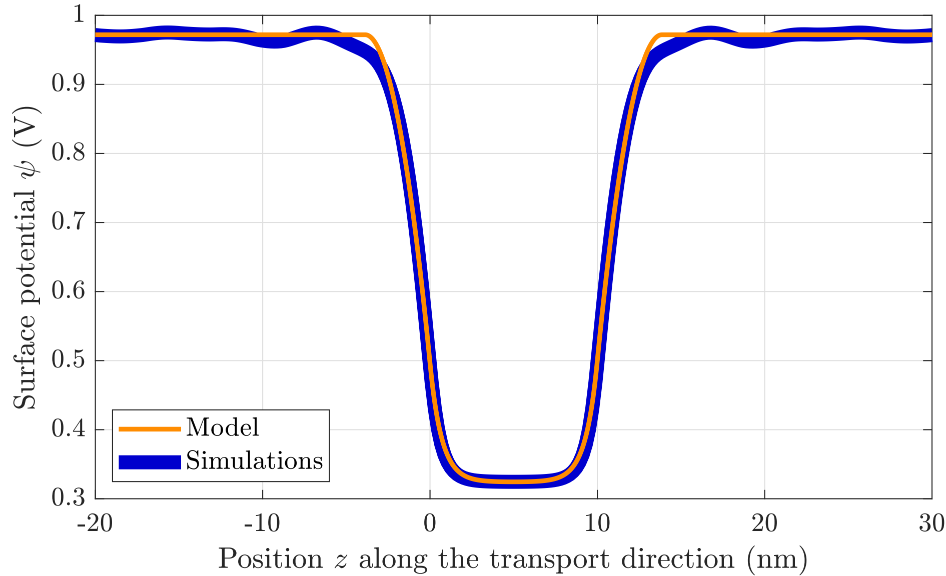
For comparison, the depletion length of a metal–silicon junction is given by [16]

$$L_{\text{metal–Si}} \approx \sqrt{\frac{2\kappa_{\mathrm{Si}}\varepsilon_0 V'_{bi}}{qN'}}, \tag{6.27}$$

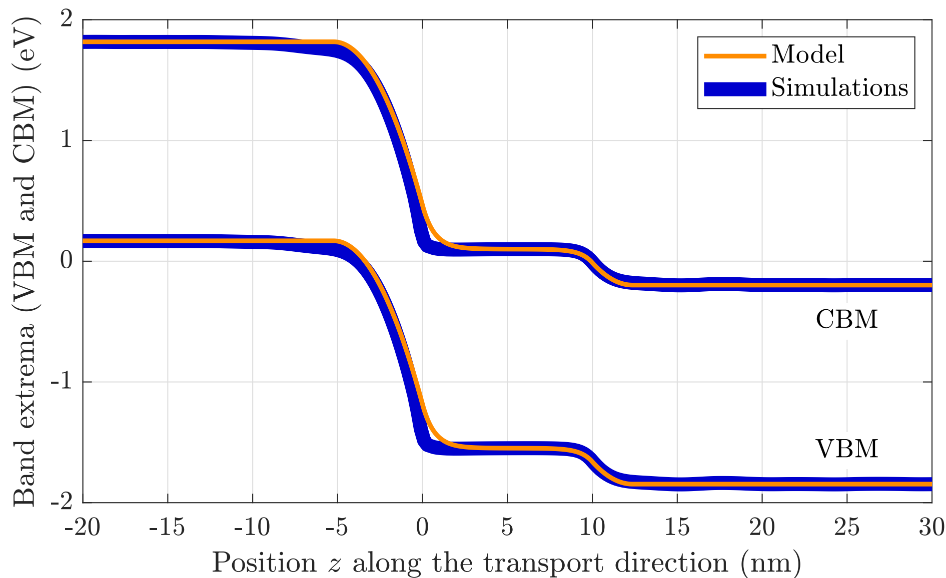
where V'_{bi} is the built-in potential between the metal and silicon and N' is the silicon doping concentration. Thus, Eq. 6.26 establishes a clear analogy between junctions in a BCE-assisted FET and metal–semiconductor junctions, due to the clear similarity between Eqs. 6.26 and 6.27. Furthermore, this establishes a clear parallel between free charges (due to chemical doping) and bound charges (due to polarization). In much the same way that the depletion length of a semiconductor junction can be reduced by increasing the doping concentration, it can also be reduced by engineering a bound charge on its surface.

6.4 Comparison to NEGF–TB Simulations

In this section, to verify the accuracy of the model developed in Sec. 6.2, it is compared to state-of-the-art numerical simulations performed in the NEGF–TB atomistic quantum transport package `Nanoskim 2.0` [Sec. 4.4]. NWs grown in [110] are considered [Fig. 4.4]. Their radii are 0.58 nm, 1.04 nm, and 1.48 nm; their bandgap energies, as computed from the TB model, are, respectively, 2.24 eV, 1.65 eV, and 1.45 eV, which is in agreement with previous theoretical and experimental reports [48, 213, 214]. All simulated devices have



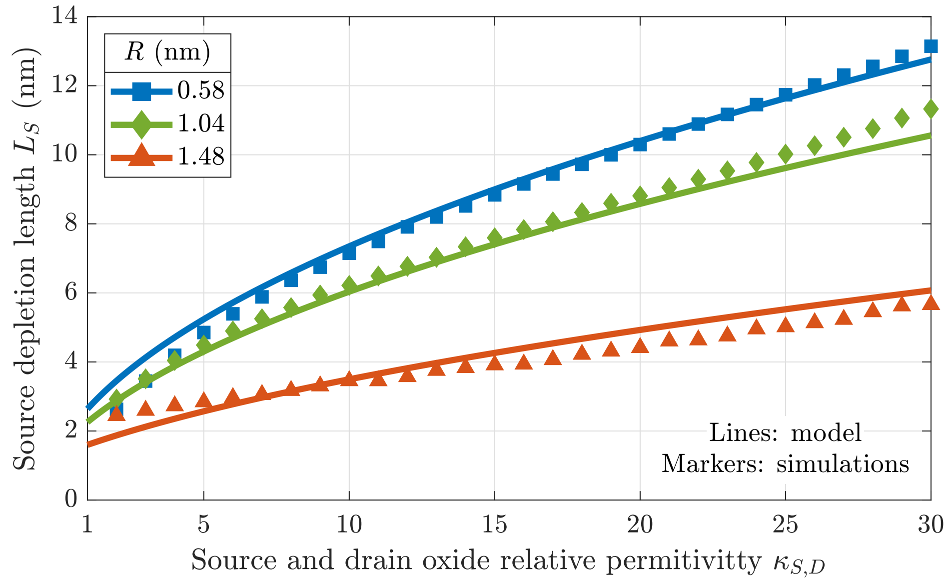
(a) MOSFET



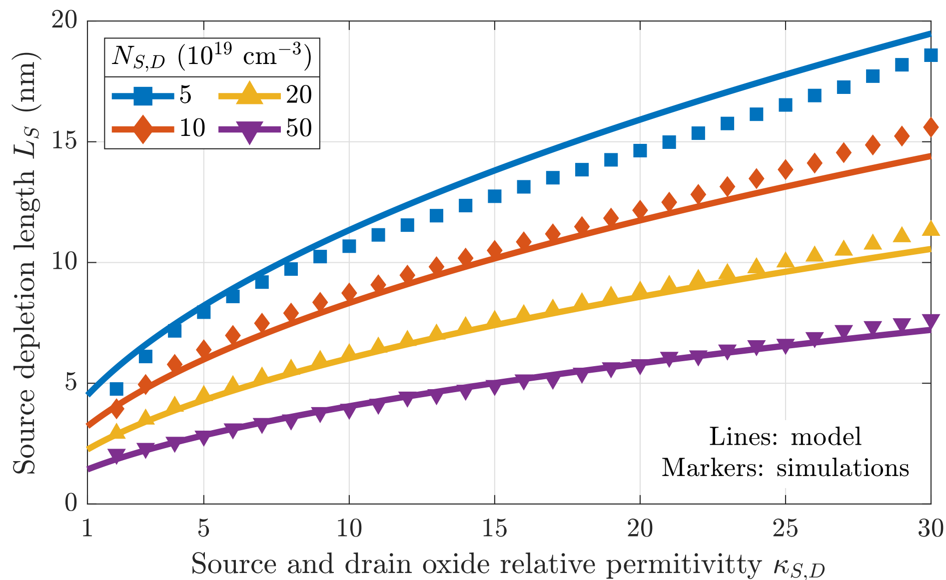
(b) TFET

Figure 6.3 – Comparison of the analytical surface potential model [Sec. 6.2] to NEGF–TB simulations for cylindrical GAA silicon NW FETs [Fig. 6.2a]. (a) Surface potential of a MOSFET with $R = 1$ nm, SiO_2 source and drain oxide ($\kappa_{S,D} = 4$), and $N_{S,D} = 2 \times 10^{20}$ cm^{-3} at gate voltage $V_{GS} = 0.5$ V and drain voltage $V_{DS} = 0$ V. (b) Band diagrams of a TFET with $R = 1$ nm, SiO_2 source and drain oxide ($\kappa_{S,D} = 4$), $N_S = -3 \times 10^{20}$ cm^{-3}, and $N_D = 2 \times 10^{20}$ cm^{-3} at $V_{GS} = 0.4$ V and $V_{DS} = 0.05$ V.

$L = 10$ nm, $N_C = 0$, HfO_2 gate oxide ($\kappa_G = 30$), and $t = 2$ nm. It is assumed that the gate metal's workfunction is 0.5 eV greater than the silicon NW's electron affinity.



(a) Various R, fixed $N_{S,D} = 2 \times 10^{20}$ cm^{-3}



(b) Various $N_{S,D}$, fixed $R = 1$ nm

Figure 6.4 – Comparison of the analytical surface potential model [Sec. 6.2] to NEGF–TB simulations for cylindrical GAA silicon NW FETs [Fig. 6.2a]. (a) Source depletion length for MOSFETs with $N_{S,D} = 2 \times 10^{20}$ cm^{-3} and varying $\kappa_{S,D}$ and R at $V_{GS} = 0.5$ V and $V_{DS} = 0$ V. (b) Source depletion length for MOSFETs with $R = 1$ nm and varying $\kappa_{S,D}$ and $N_{S,D}$ at $V_{GS} = 0.5$ V and $V_{DS} = 0$ V.

The surface potential model of Eq. 6.24 is compared to NEGF–TB simulations in Fig. 6.3a in the case of a MOSFET, and in Fig. 6.3b in the case of a TFET. The source depletion lengths of MOSFETs with various source and drain oxide relative permittivities, various NW radii, and various source and drain doping concentrations, as predicted by Eq. 6.22 and as extracted from NEGF–TB simulations, are compared in Fig. 6.4. The source depletion length L_S is extracted from the simulation data by a method similar to that shown in Fig. 5.6, with the exception that the fitting model is given by $V(z) = a(z + L_S)^2 + \phi_S$, where a and L_S are fitting parameters. Several observations can be made from the model-to-simulations comparisons shown in Figs. 6.3 and 6.4.

- The model exhibits good agreement with simulations for both MOSFETs [Fig. 6.3a] and TFETs [Fig. 6.3b]. However, the simulation data is characterized in the source and drain by Friedel oscillations [42], a quantum effect not captured in the classical surface potential model.
- The predicted depletion lengths [Eq. 6.22] agree well with those extracted from simulations for a wide range of source and drain oxide permittivities [Fig. 6.4], NW radii [Fig. 6.4a], and doping concentrations [Fig. 6.4b].
- The model fits simulations with lower accuracy for the NW with $R = 1.48$ nm [Fig. 6.4a]. This is expected since for thicker NWs, the approximation of full depletion (and, implicitly, the assumption of body depletion) [Eq. 6.15] is less accurate.
- The best-fit value of n [Eq. 6.8] does not have a strong dependence on device type (i.e. MOSFET or TFET) [Fig. 6.3], permittivity [Fig. 6.4], nor doping concentration [Fig. 6.4b], but does depend on R. For all results shown in Fig. 6.3, the value of n for devices with $R = 0.58$ nm ($R = 1.04$ nm, $R = 1.48$ nm) is set to $n = 2.00$ ($n = 2.55$, $n = 1.85$). To understand this R dependence, consider that Eq. 6.8 may be viewed as a finite multipole expansion [47]. Since the charge density in region A is approximately uniform (due to full depletion, body depletion, and uniform doping), the multipole expansion depends solely on device geometry [47].
- The model does not fit simulations to the same degree of accuracy for all values of $N_{S,D}$ [Fig. 6.4b]. Indeed, the thickness of the depletion layer in a MOS capacitor depends on

doping concentration [16]; as a result, the validity of the assumption of body depletion [Eq. 6.15] varies with $N_{S,D}$. This issue could be circumvented by making n a parameter that is $N_{S,D}$-dependent.

- The model predicts L_S with less accuracy for large values of $\kappa_{S,D}$ [Fig. 6.4]. This is due to an implicit discontinuity of the displacement field in the model on the interface between the source and gate oxides. Indeed, in Ref. [259] and Eq. 6.6, the gate oxide potential is implicitly assumed to be that of an infinite cylindrical capacitor [47] and thus z-independent; in contrast, the source oxide potential $A(r,z)$ [Eq. 6.13] is z-dependent. This unphysical discontinuity is large for high $\kappa_{S,D}$ but small for low $\kappa_{S,D}$, the regime in which BCE is typically applied.

6.5 Conclusion

Some regions of FETs, e.g. the source and drain depletion regions, can effectively be doped by surface bound charges on their interfaces with neighbouring low-κ oxides. In turn, depletion lengths can be reduced by BCE. The physics of BCE is very general, thus making BCE applicable to a variety of materials and devices. In this chapter, I derived an analytical surface potential model for cylindrical GAA BCE-assisted FETs and verified it against state-of-the-art numerical simulations. Under realistic approximations, I showed that the depletion length of such a BCE-assisted FET mimics that of a metal–silicon junction. In both cases, the depletion length scales with the square root of the built-in potential and with the inverse of the square root of the screening charge. In the case of the cylindrical GAA FET, a confined structure, this screening charge has contributions from the free charge due to chemical doping and the bound charge due to polarization. Overall, the surface potential model I derived in this chapter provides firm and intuitive theoretical grounds for BCE as well as for future compact modelling and device development. Indeed, since the drain current of a TFET has a strong dependence on tunnelling length (or depletion length), this surface potential model could be used to develop a compact model for BCE-assisted TFETs.

Chapter 7

Reducing Tunnelling Leakage with BCE

Does source-to-drain tunneling limit the ultimate scaling of MOSFETs?

— J. Wang & M. S. Lundstrom, 2002 [67]

Ever since Gordon Moore formulated his eponymous law [12], scientists and engineers have projected various "ultimate" limits of transistor downscaling. In the 1970s, some expected an FET's channel length L not to be scalable below 1 μm due to the limitations of optical lithography at the time [266]. In the 1980s, before high-κ dielectrics became CMOS-compatible materials, tunnelling leakage through the gate stack's silicon dioxide was expected to limit L to 250 nm and above [267]. More recently, the random distribution of dopant atoms in nanoscale FETs has been shown to lead to significant fluctuations of device characteristics that may be intolerable in ultra-scaled FETs [268–270]. Such technological roadblocks, pertaining to limitations of photolithography and materials integration, have been consistently overcome, so far.

In 2020, the IRDS™predicted that state-of-the-art MOSFETs on ICs will have $L > 10$ nm at least until 2034 [14]. However, scaling theory [Sec. 6.1] predicts that a significantly lower L would be possible. Indeed, for a 3 nm-wide cylindrical GAA semiconductor NW FET with 2 nm-thick HfO_2 gate oxide—which is realistic and was realized experimentally [271]—the natural length is around 1 nm. In an FET with $L \approx 1$ nm, the channel is merely ten atoms or so in length. It can thus be expected that such an FET would face scaling limits pertaining

to fundamental physics. For example, errors induced by thermal fluctuations may limit L to 5 nm and above in the von Neumann computing paradigm [272, 273]; this issue could be solved through adiabatic computing [Sec. 2.2.3]. A phenomenon considered by many experts to set the ultimate scaling limit of MOSFETs is direct source-to-drain tunnelling (DSDT) [266, 274]; it is illustrated in Fig. 2.5. The probability of quantum tunnelling through a MOSFET's channel increases exponentially as the channel is shortened [Eq. 2.12]. The working principle of MOSFETs is to modulate current with an energy barrier [Sec. 2.2]. DSDT causes this energy barrier to become permeable, making very short-channel MOSFETs inadequate electrical switches.

In 2000, Kawaura, Sakamoto, and Baba reported experimental evidence of DSDT in MOSFETs with $L = 8$ nm [275]. In 2002, Jing Wang and Mark Lundstrom reported one of the first comprehensive and rigorous theoretical studies of DSDT: "Does source-to-drain tunneling limit the ultimate scaling of MOSFETs?" Supported by NEGF simulations of DG silicon MOSFETs, they concluded that "source-to-drain tunneling does set an ultimate scaling limit [but this] limit is well below 10 nm." [67]. Since then, there have been several proposals to reduce DSDT-induced deterioration of device performance. For MOSFET applications, the conventional wisdom would favour semiconductors with low transport EMs since low masses lead to higher mobility and on-state current [16]. However, the probability of DSDT decreases exponentially with EM. Thus, several authors have proposed to engineer high tunnelling EMs in very short-channel MOSFETs to limit DSDT. This can be achieved through choice of material [276–280], choice of crystal orientation [276, 278–281], and strain [278, 280, 282]. In 2004, Wakabayashi *et al.* showed that DIBL [Fig. 2.4] may exacerbate DSDT in very short-channel MOSFETs; excellent gate control is thus required [283]. In 2012, Maassen and Guo proposed to limit DSDT through localized channel doping [49]. This year, Pandey *et al.* showed how a negative-capacitance gate stack may be used to limit DSDT-induced STS degradation [284]. These reports demonstrated suppression of DSDT down to channels no shorter than about $L = 3$ nm. In 2016, Ilatikhameneh *et al.* showed how a novel gate geometry, which exploits the anisotropy of the EM in few-layers black phosphorus, can be used to reduce DSDT to tolerable levels down to $L = 1.6$ nm in TFETs [285].

The study of FETs with such small channel lengths is not merely a theoretical exercise. Indeed, in 2016, Desai *et al.* experimentally demonstrated a molybdenum disulphide MOSFET with a gate length of just 1 nm; the gate was composed of a metallic carbon nanotube [286]. This year, Liu *et al.* experimentally demonstrated a monolayer molybdenum disulphide vertical FET with a channel length of under 1 nm, corresponding to the thickness of the monolayer.

In this chapter, I propose to use BCE as a means to greatly reduce DSDT in very short-channel MOSFETs. The tunnelling length in DSDT is proportional to the sum of the source depletion length, the channel length, and the drain depletion length. BCE enables control of the depletion lengths. Thus, through BCE, tunnelling leakage can be exponentially reduced. In Sec. 7.1, the DSDT reduction scheme and its application to MOSFETs are described. Long depletion lengths require high-κ source and drain oxides. It is shown how the BCE scheme is compatible with small load capacitance, which usually requires low-κ oxides to achieve low power consumption. In Sec. 7.2, using atomistic quantum transport simulations, the DSDT reduction scheme is tested on silicon NW MOSFETs. It is shown that BCE can reduce DSDT to acceptable levels down to $L = 1.5$ nm. This is achieved not only through longer tunnelling length but also through higher gate control. Overall, the BCE-assisted DSDT reduction scheme is widely applicable to MOSFETs and could be used in conjunction with previously discussed methods, such as EM engineering and localized doping, to design ultra-short-channel MOSFETs.

7.1 Tunnelling and Electrostatics

The potential profile of a MOSFET (such as those shown in Fig. 5.5) may be approximated to be trapezoidal: linear in the source and drain depletion regions and constant in the channel. DSDT consists not only of tunnelling through the channel but also through the depletion regions. The WKB approximation predicts that the tunnelling probabilities of an electron of energy $E = \mu_S$ (where μ_S is the source Fermi level) through the depletion regions' triangular

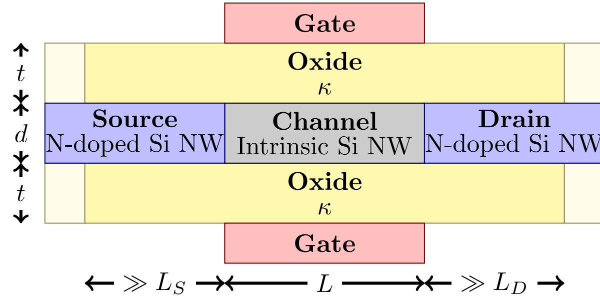


Figure 7.1 – Schematic of the n-type cylindrical GAA silicon NW MOSFET investigated in this chapter. A cross section through the NW's axis of rotational symmetry is shown. For small channel length L, direct source-to-drain tunnelling (DSDT) may occur. An oxide with permittivity κ surrounds the channel as well as the source and drain depletion regions (with lengths $L_{S,D}$). The nature of the oxide located far from these regions (light-yellow regions) does not affect DSDT.

barriers [287] or through the channel's rectangular barrier [Eq. 2.12] are

$$T_{S,C,D} \approx \exp\left(-\frac{a_{S,C,D} L_{S,C,D}}{\hbar}\sqrt{m^\star_{e,t}\phi}\right) , \tag{7.1}$$

where $L_{S,D}$ are the source and drain depletion lengths, $L_C \stackrel{\text{def}}{=} L$ is the channel length, $m^\star_{e,t}$ is the electron tunnelling EM, and $\phi \geq E$ is the channel barrier energy. The dimensionless constants $a_{S,C,D}$ depend on the geometry of the tunnelling barrier. For the channel's rectangular barrier, $a_C = 2\sqrt{2}$ [Eq. 2.12]. For other barrier geometries, this dimensionless constant must be smaller than $2\sqrt{2}$; for example, for the source and drain depletion regions' triangular barriers, $a_{S,D} = \frac{4}{3}\sqrt{2}$ [287]. Overall, the DSDT probability is the product of the tunnelling probabilities $T_{S,C,D}$ through the source, channel, and drain:

$$T_{\text{DSDT}} \approx \exp\left(-\frac{a_S L_S + a_C L + a_D L_D}{\hbar}\sqrt{m^\star_{e,t}\phi}\right) . \tag{7.2}$$

Chapters 5 and 6 show that in FETs, the depletion lengths $L_{S,D}$ can be controlled through the permittivity κ of the oxides surrounding the source and drain, i.e. through BCE. Within the regime of validity of Eq. 6.26, it can be shown that

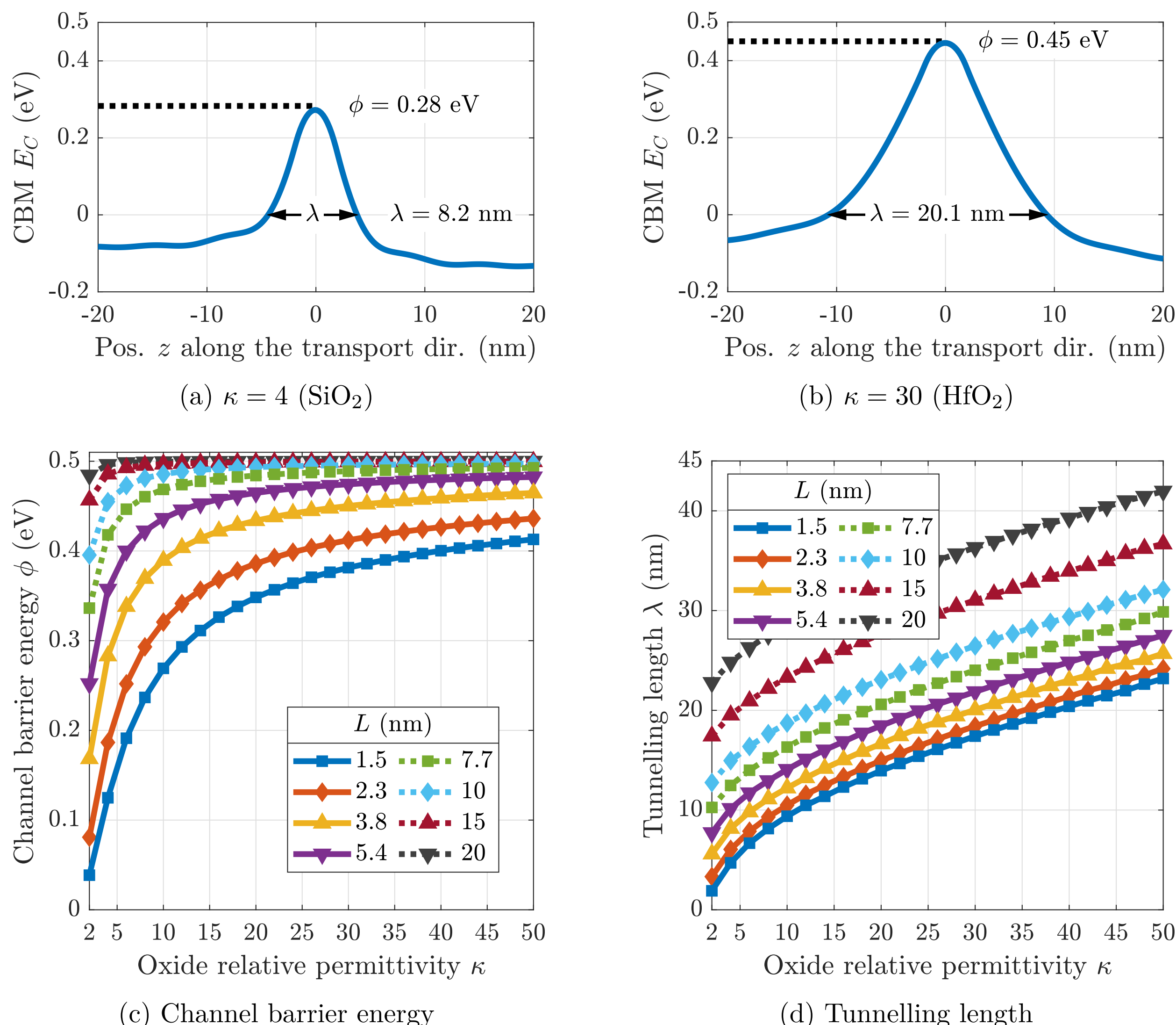
$$L_{S,D} \propto \sqrt{\kappa\phi} . \tag{7.3}$$

In Sec. 5.3, a low-κ source oxide was applied to achieve low L_S in a TFET; this enabled high BTBT probability and on-state current. Contrarily, for very short-channel MOSFETs, Eq. 7.2 indicates that high $L_{S,D}$ would be desired, as this would lead to low DSDT probability and low off-state leakage. This calls for the MOSFET design shown in Fig. 7.1, where the source, channel, and drain are all surrounded by a high-κ oxide. A high-κ gate oxide enables high gate control, which is especially important in very short-channel MOSFETs [Sec. 6.1]. A high-κ source and drain oxide results in long depletion lengths $L_{S,D}$ and low DSDT probability. Due to the exponential dependence of the DSDT probability on κ [Eqs. 7.2 and 7.3], it can be expected that this BCE-inspired scheme would be particularly effective at suppressing DSDT.

It should be noted that in modern nanoelectronics, low-κ spacer oxides are often placed between the source and gate contacts so as to minimize load capacitance, loaded charging time delay, and power consumption [Sec. 2.2.2]. This is especially important for the GAA geometry, for which the proportion of such parasitic capacitance is higher than for other gate geometries [288]. The DSDT reduction scheme described in this section is perfectly compatible with low-κ spacer oxides. Indeed, to achieve long depletion lengths, only the source and drain depletion regions, which typically are only a few nanometres in length, need to be surrounded by high-κ oxides. Outside of these regions, low-κ oxides may be used, as illustrated by the light-yellow regions in Fig. 7.1.

7.2 Atomistic Quantum Transport Simulations

Having argued how DSDT may be reduced through BCE, this section investigates the electrostatics of very short-channel MOSFETs using atomistic quantum transport simulations based on the NEGF–TB package `Nanoskim 2.0` [Sec. 4.4]. Throughout this chapter, the simulated MOSFETs have a thin body, a GAA geometry, and a thin surrounding oxide; indeed, scaling theory shows that these characteristics are *sine quibus non* for high gate control in very short-channel devices [Sec. 6.1]. Specifically, cylindrical GAA silicon NW MOSFETs are considered. All simulated MOSFETs are composed of an NW grown in [110]



(a) $\kappa = 4$ (SiO_2)

(b) $\kappa = 30$ (HfO_2)

(c) Channel barrier energy

(d) Tunnelling length

Figure 7.2 – (a)–(b) Band diagrams of very short-channel cylindrical GAA silicon NW MOSFETs [Fig. 7.1] at a drain voltage of $V_{DS} = 50$ mV and a gate voltage of $V_{GS} = 0$ V, as obtained from NEGF–TB simulations. A device with a low-κ oxide, SiO_2, and a device with a high-κ oxide, HfO_2, are investigated. The devices have an extremely short channel length of $L = 3.8$ nm. (c)–(d) Channel barrier energy ϕ and tunnelling length λ of MOSFETs with various κ and L, extracted from simulation data as illustrated in (a)–(b).

with a diameter of $d = 2$ nm [Fig. 4.4], a source and drain donor doping concentration of $N_D = 1 \times 10^{20}$ cm^{-3}, an intrinsic channel, and an oxide thickness of $t = 2$ nm [Fig. 7.1]. The gate metal's workfunction is taken to be 0.5 eV greater than the silicon NW's electron affinity.

In Figs. 7.2a and 7.2b, the potential profiles of two MOSFETs with a channel length of $L = 3.8$ nm are shown. These devices are identical in every way other than the nature of

their oxides and are subjected to the same temperature and applied voltages. Nevertheless, the MOSFET with a low-κ oxide (SiO_2) has a channel barrier energy ϕ only about 62% as large as that of the MOSFET with a high-κ oxide (HfO_2). This is a consequence of the relatively long natural length of the device with SiO_2, about 1.4 nm [Eq. 6.4], which is not short enough to ensure high gate control at $L = 3.8$ nm. As a result, thermionic emission is exponentially higher in this device [Eq. 2.15]. Furthermore, this device's tunnelling length λ is less than half of that of the device with HfO_2. This is a consequence not only of the low permittivity of the source and drain oxides but also of the low barrier energy [Eq. 6.22]. As a result, DSDT is also exponentially higher in this device [Eq. 7.2]. Thus, both dominant leakage mechanisms in ultra-short-channel MOSFETs (thermionic emission and DSDT) are negatively affected by low-κ oxides; such oxides lead to very poor switching.

In Figs. 7.2c and 7.2d, these findings are verified on a broader collection of devices with oxide permittivities κ ranging from 2 to 50 and channel lengths L ranging from 1.5 nm to 20 nm. Expectedly, the "long-channel" devices have a barrier height that is relatively unaffected by κ; their channels are much longer than the natural length, even in the low-κ case. The opposite is true for the short-channel devices. However, increasing the oxide permittivity is a very efficient way to retain high gate control in these devices. For example, the device with $L = 1.5$ nm and $\kappa = 30$ has a higher barrier energy than the device with $L = 5.4$ nm and $\kappa = 4$; in other words, replacing silicon dioxide by hafnium dioxide well compensates for reducing the channel length almost fourfold, as far as gate control is concerned. The tunnelling length also exhibits a dependence on κ which is stronger in short-channel devices. Indeed, the tunnelling length is the sum of the depletion lengths $L_{S,D}$ and the channel length L; $L_{S,D}$ are proportional to $\sqrt{\kappa}$ [Eq. 7.3] while L is κ-independent. In ultra-short-channel devices, L does not strongly contribute to λ, even for low κ. Increasing the oxide permittivity can thus very significantly increase the tunnelling length and lower the DSDT probability. For example, in the devices with $L = 1.5$ nm, increasing κ from 4 to 30 leads to λ increasing almost fourfold. While it has long been known that high-κ oxides are beneficial to gate control, Fig. 7.2 shows that these oxides are crucial in ultra-short-channel MOSFETs to reduce DSDT, an important leakage mechanism in such devices.

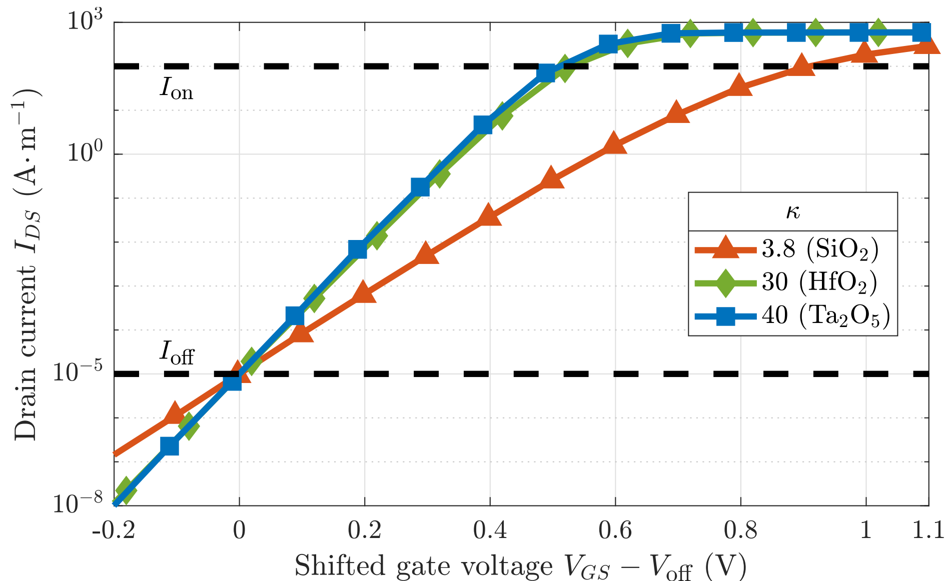


Figure 7.3 – Transfer characteristics of ultra-short-channel cylindrical GAA silicon NW MOSFETs [Fig. 7.1] at a temperature of $T = 300$ K and a drain voltage of $V_{DS} = 50$ mV, as obtained from NEGF–TB simulations. Devices with various surrounding oxide permittivities κ are investigated. The devices have an extremely short channel length of $L = 2.3$ nm. The current is normalized by the NW perimeter. The off-state voltage V_{off} is defined as the value of gate voltage V_{GS} at which the drain current I_{DS} is equal to $I_{\text{off}} \stackrel{\text{def}}{=} 10^{-5}$ A·m^{-1}.

7.3 High-Performance, Short-Channel MOSFETs

In the previous section, it was argued that in short-channel MOSFETs, leakage is much stronger when the oxide permittivity is low. This leakage leads to a higher off-state current or, equivalently, a higher STS. In turn, ultra-short-channel MOSFETs with low oxide permittivity can only operate at high power supply voltage, wasting a significant amount of heat in switching. In this section, to quantify these inferences, the transfer characteristics of various short-channel MOSFETs are obtained by NEGF–TB simulations and analyzed.

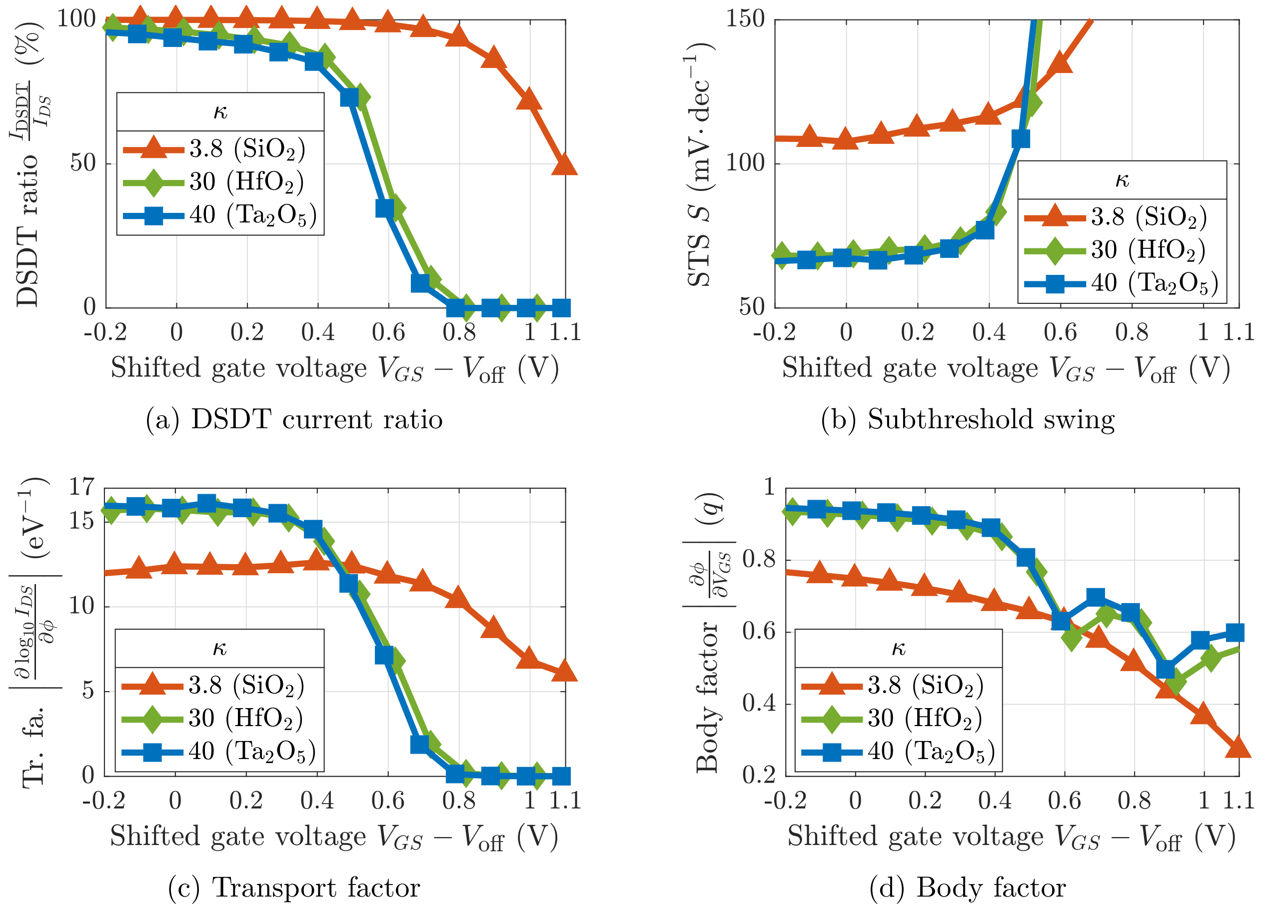
Figure 7.3 shows the transfer characteristics of three MOSFETs with a channel length of $L = 2.3$ nm, one with silicon dioxide ($\kappa = 3.8$), one with hafnium dioxide ($\kappa = 30$), and one with tantalum pentoxide ($\kappa = 40$). The former MOSFET has an off-state leakage current orders of magnitude higher than the others, as expected from its weak gate control and short tunnelling length. Thus, to ensure a fair comparison, the gate voltage V_{GS} is shown in

reference to the off-state voltage V_{off}, so that the drain current I_{DS} of the three MOSFETs equals $I_{\text{off}} \stackrel{\text{def}}{=} 10^{-5}$ A$\cdot$ m^{-1} [14] at $V_{GS} - V_{\text{off}} = 0$ V. In practice, a "shift" in gate voltage can be achieved through tuning of the gate metal workfunction [Sec. 2.2.3]. Several observations can be made from the data. First, the transfer characteristics of the MOSFET with HfO_2 (green curve) and the MOSFET with Ta_2O_5 (blue curve) nearly overlap each other. This suggests that HfO_2 may be an appropriate oxide to reduce DSDT in ultra-short-channel MOSFETs with $L \geq 2.3$ nm; replacing it by higher-κ oxides such as Ta_2O_5 may not result in significant benefits to device performance. Second, the three MOSFETs all exhibit a roughly constant slope in the subthreshold regime, despite being prone to DSDT. Indeed, not unlike thermionic leakage [Eq. 2.15], the DSDT leakage current decays exponentially with the barrier energy ϕ [Eqs. 7.2 and 7.3], leading to V_{GS}-independent STS. Third, the MOSFET with SiO_2 (red curve) has a much higher STS than the other devices; specifically, its average STS is 130 mV$\cdot$ dec^{-1}, to be compared to 75.2 and 72.4 mV$\cdot$ dec^{-1} in the MOSFETs with HfO_2 and Ta_2O_5, respectively [Table 7.1]. The STSs of the devices with high-κ oxides approach the thermal limit of 60 mV$\cdot$ dec^{-1} at room temperature, suggesting that charge transport in these devices is at least partly carried by high-energy electrons, as in thermionic emission. Fourth, the increment in gate voltage required to increase the drain current from I_{off} to $I_{\text{on}} \stackrel{\text{def}}{=} 10^2$ A$\cdot$ m^{-1} [14], which roughly corresponds to the power supply voltage V_{DD}, is slightly less than twice as high for the device with SiO_2 compared to the other devices [Table 7.1], which significantly increases power consumption [Eq. 2.33]. On the other hand, the devices with high-κ oxides have $V_{DD} < 0.55$ V, which is well within the acceptable range for the next generations of transistors, as recommended by the IRDS™ [14]. Thus, as predicted, high-κ oxides are effective tools to scale the channel lengths of MOSFETs down to the nanometre range while maintaining acceptable power consumption.

In Table 7.1, the average STS S_{av} and the power supply voltage V_{DD} of the MOSFETs of Fig. 7.3 are compared to those of MOSFETs with $L = 5.4$ nm and $L = 10$ nm. Expectedly, a longer channel length leads to lower S_{av} and V_{DD} since the DSDT probability decays exponentially with L. However, the change is far more drastic for the MOSFETs with SiO_2. Indeed, for these devices, the average STS increases by 84% when L is lowered from 10 nm

κ	3.8 (SiO_2)			30 (HfO_2)			40 (Ta_2O_5)		
L (nm)	2.3	5.4	10	2.3	5.4	10	2.3	5.4	10
S_{av} (mV· dec^{-1})	130	97.3	70.7	75.2	66.9	61.6	72.4	65.6	61.3
V_{DD} (mV)	908	681	495	526	468	431	507	459	429

Table 7.1 – Average STS S_{av} and power supply voltage V_{DD} of cylindrical GAA silicon NW MOSFETs [Fig. 7.1] with various channel lengths L and surrounding oxide permittivities κ at $T = 300$ K and $V_{DS} = 50$ mV. The off-state and on-state currents are defined as $I_{\mathrm{off}} = 10^{-5}$ A· m^{-1} and $I_{\mathrm{on}} = 10^{2}$ A· m^{-1}, respectively. The average of the STS is taken between the off state and the on state. The power supply voltage V_{DD} is defined as the increment in gate voltage required to increase the drain current from I_{off} to I_{on}.



(a) DSDT current ratio

(b) Subthreshold swing

(c) Transport factor

(d) Body factor

Figure 7.4 – Metrics pertaining to charge transport in the ultra-short-channel MOSFETs of Fig. 7.3 at $T = 300$ K and $V_{DS} = 50$ mV. The DSDT current ratio is obtained through Eq. 7.4, while the STS, transport factor, and body factors are obtained by numerical differentiation of the NEGF–TB simulation data.

to 2.3 nm; for the devices with HfO_2, this increase is of only 22%. This further corroborates the notion that high-κ dielectrics are especially important in ultra-short-channel MOSFETs.

Quantitatively, what degrades the STS of ultra-short-channel MOSFETs with low-κ oxides most strongly: poor gate control or high DSDT current? To better understand charge transport in the devices of Fig. 7.3, several relevant metrics are shown in Fig. 7.4. The DSDT current can be computed through

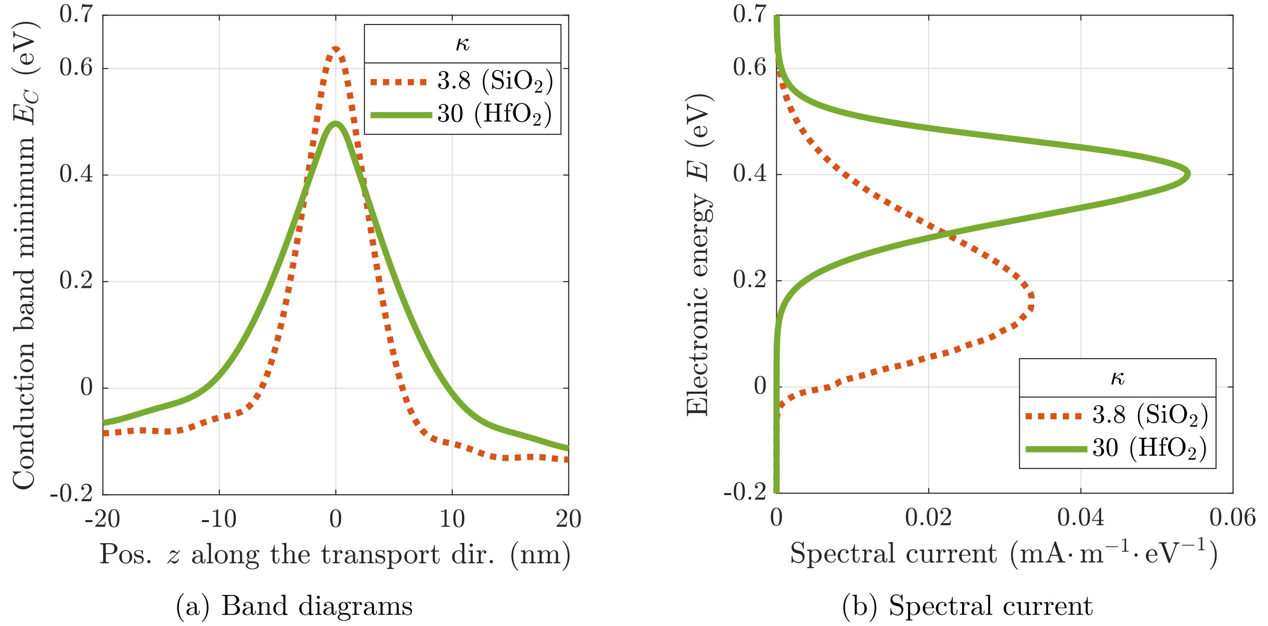
$$I_{\mathrm{DSDT}} = \frac{q}{h} \int_{E_{C,S}}^{\phi} \overline{T}(E) \left[f_S(E) - f_D(E) \right] dE \,, \tag{7.4}$$

where E is the electronic energy, $E_{C,S}$ is the value of the CBM energy in the source, $\overline{T}(E)$ is the transmission function, and $f_{S,D}(E)$ are the source and drain Fermi–Dirac distributions; this equation is identical to Eq. 4.33, with the exception that the range of integration has been restricted to that compatible with DSDT, namely to energies under the channel barrier. Figure 7.4a shows the contribution of DSDT to total current, i.e. $\frac{I_{\mathrm{DSDT}}}{I_{DS}}$, in the investigated devices. Expectedly, the MOSFET with SiO_2 exhibits a higher proportion of DSDT. Furthermore, in the subthreshold regime, DSDT is by far the dominant charge transport mechanism, even in the devices with high-κ oxides. The STS [Fig. 7.4b] can be expressed as the product of the transport factor and the body factor [Eq. 2.14]. Both factors are lower by around 20% in the MOSFET with SiO_2, compared to the MOSFETs with high-κ oxides. In other words, poor gate control and DSDT contribute to STS degradation in MOSFETs with low-κ oxides in a roughly equal manner.

It remains to investigate the nature of charge transport in these ultra-short-channel MOSFETs. What explains the significantly lower transport factor of the device with SiO_2? The transport factors of the MOSFETs with high-κ oxides are very close to the thermionic emission value of $\frac{1}{\ln 10}\frac{1}{k_B T} \approx 16.8\ \mathrm{eV}^{-1}$ at room temperature [Eq. 2.16], despite the fact that only a very small proportion of off-state current is due to thermionic emission in these devices [Fig. 7.4a]. To understand this perplexing observation, consider that over the range of energies for DSDT, one has

$$\overline{T}(E) \propto \exp\left(-\frac{\lambda}{\hbar} \sqrt{m^{\star}_{e,t} (\phi - E)} \right) \,, \tag{7.5}$$

$$f_S(E) - f_D(E) \propto \exp\left(-\frac{E - \mu_S}{k_B T} \right) \,, \tag{7.6}$$



(a) Band diagrams (b) Spectral current

Figure 7.5 – Band diagrams and spectral current (i.e. drain current per unit energy) of two of the ultra-short-channel MOSFETs of Fig. 7.3 at T = 300 K, V_{DS} = 50 mV, and $V_{GS} - V_{\text{off}} = 0$ V, as obtained from NEGF–TB simulations.

where μ_S is the source Fermi level and T is the device temperature. Equation 7.5 is derived from the WKB approximation [66], while Eq. 7.6 is derived from Eq. 2.6. The transmission function increases exponentially with E, while the Fermi–Dirac distributions decrease exponentially with E. More specifically, $\overline{T}(E)$ is close to unity around $E = \phi$ and exponentially smaller around $E = \mu_S$; $f_S(E) - f_D(E)$ is close to unity around $E = \mu_S$ and exponentially smaller around $E = \phi$. Hence, it is the competition between these two exponentials that determines the range of energies in which most of the current is carried. In the case of a low-κ oxide, λ is small, so that most current is carried at low energies close to $E = \mu_S$. In the case of a high-κ oxide, λ is large, so that most current is carried at high energies close to $E = \phi$. This is verified numerically in Fig. 7.5, where the spectral current, i.e. $\frac{q}{h}\overline{T}(E)[f_S(E) - f_D(E)]$, is graphed as a function of E for ultra-short-channel MOSFETs with SiO_2 and HfO_2. In the high-κ case, the spectral current is confined to an energy range where it is proportional to $\exp\left(-\frac{E}{k_B T}\right)$. Readers may recall that this was the initial assumption that was made to derive the transport factor for thermionic emission [Eq. 2.15]. This explains why the transport factors of ultra-short-channel MOSFETs with high-κ oxides are close to the thermionic emission value. In this case, charge transport could be described

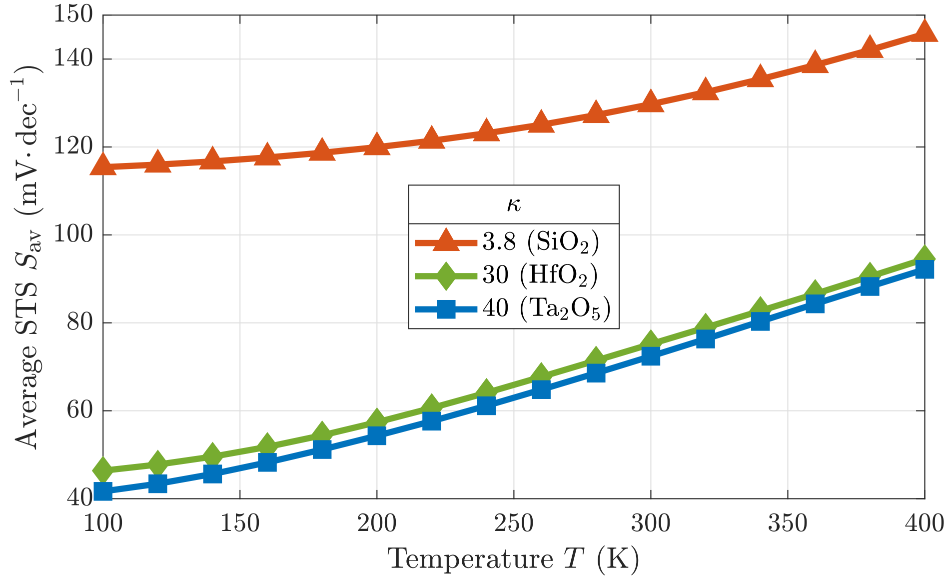


Figure 7.6 – Average STS of the ultra-short-channel MOSFETs of Fig. 7.3 at $V_{DS} = 50$ mV and various temperatures T, as obtained by the method of Table 7.1.

as "thermal DSDT." In the low-κ case, the channel barrier of an ultra-short-channel MOSFET can be approximated to be a square barrier with width $L_S + L_D$ and height ϕ. Thus, from Eqs. 6.26 and 7.2, it can be approximated that

$$I_{DS} \propto \exp\left(-\frac{2a_C}{\hbar}\sqrt{\frac{2\kappa_{\mathrm{Si}}\varepsilon_0 m_{e,t}^{\star}}{q^2 N_D}}\phi\right) \tag{7.7}$$

$$\implies \frac{\partial \log_{10} I_{DS}}{\partial \phi} = -\frac{1}{\ln 10}\frac{2a_C}{\hbar}\sqrt{\frac{2\kappa_{\mathrm{Si}}\varepsilon_0 m_{e,t}^{\star}}{q^2 N_D}}\,. \tag{7.8}$$

For the investigated MOSFETs, this results in a transport factor of $\left|\frac{\partial \log_{10} I_{DS}}{\partial \phi}\right| = 10.5$ eV^{-1}, which is close to that observed in simulations for the device with SiO_2 [Fig. 7.4c], and lower than the transport factor for thermal DSDT of 16.8 eV^{-1} at room temperature. Furthermore, since Eq. 7.8 does not exhibit a T-dependence, charge transport in the low-κ case could be described as "athermal DSDT." To summarize, ultra-short-channel MOSFETs with high-κ and low-κ oxides exhibit qualitatively different charge transport, i.e. temperature-dependent and temperature-independent DSDT, respectively; this explains the much lower transport factor in the low-κ case.

In Fig. 7.6, to confirm the nature of DSDT in ultra-short-channel MOSFETs, the average

STS S_{av} of the MOSFETs of Fig. 7.3 are graphed as a function of temperature T. The devices with HfO_2 and Ta_2O_5 exhibit a nearly linear dependence of S_{av} on T, just as for thermionic emission [Eq. 2.20]. This is befitting of the "thermal" qualifier that was previously introduced to describe DSDT in these devices. In contrast, the device with SiO_2 exhibits a weaker dependence of S_{av} on T. In this case, charge transport fits somewhere between the two limiting cases of thermal and athermal DSDT. This is consistent with the observation that the mode of the spectral current for this device is at an energy slightly higher than μ_S, as can be seen in Fig. 7.5b.

7.4 Conclusion

In this chapter, I investigated ultra-scaled MOSFETs with channel lengths as small as 1.5 nm through NEGF–TB simulations. When low-κ oxides are used, performance severely deteriorates as the channel is shortened. This is due to (1) poor gate control and (2) high DSDT probability, as explained by the short depletion lengths at the channel junctions. Both of these issues contribute to STS degradation in a roughly equal manner: in the case of SiO_2, the STS increases by around 84% when the channel is shortened from 10 nm to 2.3 nm. High-κ oxides feed two birds with one seed, in that they significantly increase both the transport and body factors of ultra-short-channel MOSFETs, leading to acceptable device performance for a channel length of just 2.3 nm. The improvement of the body factor is expected from lengthscale theory. The improvement of the transport factor is due to a change in the nature of charge transport: "thermal DSDT" in MOSFETs with high-κ oxides and "athermal DSDT" in MOSFETs with low-κ oxides. Overall, the BCE-inspired scheme to reduce DSDT that I presented in this chapter is generally applicable to short-channel MOSFETs and is compatible with low-κ spacer oxides, which are often used by the industry to reduce load capacitance. Furthermore, this scheme can be used with other common methods to reduce DSDT, such as EM engineering, and may pave a way toward ultra-scaled MOSFETs. While DSDT can be reduced with the industry-standard HfO_2, recent reports of ultra-high-κ dielectrics, e.g.

cubic beryllium oxide with $\kappa \approx 275$ [289], offer promising prospects toward an even greater suppression of DSDT in ultra-short-channel MOSFETs.

Chapter 8

Conclusion and Outlook

Moore's Law is a violation of Murphy's Law. Everything gets better and better.

— G. E. Moore, 2005 [290]

This thesis begun with a simple question that Gordon Moore asked nearly six decades ago: "[w]ill it be possible to remove the heat generated by [a large number of] components in a single silicon chip?" Since then, the number of such components on chips, mostly transistors, has increased by around seven orders of magnitude. The physical dimensions of these components have shrunk by more than three orders of magnitude. This exponential downscaling came with fundamental changes to device physics as well as a multiplication of parasitics and apparent roadblocks that, against all odds, have been consistently overcome. In Moore's own words: "Moore's Law is a violation of Murphy's Law. Everything gets better and better." It is thanks to the zeal of industrial and academic researchers, government support, and consumer demand that this could be achieved. This thesis, which investigates novel transistor materials, geometries, and physics, is part of this effort toward denser nanoelectronics.

Future generations of transistors will need a gate-all-around geometry since this enables lower subthreshold swing, lower power consumption, and more aggressive downscaling. To this end, the channels of these transistors will be composed of low-dimensional semiconductors such as molybdenum disulphide or silicon nanowires (NWs). Furthermore, the industry may adopt steep-slope field-effect transistors (FETs) like tunnel field-effect transistors (TFETs),

which require short depletion lengths to achieve high on-state current and to be run at high clock frequency. Problematically, low-dimensional materials have relatively few free charges, which redistribute themselves through the screening effect to shorten depletion regions. Screening is typically strengthened by increasing the chemical doping concentration, which is in practice limited by the solid solubility limits of dopants and bandgap narrowing.

The central contribution to knowledge of my thesis is to introduce bound-charge engineering (BCE), a method to engineer bound charges to our advantage in devices. Specifically, using Maxwell's equations, I showed that when an external electric field traverses the interface between two materials, a bound charge forms on that interface. Furthermore, the amount of bound charge is proportional to the electric field as well as the difference between the permittivities of the two materials. The permittivity κ is a material property that quantifies how much a material polarizes in response to an external electric field. Thus, by tuning the electric field and choosing materials with suitable permittivities, the surface bound charge can be controlled. This is readily applicable to semiconductor–oxide interfaces in FETs, in which BCE can be used to control the size of depletion regions of junctions with the channel. In much the same way that semiconductors can be doped by chemical dopants, they can effectively be doped by surface bound charges, thereby alleviating the screening problems in low-dimensional semiconductors. Low-κ oxides lead to short depletion lengths, while high-κ oxides lead to long depletion lengths. Ideally, these oxides are used in conjunction: a high-κ oxide around the channel for high gate control, and low-κ oxides around the source and drain for short depletion lengths.

To substantiate the principle of BCE, I performed self-consistent atomistic quantum transport simulations based on the tight-binding (TB) model and the nonequilibrium Green's function (NEGF) formalism, the state of the art for nanotransistor simulation. Furthermore, I developed a general analytical surface potential model for BCE-assisted FETs, thereby improving our theoretical understanding of BCE. I applied BCE to improve the performance of silicon NW TFETs. Specifically, through NEGF–TB simulations, I showed that a TFET combining silicon dioxide and hafnium dioxide, industry-standard low-κ and high-κ oxides, respectively, had lower subthreshold swing and on-state current orders of magnitude higher

than single-oxide TFETs. Furthermore, I applied BCE to reduce direct source-to-drain tunnelling, the dominant leakage mechanism in ultra-short-channel metal–oxide–semiconductor FETs (MOSFETs), down to acceptable levels in devices with channel lengths as small as 1.5 nm. BCE can be used in conjunction with known methods to improve the performance of TFETs and ultra-short-channel MOSFETs (e.g. effective mass engineering). It may thereby pave a way toward improved low-power ultra-scaled FETs.

Several potential future projects may be interesting to pursue. First and foremost, the devices introduced in this thesis should be realized experimentally. Indeed, the investigations of this thesis were purely theoretical; only a solid, real-world realization of BCE-assisted devices can prove beyond doubt the concept of BCE. In the case of BCE-assisted TFETs, the most challenging aspect of an experimental realization may be the nanometric alignment of the tunnelling (source–channel) junction with the junction between the low-κ and high-κ oxides. A very recent report by Convertino *et al.* shows that this is possible in III–V-semiconductors heterojunction TFETs, although the effect of BCE is not directly investigated [291]. Second, TFETs are typically plagued by ambipolar conduction, which refers to band-to-band tunnelling at the channel–drain junction in the deep subthreshold regime. This leakage mechanism could be alleviated with a high-κ drain oxide since BCE predicts that such an oxide would increase the tunnelling length of the channel–drain junction. Finally, BCE is a very general idea established by basic laws of electromagnetism. Thus, in principle, it is not limited to applications in the fields of nanoelectronics and transistor design. Therefore, BCE could potentially be applied in other fields of research where bound charges and screening may be important, such as molecular electronics, electrochemistry, and artificial photosynthesis.

Appendix A

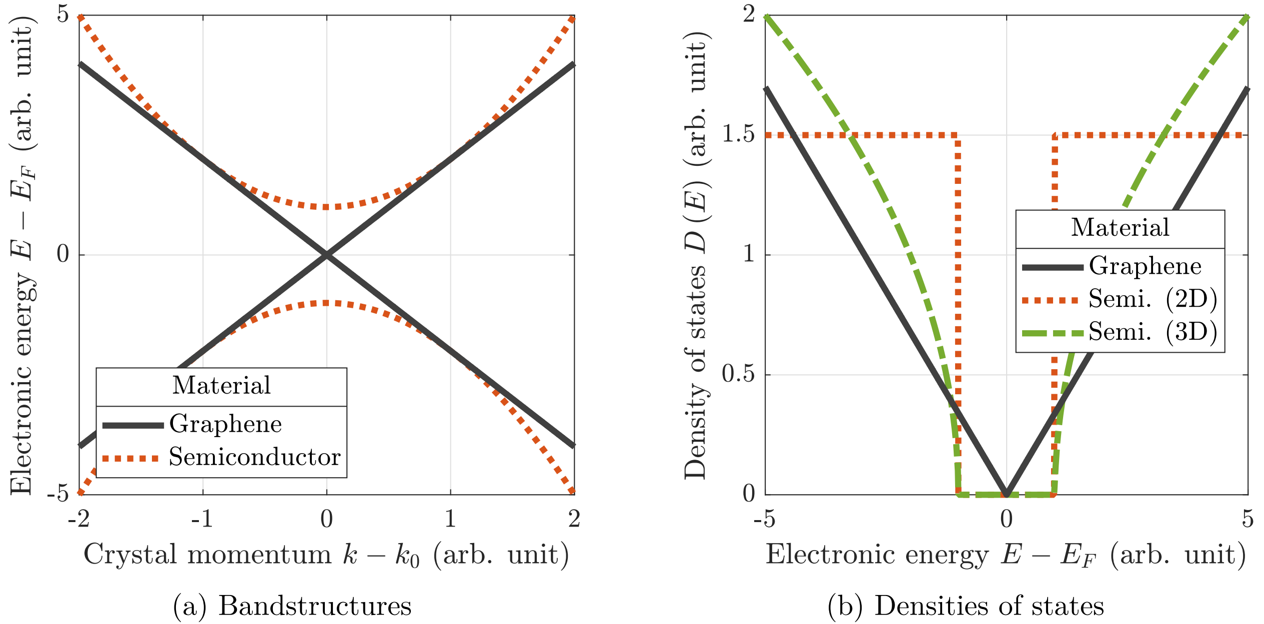
More Steep-Slope Field-Effect Transistors

While the TFET and the NCFET, which are described in Chapter 3, are arguably the two steep-slope FETs most likely to replace the MOSFET for low-power logic, many other steep-slope FETs have been proposed in the literature. Several such devices are reviewed in this chapter. They can be classified as FETs with large transport factors [Sec. A.1], FETs with large body factors [Sec. A.2], and FETs with large transport and body factors [Sec. A.3].

A.1 FETs with Large Transport Factors

A.1.1 The Dirac-Source FET

Graphene is a material consisting of a single layer of carbon atoms arranged in a honeycomb lattice. Since graphite consists of stacked sheets of graphene bound by van der Waals (weak) interactions [292], graphene can be easily obtained from graphite through mechanical exfoliation [293]. It is thus likely that graphene has been unintentionally produced for centuries using pencils. The existence of graphene has been theorized as early as 1947 by Wallace [294], but was first formally isolated and formally characterized in 2004 by Novoselov, Geim, *et al.* [293], who thereupon opened the field of 2D materials research in its modern form. Graphene has several exceptional electronic properties, including charge carriers mobilities several orders of magnitude higher than those of silicon [295]. Because of this, graphene has been considered to be a potential surrogate to silicon for future generations of MOSFETs [296].



(a) Bandstructures

(b) Densities of states

Figure A.1 – Bandstructures and densities of states of graphene, a typical 2D semiconductor, and a typical 3D semiconductor; all materials are assumed to be intrinsic. Bandstructures show electronic energies E relative to Fermi energies E_F as a function of crystal momenta k around minima k_0.

However, graphene has little to no bandgap [297], leading to poor switching [298]. This makes the adoption of graphene toward MOSFET applications ineffective.

Whereas conventional semiconductors have a gapped, parabolic dispersion, graphene has a gapless, linear dispersion [293, 294] [Fig. A.1a]; graphene should thus be classified as a semimetal. The energy at which the valence and conduction bands merge is called the Dirac point energy E_D. The unique dispersion relation of graphene results in [299]:

- near-zero electron and hole EMs;
- a near-zero DOS around E_D [Fig. A.1b].

Relativistic effects should thus be included in a proper theoretical description of graphene, such as quantum electrodynamics [159]. An important result from quantum electrodynamics is the Klein paradox, namely, the prediction that massless Dirac fermions (such as electrons and holes in graphene) hitting a large potential barrier will tunnel through the barrier with near-unit probability, regardless of the barrier height and width [300–302]. This is in contrast with the tunnelling of massive particles, for which the tunnelling probability

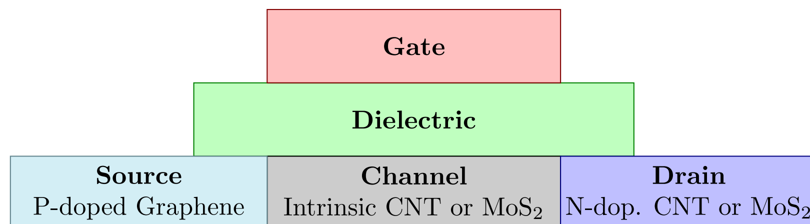


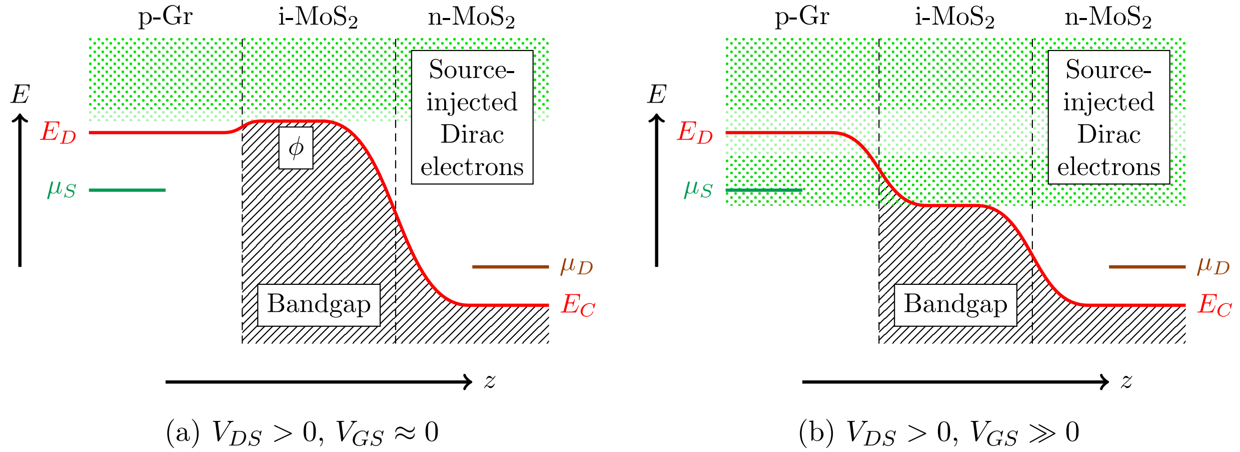
Figure A.2 – Schematic of a typical n-type enhancement-mode Dirac-source field-effect transistor (DSFET). The source is p-doped graphene, the channel is intrinsic, and the drain is n-doped; the channel and drain are semiconducting and typically composed of a carbon nanotube (CNT) or few-layers monolayer molybdenum disulphide (MoS_2). The main charge transport mechanism is the thermionic emission of Dirac electrons.

decays exponentially with barrier height and width [Eq. 2.12]. Klein tunnelling has been experimentally demonstrated in graphene p–n junctions in 2009 [303].

A significant limitation of TFETs is their low on-state current, which is due to the limitations of BTBT. To resolve this issue, one could imagine a TFET-like device that operates on Klein tunnelling rather than conventional BTBT. This is the motivation behind the Dirac-source field-effect transistor (DSFET), a transistor first proposed, experimentally demonstrated [251], and theoretically investigated [247] in 2018 by Qiu, Liu, *et al.*

The structure of a DSFET is shown in Fig. A.2. In an n-type enhancement mode DSFET, the source is p-doped graphene. The channel and drain are, respectively, intrinsic and n-doped semiconductors; DSFETs with channel and drain composed of carbon nanotubes [251,304,305] and few-layers MoS_2 [306–308] have been experimentally demonstrated.

Band diagrams of a DSFET are shown in Fig. A.3. In the off state [Fig. A.3a], the channel barrier energy ϕ is greater than or equal to the Dirac point energy E_D, so that electrons from the graphene source are thermionically injected, as in a MOSFET. In a MOSFET, the DOS of thermionically injected electrons is constant [Fig. A.1b]. In a DSFET, this DOS is proportional to $(E - E_D)$; in particular, it is 0 at energy $E = E_D$ [Fig. A.1b]. As a result, very few electrons are thermionically injected in a DSFET in the off state. In the on state, since the source Fermi level μ_S is well within the graphene valence band, far from the energy range of low DOS, many electrons are injected into the channel, leading to relatively high drain current.



(a) $V_{DS} > 0$, $V_{GS} \approx 0$ (b) $V_{DS} > 0$, $V_{GS} \gg 0$

Figure A.3 – Band diagrams of a typical n-type enhancement-mode MoS_2 DSFET off state (a) and on state (b). In the graphene source, the CBM E_C equals the Dirac point energy E_D. Dirac electrons are thermionically injected from the source (green-dotted region); their DOS is very low around E_D (very light-green-dotted region). In the off state, very few electrons are injected. In the on state, many electrons are injected.

It should be noted that although DSFETs and TFETs have somewhat similar structures (both have p–i–n doping profiles for n-type devices), their charge transport mechanisms are quite different. Both devices exhibit energy filtering and tunnelling at the source–channel junction. However, energy filtering is stronger in TFETs (thanks to the source bandgap) and the tunnelling barrier is essentially transparent in DSFETs (thanks to Klein tunnelling). Charge transport in DSFETs could thus be described as a thermionic injection of Dirac electrons. Experimentally, this is evidenced by the linear temperature dependence of the STS in DSFETs [251, 307], a characteristic not shared with TFETs [41].

To date, there have been at least half a dozen experimental reports demonstrating various DSFETs [251, 304–308]. In particular, Xiao *et al.* demonstrated in 2020 an n-type CNT DSFET with on-state current at least as high as $I_{\text{on}} = 2.6$ A$\cdot$ m^{-1}, minimum STS of 37 mV$\cdot$ dec^{-1} at room temperature, and sub-60 mV$\cdot$ dec^{-1} STS sustained over 4 decades of drain current at drain voltage $V_{DS} = 0.1$ V [305].

Overall, the DSFET is a promising steep-slope FET that, to some extent, rejuvenated interest in graphene for applications toward energy-efficient FETs. A potential limitation of DSFETs may be the relatively high doping required in the graphene source, which in all

experimental reports appears to be achieved actively using a control/back gate, rather than with passive methods such as chemical doping; the external voltages required to sustain such doping require some amount of power. Another potential limitation may be the somewhat low on-state current of DSFETs, which appears to trail that of silicon MOSFETs by around one order of magnitude or more in experimental reports [251]. Finally, graphene, which is required in DSFETs, is costly to mass-produce [309].

A.1.2 The Cold-Source FET

The principle of energy filtering, that is, the suppression of high-energy charge carriers to reduce thermionic leakage, was the source of inspiration for several steep-slope FETs, including the TFET and the DSFET. Both the TFET and the DSFET suffer from certain drawbacks that can typically be alleviated but not eliminated, as discussed in previous sections. In an effort to avert such drawbacks, one may pursue an energy-filtering FET with the following characteristics:

- MOSFET-like thermionic emission as the main on-state charge transport mechanism to reach MOSFET-like on-state current and avoid ambipolarity;
- a structure involving solely silicon in the semiconducting parts of the device, since silicon is the material of choice for the semiconductor industry.

These requirements lead to the conceptualization of a MOSFET with a "cold" silicon source deprived of "hot" (i.e. high-energy) electrons. Such a cold source cannot simply be p-doped silicon, as this would result in a TFET.

In 2018, this lead Liu *et al.* to propose and theoretically investigate a cold-source FET (CSFET) [57] with a cold source composed of degenerately p-doped silicon, metal, and degenerately n-doped silicon (in the n-type enhancement-mode case), as illustrated in Fig. A.4. The CSFET can be thought of as a low-pass filter for electrons (the p-doped silicon source) connected in series to an n-type enhancement-mode MOSFET. These two components are bridged by a metal layer forming Ohmic contacts both with p-doped and n-doped silicon in order to minimally impede the flow of electrons. Through first-principles

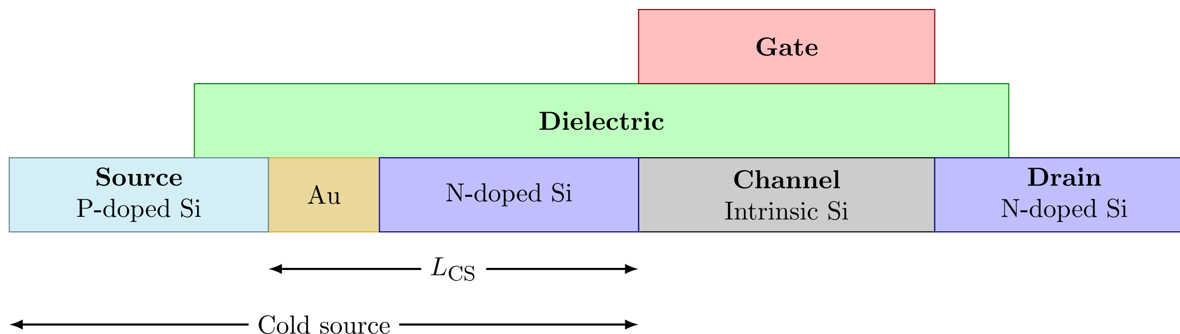
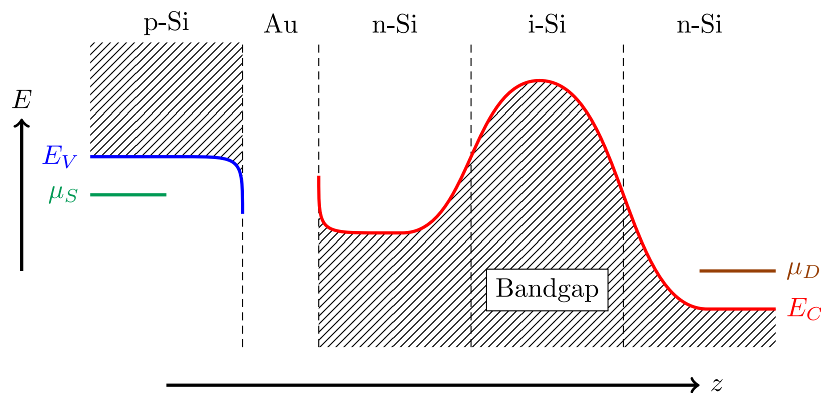


Figure A.4 – Schematic of a typical n-type enhancement-mode silicon cold-source field-effect transistor (CSFET). The source is composed of p-doped and n-doped layers bridged by a thin metal layer, typically gold (Au). The channel is intrinsic and the drain is n-doped. The main charge transport mechanism is the thermionic emission of cold electrons. Rethermalization in the cold source's metal and n-Si layers, with thickness L_{CS}, may lead to STS degradation.

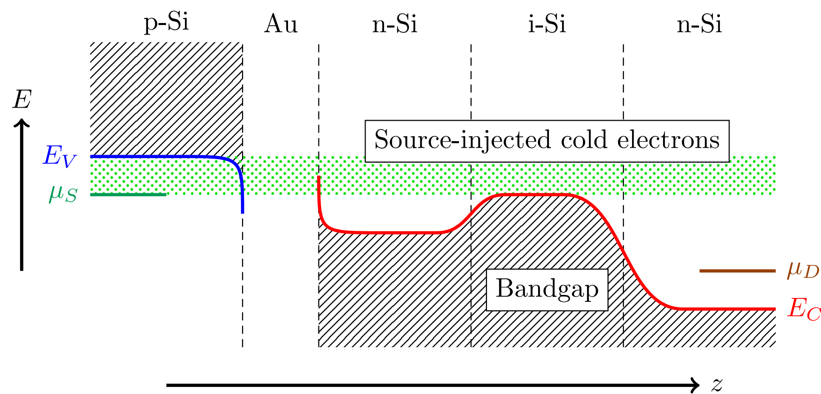
calculations investigating cold sources with various bridging metals, gold was found to be the best candidate in terms of cold-source transport properties [57].

Band diagrams of a CSFET are shown in Fig. A.5. Due to their degenerate doping, the p-doped and n-doped layers of the cold source form a broken gap band alignment, which restricts CSFET charge transport to energies between the n-doped layer's CBM and the p-doped layer's VBM. In the off state [Fig. A.5a], the channel barrier is above this energy window, resulting in no thermionic current. Instead, in the ballistic limit, off-state leakage is dominated by direct source-to-drain tunnelling. In the on state [Fig. A.5b], the channel barrier falls within the cold-source energy window, resulting in thermionic emission and high on-state current. Since most source-injected electrons have energies close to the source Fermi level μ_S, the somewhat narrow cold source energy window of a CSFET does not strongly limit its on-state current compared to a MOSFET. Indeed, the on-state current of a CSFET has been theoretically shown to be on the order of several thousands of $\mathrm{A \cdot m^{-1}}$, which is comparable to that of a MOSFET [57].

The energy-filtering quality of the cold source requires ballistic transport in the cold source's metal and n-doped silicon; in other words, electrons should traverse the cold source with little to no scattering. Indeed, when scattering is present in the cold source, some electrons may gain energy through inelastic scattering (notably, through electron–phonon



(a) $V_{DS} > 0$, $V_{GS} \approx 0$



(b) $V_{DS} > 0$, $V_{GS} \gg 0$

Figure A.5 – Band diagrams of a typical n-type enhancement-mode silicon CSFET in the off state (a) and on state (b). Charge transport is restricted to an energy window bounded below by the cold source n-Si layer's CBM and above by the p-Si layer's VBM. In the off state, essentially no electrons are injected. In the on state, cold electrons are thermionically injected from the cold source.

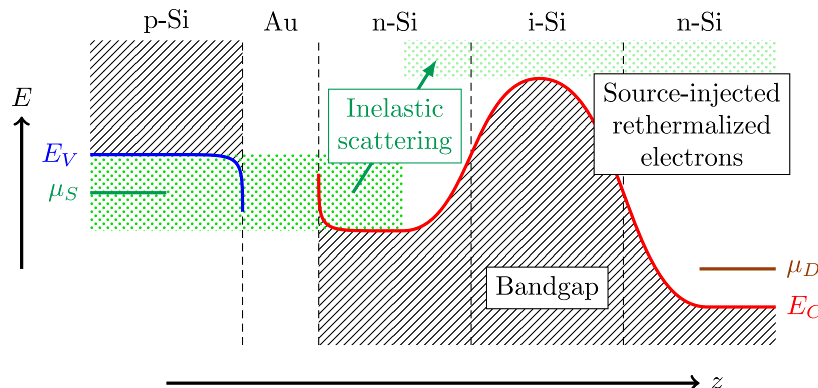


Figure A.6 – Band diagrams of a typical n-type enhancement-mode silicon CSFET in the off state. Inelastic scattering in the cold source may increase the energy of some electrons (green arrow), leading to partial rethermalization and thermionic leakage (very light-green dotted region).

scattering). If the energy gain is large enough, such electrons can undergo thermionic emission over the channel barrier, leading to thermionic leakage and STS degradation; this process is known as rethermalization and is illustrated in Fig. A.6. In the limiting case of diffusive transport, where the electrons are, everywhere in the device, locally described by a Fermi–Dirac distribution, thermionic leakage and STS are comparable to that of a MOSFET [5]. Assuming electron–phonon scattering to be the dominant form of inelastic scattering, the rate of rethermalization is proportional to $\frac{L_{\mathrm{CS}}}{\lambda_{\mathrm{e-ph}}}$, where L_{CS} is the thickness of the cold source's metal and n-Si layers [Fig. A.4] and $\lambda_{\mathrm{e-ph}}$ is the electron–phonon-scattering-limited electron mean free path [310]. Rethermalization could thus be reduced by reducing the physical dimensions of the cold source L_{CS} or by increasing the mean free path $\lambda_{\mathrm{e-ph}}$ (which can be achieved by temperature reduction or substitution of the cold source material by one with long $\lambda_{\mathrm{e-ph}}$).

The CSFET has not been realized experimentally. However, together with the DSFET, the CSFET rejuvenated interest in energy-filtering steep-slope FETs. Since 2018, several new CSFET designs have been proposed. These devices somewhat generalize the "CSFET" nomenclature: their cold sources are composed of van der Waals heterostructures with broken gap band alignment and do not include any metal [311–313]. In simulations, these van-der-Waals-heterojunctions-based CSFETs have minimal STS around 25 $\mathrm{mV \cdot dec^{-1}}$ [312, 313],

to be compared with around 40 mV·dec^{-1} for Si-based CSFETs [57], and may exhibit less rethermalization [312].

It should be noted that the "CSFET" nomenclature is often generalized to include DSFETs. More generally, any energy-filtering FET (including TFETs) could be described as a "cold source FET" since energy filtering and removal of hot carriers (that is, the cold-source effect) are equivalent.

A.1.3 The Superlattice FET and Bandstructure Engineering

Most semiconductors exhibit crystalline order: their atoms are orderly and predictably arranged, forming a periodic lattice. Electronic quantum states in crystals satisfy Bloch's theorem [156] and can therefore be described with a vector of continuous quantum numbers: the crystal momentum $\mathbf{k}$ [42]. It is, among other things, the configuration of the crystal (the distance between atoms and their relative positions) that determines the ranges of allowed energies for electrons as a function of $\mathbf{k}$, i.e. the energy bands. In contrast, some ranges of energies, the bandgaps, are forbidden; the TFET and CSFET exploit such bandgaps as means for energy filtering.

Semiconductor structures may be characterized by a second, higher layer of spatial periodicity, forming a so-called "superlattice." A superlattice can be obtained by periodically stacking, along one dimension of space, semiconductor layers with distinct material composition, doping, and/or geometry. Just like a crystal lattice, a superlattice may be described by Bloch's theorem, band theory, and exhibits energy bands and bandgaps. The spatial period of a superlattice must be greater than the underlying crystal lattice constant. Consequently, its energy bands and bandgaps are narrower (typically, a few hundreds of meV or less) than those of the crystal (typically, a few eV). The bands and bandgaps of a superlattice are thus often called, respectively, minibands and minibandgaps. Synthetic semiconductor superlattices with minibandgaps were proposed [314], theoretically investigated [315], and experimentally realized [316] by Esaki, Tsu, and Chang in 1970, 1973 and 1974, respectively.

Minibandgaps can be harnessed for energy filtering by replacing the source of a MOSFET by an appropriate superlattice. This device is known as the superlattice FET (SLFET) and was first proposed and patented by Björk *et al.* [317]. The SLFET has been the object of computational investigations by Gnani *et al.*, who predicted in 2011 that an SLFET with an InGaAs–InAlAs superlattice has a minimal STS of 13 $\text{mV}\cdot\text{dec}^{-1}$ and an on-state current $I_{\text{on}} = 4.5 \times 10^3\ \text{A}\cdot\text{m}^{-1}$ at a power supply voltage $V_{DD} = 0.4$ V [318]. In contrast to TFETs, charge transport in SLFETs does not involve BTBT, thereby explaining the exceptionally high on-state current of these devices. However, to achieve steep slopes, SLFETs require some fine-tuning of the configuration of their superlattice. Such fine-tuning is difficult to achieve in real-world devices; the SLFET has not been realized experimentally.

The SLFET is a prime example of an FET with bandstructure engineering: its source is engineered to limit (or even suppress) thermionic leakage by limiting (or even suppressing) energy states that could lead to such leakage. In recent years, several FETs with bandstructure engineering boasting sub-$\frac{k_B T \ln 10}{q}$ STS were theoretically and computationally investigated. These include FETs with source composed of MoS_2 nanoribbons [319], 2D-transition-metal-dichalcogenide "cold" metals [320], graphene nanoribbons [321, 322], and C_{31} [323] (a 2D allotrope of carbon [324]). These investigations demonstrate the power of the principle of energy filtering.

A.1.4 The Impact–Ionization FET

In semiconductors, impact–ionization refers to a physical process whereby a high-energy charge carrier (an electron or a hole) loses some of its kinetic energy to the creation of an electron–hole pair [16]; this process may be accompanied by the creation or annihilation of a phonon for momentum conservation. Exciting an electron from the valence band to the conduction band requires an energy greater than or equal to the semiconductor bandgap energy E_G. Correspondingly, a charge carrier must have a kinetic energy greater than or equal to E_G to impact–ionize an electron–hole pair. Due to thermal fluctuations, this process constantly occurs in semiconductors, even at equilibrium, albeit at a minute rate. The rate

of impact–ionization significantly increases in a region with high electric field due to:

- the acceleration (and corresponding increase in kinetic energy) of charge carriers by the field;
- the onset of a positive feedback loop where charge carriers generated through impact–ionization can generate even more charge carriers through impact–ionization.

If the field is large enough, an enormous number of charge carriers can be generated from a single charge carrier, resulting in so-called "avalanche breakdown." Electrons and holes are then swept away in opposite directions by the electric field, giving rise to a large electrical current. Metaphorically, this process is analogous to a snowball rolling down a mountain, becoming bigger, and eventually creating an avalanche. Alternatively, avalanche breakdown is analogous to the nuclear chain reaction on which an atomic bomb operates.

Quantitatively, given an external electric field $\mathbf{E}_{\text{ext}}$, the impact–ionization rate of a solitary charge carrier per unit distance travelled is given, in the effective driving field model [325, 326], by

$$\alpha \propto \exp\left(-\frac{E_{\text{crit}}}{E_{\text{ext}}}\right) , \tag{A.1}$$

where E_{crit} is the critical electric field, namely, the value of E_{ext} at the onset of avalanche breakdown, and is given by

$$E_{\text{crit}} = \frac{E_G}{q\lambda_{\text{op ph}}} , \tag{A.2}$$

where $\lambda_{\text{op ph}}$ is the optical phonon mean free path.

A graph of the impact–ionization rate in Eq. A.1 is shown in Fig. A.7. Importantly, it exhibits a super-exponential dependence on E_{ext}. One could thus envision an FET engineered in such a way that the electric field in part of its channel is controlled by the applied gate voltage V_{GS}; a low V_{GS} would result in a low field and a high V_{GS} would result in high field. This would amount to a device in which impact–ionization is switched by V_{GS}. Thanks to the super-exponential form of α, this device would have a very steep slope. This is the idea behind the impact–ionization field-effect transistor (IIFET), a device first proposed and experimentally realized in 2002 by Gopalakrishnan, Griffin, and Plummer [327–329], who reported a p-type silicon IIFET with a minimum STS of 10 $\text{mV}\cdot\text{dec}^{-1}$ and MOSFET-like

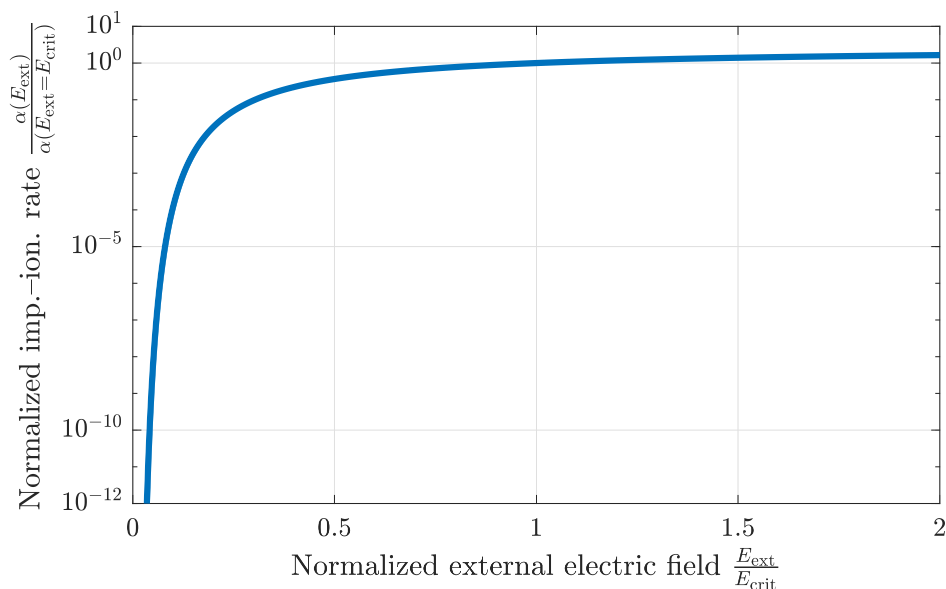


Figure A.7 – Impact–ionization rate α per unit distance travelled of a solitary charge carrier in a semiconductor [Eq. A.1] as a function of the external electric field E_{ext}. The axis of ordinates is normalized by the critical electric field E_{crit}. The axis of abscissas is normalized by $\alpha\left(E_{\text{ext}} = E_{\text{crit}}\right)$.

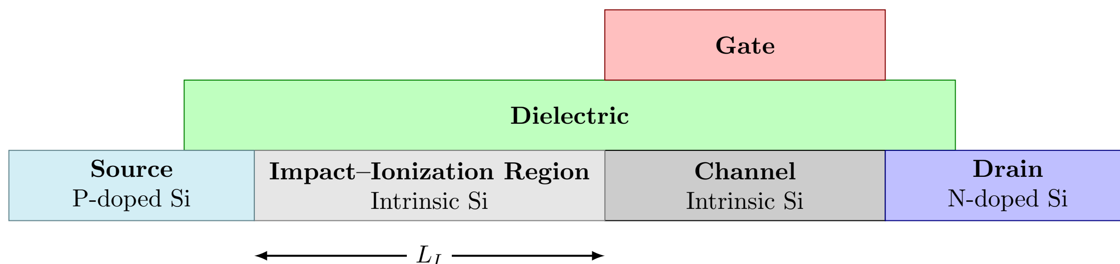


Figure A.8 – Schematic of a typical n-type enhancement-mode silicon impact–ionization field-effect transistor (IIFET). The source is p-doped, the channel is intrinsic, and the drain is n-doped. There is an ungated, intrinsic region with length L_I between the source and channel in which impact–ionization is modulated. The main charge transport mechanism is impact–ionization avalanche breakdown that arises when the electric field in the impact–ionization region exceeds a critical value.

on-state current at room temperature. The IIFET was originally called and is still commonly referred to as “I-MOS.”

The structure of an IIFET is similar to that of a TFET, with the exception that an ungated, intrinsic semiconductor layer, which can be called the “impact–ionization region,” with length L_I, is placed between the source and channel, as illustrated in Fig. A.8. The electric field in this region, and, in turn, the impact–ionization rate, are controlled by V_{GS}.

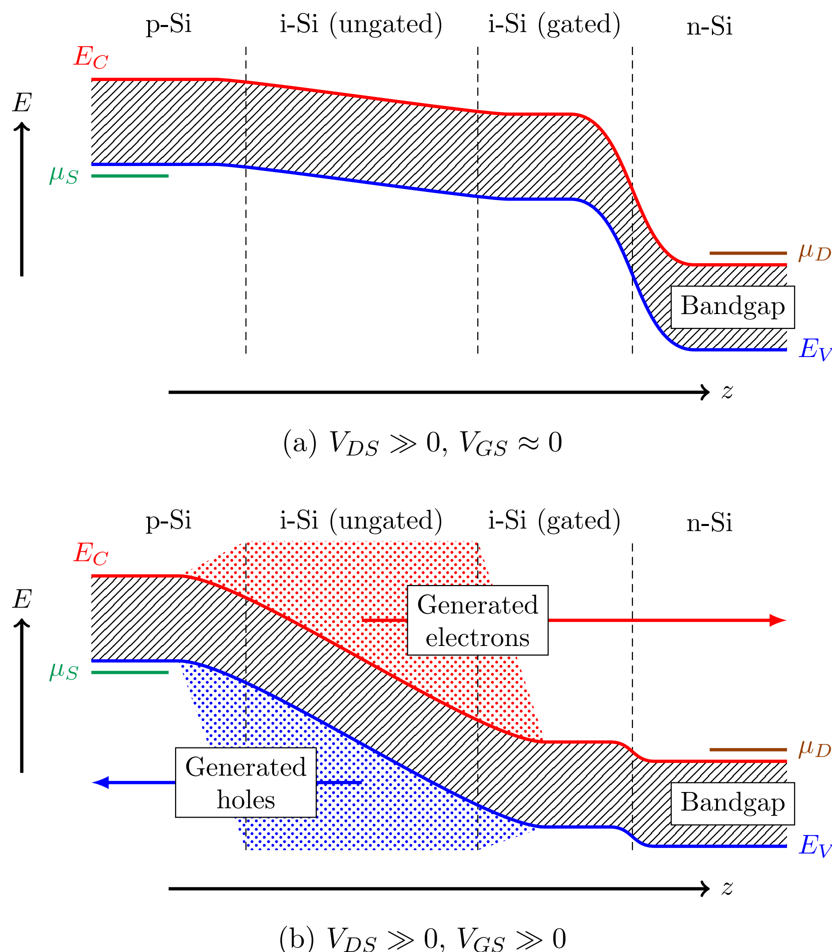


Figure A.9 – Band diagrams of a typical n-type enhancement-mode silicon IIFET in the off state (a) and on state (b); note that $V_{DS} \geq \frac{E_G}{q}$. In the off state, the electric field in the ungated intrinsic silicon region is small; correspondingly, the impact–ionization rate is small. In the on state, this electric field is large, resulting in a large impact–ionization rate, avalanche breakdown, and large drain current.

Indeed, the electric potentials in the p-doped source and in the (gated) channel are set by the source doping concentration and V_{GS}.

Off-state and on-state band diagrams of the IIFET are shown in Fig. A.9. In the off state [Fig. A.9a], $E_{\text{ext}} \ll E_{\text{crit}}$; correspondingly, the drain current is extremely small, with minute leakage caused by thermionic emission of holes and electrons and by BTBT at the channel–drain junction. In the on state [Fig. A.9b], $E_{\text{ext}} \gtrapprox E_{\text{crit}}$, resulting in substantial impact–ionization rate and avalanche breakdown. Generated electrons are swept toward the drain by the electric field, while generated holes are swept toward the source; this results in

substantial drain current.

To ensure that the generated electrons do not encounter a potential barrier in the drain, the drain potential must be smaller than or equal to the channel potential [Fig. A.9b]. Thus, assuming an ideal body factor of $-e$ [Eq. 2.19], the potential drop in the impact–ionization region must be greater than or equal to V_{DS}. Correspondingly, due to a lack of doping, the magnitude of the electric field in this region is thus roughly constant and equal to

$$E_{\text{ext}} \approx \frac{V_{DS}}{L_I}\,. \tag{A.3}$$

Combining Eqs. A.1–A.3, the onset of impact–ionization avalanche and turn-on drain voltage of the IIFET is

$$V_{DS,\text{avalanche}} \approx \frac{L_I}{\lambda_{\text{op ph}}}\frac{E_G}{q}\,. \tag{A.4}$$

Equation A.4 exhibits a possible limitation of IIFETs. Despite a very low STS, the IIFET may not be compatible with operation at low power supply voltage V_{DD}, since V_{DD} should satisfy $V_{DD} \geq V_{DS,\text{avalanche}}$. The avalanche breakdown voltage $V_{DS,\text{avalanche}}$ can be reduced by reducing L_I. However, ballistic transport arises as L_I is downscaled below $\lambda_{\text{op ph}}$ [69,231,330]; specifically, charge carriers stop scattering with phonons and, *a fortiori*, stop impact–ionizing. A best-case scenario of $\frac{L_I}{\lambda_{\text{op ph}}} \approx 1$ can thereby be inferred. A typical semiconductor has a bandgap energy $E_G \geq 0.7$ eV, which sets the minimum V_{DD} for an IIFET to around 0.7 V, which is similar to the power supply voltage for state-of-the-art MOSFETs [14].

This limitation could be solved using the recently observed phenomenon of ballistic impact–ionization [151], which has been used by Gao *et al.* to experimentally realize an IIFET based on a heterostructure of graphene, black phosphorus, and indium selenide with an on-state current $I_{\text{on}} = 100$ A$\cdot$ m^{-1} (the upper bound for indium selenide transistors [331]), minimal STS of 1 mV$\cdot$ dec^{-1} or less, and sub-40 mV$\cdot$ dec^{-1} STS sustained over 5 decades of drain current (measurements were performed at a temperature of 200 K) at a minimum drain voltage of $V_{DS} = 0.6$ V [152].

Among the steep-slope FETs reviewed in this chapter, the IIFET is one of the devices

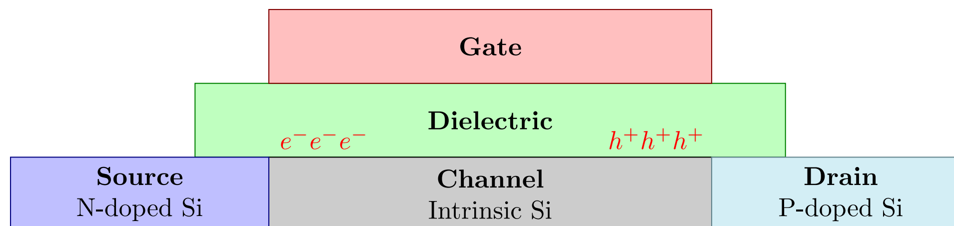


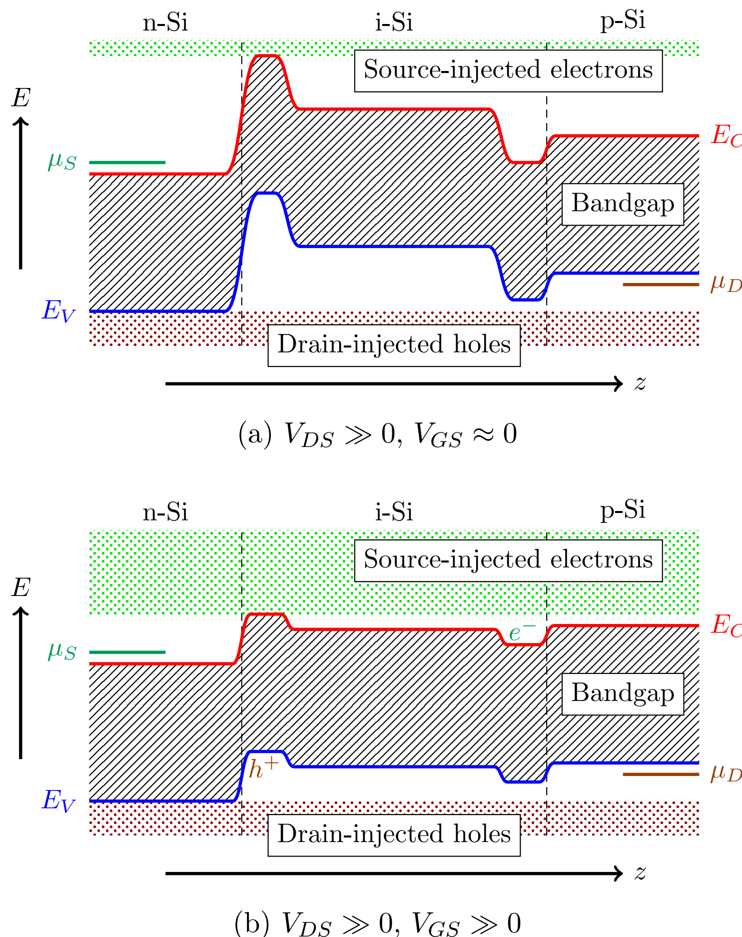
Figure A.10 – Schematic of a typical n-type enhancement-mode silicon feedback field-effect transistor (FBFET). The source is n-doped, the channel is intrinsic, and the drain is p-doped. Electrons (e^-) are trapped in the dielectric near the source–channel junction; holes (h^+) are trapped in the dielectric near the channel–drain junction. The main charge transport mechanism is thermionic emission modulated by a charge–potential positive-feedback mechanism.

with transfer characteristics closest to that of an "ideal" FET [Fig. 3.1], in the sense that it exhibits extremely sharp switching and does not suffer from low on-state current, unlike most energy-filtering FETs. Nevertheless, poor scaling of the power supply voltage remains a major concern for IIFETs, which calls for further research and innovations.

A.1.5 The Feedback FET and Band-Modulation FETs

A major cause for transistor ageing is charge trapping [332]. In an FET, due to thermal fluctuations (hot-carrier injection) and quantum fluctuations (Fowler—Nordheim tunnelling), charge carriers may get trapped in the gate dielectric, which can dramatically increase the threshold voltage. Over several years, charge trapping leads to degradation of processor performance and, ultimately, to so-called "threshold collapse." In contrast, some devices harness charge trapping. For example, in flash memory, a ubiquitous computer memory storage medium, bits of information are distinguished by the presence or absence of charge on a floating gate [333]. This floating gate is completely surrounded by insulators; the charge that it stores is, therefore, a trapped charge.

In 2008, Padilla, Yeung, Shin, Hu, and Liu introduced the feedback FET (FBFET), an FET that harnesses charges trapped in its gate dielectric and a positive-feedback mechanism to achieve a steep subthreshold slope [334]. The structure of a typical FBFET is shown in Fig. A.10. Its structure is similar to that of a TFET, except that the source (drain) of an



(a) $V_{DS} \gg 0$, $V_{GS} \approx 0$

(b) $V_{DS} \gg 0$, $V_{GS} \gg 0$

Figure A.11 – Band diagrams of a typical n-type enhancement-mode silicon FBFET in the off state (a) and on state (b); note that $V_{DS} \geq \frac{E_G}{q}$. In the off state, the large potential barriers near the channel's junctions lead to minute thermionic emission of electrons (green-dotted region) and holes (brown-dotted region). For sufficiently high V_{GS}, a charge–potential positive-feedback mechanism, which involves bound electrons (e^-) and holes (h^+), sharply increases the drain current.

FBFET is doped with polarity opposite to that of the source (drain) of a TFET. In addition, electrons (holes) are trapped in the gate oxide near the source–channel (channel–drain) by application of a very high gate-to-source (gate-to-drain) voltage; this is a preconditioning step that is required for an FBFET's feedback mechanism and steep slope [335].

Band diagrams of an FBFET are shown in Fig. A.11. In the off state, the electrons (holes) trapped in gate oxide result in a significant potential barrier for free electrons (holes) near the source–channel (channel–drain) junction, resulting in minute thermionic emission of electrons (holes) [Fig. A.11a]. When the gate voltage V_{GS} is increased beyond a certain

threshold [Fig. A.11b], a positive-feedback mechanism is initiated. Specifically:

(1) a higher V_{GS} reduces the potential barrier for electrons near the source–channel junction;
(2) this results in increased thermionic emission of source-injected electrons;
(3) this results in an increased number of electrons in the channel;
(4) this results in bound electrons in the conduction band quantum well near the channel–drain junction;
(5) this results in a shallower quantum well in the conduction band near the channel–drain junction;
(6) this results in a reduced potential barrier for holes near the channel–drain junction;
(7) this results in increased thermionic emission of drain-injected holes;
(8) this results in an increased number of holes in the channel;
(9) this results in bound holes in the valence band quantum well near the source–channel junction;
(10) this results in a shallower quantum well in the valence band near the channel–drain junction;
(11) this results in a reduced potential barrier for electrons near the source–channel junction;
(12) this results in increased thermionic emission of source-injected electrons, as in step (2).

This triggers a collapse of the potential barrier for source-injected electrons and drain-injected holes, leading to an extremely sharp increase of drain current. Due to the clever interplay between charges and energy bands on which it operates, the FBFET is sometimes referred to as a "band-modulation FET." Since the invention of the FBFET, several other band-modulation devices have been proposed and demonstrated. These include the Z^2-FET (for zero STS, zero impact–ionization) [336], which only requires trapped charge on one side of the channel, and the Z^3-FET (for zero front gate, zero STS, zero impact–ionization) [337], which does not require trapped charge but achieves band modulation with additional gates instead.

The FBFET suffers from two drawbacks. The first is the large drain voltage V_{DS} that the device requires to achieve a steep slope. Indeed, the positive feedback loop that defines the

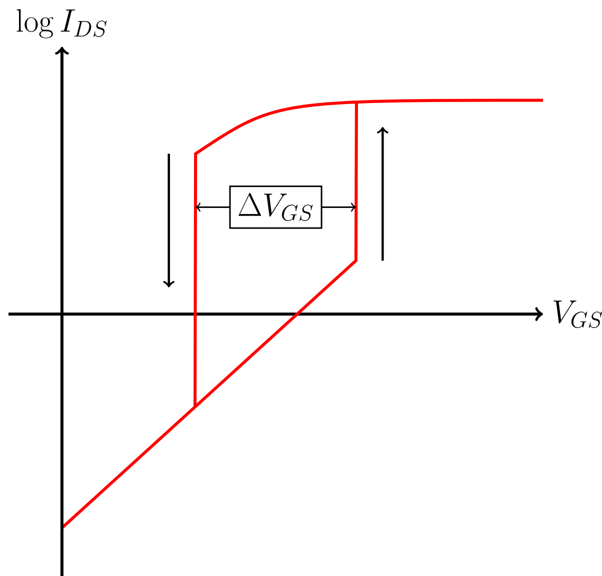


Figure A.12 – Transfer characteristics of a typical FET exhibiting hysteresis. The turn-on voltage is different than the turn-off voltage. Hysteresis is often quantified as the difference between the turn-on and turn-off voltages ΔV_{GS}.

FBFET requires comparable potential barriers for source-injected electrons and drain-injected holes; therefore, to achieve a steep slope,

$$V_{DS} \gtrapprox \frac{E_G}{q}, \tag{A.5}$$

where E_G is the semiconductor bandgap energy. This bound is similar to that required in IIFETs to trigger the impact–ionization avalanche [Eq. A.4]. It severely limits the scaling of the power supply voltage V_{DD} in FBFET-based ICs.

The second drawback is hysteretic switching, meaning that the FBFET turns on (as V_{GS} is swept forward) at a different value of V_{GS} than it turns off (as V_{GS} is swept backward). Indeed, bound electrons and holes accumulate in quantum wells as the FBFET is switched on. These bound charges and associated quantum wells remain as the device is switched off up to the point where the injection of electrons is too low to sustain them. Hysteresis is graphically illustrated in Fig. A.12; this phenomenon hinders the operation of logic gates and may limit the scaling of V_{DD}.

In 2009, Yeung *et al.* experimentally realized a silicon FBFET with on-state current

$I_{\text{on}} = 2 \times 10^1$ A·m^{-1} and minimum STS of 0.35 mV·dec^{-1} at $V_{DS} = 1.25$ V. Due to the high V_{DS} and relatively low I_{on} that the FBFET exhibits, it may not be an ideal steep-slope FET for logic applications. However, the principle of band modulation is powerful and has given rise to a wealth of research. For example, band-modulation FETs have been shown to be promising candidates for applications toward random-access memory [338], electrostatic discharge protection [339, 340], and neuromorphic computing [341].

A.2 FETs with Large Body Factors

A.2.1 The Nanoelectromechanical FET

Microelectromechanical systems (MEMS) and nanoelectromechanical systems (NEMS) are classes of devices that integrate electrical and mechanical functionalities on the microscale and nanoscale, respectively. Such systems are typically fabricated using processes similar to those used to fabricate MOSFETs and other electronics on ICs. As a result, MEMS research and manufacturing only took off in the 1970s [342], although the potential of MEMS was foreseen years before [343]. MEMS and NEMS may include parts that move in response to external stimuli and convert such mechanical signals into electrical signals. This leads to numerous applications in metrology, e.g. for accelerometers, barometers, and atomic-force microscopes [344].

Conversely, MEMS and NEMS may convert an electrical signal into a mechanical signal. Based on this principle, in 2002, Ionescu *et al.* proposed a MOSFET-like device with a suspended metal membrane as the gate [345]. This device, originally called "suspended-gate MOSFET" [345, 346] and now more commonly referred to as the "nanoelectromechanical FET" (NEMFET) [347], is illustrated in Fig. A.13. Its structure is identical to that of a MOSFET, with the exception that the metal gate is free to move; its position is dictated by the balance of mechanical forces, which tend to keep it away from the transistor channel,

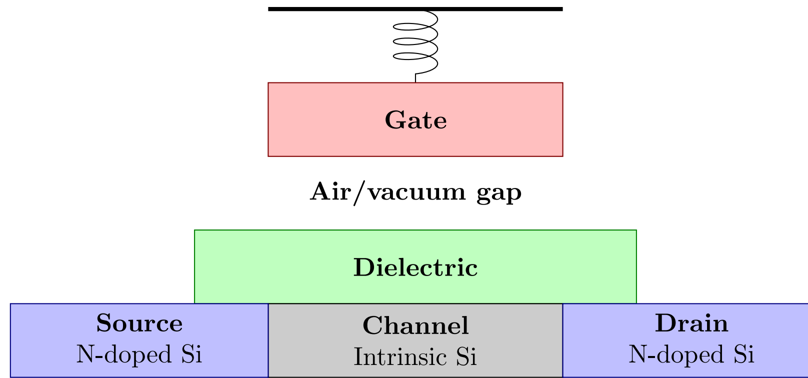


Figure A.13 – Schematic of a typical n-type enhancement-mode silicon nanoelectromechanical field-effect transistor (NEMFET). The source is n-doped, the channel is intrinsic, and the drain is n-doped. The gate is suspended and moves in response to the applied gate voltage. In the off state, the gate and dielectric are separated by an air/vacuum gap; the gate capacitance is low, and the drain current is low. In the on state, the gate makes physical contact with the dielectric; the gate capacitance is high, and the drain current is high. The main charge transport mechanism is thermionic emission.

and electrostatic forces, which tend to bring it closer to the transistor channel for sufficiently high gate voltage, in the enhancement-mode case. Therefore, the gate voltage serves two purposes in a NEMFET. The first is to control the inversion charge through the field effect, just as in any other FET. The second is to provide sufficient electrostatic force to pull the gate electrode down. For low gate voltage (in the off state), the gate sits far from the rest of the device, leaving a sizeable gap between itself and the gate dielectric. This makes the off-state gate capacitance C_{off} very low. Conversely, for high gate voltage (in the on state), the gate makes physical contact with the gate dielectric, making the on-state gate capacitance C_{on} very high. In Ref. [345], a NEMFET with $\frac{C_{\text{on}}}{C_{\text{off}}} > 100$ is reported. The sharp increase of gate capacitance as a function of gate voltage yields a body factor $\left|\frac{\partial \phi}{\partial V_{GS}}\right| > q$ breaking from Gauss's tyranny [Eq. 3.9]. Correspondingly, the STS is low. Furthermore, since NEMFETs have thermionic emission as their main charge transport mechanism, their on-state current is fairly high.

A drawback of NEMFETs is the relatively low maximal frequency at which they may operate. Indeed, the mechanical resonant frequency of the metal gate is given by

$$f_m = \frac{1}{2\pi}\sqrt{\frac{k}{m}}\,, \tag{A.6}$$

where k is the spring constant of the gate membrane and m is its mass; f_m sets the maximum frequency at which a NEMFET can be switched. For typical NEMFETs, f_m is in the MHz range, although ultra-scaled NEMFETs may have f_m in the GHz range (due in part to their ultra-scaled gate mass) [347]. Another drawback of NEMFETs is the structural damage that accumulates as their gates move up and down. To mitigate this issue, NEMFETs in which the moving part is not the metal gate but a structurally sounder semiconductor channel, such as a silicon–germanium nanowire [348], have been experimentally demonstrated. Finally, NEMFETs tend to exhibit significant hysteresis and generally exhibit a trade-off between STS and hysteresis [349], much like NCFETs.

In 2014, Kim, Chen, Kwon, and Xiang experimentally realized a silicon–germanium-nanowire-based NEMFET with minimal STS of 6 $\text{mV}\cdot\text{dec}^{-1}$ at room temperature, on-state current $I_{\text{on}} \approx 10^2$ $\text{A}\cdot\text{m}^{-1}$ at drain voltage $V_{DS} = 1$ V, and hysteresis of $\Delta V_{GS} = 1.6\text{V}$. This NEMFET reliably switched on and off for up to 130 cycles and may operate at a frequency $f = 125$ MHz [348].

A.2.2 The Piezoelectric FET

Some materials, known as piezoelectric materials, develop spontaneous polarization (and internal electric field) when mechanically strained. Conversely, piezoelectric materials can be strained by an external electric field. Semiconductor straining has been a key enabler of transistor scaling since the mobility (and on-state current) of semiconductor devices can be increased with strain [Chapter 1]. Typically, strain engineering is achieved by contacting the semiconductor channel with a strain-inducing material with an appropriate lattice constant (for instance, by using capping layers or by epitaxial growth of silicon on silicon–germanium) [16]. The amount of strain induced by such techniques does not exhibit a strong dependence on the applied gate voltage V_{GS}.

In addition to mobility, strain also modulates the bandgap energy E_G and electron affinity χ of semiconductors. This is because strain affects interatomic distances, crystal

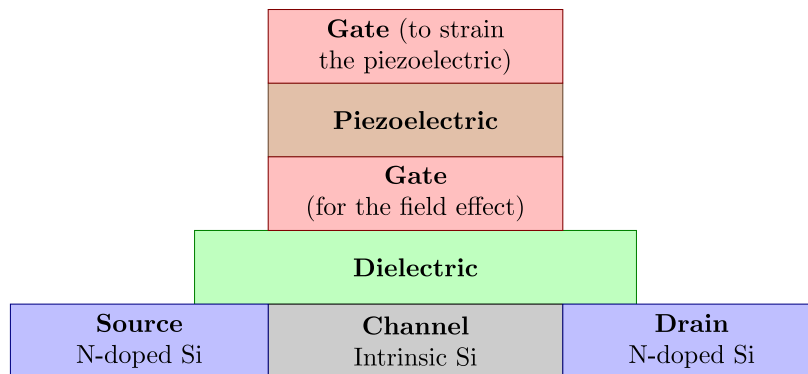


Figure A.14 – Schematic of a typical n-type enhancement-mode silicon piezoelectric field-effect transistor (PEFET). The source is n-doped, the channel is intrinsic, and the drain is n-doped. The voltage on the upper gate modulates the strain of the piezoelectric layer and, in turn, the semiconductor strain. The main charge transport mechanism is thermionic emission.

lattice symmetry, and therefore bandstructure. Since the channel barrier energy ϕ of an FET depends on χ [Eq. 2.36] and E_G (for p-type FETs), one could imagine an FET in which ϕ is modulated by V_{GS} not only electrostatically, through the field effect, but also mechanically, through a V_{GS}-induced strain. This is the motivation behind the piezoelectric FET (PEFET), a steep-slope FET proposed by Hemert and Hueting in 2012 [350, 351] and independently by Jana, Snider, and Jena in 2013 [352, 353]. The PEFET is sometimes referred to as the "π-FET."

The structure of a PEFET is identical to that of a MOSFET, with the exception that a piezoelectric layer is inserted in the gate stack. Several gate stack configurations are possible, the most basic of which consists of replacing the dielectric layer of a MOSFET by a piezoelectric layer. However, analytical models of PEFETs have shown that this configuration is unlikely to result in sub-$\frac{k_B T \ln 10}{q}$ STS [153]. Instead, a metal–piezoelectric–metal–dielectric–semiconductor gate stack offers greater potential; the corresponding PEFET is illustrated in Fig. A.14. The lower gate serves the same purpose as the gate of a MOSFET: it modulates the channel conductivity electrostatically, through the field effect. On the other hand, a voltage on the upper gate strains the piezoelectric layer through the piezoelectric effect; this strain carries over to the other layers of the gate stack, including the semiconductor channel. This strain on the channel is used to modulate E_G, χ, ϕ, as explained previously. Thus, in a PEFET, thermionic emission is modulated (1) electrostatically, through the field effect,

and (2) mechanically, through the piezoelectric effect. It is this dual modulation that enables PEFETs to break Gauss's tyranny [Eq. 3.9] and achieve low STS.

To switch the PEFET shown in Fig. A.14, not one but two gates need to be charged and discharged. This has a significant impact on dynamic power. As a result, it has been argued that it is unlikely that PEFETs can be used to reduce dynamic power in processors, although they can be used to reduce leakage power [153].

So far, a PEFET with sub-$\frac{k_B T \ln 10}{q}$ STS has not been realized experimentally. However, it has been shown experimentally that the presence of a piezoelectric layer in the gate stack of a silicon FinFET can reduce its STS by around 5 $\mathrm{mV \cdot dec^{-1}}$ [354]. Furthermore, simulations indicate that the STS of a piezoelectric FinFET could be as low as 40 $\mathrm{mV \cdot dec^{-1}}$ at room temperature [355].

A.3 FETs with Large Transport and Body Factors

So far, the steep-slope FETs investigated in this chapter had either a large transport factor or a large body factor. A MOSFET's transport factor is bounded by thermodynamics, while its body factor is bounded by electrostatics. A large transport factor requires adjustments to charge transport. In terms of device architecture, this requires alterations to the semiconductor source and/or channel (for instance, by replacing the silicon source by a graphene source, as in a DSFET). On the other hand, a large body factor requires adjustments to field-effect electrostatics. In terms of device architecture, this requires alterations to the gate stack (for instance, by inserting a ferroelectric layer in the gate stack, as in an NCFET). Large transport and body factors are thus achieved through separate and independent means, both in terms of device physics and architecture.

The subthreshold slope of an FET is given by the product of its transport and body factors [Eq. 3.1]. Thus, in principle, an FET that combines the physics and architecture of

a large-transport-factor FET and a large-body-factor FET should have an STS lower than that of both of its parents, in a two-front war on Boltzmann's tyranny and Gauss's tyranny [Eq. 3.9]. Examples of such "hybrid" devices may include a nanoelectromechanical feedback FET or a piezoelectric superlattice FET. Such amalgamations may prove to be an effective way to obtain FETs with ultra-scaled STS and processors with ultra-scaled power supply voltage. To date, hybrid FETs have not been extensively investigated, although some reports can be found in the literature.

An early example of a hybrid FET is the "hysteretic ferroelectric tunnel FET," a device experimentally realized by Ionescu *et al.* in 2010 [356]. This device combines the functionalities of a TFET and a ferroelectric FET (an NCFET without a dielectric layer in its gate stack). By harnessing energy filtering, this device achieved an STS 29 $\mathrm{mV \cdot dec^{-1}}$ lower than a ferroelectric FET with similar material composition and geometry. An important characteristic of the ferroelectric tunnel FET (and, generally, any other hybrid FET) is that it inherited the shortcomings of both its parents: low on-state current and ambipolarity from its TFET father and hysteresis from its ferroelectric FET mother. All steep-slope FETs have numerous shortcomings that, to date, have hindered their adoption by the semiconductor industry. Thus, despite exceptionally low STS, hybrid FETs may suffer from particularly serious barriers to adoption. This is exacerbated by the complex fabrication processes required to combine large body and transport factors in a single device. However, this argument would become moot if further research and development eliminated all significant shortcomings of a large-transport-factor FET and a large-body-factor FET, which could thereby be combined to form an extraordinary transistor. Thus, research on hybrid FETs remains important.

Arguably, the most obvious hybrid FET is a negative-capacitance tunnel FET; indeed, the TFET (NCFET) is the most studied large-transport-factor (large-body-factor) FET. This device has been investigated in numerous theoretical [357–359] and experimental [360] reports; an STS of 18 $\mathrm{mV \cdot dec^{-1}}$ was demonstrated in simulations in Ref. [358]. Furthermore, I have theoretically studied a negative-capacitance Dirac-source FET and a negative-capacitance cold-source FET, and showed that negative capacitance can reduce the threshold voltage of a DSFET and a CSFET by 30% [7]; however, this work will not be further discussed in this

thesis.

A particularly interesting hybrid FET is the piezoelectric tunnel FET [361, 362]. Indeed, the current of a TFET decays exponentially with bandgap energy [Eq. 3.7]. On the other hand, through the converse piezoelectric effect, the bandgap in a PEFET's channel is narrower in the on state than in the off state. Thus, a significant boost not only to STS but also to on-state current can be achieved by adding piezoelectric functionality to a TFET. In 2018, Saeidi *et al.* experimentally demonstrated that the converse piezoelectric effect increases the on-state current of a TFET by a factor of 100. This device is thus better than the sum of its parts; it is a unique example of a hybrid FET with less significant shortcomings than its parents.

Appendix B

Validation of Simulations

This Appendix presents two "sanity checks" regarding the NEGF–TB algorithm described in Sec. 4.4 to obtain the data pertaining to BCE presented in Chapter 5. In Sec. B.1, it is established that the density of real-space mesh points used to numerically solve Poisson's equation is sufficient to fully capture the electrostatics of the simulated systems. Section B.2 investigates the extent to which the smoothness of the position-dependent relative permittivity affects the simulation results.

B.1 Density of Mesh Points

Poisson's equation [Eq. 4.38] is solved numerically using the finite difference method—on a cubic, uniform mesh. In most of the simulations presented in this thesis, a real-space mesh point density of $n = 0.1944$ Å^{-3} is used. To confirm that this mesh point density is sufficient, in Fig. B.1, potentials computed on this sparser mesh with $n = 0.1944$ Å^{-3} (V_s, red curves) and potentials computed on a denser mesh with $n = 1.5551$ Å^{-3} (V_d, blue curves) are compared. Overall, as can be seen from Figs. B.1a and B.1b, V_s and V_d very closely overlap each other. In particular, increasing the mesh point density for Poisson's equation does not affect the depletion length ℓ at the source–channel and channel–drain interfaces, and thereby does not affect the conclusions drawn in Chapter 5 regarding BCE. Quantitatively, as can be seen from Figs. B.1c and B.1d, the error induced by the use of the sparser mesh compared to the denser mesh is always lower than about 10 mV, and is typically on the order of 1 mV

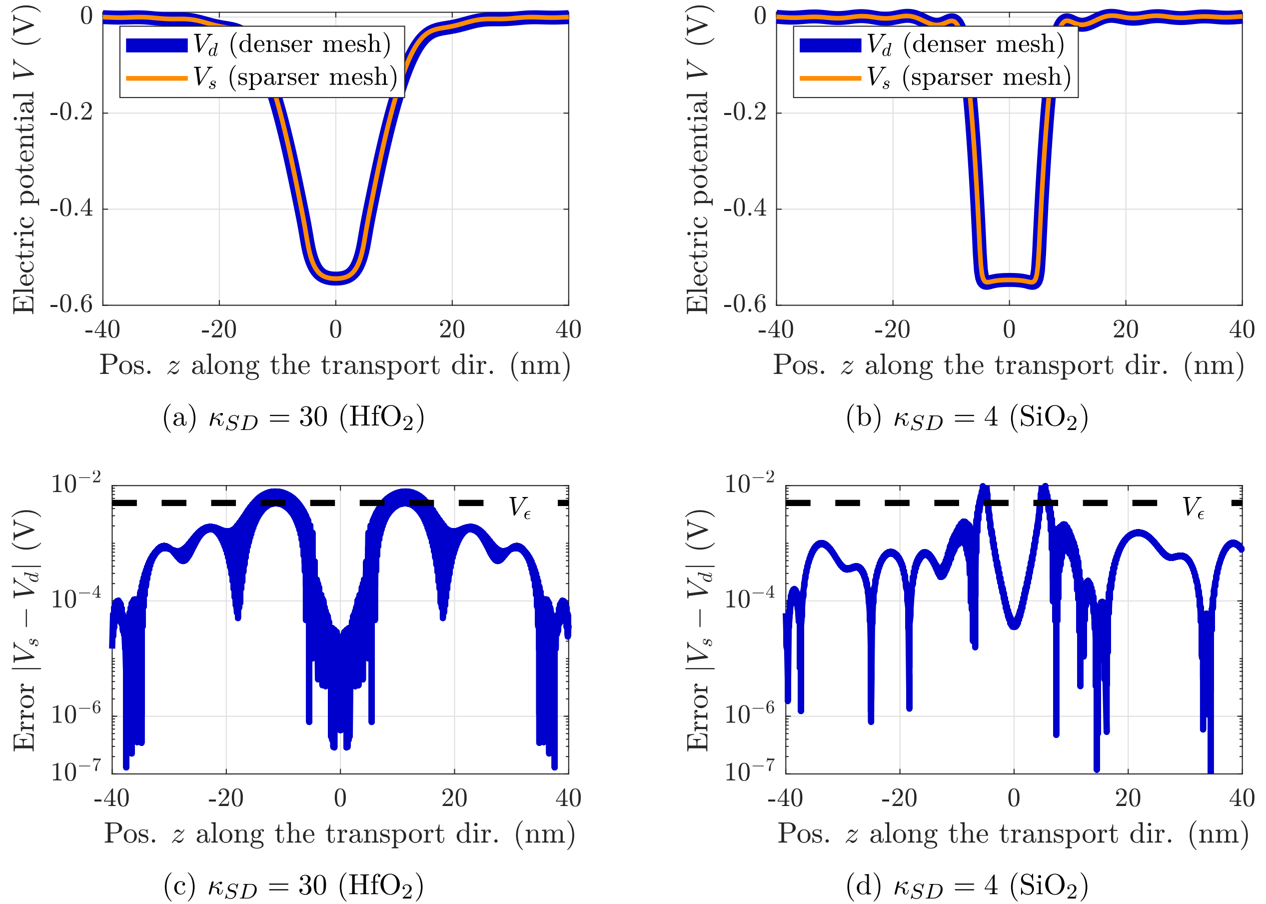


Figure B.1 – (a) and (c) Electric potential V as a function of position z along the transport direction from the channel centre in n-type $d = 1$ nm-wide cylindrical GAA silicon NW MOSFETs [Fig. 5.3b], as obtained from NEGF–TB simulations. The source and drain doping concentration is $N_D = 2 \times 10^{20}$ cm^{-3} and the gate oxide is HfO_2 ($\kappa_G = 30$). The drain and gate voltages are, respectively, set to $V_{DS} = 0$ V and $V_{GS} = 0.5$ V. In (a), the source and drain oxide is HfO_2 ($\kappa_{SD} = 30$); in (c), it is SiO_2 ($\kappa_{SD} = 4$). Potentials as computed with mesh points densities of $n = 0.1944$ Å^{-3} (V_s) and $n = 1.5551$ Å^{-3} (V_d) are compared. (b) and (d) Absolute values of the differences between V_s and V_d in (a) and (c), respectively. The convergence threshold of $V_\epsilon = 5$ mV [Eq. 4.42] is shown as a red line.

or less—well within the convergence threshold V_ϵ of the iterative algorithm described in Sec. 4.4 (black line). It can be concluded that using a mesh point density of $n = 0.1944$ Å^{-3} to numerically solve Poisson's equation is sufficient to fully capture the electrostatics of the investigated systems and the physics of BCE.

B.2 Smoothness of the Permittivity

BCE deals with systems with position-dependent permittivity $\kappa(\mathbf{r})$. Chapter 5 investigates systems with permittivity shown in Fig. 5.3b, namely

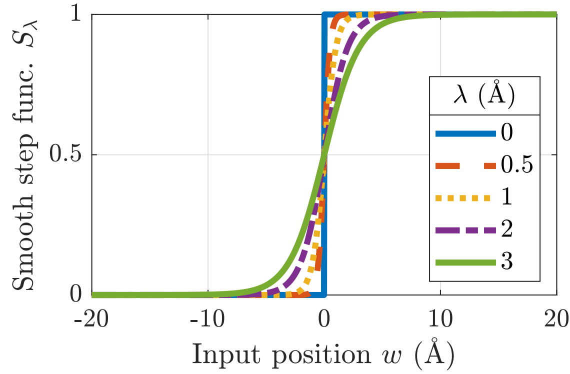
$$\kappa(\mathbf{r}) = \begin{cases} \kappa_{\mathrm{Si}} & \text{for } \sqrt{x^2+y^2} < \frac{d}{2} \\ \kappa_G & \text{for } \sqrt{x^2+y^2} > \frac{d}{2} \text{ and } |z| < \frac{L}{2} \\ \kappa_{SD} & \text{for } \sqrt{x^2+y^2} > \frac{d}{2} \text{ and } |z| > \frac{L}{2} \end{cases}, \tag{B.1}$$

where (x, y, z) are the Cartesian coordinates of the position vector $\mathbf{r}$; the origin $(0, 0, 0)$ is located at the centre of the transistor channel. The terms κ_{Si}, κ_G, and κ_{SD} are the relative permittivities of the silicon NW, gate oxide, and source and drain oxide, respectively; d is the silicon NW diameter, and L is the channel length. Alternatively, Eq. B.1 may be expressed as

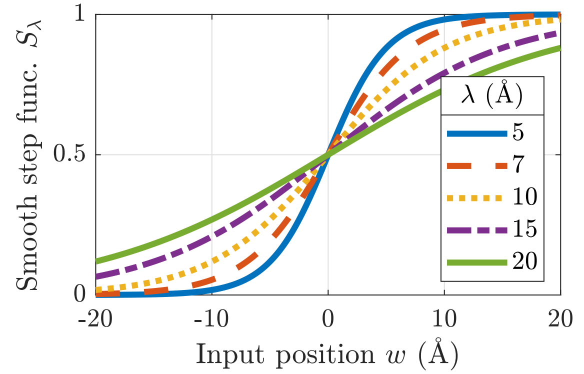
$$\begin{aligned} \kappa(\mathbf{r}) = {} & \Theta\left(\frac{d}{2} - \sqrt{x^2+y^2}\right)\kappa_{\mathrm{Si}} \\ & +\Theta\left(\sqrt{x^2+y^2} - \frac{d}{2}\right)\Theta\left(\frac{L}{2} - |z|\right)\kappa_G \\ & +\Theta\left(\sqrt{x^2+y^2} - \frac{d}{2}\right)\Theta\left(|z| - \frac{L}{2}\right)\kappa_{SD}\,, \end{aligned} \tag{B.2}$$

where, in the simulations presented in Chapter 5, $\Theta(w)$ is taken to be equal to the Heaviside step function:

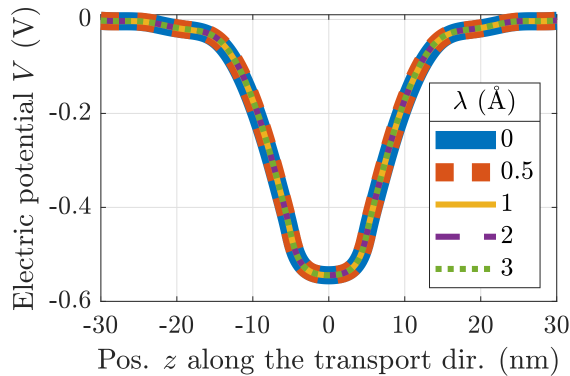
$$H(w) = \begin{cases} 0 & \text{for } w < 0 \\ 1 & \text{for } w > 0 \end{cases}. \tag{B.3}$$



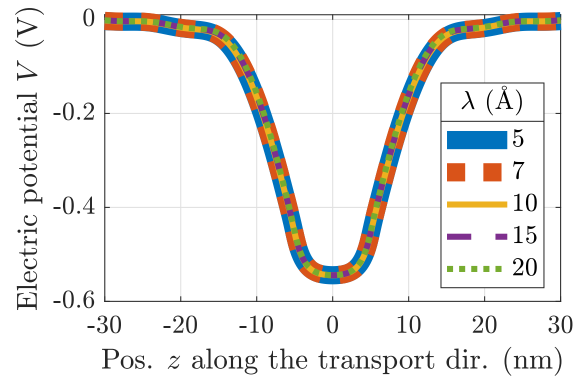
(a) Smooth step function



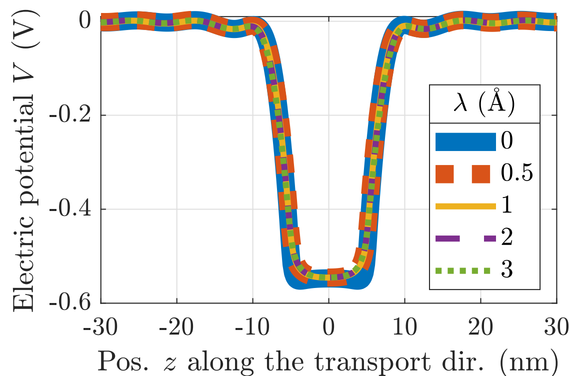
(b) Smooth step function



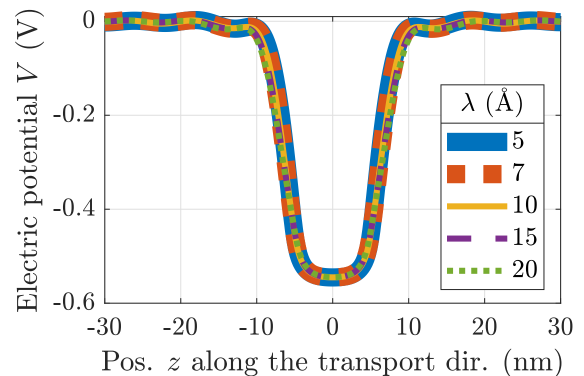
(c) $\kappa_{SD} = 30$ (HfO_2)



(d) $\kappa_{SD} = 30$ (HfO_2)



(e) $\kappa_{SD} = 4$ (SiO_2)



(f) $\kappa_{SD} = 4$ (SiO_2)

Figure B.2 – (a)–(b) Graphs of the “smooth step functions” S_λ [Eq. B.4] for values of λ ranging from 0 Å to 20 Å. (c)–(f) Electric potential V as a function of position z along the transport direction from the channel centre in n-type $d = 1$ nm-wide cylindrical GAA silicon NW MOSFETs [Fig. 5.3b], as obtained from NEGF–TB simulations. The source and drain doping concentration is $N_D = 2 \times 10^{20}$ cm^{-3} and the gate oxide is HfO_2 ($\kappa_G = 30$). The drain and gate voltages are, respectively, set to $V_{DS} = 0$ V and $V_{GS} = 0.5$ V. In (c)–(d), the source and drain oxide relative permittivity is $\kappa_{SD} = 30$ (HfO_2); in (e)–(f), it is $\kappa_{SD} = 4$ (SiO_2). Potentials as computed with position-dependent relative permittivities $\kappa(\mathbf{r})$ [Eq. B.2] for various values of λ are compared.

Alternatively, one might consider a smoother permittivity, where instead, $\Theta(w)$ is taken to be equal to

$$S_\lambda(w) = \frac{1}{2}\left[1 + \tanh\left(\frac{w}{\lambda}\right)\right], \tag{B.4}$$

where λ represents a lengthscale over which $\kappa(\mathbf{r})$ varies near interfaces between two materials. Graphs of the functions S_λ for various values of λ are plotted in Figs. B.2a and B.2b. In the limiting case of $\lambda \to 0$, S_λ approaches the Heaviside step function H.

Figure B.2 investigates the effect of varying the smoothness of $\kappa(\mathbf{r})$ [Eq. B.2] through tuning of the parameter λ [Eq. B.4]. Overall, Figs. B.2c–B.2f show that the electric potential profile $V(z)$ is not strongly affected by variations in λ. For example, increasing λ from 0 Å to 20 Å leads to variations in $V(z)$ well below the convergence threshold of $V_\epsilon = 5$ mV [Eq. 4.42] for most values of z. This excludes values of z within a few λ of the source–channel and channel–drain interfaces, where potential variations may be as large as 0.1 V. These variations are expected and do not lead to strong variations in charge transport. Moreover, the length ℓ of the depletion regions at the source–channel and channel–drain interfaces appears to be independent of λ, even when λ is as large as the oxide thickness. It can be concluded that, on both qualitative and quantitative levels, the smoothness of the position-dependent permittivity $\kappa(\mathbf{r})$ does not strongly affect the physics of the devices investigated in Chapter 5.

Bibliography

[1] R. J. Prentki, M. Harb, L. Liu, and H. Guo, "Nanowire transistors with bound-charge engineering," *Physical Review Letters*, vol. 125, no. 24, p. 247704, Dec. 2020. DOI: 10.1103/PhysRevLett.125.247704

[2] R. J. Prentki, "A surface potential model for field-effect transistors with bound-charge engineering," *IEEE Transactions on Electron Devices*, vol. 68, no. 9, pp. 4625–4629, July 2021. DOI: 10.1109/TED.2021.3096779

[3] R. J. Prentki and H. Guo, "Ultra-short-channel transistors with minimal tunnelling leakage," manuscript under preparation for submission to a peer-reviewed journal. 2022.

[4] R. J. Prentki, F. Liu, and H. Guo, "Modeling of ballistic monolayer black phosphorus MOSFETs," *IEEE Transactions on Electron Devices*, vol. 66, no. 8, pp. 3668–3674, July 2019. DOI: 10.1109/TED.2019.2924170

[5] W. Gan, R. J. Prentki, F. Liu, J. Bu, K. Luo, Q. Zhang, H. Zhu, W. Wang, T. Ye, H. Yin, Z. Wu, and H. Guo, "Design and simulation of steep-slope silicon cold source FETs with effective carrier distribution model," *IEEE Transactions on Electron Devices*, vol. 67, no. 6, pp. 2243–2248, May 2020. DOI: 10.1109/TED.2020.2988855

[6] W. Gan, R. J. Prentki, F. Liu, J. Bu, Q. Zhang, H. Zhu, H. Yin, W. Wang, T. Ye, Z. Wu, and H. Guo, "A multi-physics TCAD framework for fast and accurate simulation of steep-slope Si-based cold source FET," in *International Symposium on VLSI Technology, Systems and Applications (VLSI-TSA)*, pp. 66–67, Aug. 2020. DOI: 10.1109/VLSI-TSA48913.2020.9203735

[7] S. Guo, R. J. Prentki, K. Jin, C.-l. Chen, and H. Guo, "Negative-capacitance FET with a cold source," *IEEE Transactions on Electron Devices*, vol. 68, no. 2, pp. 911–918, Dec. 2020. DOI: 10.1109/TED.2020.3041216

[8] W. Gan, R. J. Prentki, K. Luo, J. Huo, W. Huang, Q. Huo, J. Bu, R. Cao, Y. Lu, H. Yin, H. Guo, and Z. Wu, "Multi-physics evaluation of silicon steep-slope cold source FET," in *IEEE Electron Devices Technology & Manufacturing Conference (EDTM)*, Apr. 2021. DOI: 10.1109/EDTM50988.2021.9420963

[9] W. Gan, K. Luo, G. Qi, R. J. Prentki, F. Liu, J. Huo, W. Huang, J. Bu, Q. Zhang, H. Yin, H. Guo, Y. Lu, and Z. Wu, "A multiscale simulation framework for steep-slope Si nanowire cold source FET," *IEEE Transactions on Electron Devices*, vol. 68, no. 11, pp. 5455–5461, June 2021. DOI: 10.1109/TED.2021.3083602

[10] E. K. Carlson, "Harnessing bound charge in semiconductors," *Physics*, vol. 13, p. s155, Dec. 2020. DOI: 10.1103/Physics.13.s155

[11] *Le Système international d'unités*, 9th ed. Bureau International des Poids et Mesures, 2019. URL: https://www.bipm.org/en/publications/si-brochure

[12] G. E. Moore, "Cramming more components onto integrated circuits," *Electronics*, vol. 38, no. 8, p. 114, Apr. 1965. DOI: 10.1109/N-SSC.2006.4785860

[13] R. H. Dennard, F. H. Gaensslen, V. L. Rideout, E. Bassous, and A. R. LeBlanc, "Design of ion-implanted MOSFET's with very small physical dimensions," *IEEE Journal of Solid-State Circuits*, vol. 9, no. 5, pp. 256–268, Oct. 1974. DOI: 10.1109/JSSC.1974.1050511

[14] *International Roadmap for Devices and Systems.* IEEE, 2020. URL: https://irds.ieee.org/editions

[15] T. N. Theis and P. M. Solomon, "In quest of the 'next switch': Prospects for greatly reduced power dissipation in a successor to the silicon field-effect transistor," *Proceedings of the IEEE*, vol. 98, no. 12, pp. 2005–2014, Oct. 2010. DOI: 10.1109/JPROC.2010.2066531

[16] S. M. Sze and K. K. Ng, *Physics of Semiconductor Devices.* John Wiley & Sons, 2006. DOI: 10.1002/0470068329

[17] G. D. Wilk, R. M. Wallace, and J. M. Anthony, "High-κ gate dielectrics: Current status and materials properties considerations," *Journal of Applied Physics*, vol. 89, no. 10, pp. 5243–5275, May 2001. DOI: 10.1063/1.1361065

[18] G. D. Wilk and R. M. Wallace, "Electrical properties of hafnium silicate gate dielectrics deposited directly on silicon," *Applied Physics Letters*, vol. 74, no. 19, pp. 2854–2856, May 1999. DOI: 10.1063/1.124036

[19] G. D. Wilk, R. M. Wallace, and J. M. Anthony, "Hafnium and zirconium silicates for advanced gate dielectrics," *Journal of Applied Physics*, vol. 87, no. 1, pp. 484–492, Jan. 2000. DOI: 10.1063/1.371888

[20] L. Risch, "Pushing CMOS beyond the roadmap," *Solid-State Electronics*, vol. 50, no. 4, pp. 527–535, Apr. 2006. DOI: 10.1016/j.sse.2006.03.026

[21] H.-S. P. Wong, K. K. Chan, and Y. Taur, "Self-aligned (top and bottom) double-gate MOSFET with a 25 nm thick silicon channel," in *International Electron Devices Meeting (IEDM) Technical Digest*, pp. 427–430, Dec. 1997. DOI: 10.1109/IEDM.1997.650416

[22] D. Hisamoto, T. Kaga, Y. Kawamoto, and E. Takeda, "A fully depleted lean-channel transistor (DELTA)—A novel vertical ultra thin SOI MOSFET," in *International Electron Devices Meeting (IEDM) Technical Digest*, pp. 833–836, Dec. 1989. DOI: 10.1109/IEDM.1989.74182

[23] H. Takato, K. Sunouchi, N. Okabe, A. Nitayama, K. Hieda, F. Horiguchi, and F. Masuoka, "High performance CMOS surrounding gate transistor (SGT) for ultra high density LSIs," in *International Electron Devices Meeting (IEDM) Technical Digest*, pp. 222–225, Dec. 1988. DOI: 10.1109/IEDM.1988.32796

[24] Y. Sun, S. E. Thompson, and T. Nishida, "Physics of strain effects in semiconductors and metal–oxide–semiconductor field-effect transistors," *Journal of Applied Physics*, vol. 101, no. 10, p. 104503, May 2007. DOI: 10.1063/1.2730561

[25] S. W. Bedell, A. Khakifirooz, and D. K. Sadana, "Strain scaling for CMOS," *MRS Bulletin*, vol. 39, no. 2, pp. 131–137, Feb. 2014. DOI: 10.1557/mrs.2014.5

[26] H. M. Manasevit, I. S. Gergis, and A. B. Jones, "Electron mobility enhancement in epitaxial multilayer Si-$Si_{1-x}Ge_x$ alloy films on (100) Si," *Applied Physics Letters*, vol. 41, no. 5, pp. 464–466, Sep. 1982. DOI: 10.1063/1.93533

[27] W. Shockley and J. Bardeen, "Energy bands and mobilities in monatomic semiconductors," *Physical Review*, vol. 77, no. 3, pp. 407–408, Feb. 1950. DOI: 10.1103/PhysRev.77.407

[28] J. Bardeen and W. Shockley, "Deformation potentials and mobilities in non-polar crystals," *Physical Review*, vol. 80, no. 1, pp. 72–80, Oct. 1950. DOI: 10.1103/PhysRev.80.72

[29] A. N. Ziogas, T. Ben-Nun, G. I. Fernández, T. Schneider, M. Luisier, and T. Hoefler, "Optimizing the data movement in quantum transport simulations via data-centric parallel programming," in *SC '19: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, p. 78, Nov. 2019. DOI: 10.1145/3295500.3356200

[30] "What percentage of internet traffic is mobile?" URL: https://www.oberlo.com/statistics/mobile-internet-traffic

[31] L. Chua, "Memristor—The missing circuit element," *IEEE Transactions on Circuit Theory*, vol. 18, no. 5, pp. 507–519, Sep. 1971. DOI: 10.1109/TCT.1971.1083337

[32] A. K. Maan, D. A. Jayadevi, and A. P. James, "A survey of memristive threshold logic circuits," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 8, pp. 1734–1746, May 2016. DOI: 10.1109/TNNLS.2016.2547842

[33] C. Mead, "Neuromorphic electronic systems," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1629–1636, Oct. 1990. DOI: 10.1109/5.58356

[34] J. D. Kendall and S. Kumar, "The building blocks of a brain-inspired computer," *Applied Physics Reviews*, vol. 7, no. 1, p. 011305, Jan. 2020. DOI: 10.1063/1.5129306

[35] D. Loss and D. P. DiVincenzo, "Quantum computation with quantum dots," *Physical Review A*, vol. 57, no. 1, pp. 120–126, Jan. 1998. DOI: 10.1103/PhysRevA.57.120

[36] P. Benioff, "The computer as a physical system: A microscopic quantum mechanical Hamiltonian model of computers as represented by Turing machines," *Journal of Statistical Physics*, vol. 22, no. 5, pp. 563–591, May 1980. DOI: 10.1007/BF01011339

[37] R. P. Feynman, "Simulating physics with computers," *International Journal of Theoretical Physics*, vol. 21, no. 6–7, pp. 467–488, June 1982. DOI: 10.1007/BF02650179

[38] A. Papinikolau, D. Soudris, and R. Radojcic, *Three Dimensional System Integration*. Springer, 2011. DOI: 10.1007/978-1-4419-0962-6

[39] D. Shamiryan, T. Abell, F. Iacopi, and K. Maex, "Low-k dielectric materials," *Materials Today*, vol. 7, no. 1, pp. 34–39, Jan. 2004. DOI: 10.1016/S1369-7021(04)00053-7

[40] G. Fiori, F. Bonaccorso, G. Iannaccone, T. Palacios, D. Neumaier, A. Seabaugh, S. K. Banerjee, and L. Colombo, "Electronics based on two-dimensional materials," *Nature Nanotechnology*, vol. 9, pp. 768–779, Oct. 2014. DOI: 10.1038/nnano.2014.207

[41] A. M. Ionescu and H. Riel, "Tunnel field-effect transistors as energy-efficient electronic switches," *Nature*, vol. 479, pp. 329–337, Nov. 2011. DOI: 10.1038/nature10679

[42] N. W. Ashcroft and N. D. Mermin, *Solid State Physics*. Brooks Cole, 1976.

[43] R. J. Van Overstraeten and R. P. Mertens, "Heavy doping effects in silicon," *Solid-State Electronics*, vol. 30, no. 11, pp. 1077–1087, Nov. 1987. DOI: 10.1016/0038-1101(87)90070-0

[44] S. Tiwari and S. L. Wright, "Material properties of p-type GaAs at large dopings," *Applied Physics Letters*, vol. 56, no. 6, pp. 563–565, Feb. 1990. DOI: 10.1063/1.102745

[45] S. C. Jain and D. J. Roulston, "A simple expression for band gap narrowing (BGN) in heavily doped Si, Ge, GaAs and Ge_xSi_{1-x} strained layers," *Solid-State Electronics*, vol. 34, no. 5, pp. 453–465, May 1991. DOI: 10.1016/0038-1101(91)90149-S

[46] J. D. Plummer and P. B. Griffin, "Material and process limits in silicon VLSI technology," *Proceedings of the IEEE*, vol. 89, no. 3, pp. 240–258, Mar. 2001. DOI: 10.1109/5.915373

[47] D. J. Griffiths, *Introduction to Electrodynamics*, 4th ed. Pearson, 2013. DOI: 10.1017/9781108333511

[48] X. Zhao, C. M. Wei, L. Yang, and M. Y. Chou, "Quantum confinement and electronic properties of silicon nanowires," *Physical Review Letters*, vol. 92, no. 23, p. 236805, June 2004. DOI: 10.1103/PhysRevLett.92.236805

[49] J. Maassen and H. Guo, "Suppressing leakage by localized doping in Si nanotransistor channels," *Physical Review Letters*, vol. 109, no. 26, p. 266803, Dec. 2012. DOI: 10.1103/PhysRevLett.109.266803

[50] L. Li, Y. Yu, G. J. Ye, Q. Ge, X. Ou, H. Wu, D. Feng, X. H. Chen, and Y. Zhang, "Black phosphorus field-effect transistors," *Nature Nanotechnology*, vol. 9, pp. 372–377, Mar. 2014. DOI: 10.1038/nnano.2014.35

[51] S. P. Koenig, R. A. Doganov, H. Schmidt, A. H. Castro Neto, and B. Özyilmaz, "Electric field effect in ultrathin black phosphorus," *Applied Physics Letters*, vol. 104, no. 10, p. 103106, Mar. 2014. DOI: 10.1063/1.4868132

[52] K. F. Mak, C. Lee, J. Hone, J. Shan, and T. F. Heinz, "Atomically thin MoS_2: A new direct-gap semiconductor," *Physical Review Letters*, vol. 105, no. 13, p. 136805, Sep. 2010. DOI: 10.1103/PhysRevLett.105.136805

[53] M. Menon and D. Srivastava, "Carbon nanotube 'T junctions': Nanoscale metal–semiconductor–metal contact devices," *Physical Review Letters*, vol. 79, no. 22, pp. 4453–4456, Dec. 1997. DOI: 10.1103/PhysRevLett.79.4453

[54] J. Appenzeller, Y.-M. Lin, J. Knoch, and P. Avouris, "Band-to-band tunneling in carbon nanotube field-effect transistors," *Physical Review Letters*, vol. 93, no. 19, p. 196805, Nov. 2004. DOI: 10.1103/PhysRevLett.93.196805

[55] X. Wang, Y. Ouyang, X. Li, H. Wang, J. Guo, and H. Dai, "Room-temperature all-semiconducting sub-10-nm graphene nanoribbon field-effect transistors," *Physical Review Letters*, vol. 100, no. 20, p. 206803, May 2008. DOI: 10.1103/PhysRevLett.100.206803

[56] N. Loubet, T. Hook, P. Montanini, C.-W. Yeung, S. Kanakasabapathy, M. Guillom, T. Yamashita, J. Zhang, X. Miao, J. Wang, A. Young, R. Chao, M. Kang, Z. Liu, S. Fan, B. Hamieh, S. Sieg, Y. Mignot, W. Xu, S.-C. Seo, J. Yoo, S. Mochizuki, M. Sankarapandian, O. Kwon, A. Carr, A. Greene, Y. Park, J. Frougier, R. Galatage, R. Bao, J. Shearer, R. Conti, H. Song, D. Lee, D. Kong, Y. Xu, A. Arceo, Z. Bi, P. Xu, R. Muthinti, J. Li, R. Wong, D. Brown, P. Oldiges, R. Robison, J. Arnold, N. Felix, S. Skordas, J. Gaudiello, T. Standaert, H. Jagannathan, D. Corliss, M.-H. Na, A. Knorr, T. Wu, D. Gupta, S. Lian, R. Divakaruni, T. Gow, C. Labelle, S. Lee, V. Paruchuri, H. Bu, and M. Khare, "Stacked nanosheet gate-all-around transistor to enable scaling

beyond FinFET," in *Symposium on VLSI Technology*, pp. T230–T231, Aug. 2017. DOI: 10.23919/VLSIT.2017.7998183

[57] F. Liu, C. Qiu, Z. Zhang, L.-M. Peng, J. Wang, Z. Wu, and H. Guo, "First principles simulation of energy efficient switching by source density of states engineering," in *International Electron Devices Meeting (IEDM) Technical Digest*, p. 33.2, Dec. 2018. DOI: 10.1109/IEDM.2018.8614597

[58] R.-H. Yan, A. Ourmazd, and K. F. Lee, "Scaling the Si MOSFET: From bulk to SOI to bulk," *IEEE Transactions on Electron Devices*, vol. 39, no. 7, pp. 1704–1710, July 1992. DOI: 10.1109/16.141237

[59] H. Bruus and K. Flensberg, *Many-Body Quantum Theory in Condensed Matter Physics: An Introduction*. Oxford University Press, 2004.

[60] H. Kroemer, "Quasi-electric fields and band offsets: Teaching electrons new tricks," in *Nobel Lecture*, Dec. 2000. DOI: 10.1142/S0217979202010245

[61] A. Ortiz-Conde, F. G. Sánchez, J. J. Liou, A. Cerdeira, M. Estrada, and Y. Yue, "A review of recent MOSFET threshold voltage extraction methods," *Microelectronics reliability*, vol. 42, no. 4, pp. 583–596, Apr. 2002. DOI: 10.1016/S0026-2714(02)00027-6

[62] R. J. Prentki, "Compact modeling of ballistic monolayer black phosphorus metal–oxide–semiconductor field-effect transistors," Master's thesis, McGill University, Aug. 2017. URL: https://escholarship.mcgill.ca/concern/theses/b2773z17n

[63] S. Veeraraghavan and J. G. Fossum, "Short-channel effects in SOI MOSFETs," *IEEE Transactions on Electron Devices*, vol. 36, no. 3, pp. 522–528, Mar. 1989. DOI: 10.1109/16.19963

[64] J. Wu, J. Min, and Y. Taur, "Short-channel effects in tunnel FETs," *IEEE Transactions on Electron Devices*, vol. 62, no. 9, pp. 3019–3024, Aug. 2015. DOI: 10.1109/TED.2015.2458977

[65] K. K. Young, "Short-channel effect in fully depleted SOI MOSFET's," *IEEE Transactions on Electron Devices*, vol. 36, no. 2, pp. 399–402, Feb. 1989. DOI: 10.1109/16.19942

[66] D. J. Griffiths and D. F. Schroeter, *Introduction to Quantum Mechanics*, 3rd ed. Cambridge University Press, 2018. DOI: 10.1017/9781316995433

[67] J. Wang and M. Lundstrom, "Does source-to-drain tunneling limit the ultimate scaling of MOSFETs?" in *International Electron Devices Meeting (IEDM) Technical Digest*, pp. 707–710, Dec. 2002. DOI: 10.1109/IEDM.2002.1175936

[68] K. P. Cheung, "On the 60 mV/dec @300 K limit for MOSFET subthreshold swing," in *Proceedings of 2010 International Symposium on VLSI Technology, System and Application*, pp. 72–73, Apr. 2010. DOI: 10.1109/VTSA.2010.5488941

[69] A. Rahman, J. Guo, S. Datta, and M. S. Lundstrom, "Theory of ballistic nanotransistors," *IEEE Transactions on Electron Devices*, vol. 50, no. 9, pp. 1853–1864, Aug. 2003. DOI: 10.1109/TED.2003.815366

[70] V. V. Zhirnov and R. K. Cavin, "Negative capacitance to the rescue?" *Nature Nanotechnology*, vol. 3, pp. 77–78, Feb. 2008. DOI: 10.1038/nnano.2008.18

[71] R. Chau, S. Datta, M. Doczy, B. Doyle, B. Jin, J. Kavalieros, A. Majumdar, M. Metz, and M. Radosavljevic, "Benchmarking nanotechnology for high-performance and low-power logic transistor applications," *IEEE Transactions on Nanotechnology*, vol. 4, no. 2, pp. 153–158, Mar. 2005. DOI: 10.1109/TNANO.2004.842073

[72] M. César, D. Liu, D. Gall, and H. Guo, "Calculated resistances of single grain boundaries in copper," *Physical Review Applied*, vol. 2, no. 4, p. 044007, Oct. 2014. DOI: 10.1103/PhysRevApplied.2.044007

[73] M. César, D. Gall, and H. Guo, "Reducing grain-boundary resistivity of copper nanowires by doping," *Physical Review Applied*, vol. 5, no. 5, p. 054018, May 2016. DOI: 10.1103/PhysRevApplied.5.054018

[74] H. J. M. Veendrick, "Short-circuit dissipation of static CMOS circuitry and its impact on the design of buffer circuits," *IEEE Journal of Solid-State Circuits*, vol. 19, no. 4, pp. 468–473, Aug. 1984. DOI: 10.1109/JSSC.1984.1052168

[75] K. Nose and T. Sakurai, "Analysis and future trend of short-circuit power," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 19, no. 9, pp. 1023–1030, Sep. 2000. DOI: 10.1109/43.863642

[76] R. Jammy, "More Moore or more than Moore?" in *3rd Annual SEMATECH Symposium Taiwan 2010*, pp. 173–185, Sep. 2010.

[77] Y.-C. Yeo, T.-J. King, and C. Hu, “MOSFET gate leakage modeling and selection guide for alternative gate dielectrics based on leakage considerations,” *IEEE Transactions on Electron Devices*, vol. 50, no. 4, pp. 1027–1035, June 2003. DOI: 10.1109/TED.2003.812504

[78] K. Nose and T. Sakurai, “Optimization of V_{DD} and V_{TH} for low-power and high speed applications,” in *Proceedings of the 2000 Asia and South Pacific Design Automation Conference*, pp. 469–474, Jan. 2000. DOI: 10.1109/ASPDAC.2000.835145

[79] G. Yeap, S. S. Lin, Y. M. Chen, H. L. Shang, P. W. Wang, H. C. Lin, Y. C. Peng, J. Y. Sheu, M. Wang, X. Chen, B. R. Yang, C. P. Lin, F. C. Yang, Y. K. Leung, D. W. Lin, C. P. Chen, K. F. Yu, D. H. Chen, C. Y. Chang, H. K. Chen, P. Hung, C. S. Hou, Y. K. Cheng, J. Chang, L. Yuan, C. K. Lin, C. C. Chen, Y. C. Yeo, M. H. Tsai, H. T. Lin, C. O. Chui, K. B. Huang, W. Chang, H. J. Lin, K. W. Chen, R. Chen, S. H. Sun, Q. Fu, H. T. Yang, H. T. Chiang, C. C. Yeh, T. L. Lee, C. H. Wang, S. L. Shue, C. W. Wu, R. Lu, W. R. Lin, J. Wu, F. Lai, Y. H. Wu, B. Z. Tien, Y. C. Huang, L. C. Lu, J. He, Y. Ku, J. Lin, M. Cao, T. S. Chang, and S. M. Jang, “5nm CMOS production technology platform featuring full-fledged EUV, and high mobility channel FinFETs with densest 0.021 μm^2 SRAM cells for mobile SoC and high performance computing applications,” in *International Electron Devices Meeting (IEDM) Technical Digest*, p. 36.7, Dec. 2019. DOI: 10.1109/IEDM19573.2019.8993577

[80] J. C. Liu, S. Mukhopadhyay, A. Kundu, S. H. Chen, H. C. Wang, D. S. Huang, J. H. Lee, M. I. Wang, R. Lu, S. S. Lin, Y. M. Chen, H. L. Shang, P. W. Wang, H. C. Lin, G. Yeap, and J. He, “A reliability enhanced 5nm CMOS technology featuring 5th generation FinFET with fully-developed EUV and high mobility channel for mobile SoC and high performance computing application,” in *International Electron Devices Meeting (IEDM) Technical Digest*, p. 9.2, Dec. 2020. DOI: 10.1109/IEDM13553.2020.9372009

[81] C. Adelmann, L. G. Wen, A. P. Peter, Y. K. Siew, K. Croes, J. Swerts, M. Popovici, K. Sankaran, G. Pourtois, S. Van Elshocht, J. Bömmels, and Z. Tőkei, “Alternative metals for advanced interconnects,” in *IEEE International Interconnect Technology Conference*, pp. 173–176, May 2014. DOI: 10.1109/IITC.2014.6831863

[82] L. G. Wen, C. Adelmann, O. V. Pedreira, S. Dutta, M. Popovici, B. Briggs, N. Heylen, K. Vanstreels, C. J. Wilson, S. Van Elshocht, K. Croes, J. Bömmels, and Z. Tőkei, "Ruthenium metallization for advanced interconnects," in *2016 IEEE International Interconnect Technology Conference/Advanced Metallization Conference (IITC/AMC)*, pp. 34–36, May 2016. DOI: 10.1109/IITC-AMC.2016.7507651

[83] S. Dutta, S. Kundu, A. Gupta, G. Jamieson, J. F. G. Granados, J. Bömmels, C. J. Wilson, Z. Tőkei, and C. Adelmann, "Highly scaled ruthenium interconnects," *IEEE Electron Device Letters*, vol. 38, no. 7, pp. 949–951, May 2017. DOI: 10.1109/LED.2017.2709248

[84] R. D. Miller, "In search of low-k dielectrics," *Science*, vol. 286, no. 5439, pp. 421–423, Oct. 1999. DOI: 10.1126/science.286.5439.421

[85] J. Reynard, C. Verove, E. Sabouret, P. Motte, B. Descouts, C. Chaton, J. Michailos, and K. Barla, "Integration of fluorine-doped silicon oxide in copper pilot line for 0.12-μm technology," *Microelectronic Engineering*, vol. 60, no. 1-2, pp. 113–118, Jan. 2002. DOI: 10.1016/S0167-9317(01)00586-X

[86] B. D. Hatton, K. Landskron, W. J. Hunks, M. R. Bennett, D. Shukaris, D. D. Perovic, and G. A. Ozin, "Materials chemistry for low-k materials," *Materials Today*, vol. 9, no. 3, pp. 22–31, Mar. 2006. DOI: 10.1016/S1369-7021(06)71387-6

[87] Y. Xiao, Z. Ma, C. Prawoto, C. Zhou, and M. Chan, "Ultralow-k dielectric with structured pores for interconnect delay reduction," *IEEE Transactions on Electron Devices*, vol. 67, no. 5, pp. 2071–2075, Apr. 2020. DOI: 10.1109/TED.2020.2983230

[88] S. Natarajan, M. Agostinelli, S. Akbar, M. Bost, A. Bowonder, V. Chikarmane, S. Chouksey, A. Dasgupta, K. Fischer, Q. Fu, T. Ghani, M. Giles, S. Govindaraju, R. Grover, W. Han, D. Hanken, E. Haralson, M. Haran, M. Heckscher, R. Heussner, P. Jain, R. James, R. Jhaveri, I. Jin, H. Kam, E. Karl, C. Kenyon, M. Liu, Y. Luo, R. Mehandru, S. Morarka, L. Neiberg, P. Packan, A. Paliwal, C. Parker, P. Patel, R. Patel, C. Pelto, L. Pipes, P. Plekhanov, M. Prince, S. Rajamani, J. Sandford, B. Sell, S. Sivakumar, P. Smith, B. Song, K. Tone, T. Troeger, J. Wiedemer, M. Yang, and K. Zhang, "A 14nm logic technology featuring 2nd-generation FinFET, air-gapped interconnects, self-aligned

double patterning and a 0.0588 μm^2 SRAM cell size," in *International Electron Devices Meeting (IEDM) Technical Digest*, p. 3.7, Dec. 2014. DOI: 10.1109/IEDM.2014.7046976

[89] C. Prawoto, Z. Ma, Y. Xiao, S. Raju, C. Zhou, and M. Chan, "Interconnect technology with h-BN-capped air-gaps," *IEEE Electron Device Letters*, vol. 40, no. 11, pp. 1876–1879, Sep. 2019. DOI: 10.1109/LED.2019.2944418

[90] C. H. Bennett, "Logical reversibility of computation," *IBM Journal of Research and Development*, vol. 17, no. 6, pp. 525–532, Nov. 1973. DOI: 10.1147/rd.176.0525

[91] C. H. Bennett, "The thermodynamics of computation—A review," *International Journal of Theoretical Physics*, vol. 21, no. 12, pp. 905–940, Dec. 1982. DOI: 10.1007/BF02084158

[92] M. Saeedi and I. L. Markov, "Synthesis and optimization of reversible circuits—A survey," *ACM Computing Surveys (CSUR)*, vol. 45, no. 2, p. 21, Feb. 2013. DOI: 10.1145/2431211.2431220

[93] I. K. Hänninen, C. O. Campos-Aguillón, R. Celis-Cordova, and G. L. Snider, "Design and fabrication of a microprocessor using adiabatic CMOS and Bennett clocking," in *International Conference on Reversible Computation*, pp. 173–185, June 2015. DOI: 10.1007/978-3-319-20860-2_11

[94] B. Behin-Aein, S. Salahuddin, and S. Datta, "Switching energy of ferromagnetic logic bits," *IEEE Transactions on Nanotechnology*, vol. 8, no. 4, pp. 505–514, Mar. 2009. DOI: 10.1109/TNANO.2009.2016657

[95] R. Landauer, "Irreversibility and heat generation in the computing process," *IBM Journal of Research and Development*, vol. 5, no. 3, pp. 183–191, July 1961. DOI: 10.1147/rd.53.0183

[96] A. Bérut, A. Arakelyan, A. Petrosyan, S. Ciliberto, R. Dillenschneider, and E. Lutz, "Experimental verification of Landauer's principle linking information and thermodynamics," *Nature*, vol. 483, pp. 187–189, Mar. 2012. DOI: 10.1038/nature10872

[97] Y. Jun, M. Gavrilov, and J. Bechhoefer, "High-precision test of Landauer's principle in a feedback trap," *Physical Review Letters*, vol. 113, no. 19, p. 190601, Nov. 2014. DOI: 10.1103/PhysRevLett.113.190601

[98] J. Hong, B. Lambson, S. Dhuey, and J. Bokor, "Experimental test of Landauer's principle in single-bit operations on nanomagnetic memory bits," *Science Advances*, vol. 2, no. 3, p. e1501492, Mar. 2016. DOI: 10.1126/sciadv.1501492

[99] J. Koomey, S. Berard, M. Sanchez, and H. Wong, "Implications of historical trends in the electrical efficiency of computing," *IEEE Annals of the History of Computing*, vol. 33, no. 3, pp. 46–54, Mar. 2010. DOI: 10.1109/MAHC.2010.28

[100] C. Enz, A. Beckers, and F. Jazaeri, "Cryo-CMOS compact modeling," in *International Electron Devices Meeting (IEDM) Technical Digest*, p. 25.3, Dec. 2020. DOI: 10.1109/IEDM13553.2020.9371894

[101] F. Tavernier, A. Gatti, and C. Barretto, "Chip design for future gravitational wave detectors," in *International Electron Devices Meeting (IEDM) Technical Digest*, p. 25.4, Dec. 2020. DOI: 10.1109/IEDM13553.2020.9372071

[102] M. S. Lundstrom and C. Jeong, *Near-Equilibrium Transport: Fundamentals and Applications.* World Scientific Publishing Company, 2012. DOI: 10.1142/7975

[103] C. Zener, "A theory of the electrical breakdown of solid dielectrics," *Proceedings of the Royal Society of London A*, vol. 145, no. 855, pp. 523–529, July 1934. DOI: 10.1098/rspa.1934.0116

[104] O. M. Stuetzer, "Junction fieldistors," *Proceedings of the IRE*, vol. 40, no. 11, pp. 1377–1381, Nov. 1952. DOI: 10.1109/JRPROC.1952.273965

[105] L. Esaki, "New phenomenon in narrow germanium p–n junctions," *Physical Review*, vol. 109, no. 2, pp. 603–604, Oct. 1958. DOI: 10.1103/PhysRev.109.603

[106] S. R. Hofstein and G. Warfield, "The insulated gate tunnel junction triode," *IEEE Transactions on Electron Devices*, vol. 12, no. 2, pp. 66–76, Feb. 1965. DOI: 10.1109/T-ED.1965.15455

[107] J. J. Quinn, G. Kawamoto, and B. D. McCombe, "Subband spectroscopy by surface channel tunneling," *Surface Science*, vol. 73, pp. 190–196, May 1978. DOI: 10.1016/0039-6028(78)90489-2

[108] S. Banerjee, W. Richardson, J. Coleman, and A. Chatterjee, “A new three-terminal tunnel device,” *IEEE Electron Device Letters*, vol. 8, no. 8, pp. 347–349, Aug. 1987. DOI: 10.1109/EDL.1987.26655

[109] A. C. Seabaugh and Q. Zhang, “Low-voltage tunnel transistors for beyond CMOS logic,” *Proceedings of the IEEE*, vol. 98, no. 12, pp. 2095–2110, Oct. 2010. DOI: 10.1109/JPROC.2010.2070470

[110] J. Knoch, S. Mantl, and J. Appenzeller, “Impact of the dimensionality on the performance of tunneling FETs: Bulk versus one-dimensional devices,” *Solid-State Electronics*, vol. 51, no. 4, pp. 572–578, Apr. 2007. DOI: 10.1016/j.sse.2007.02.001

[111] M. Luisier and G. Klimeck, “Simulation of nanowire tunneling transistors: From the Wentzel–Kramers–Brillouin approximation to full-band phonon-assisted tunneling,” *Journal of Applied Physics*, vol. 107, no. 8, p. 084507, Apr. 2010. DOI: 10.1063/1.3386521

[112] K. Boucart, W. Riess, and A. M. Ionescu, “Lateral strain profile as key technology booster for all-silicon tunnel FETs,” *IEEE Electron Device Letters*, vol. 30, no. 6, pp. 656–658, Apr. 2009. DOI: 10.1109/LED.2009.2018127

[113] T. Krishnamohan, D. Kim, S. Raghunathan, and K. Saraswat, “Double-gate strained-Ge heterostructure tunneling FET (TFET) with record high drive currents and $\ll$ 60 mV/dec subthreshold slope,” in *International Electron Devices Meeting (IEDM) Technical Digest*, Dec. 2008. DOI: 10.1109/IEDM.2008.4796839

[114] O. M. Nayfeh, C. N. Chleirigh, J. Hennessy, L. Gomez, J. L. Hoyt, and D. A. Antoniadis, “Design of tunneling field-effect transistors using strained-silicon/strained-germanium type-II staggered heterojunctions,” *IEEE Electron Device Letters*, vol. 29, no. 9, pp. 1074–1077, Sep. 2008. DOI: 10.1109/LED.2008.2000970

[115] F. Conzatti, M. G. Pala, D. Esseni, E. Bano, and L. Selmi, “Strain-induced performance improvements in InAs nanowire tunnel FETs,” *IEEE Transactions on Electron Devices*, vol. 59, no. 8, pp. 2085–2092, June 2012. DOI: 10.1109/TED.2012.2200253

[116] M. Visciarelli, E. Gnani, A. Gnudi, S. Reggiani, and G. Baccarani, “Impact of strain on tunneling current and threshold voltage in III–V nanowire TFETs,” *IEEE Electron Device Letters*, vol. 37, no. 5, pp. 560–563, Mar. 2016. DOI: 10.1109/LED.2016.2539389

[117] Q. Shi, L. Zhang, Y. Zhu, L. Liu, M. Chan, and H. Guo, "Atomic disorder scattering in emerging transistors by parameter-free first principle modeling," in *International Electron Devices Meeting (IEDM) Technical Digest*, p. 30.6, Dec. 2014. DOI: 10.1109/IEDM.2014.7047144

[118] M. Luisier and G. Klimeck, "Performance comparisons of tunneling field-effect transistors made of InSb, carbon, and GaSb–InAs broken gap heterostructures," in *International Electron Devices Meeting (IEDM) Technical Digest*, p. 37.6, Dec. 2009. DOI: 10.1109/IEDM.2009.5424280

[119] D. K. Mohata, R. Bijesh, S. Mujumdar, C. Eaton, R. Engel-Herbert, T. Mayer, V. Narayanan, J. M. Fastenau, D. Loubychev, A. K. Liu, and S. Datta, "Demonstration of MOSFET-like on-current performance in arsenide/antimonide tunnel FETs with staggered hetero-junctions for 300 mV logic applications," in *International Electron Devices Meeting (IEDM) Technical Digest*, p. 33.5, Dec. 2011. DOI: 10.1109/IEDM.2011.6131665

[120] P. Long, J. Z. Huang, M. Povolotskyi, G. Klimeck, and M. J. W. Rodwell, "High-current tunneling FETs with $(1\bar{1}0)$ orientation and a channel heterojunction," *IEEE Electron Device Letters*, vol. 37, no. 3, pp. 345–348, Jan. 2016. DOI: 10.1109/LED.2016.2523269

[121] A. Vladimirescu, A. Amara, and C. Anghel, "An analysis on the ambipolar current in Si double-gate tunnel FETs," *Solid-State Electronics*, vol. 70, pp. 67–72, Jan. 2012. DOI: 10.1016/j.sse.2011.11.009

[122] K. R. Pasupathy, T. S. Manivannan, and G. Lakshminarayanan, "A review of engineering techniques to suppress ambipolarity in tunnel FET," *Silicon*, Feb. 2021. DOI: 10.1007/s12633-021-01018-2

[123] V. Nagavarapu, R. Jhaveri, and J. C. S. Woo, "The tunnel source (PNPN) n-MOSFET: A novel high performance transistor," *IEEE Transactions on Electron Devices*, vol. 55, no. 4, pp. 1013–1019, Mar. 2008. DOI: 10.1109/TED.2008.916711

[124] R. Jhaveri, V. Nagavarapu, and J. C. S. Woo, "Effect of pocket doping and annealing schemes on the source-pocket tunnel field-effect transistor," *IEEE Transactions on Electron Devices*, vol. 58, no. 1, pp. 80–86, Nov. 2010. DOI: 10.1109/TED.2010.2089525

[125] F. Mayer, C. Le Royer, J.-F. Damlencourt, K. Romanjek, F. Andrieu, C. Tabone, B. Previtali, and S. Deleonibus, "Impact of SOI, $Si_{1-x}Ge_xOI$ and GeOI substrates on CMOS compatible tunnel FET performance," in *International Electron Devices Meeting (IEDM) Technical Digest*, Dec. 2008. DOI: 10.1109/IEDM.2008.4796641

[126] E. Memisevic, J. Svensson, M. Hellenbrand, E. Lind, and L.-E. Wernersson, "Vertical InAs/GaAsSb/GaSb tunneling field-effect transistor on Si with S = 48 mV/decade and I_{on} = 10 μA/μm for I_{off} = 1 nA/μm at V_{DS} = 0.3 V," in *International Electron Devices Meeting (IEDM) Technical Digest*, p. 19.1, Dec. 2016. DOI: 10.1109/IEDM.2016.7838450

[127] K. Tomioka, H. Gamo, J. Motohisa, and T. Fukui, "Vertical gate-all-around tunnel FETs using InGaAs nanowire/Si with core-multishell structure," in *International Electron Devices Meeting (IEDM) Technical Digest*, p. 21.1, Dec. 2020. DOI: 10.1109/IEDM13553.2020.9371991

[128] M. Si, C.-J. Su, C. Jiang, N. J. Conrad, H. Zhou, K. D. Maize, G. Qiu, C.-T. Wu, A. Shakouri, M. A. Alam, and P. D. Ye, "Steep-slope hysteresis-free negative capacitance MoS_2 transistors," *Nature Nanotechnology*, vol. 13, pp. 24–28, Dec. 2017. DOI: 10.1038/s41565-017-0010-1

[129] V. C. Lo, "Simulation of thickness effect in thin ferroelectric films using Landau–Khalatnikov theory," *Journal of Applied Physics*, vol. 94, no. 5, pp. 3353–3359, Aug. 2003. DOI: 10.1063/1.1598275

[130] A. I. Khan, K. Chatterjee, B. Wang, S. Drapcho, L. You, C. Serrao, S. R. Bakaul, R. Ramesh, and S. Salahuddin, "Negative capacitance in a ferroelectric capacitor," *Nature Materials*, vol. 14, pp. 182–186, Dec. 2015. DOI: 10.1038/nmat4148

[131] A. Islam Khan, D. Bhowmik, P. Yu, S. Joo Kim, X. Pan, R. Ramesh, and S. Salahuddin, "Experimental evidence of ferroelectric negative capacitance in nanoscale heterostructures," *Applied Physics Letters*, vol. 99, no. 11, p. 113501, Sep. 2011. DOI: 10.1063/1.3634072

[132] W. Gao, A. Khan, X. Marti, C. Nelson, C. Serrao, J. Ravichandran, R. Ramesh, and S. Salahuddin, "Room-temperature negative capacitance in a ferroelectric–dielectric

superlattice heterostructure," *Nano Letters*, vol. 14, no. 10, pp. 5814–5819, Sep. 2014. DOI: 10.1021/nl502691u

[133] S. Salahuddin and S. Datta, "Use of negative capacitance to provide voltage amplification for low power nanoscale devices," *Nano Letters*, vol. 8, no. 2, pp. 405–410, Dec. 2007. DOI: 10.1021/nl071804g

[134] S. Salahuddin and S. Datta, "Can the subthreshold swing in a classical FET be lowered below 60 mV/decade?" in *International Electron Devices Meeting (IEDM) Technical Digest*, Dec. 2008. DOI: 10.1109/IEDM.2008.4796789

[135] J. H. Park, G. S. Jang, H. Y. Kim, K. H. Seok, H. J. Chae, S. K. Lee, and S. K. Joo, "Sub-kT/q subthreshold-slope using negative capacitance in low-temperature polycrystalline-silicon thin-film transistor," *Scientific Reports*, vol. 6, p. 24734, Apr. 2016. DOI: 10.1038/srep24734

[136] M. H. Park, Y. H. Lee, H. J. Kim, Y. J. Kim, T. Moon, K. D. Kim, J. Mueller, A. Kersch, U. Schroeder, T. Mikolajick, and C. S. Hwang, "Ferroelectricity and antiferroelectricity of doped thin HfO_2-based films," *Advanced Materials*, vol. 27, no. 11, pp. 1811–1831, Feb. 2015. DOI: 10.1002/adma.201404531

[137] J. Wang, D. Wang, Q. Li, A. Zhang, D. Gao, M. Guo, J. Feng, Z. Fan, D. Chen, M. Qin, M. Zeng, X. Gao, G. Zhou, X. Lu, and J.-M. Liu, "Excellent ferroelectric properties of $Hf_{0.5}Zr_{0.5}O_2$ thin films induced by Al_2O_3 dielectric layer," *IEEE Electron Device Letters*, vol. 40, no. 12, pp. 1937–1940, Nov. 2019. DOI: 10.1109/LED.2019.2950916

[138] M. H. Lee, S.-T. Fan, C.-H. Tang, P.-G. Chen, Y.-C. Chou, H.-H. Chen, J.-Y. Kuo, M.-J. Xie, S.-N. Liu, M.-H. Liao, C.-A. Jong, K.-S. Li, M.-C. Chen, and C. W. Liu, "Physical thickness 1.x nm ferroelectric HfZrOx negative capacitance FETs," in *International Electron Devices Meeting (IEDM) Technical Digest*, p. 12.1, Dec. 2016. DOI: 10.1109/IEDM.2016.7838400

[139] E. Ko, J. W. Lee, and C. Shin, "Negative capacitance FinFET with sub-20-mV/decade subthreshold slope and minimal hysteresis of 0.48 V," *IEEE Electron Device Letters*, vol. 38, no. 4, pp. 418–421, Feb. 2017. DOI: 10.1109/LED.2017.2672967

[140] D. Kwon, S. Cheema, N. Shanker, K. Chatterjee, Y.-H. Liao, A. J. Tan, C. Hu, and S. Salahuddin, “Negative capacitance FET with 1.8-nm-thick Zr-doped HfO_2 oxide,” *IEEE Electron Device Letters*, vol. 40, no. 6, pp. 993–996, Apr. 2019. DOI: 10.1109/LED.2019.2912413

[141] M. Hoffmann, S. Slesazeck, U. Schroeder, and T. Mikolajick, “What’s next for negative capacitance electronics?” *Nature Electronics*, vol. 3, pp. 504–506, Sep. 2020. DOI: 10.1038/s41928-020-00474-9

[142] J. A. Kittl, B. Obradovic, D. Reddy, T. Rakshit, R. M. Hatcher, and M. S. Rodder, “On the validity and applicability of models of negative capacitance and implications for MOS applications,” *Applied Physics Letters*, vol. 113, no. 4, p. 042904, July 2018. DOI: 10.1063/1.5036984

[143] Z. Liu, M. A. Bhuiyan, and T. P. Ma, “A critical examination of ‘quasi-static negative capacitance’ (QSNC) theory,” in *International Electron Devices Meeting (IEDM) Technical Digest*, p. 31.2, Dec. 2018. DOI: 10.1109/IEDM.2018.8614614

[144] Z. Liu, H. Jiang, B. Ordway, and T. P. Ma, “Unveiling the apparent ‘negative capacitance’ effects resulting from pulse measurements of ferroelectric–dielectric bilayer capacitors,” *IEEE Electron Device Letters*, vol. 41, no. 10, pp. 1492–1495, Sep. 2020. DOI: 10.1109/LED.2020.3020857

[145] M. A. Alam, M. Si, and P. D. Ye, “A critical review of recent progress on negative capacitance field-effect transistors,” *Applied Physics Letters*, vol. 114, no. 9, p. 090401, Mar. 2019. DOI: 10.1063/1.5092684

[146] W. Cao and K. Banerjee, “Is negative capacitance FET a steep-slope logic switch?” *Nature Communications*, vol. 11, p. 196, Jan. 2020. DOI: 10.1038/s41467-019-13797-9

[147] P. Wu and J. Appenzeller, “Artificial sub-60 millivolts/decade switching in a metal–insulator–metal–insulator–semiconductor transistor without a ferroelectric component,” *ACS Nano*, vol. 15, no. 3, pp. 5158–5164, Mar. 2021. DOI: 10.1021/acsnano.0c10344

[148] Q. Huang, R. Huang, Y. Pan, S. Tan, and Y. Wang, “Resistive-gate field-effect transistor: A novel steep-slope device based on a metal–insulator–metal–oxide gate

stack," *IEEE Electron Device Letters*, vol. 35, no. 8, pp. 877–879, June 2014. DOI: 10.1109/LED.2014.2327219

[149] N. Shukla, A. V. Thathachary, A. Agrawal, H. Paik, A. Aziz, D. G. Schlom, S. K. Gupta, R. Engel-Herbert, and S. Datta, "A steep-slope transistor based on abrupt electronic phase transition," *Nature Communications*, vol. 6, p. 7812, Aug. 2015. DOI: 10.1038/ncomms8812

[150] M. Nadeem, I. Di Bernardo, X. Wang, M. S. Fuhrer, and D. Culcer, "Overcoming Boltzmann's tyranny in a transistor via the topological quantum field effect," *Nano Letters*, vol. 21, no. 7, pp. 3155–3161, Mar. 2021. DOI: 10.1021/acs.nanolett.1c00378

[151] A. Gao, J. Lai, Y. Wang, Z. Zhu, J. Zeng, G. Yu, N. Wang, W. Chen, T. Cao, W. Hu, D. Sun, X. Chen, F. Miao, Y. Shi, and X. Wang, "Observation of ballistic avalanche phenomena in nanoscale vertical InSe/BP heterostructures," *Nature Nanotechnology*, vol. 14, pp. 217–222, Jan. 2019. DOI: 10.1038/s41565-018-0348-z

[152] A. Gao, Z. Zhang, L. Li, B. Zheng, C. Wang, Y. Wang, T. Cao, Y. Wang, S.-J. Liang, F. Miao, Y. Shi, and X. Wang, "Robust impact–ionization field-effect transistor based on nanoscale vertical graphene/black phosphorus/indium selenide heterostructures," *ACS Nano*, vol. 14, no. 1, pp. 434–441, Dec. 2019. DOI: 10.1021/acsnano.9b06140

[153] R. J. E. Hueting, T. Van Hemert, B. Kaleli, R. A. M. Wolters, and J. Schmitz, "On device architectures, subthreshold swing, and power consumption of the piezoelectric field-effect transistor (π-FET)," *IEEE Journal of the Electron Devices Society*, vol. 3, no. 3, pp. 149–157, Mar. 2015. DOI: 10.1109/JEDS.2015.2409303

[154] A. Beckers, M. Thewissen, and B. Sorée, "Energy filtering in silicon nanowires and nanosheets using a geometric superlattice and its use for steep-slope transistors," *Journal of Applied Physics*, vol. 124, no. 14, p. 144304, Oct. 2018. DOI: 10.1063/1.5043543

[155] W. E. Pauli, letter to R. E. Peierls translated from German, published in *Scientific Correspondence with Bohr, Einstein, Heisenberg and others. Volume II: 1930–1939*, Sep. 1931. DOI: 10.1007/978-3-540-78801-0

[156] F. Bloch, "Über die quantenmechanik der elektronen in kristallgittern," *Zeitschrift für Physik*, vol. 52, no. 7, pp. 555–600, July 1929. DOI: 10.1007/BF01339455

[157] M. Harb, V. Michaud-Rioux, Y. Zhu, L. Liu, L. Zhang, and H. Guo, "Quantum transport modelling of silicon nanobeams using heterogeneous computing scheme," *Journal of Applied Physics*, vol. 119, no. 12, p. 124304, Mar. 2016. DOI: 10.1063/1.4944649

[158] M. A. Harb, "Scattering effects in atomistic quantum transport simulations," Ph.D. dissertation, McGill University, Mar. 2019. URL: https://escholarship.mcgill.ca/concern/theses/8336h5962

[159] M. Peskin and D. Schroder, *An Introduction to Quantum Field Theory.* CRC Press, 1995. DOI: 10.1201/9780429503559

[160] M. Artin, *Algebra.* Prentice Hall, 1991.

[161] J. Carrier, L. Greengard, and V. Rokhlin, "A fast adaptive multipole algorithm for particle simulations," *SIAM Journal on Scientific and Statistical Computing*, vol. 9, no. 4, pp. 669–686, July 1988. DOI: 10.1137/0909044

[162] J. Wang, E. Polizzi, and M. Lundstrom, "A three-dimensional quantum simulation of silicon nanowire transistors with the effective-mass approximation," *Journal of Applied Physics*, vol. 96, no. 4, pp. 2192–2203, Aug. 2004. DOI: 10.1063/1.1769089

[163] W. Wang, T.-M. Hwang, W.-W. Lin, and J.-L. Liu, "Numerical methods for semiconductor heterostructures with band nonparabolicity," *Journal of Computational Physics*, vol. 190, no. 1, pp. 141–158, Sep. 2003. DOI: 10.1016/S0021-9991(03)00268-7

[164] J. Wang, A. Rahman, A. Ghosh, G. Klimeck, and M. Lundstrom, "On the validity of the parabolic effective-mass approximation for the I–V calculation of silicon nanowire transistors," *IEEE Transactions on Electron Devices*, vol. 52, no. 7, pp. 1589–1595, June 2005. DOI: 10.1109/TED.2005.850945

[165] J. S. Townsend, *A Modern Approach to Quantum Mechanics*, 2nd ed. University Science Books, 2013.

[166] Y. Y. Peter and M. Cardona, *Fundamentals of Semiconductors: Physics and Materials Properties*, 4th ed. Springer, 2010. DOI: 10.1007/978-3-642-00710-1

[167] P. Hohenberg and W. Kohn, "Inhomogeneous electron gas," *Physical Review*, vol. 136, no. 3B, pp. B864–B871, Nov. 1964. DOI: 10.1103/PhysRev.136.B864

[168] W. Kohn and L. Sham, “Quantum density oscillations in an inhomogeneous electron gas,” *Physical Review*, vol. 137, no. 6A, pp. A1697–A1705, Mar. 1965. DOI: 10.1103/PhysRev.137.A1697

[169] W. Kohn and L. J. Sham, “Self-consistent equations including exchange and correlation effects,” *Physical Review*, vol. 140, no. 4A, pp. A1133–A1138, Nov. 1965. DOI: 10.1103/PhysRev.140.A1133

[170] M. Gell-Mann and K. A. Brueckner, “Correlation energy of an electron gas at high density,” *Physical Review*, vol. 106, no. 2, pp. 364–368, Apr. 1957. DOI: 10.1103/PhysRev.106.364

[171] D. R. Hamann, M. Schlüter, and C. Chiang, “Norm-conserving pseudopotentials,” *Physical Review Letters*, vol. 43, no. 20, pp. 1494–1497, Nov. 1979. DOI: 10.1103/PhysRevLett.43.1494

[172] D. V. Schroeder, *An Introduction to Thermal Physics.* Addison Wesley Longman, 1999. DOI: 10.1093/oso/9780192895547.001.0001

[173] S. Datta, *Electronic Transport in Mesoscopic Systems.* Cambridge University Press, 1997. DOI: 10.1017/CBO9780511805776

[174] H. Haug and A.-P. Jauho, *Quantum Kinetics in Transport and Optics of Semiconductors*, 2nd ed. Springer, 2008. DOI: 10.1007/978-3-540-73564-9

[175] C. A. Stafford, D. M. Cardamone, and S. Mazumdar, “The quantum interference effect transistor,” *Nanotechnology*, vol. 18, no. 42, p. 424014, Sep. 2007. DOI: 10.1088/0957-4484/18/42/424014

[176] M. C. Vecchi and M. Rudan, “Modeling electron and hole transport with full-band structure effects by means of the spherical-harmonics expansion of the BTE,” *IEEE Transactions on Electron Devices*, vol. 45, no. 1, pp. 230–238, Jan. 1998. DOI: 10.1109/16.658836

[177] Z. Stanojevic, O. Baumgartner, F. Mitterbauer, H. Demel, C. Kernstock, M. Karner, V. Eyert, A. France-Lanord, P. Saxe, C. Freeman, and E. Wimmer, “Physical modeling—A new paradigm in device simulation,” in *International Electron Devices Meeting (IEDM) Technical Digest*, p. 5.1, Dec. 2015. DOI: 10.1109/IEDM.2015.7409631

[178] M. Shin, "Full-quantum simulation of hole transport and band-to-band tunneling in nanowires using the $k \cdot p$ method," *Journal of Applied Physics*, vol. 106, no. 5, p. 054505, Sep. 2009. DOI: 10.1063/1.3208067

[179] M. B. Nardelli, "Electronic transport in extended systems: Application to carbon nanotubes," *Physical Review B*, vol. 60, no. 11, pp. 7828–7833, Sep. 1999. DOI: 10.1103/PhysRevB.60.7828

[180] J. Taylor, H. Guo, and J. Wang, "*Ab initio* modeling of quantum transport properties of molecular electronic devices," *Physical Review B*, vol. 63, no. 24, p. 245407, June 2001. DOI: 10.1103/PhysRevB.63.245407

[181] J. Taylor, H. Guo, and J. Wang, "*Ab initio* modeling of open systems: Charge transfer, electron conduction, and molecular switching of a C_{60} device," *Physical Review B*, vol. 63, no. 12, p. 121104, Mar. 2001. DOI: 10.1103/PhysRevB.63.121104

[182] J. Maassen, M. Harb, V. Michaud-Rioux, Y. Zhu, and H. Guo, "Quantum transport modeling from first principles," *Proceedings of the IEEE*, vol. 101, no. 2, pp. 518–530, Aug. 2012. DOI: 10.1109/JPROC.2012.2197810

[183] Q. An, C. Hu, G. Yu, and H. Guo, "Spin-polarized quantum transport in Si dangling bond wires," *Nanoscale*, vol. 12, no. 10, pp. 6079–6088, Feb. 2020. DOI: 10.1039/D0NR00037J

[184] A. Afzalian, M. Passlack, and Y.-C. Yeo, "Scaling perspective for III–V broken gap nanowire TFETs: An atomistic study using a fast tight-binding mode-space NEGF model," in *International Electron Devices Meeting (IEDM) Technical Digest*, p. 30.1, Dec. 2016. DOI: 10.1109/IEDM.2016.7838510

[185] Y. M. Niquet, D. Rideau, C. Tavernier, H. Jaouen, and X. Blase, "Onsite matrix elements of the tight-binding Hamiltonian of a strained crystal: Application to silicon, germanium, and their alloys," *Physical Review B*, vol. 79, no. 24, p. 245201, June 2009. DOI: 10.1103/PhysRevB.79.245201

[186] J. C. Slater and G. F. Koster, "Simplified LCAO method for the periodic potential problem," *Physical Review*, vol. 94, no. 6, pp. 1498–1524, June 1954. DOI: 10.1103/PhysRev.94.1498

[187] M. Born and R. Oppenheimer, "Zur quantentheorie der molekeln," *Annalen der Physik*, vol. 389, no. 20, pp. 457–484, 1927. DOI: 10.1002/andp.19273892002

[188] P.-O. Löwdin, "On the non-orthogonality problem connected with the use of atomic wave functions in the theory of molecules and crystals," *The Journal of Chemical Physics*, vol. 18, no. 3, pp. 365–375, Mar. 1950. DOI: 10.1063/1.1747632

[189] P. Vogl, H. P. Hjalmarson, and J. D. Dow, "A semi-empirical tight-binding theory of the electronic structure of semiconductors," *Journal of Physics and Chemistry of Solids*, vol. 44, no. 5, pp. 365–378, May 1983. DOI: 10.1016/0022-3697(83)90064-1

[190] J.-M. Jancu, R. Scholz, F. Beltram, and F. Bassani, "Empirical $spds^{\star}$ tight-binding calculation for cubic semiconductors: General method and material parameters," *Physical Review B*, vol. 57, no. 11, pp. 6493–6507, Mar. 1998. DOI: 10.1103/PhysRevB.57.6493

[191] T. B. Boykin, G. Klimeck, and F. Oyafuso, "Valence band effective-mass expressions in the $sp^3d^5s^{\star}$ empirical tight-binding model applied to a Si and Ge parametrization," *Physical Review B*, vol. 69, no. 11, p. 115201, Mar. 2004. DOI: 10.1103/PhysRevB.69.115201

[192] D. A. Papaconstantopoulos and M. J. Mehl, "The Slater–Koster tight-binding method: a computationally efficient and accurate approach," *Journal of Physics: Condensed Matter*, vol. 15, no. 10, pp. R413–R440, Mar. 2003. DOI: 10.1088/0953-8984/15/10/201

[193] O. F. Sankey and D. J. Niklewski, "*Ab initio* multicenter tight-binding model for molecular-dynamics simulations and other applications in covalent systems," *Physical Review B*, vol. 40, no. 6, pp. 3979–3995, Aug. 1989. DOI: 10.1103/PhysRevB.40.3979

[194] J. L. Mercer Jr. and M. Y. Chou, "Tight-binding model with intra-atomic matrix elements," *Physical Review B*, vol. 49, no. 12, pp. 8506–8509, Mar. 1994. DOI: 10.1103/PhysRevB.49.8506

[195] D. J. Chadi and M. L. Cohen, "Tight-binding calculations of the valence bands of diamond and zincblende crystals," *Physica Status Solidi B*, vol. 68, no. 1, pp. 405–419, Mar. 1975. DOI: 10.1002/pssb.2220680140

[196] S. Reich, J. Maultzsch, C. Thomsen, and P. Ordejon, "Tight-binding description of graphene," *Physical Review B*, vol. 66, no. 3, p. 035412, July 2002. DOI: 10.1103/PhysRevB.66.035412

[197] G. Kresse and J. Hafner, "*Ab initio* molecular dynamics for liquid metals," *Physical Review B*, vol. 47, no. 1, pp. 558–561, Jan. 1993. DOI: 10.1103/PhysRevB.47.558

[198] G. Kresse and J. Hafner, "*Ab initio* molecular-dynamics simulation of the liquid-metal–amorphous-semiconductor transition in germanium," *Physical Review B*, vol. 49, no. 20, pp. 14 251–14 269, May 1994. DOI: 10.1103/PhysRevB.49.14251

[199] G. Kresse and J. Furthmüller, "Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set," *Computational Materials Science*, vol. 6, no. 1, pp. 15–50, July 1996. DOI: 10.1016/0927-0256(96)00008-0

[200] G. Kresse and J. Furthmüller, "Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set," *Physical Review B*, vol. 54, no. 16, pp. 11 169–11 186, Oct. 1996. DOI: 10.1103/PhysRevB.54.11169

[201] K. Seino, J.-M. Wagner, and F. Bechstedt, "Quasiparticle effect on electron confinement in Si/SiO_2 quantum-well structures," *Applied Physics Letters*, vol. 90, no. 25, p. 253109, June 2007. DOI: 10.1063/1.2750526

[202] Y. He, Y. Tan, Z. Jiang, M. Povolotskyi, G. Klimeck, and T. Kubis, "Surface passivation in empirical tight binding," *IEEE Transactions on Electron Devices*, vol. 63, no. 3, pp. 954–958, Feb. 2016. DOI: 10.1109/TED.2016.2519042

[203] Y. P. Tan, M. Povolotskyi, T. Kubis, T. B. Boykin, and G. Klimeck, "Tight-binding analysis of Si and GaAs ultrathin bodies with subatomic wave-function resolution," *Physical Review B*, vol. 92, no. 8, p. 085301, Aug. 2015. DOI: 10.1103/PhysRevB.92.085301

[204] S. Lee, F. Oyafuso, P. Von Allmen, and G. Klimeck, "Boundary conditions for the electronic structure of finite-extent embedded semiconductor nanostructures," *Physical Review B*, vol. 69, no. 4, p. 045316, Jan. 2004. DOI: 10.1103/PhysRevB.69.045316

[205] Y. Cui, Z. Zhong, D. Wang, W. U. Wang, and C. M. Lieber, "High performance silicon nanowire field effect transistors," *Nano Letters*, vol. 3, no. 2, pp. 149–152, Jan. 2003. DOI: 10.1021/nl025875l

[206] S. B. Samavedam, J. Ryckaert, E. Beyne, K. Ronse, N. Horiguchi, Z. Tokei, I. Radu, M. G. Bardon, M. H. Na, A. Spessot, and S. Biesemans, "Future logic scaling: Towards

atomic channels and deconstructed chips," in *International Electron Devices Meeting (IEDM) Technical Digest*, p. 1.1, Dec. 2020. DOI: 10.1109/IEDM13553.2020.9372023

[207] R. S. Wagner and W. C. Ellis, "Vapor–liquid–solid mechanism of single crystal growth," *Applied Physics Letters*, vol. 4, no. 5, pp. 89–90, Mar. 1964. DOI: 10.1063/1.1753975

[208] A. M. Morales and C. M. Lieber, "A laser ablation method for the synthesis of crystalline semiconductor nanowires," *Science*, vol. 279, no. 5348, pp. 208–211, Jan. 1998. DOI: 10.1126/science.279.5348.208

[209] Y. Wu, Y. Cui, L. Huynh, C. J. Barrelet, D. C. Bell, and C. M. Lieber, "Controlled growth and structures of molecular-scale silicon nanowires," *Nano Letters*, vol. 4, no. 3, pp. 433–436, Feb. 2004. DOI: 10.1021/nl035162i

[210] M. T. Björk, J. Knoch, H. Schmid, H. Riel, and W. Riess, "Silicon nanowire tunneling field-effect transistors," *Applied Physics Letters*, vol. 92, no. 19, p. 193504, May 2008. DOI: 10.1063/1.2928227

[211] W. Liu, K. Zhang, H. Xiao, L. Meng, J. Li, G. M. Stocks, and J. Zhong, "Surface reconstruction and core distortion of silicon and germanium nanowires," *Nanotechnology*, vol. 18, no. 21, p. 215703, Apr. 2007. DOI: 10.1088/0957-4484/18/21/215703

[212] L. Yang, L. Pei, S. Hai-Bo, C. Dan, D. Qian-Min, and W. Le, "Vacancy effect on the doping of silicon nanowires: A first-principles study," *Chinese Physics B*, vol. 23, no. 6, p. 067304, Apr. 2014. DOI: 10.1088/1674-1056/23/6/067304

[213] D. D. D. Ma, C. S. Lee, F. C. K. Au, S. Y. Tong, and S. T. Lee, "Small-diameter silicon nanowire surfaces," *Science*, vol. 299, no. 5614, pp. 1874–1877, Mar. 2003. DOI: 10.1126/science.1080313

[214] N. S. Mohammad, "Understanding quantum confinement in nanowires: basics, applications and possible laws," *Journal of Physics: Condensed Matter*, vol. 26, no. 42, p. 423202, Sep. 2014. DOI: 10.1088/0953-8984/26/42/423202

[215] A. J. Read, R. J. Needs, K. J. Nash, L. T. Canham, P. D. J. Calcott, and A. Qteish, "First-principles calculations of the electronic properties of silicon quantum wires," *Physical Review Letters*, vol. 69, no. 8, pp. 1232–1235, Aug. 1992. DOI: 10.1103/PhysRevLett.69.1232

[216] C. Harris and E. P. O'Reilly, "Nature of the band gap of silicon and germanium nanowires," *Physica E: Low-dimensional Systems and Nanostructures*, vol. 32, no. 1–2, pp. 341–345, May 2006. DOI: 10.1016/j.physe.2005.12.094

[217] M. S. Lundstrom and D. A. Antoniadis, "Compact models and the physics of nanoscale FETs," *IEEE Transactions on Electron Devices*, vol. 61, no. 2, pp. 225–233, Oct. 2013. DOI: 10.1109/TED.2013.2283253

[218] O. Moutanabbir, D. Isheim, H. Blumtritt, S. Senz, E. Pippel, and D. N. Seidman, "Colossal injection of catalyst atoms into silicon nanowires," *Nature*, vol. 496, pp. 78–82, Apr. 2013. DOI: 10.1038/nature11999

[219] W. Chen, L. Yu, S. Misra, Z. Fan, P. Pareige, G. Patriarche, S. Bouchoule, and P. R. i Cabarrocas, "Incorporation and redistribution of impurities into silicon nanowires during metal-particle-assisted growth," *Nature Communications*, vol. 5, no. 4134, June 2014. DOI: 10.1038/ncomms5134

[220] M. Luisier, A. Schenk, and W. Fichtner, "Atomistic treatment of interface roughness in Si nanowire transistors with different channel orientations," *Applied Physics Letters*, vol. 90, no. 10, p. 102103, Mar. 2007. DOI: 10.1063/1.2711275

[221] J. D. Jackson, *Classical Electrodynamics*, 3rd ed. Wiley, 1999.

[222] L. V. Keldysh and P. N. Lebedev, "Diagram technique for nonequilibrium processes," *Soviet Physics Journal of Experimental and Theoretical Physics*, vol. 20, no. 4, pp. 1018–1026, Apr. 1965.

[223] L. P. Kadanoff and G. Baym, *Quantum Statistical Mechanics.* CRC Press, 1962. DOI: 10.1201/9780429493218

[224] S. Datta, *Quantum Transport: Atom to Transistor.* Cambridge University Press, 2005. DOI: 10.1017/CBO9781139164313

[225] R. Landauer, "Spatial variation of currents and fields due to localized scatterers in metallic conduction," *IBM Journal of Research and Development*, vol. 1, no. 3, pp. 223–231, July 1957. DOI: 10.1147/rd.13.0223

[226] M. Büttiker, "Symmetry of electrical conduction," *IBM Journal of Research and Development*, vol. 32, no. 3, pp. 317–334, May 1988. DOI: 10.1147/rd.323.0317

[227] R. Landauer, "Conductance from transmission: Common sense points," *Physica Scripta*, vol. 1992, no. T42, pp. 110–114, Jan. 1992. DOI: 10.1088/0031-8949/1992/t42/020

[228] D. V. Averin and K. K. Likharev, "Coulomb blockade of single-electron tunneling, and coherent oscillations in small tunnel junctions," *Journal of Low Temperature Physics*, vol. 62, no. 3, pp. 345–373, Feb. 1986. DOI: 10.1007/BF00683469

[229] S. Datta, "The NEGF approach to nano-device simulation," June 2014. URL: https://nanohub.org/wiki/NEGF

[230] C. Jeong, R. Kim, M. Luisier, S. Datta, and M. Lundstrom, "On Landauer versus Boltzmann and full band versus effective mass evaluation of thermoelectric transport coefficients," *Journal of Applied Physics*, vol. 107, no. 2, p. 023707, Jan. 2010. DOI: 10.1063/1.3291120

[231] K. Natori, "Ballistic metal–oxide–semiconductor field effect transistor," *Journal of Applied Physics*, vol. 76, no. 8, pp. 4879–4890, Oct. 1994. DOI: 10.1063/1.357263

[232] A. Rahman and M. S. Lundstrom, "A compact scattering model for the nanoscale double-gate MOSFET," *IEEE Transactions on Electron Devices*, vol. 49, no. 3, pp. 481–489, Aug. 2002. DOI: 10.1109/16.987120

[233] M. P. López Sancho, J. M. López Sancho, and J. Rubio, "Quick iterative scheme for the calculation of transfer matrices: application to Mo (100)," *Journal of Physics F: Metal Physics*, vol. 14, no. 5, pp. 1205–1215, May 1984. DOI: 10.1088/0305-4608/14/5/016

[234] T. Frederiksen, M. Brandbyge, N. Lorente, and A.-P. Jauho, "Inelastic scattering and local heating in atomic gold wires," *Physical Review Letters*, vol. 93, no. 25, p. 256601, Dec. 2004. DOI: 10.1103/PhysRevLett.93.256601

[235] T. Frederiksen, M. Paulsson, M. Brandbyge, and A.-P. Jauho, "Inelastic transport theory from first principles: Methodology and application to nanoscale devices," *Physical Review B*, vol. 75, no. 20, p. 205413, May 2007. DOI: 10.1103/PhysRevB.75.205413

[236] N. Sergueev, D. Roubtsov, and H. Guo, "*Ab initio* analysis of electron–phonon coupling in molecular devices," *Physical Review Letters*, vol. 95, no. 14, p. 146803, Sep. 2005. DOI: 10.1103/PhysRevLett.95.146803

[237] A. S. Banerjee, P. Suryanarayana, and J. E. Pask, “Periodic Pulay method for robust and efficient convergence acceleration of self-consistent field iterations,” *Chemical Physics Letters*, vol. 647, pp. 31–35, Mar. 2016. DOI: 10.1016/j.cplett.2016.01.033

[238] J. Enkovaara, C. Rostgaard, J. J. Mortensen, J. Chen, M. Dułak, L. Ferrighi, J. Gavnholt, C. Glinsvad, V. Haikola, H. A. Hansen, H. H. Kristoffersen, M. Kuisma, A. H. Larsen, L. Lehtovaara, M. Ljungberg, O. Lopez-Acevedo, P. G. Moses, J. Ojanen, T. Olsen, V. Petzold, N. A. Romero, J. Stausholm-Møller, M. Strange, G. A. Tritsaris, M. Vanin, M. Walter, B. Hammer, H. Häkkinen, G. K. H. Madsen, R. M. Nieminen, J. K. Nørskov, M. Puska, T. T. Rantala, J. Schiøtz, K. S. Thygesen, and K. W. Jacobsen, “Electronic structure calculations with GPAW: a real-space implementation of the projector augmented-wave method,” *Journal of Physics: Condensed Matter*, vol. 22, no. 25, p. 253202, June 2010. DOI: 10.1088/0953-8984/22/25/253202

[239] S. S. Iyer, G. L. Patton, J. M. C. Stork, B. S. Meyerson, and D. L. Harame, “Heterojunction bipolar transistors using Si-Ge alloys,” *IEEE Transactions on Electron Devices*, vol. 36, no. 10, pp. 2043–2064, Oct. 1989. DOI: 10.1109/16.40887

[240] S. E. Thompson, M. Armstrong, C. Auth, M. Alavi, M. Buehler, R. Chau, S. Cea, T. Ghani, G. Glass, T. Hoffman, C.-H. Jan, C. Kenyon, J. Klaus, K. Kuhn, Z. Ma, B. Mcintyre, K. Mistry, A. Murthy, B. Obradovic, R. Nagisetty, P. Nguyen, S. Sivakumar, R. Shaheed, L. Shifren, B. Tufts, S. Tyagi, M. Bohr, and Y. El-Mansy, “A 90-nm logic technology featuring strained-silicon,” *IEEE Transactions on Electron Devices*, vol. 51, no. 11, pp. 1790–1797, Oct. 2004. DOI: 10.1109/TED.2004.836648

[241] K.-H. Kao, A. S. Verhulst, W. G. Vandenberghe, B. Soree, G. Groeseneken, and K. De Meyer, “Direct and indirect band-to-band tunneling in germanium-based TFETs,” *IEEE Transactions on Electron Devices*, vol. 59, no. 2, pp. 292–301, Dec. 2011. DOI: 10.1109/TED.2011.2175228

[242] J. Knoch and J. Appenzeller, “Modeling of high-performance p-type III–V heterojunction tunnel FETs,” *IEEE Electron Device Letters*, vol. 31, no. 4, pp. 305–307, Mar. 2010. DOI: 10.1109/LED.2010.2041180

[243] S. S. Sylvia, M. A. Khayer, K. Alam, and R. K. Lake, "Doping, tunnel barriers, and cold carriers in InAs and InSb nanowire tunnel transistors," *IEEE Transactions on Electron Devices*, vol. 59, no. 11, pp. 2996–3001, Sep. 2012. DOI: 10.1109/TED.2012.2212442

[244] H. Liu and D. Y. Peide, "MoS_2 dual-gate MOSFET with atomic-layer-deposited Al_2O_3 as top-gate dielectric," *IEEE Electron Device Letters*, vol. 33, no. 4, pp. 546–548, Mar. 2012. DOI: 10.1109/LED.2012.2184520

[245] L. Liu, Y. Lu, and J. Guo, "On monolayer MoS_2 field-effect transistors at the scaling limit," *IEEE Transactions on Electron Devices*, vol. 60, no. 12, pp. 4133–4139, Oct. 2013. DOI: 10.1109/TED.2013.2284591

[246] W. Cao, J. Kang, D. Sarkar, W. Liu, and K. Banerjee, "2D semiconductor FETs—Projections and design for sub-10 nm VLSI," *IEEE Transactions on Electron Devices*, vol. 62, no. 11, pp. 3459–3469, Sep. 2015. DOI: 10.1109/TED.2015.2443039

[247] F. Liu, C. Qiu, Z. Zhang, L.-M. Peng, J. Wang, and H. Guo, "Dirac electrons at the source: Breaking the 60-mV/decade switching limit," *IEEE Transactions on Electron Devices*, vol. 65, no. 7, pp. 2736–2743, July 2018. DOI: 10.1109/TED.2018.2836387

[248] K.-T. Lam, Z. Dong, and J. Guo, "Performance limits projection of black phosphorous field-effect transistors," *IEEE Electron Device Letters*, vol. 35, no. 9, pp. 963–965, July 2014. DOI: 10.1109/LED.2014.2333368

[249] F. Liu, Y. Wang, X. Liu, J. Wang, and H. Guo, "Ballistic transport in monolayer black phosphorus transistors," *IEEE Transactions on Electron Devices*, vol. 61, no. 11, pp. 3871–3876, Sep. 2014. DOI: 10.1109/TED.2014.2353213

[250] G. Fiori and G. Iannaccone, "Simulation of graphene nanoribbon field-effect transistors," *IEEE Electron Device Letters*, vol. 28, no. 8, pp. 760–762, July 2007. DOI: 10.1109/LED.2007.901680

[251] C. Qiu, F. Liu, L. Xu, B. Deng, M. Xiao, J. Si, L. Lin, Z. Zhang, J. Wang, H. Guo, H. Peng, and L.-M. Peng, "Dirac-source field-effect transistors as energy-efficient, high-performance electronic switches," *Science*, vol. 361, no. 6400, pp. 387–392, July 2018. DOI: 10.1126/science.aap9195

[252] N. Singh, A. Agarwal, L. K. Bera, T. Y. Liow, R. Yang, S. C. Rustagi, C. H. Tung, R. Kumar, G. Q. Lo, N. Balasubramanian, and D.-L. Kwong, “High-performance fully depleted silicon nanowire (diameter $\leq$ 5 nm) gate-all-around CMOS devices,” *IEEE Electron Device Letters*, vol. 27, no. 5, pp. 383–386, May 2006. DOI: 10.1109/LED.2006.873381

[253] K. Tachi, M. Casse, D. Jang, C. Dupré, A. Hubert, N. Vulliet, V. Maffini-Alvaro, C. Vizioz, C. Carabasse, V. Delaye, J. M. Hartmann, G. Ghibaudo, H. Iwai, S. Cristoloveanu, O. Faynot, and T. Ernst, “Relationship between mobility and high-k interface properties in advanced Si and SiGe nanowires,” in *International Electron Devices Meeting (IEDM) Technical Digest*, p. 12.7, Dec. 2009. DOI: 10.1109/IEDM.2009.5424360

[254] Y. Zhai, L. Mathew, R. Rao, M. Palard, S. Chopra, J. G. Ekerdt, L. F. Register, and S. K. Banerjee, “High-performance vertical gate-all-around silicon nanowire FET with high-κ/metal gate,” *IEEE Transactions on Electron Devices*, vol. 61, no. 11, pp. 3896–3900, Sep. 2014. DOI: 10.1109/TED.2014.2353658

[255] S. Holten and H. Kliem, “Increased dielectric permittivity of SiO_2 thin films,” *Journal of Applied Physics*, vol. 90, no. 4, pp. 1941–1949, Aug. 2001. DOI: 10.1063/1.1384484

[256] G. Gramse, I. Casuso, J. Toset, L. Fumagalli, and G. Gomila, “Quantitative dielectric constant measurement of thin films by DC electrostatic force microscopy,” *Nanotechnology*, vol. 20, no. 39, p. 395702, Sep. 2009. DOI: 10.1088/0957-4484/20/39/395702

[257] S. Hong, C.-S. Lee, M.-H. Lee, Y. Lee, K. Y. Ma, G. Kim, S. I. Yoon, K. Ihm, K.-J. Kim, T. J. Shin, S. W. Kim, E.-c. Jeon, H. Jeon, J.-Y. Kim, H.-I. Lee, Z. Lee, A. Antidormi, S. Roche, M. Chhowalla, H.-J. Shin, and H. S. Shin, “Ultralow-dielectric-constant amorphous boron nitride,” *Nature*, vol. 582, pp. 511–514, June 2020. DOI: 10.1038/s41586-020-2375-9

[258] K. Suzuki, T. Tanaka, Y. Tosaka, H. Horie, and Y. Arimoto, “Scaling theory for double-gate SOI MOSFET’s,” *IEEE Transactions on Electron Devices*, vol. 40, no. 12, pp. 2326–2329, Dec. 1993. DOI: 10.1109/16.249482

[259] C. P. Auth and J. D. Plummer, “Scaling theory for cylindrical, fully-depleted, surrounding-gate MOSFET’s,” *IEEE Electron Device Letters*, vol. 18, no. 2, pp. 74–76, Feb. 1997. DOI: 10.1109/55.553049

[260] D. J. Frank, Y. Taur, and H.-S. P. Wong, "Generalized scale length for two-dimensional effects in MOSFET's," *IEEE Electron Device Letters*, vol. 19, no. 10, pp. 385–387, Oct. 1998. DOI: 10.1109/55.720194

[261] S.-H. Oh, D. Monroe, and J. M. Hergenrother, "Analytic description of short-channel effects in fully-depleted double-gate and cylindrical, surrounding-gate MOSFETs," *IEEE Electron Device Letters*, vol. 21, no. 9, pp. 445–447, Sep. 2000. DOI: 10.1109/55.863106

[262] L. Zhang, X. Lin, J. He, and M. Chan, "An analytical charge model for double-gate tunnel FETs," *IEEE Transactions on Electron Devices*, vol. 59, no. 12, pp. 3217–3223, Nov. 2012. DOI: 10.1109/TED.2012.2217145

[263] H. R. T. Khaveh and S. Mohammadi, "Potential and drain current modeling of gate-all-around tunnel FETs considering the junctions depletion regions and the channel mobile charge carriers," *IEEE Transactions on Electron Devices*, vol. 63, no. 12, pp. 5021–5029, Nov. 2016. DOI: 10.1109/TED.2016.2619761

[264] D. Keighobadi, S. Mohammadi, and M. Fathipour, "An analytical drain current model for the cylindrical channel gate-all-around heterojunction tunnel FETs," *IEEE Transactions on Electron Devices*, vol. 66, no. 8, pp. 3646–3651, June 2019. DOI: 10.1109/TED.2019.2922232

[265] R. Rurali, "Colloquium: Structural, electronic, and transport properties of silicon nanowires," *Reviews of Modern Physics*, vol. 82, no. 1, pp. 427–449, Feb. 2010. DOI: 10.1103/RevModPhys.82.427

[266] H. Iwai, "CMOS scaling for sub-90 nm to sub-10 nm," in *Proceedings of the 17th International Conference on VLSI Design*, pp. 30–35, Jan. 2004. DOI: 10.1109/ICVD.2004.1260899

[267] C. Mead and L. Conway, *Introduction to VLSI Systems.* Addison–Wesley, 1979.

[268] Y. Taur, D. A. Buchanan, W. Chen, D. J. Frank, K. E. Ismail, S.-H. Lo, G. A. Sai-Halasz, R. G. Viswanathan, H.-J. C. Wann, S. J. Wind, and H.-S. Wong, "CMOS scaling into the nanometer regime," *Proceedings of the IEEE*, vol. 85, no. 4, pp. 486–504, Apr. 1997. DOI: 10.1109/5.573737

[269] A. Asenov, "Random dopant induced threshold voltage lowering and fluctuations in sub-0.1 µm MOSFET's: A 3-D 'atomistic' simulation study," *IEEE Transactions on Electron Devices*, vol. 45, no. 12, pp. 2505–2513, Dec. 1998. DOI: 10.1109/16.735728

[270] Q. Shi, H. Guo, Y. Zhu, and L. Liu, "Analyzing variability in short-channel quantum transport from atomistic first principles," *Physical Review Applied*, vol. 3, no. 6, p. 064008, June 2015. DOI: 10.1103/PhysRevApplied.3.064008

[271] Y. Kang, S. Xu, K. Han, E. Y.-J. Kong, Z. Song, S. Luo, A. Kumar, C. Wang, W. Fan, G. Liang, and X. Gong, "$Ge_{0.95}Sn_{0.05}$ gate-all-around p-channel metal–oxide–semiconductor field-effect transistors with sub-3 nm nanowire width," *Nano Letters*, vol. 21, no. 13, pp. 5555–5563, June 2021. DOI: 10.1021/acs.nanolett.1c00934

[272] V. V. Zhirnov, R. K. Cavin, J. A. Hutchby, and G. I. Bourianoff, "Limits to binary logic switch scaling—A gedanken model," *Proceedings of the IEEE*, vol. 91, no. 11, pp. 1934–1939, Nov. 2003. DOI: 10.1109/JPROC.2003.818324

[273] D. Mamaluy and X. Gao, "The fundamental downscaling limit of field effect transistors," *Applied Physics Letters*, vol. 106, no. 19, p. 193503, May 2015. DOI: 10.1063/1.4919871

[274] Y.-B. Kim, "Challenges for nanoscale MOSFETs and emerging nanoelectronics," *Transactions on Electrical and Electronic Materials*, vol. 11, no. 3, pp. 93–105, June 2010. DOI: 10.4313/TEEM.2010.11.3.093

[275] H. Kawaura, T. Sakamoto, and T. Baba, "Observation of source-to-drain direct tunneling current in 8 nm gate electrically variable shallow junction metal–oxide–semiconductor field-effect transistors," *Applied Physics Letters*, vol. 76, no. 25, pp. 3810–3812, June 2000. DOI: 10.1063/1.126789

[276] Q. Rafhay, R. Clerc, G. Ghibaudo, and G. Pananakakis, "Impact of source-to-drain tunnelling on the scalability of arbitrary oriented alternative channel material nMOSFETs," *Solid-State Electronics*, vol. 52, no. 10, pp. 1474–1481, Oct. 2008. DOI: 10.1016/j.sse.2008.06.035

[277] S. S. Sylvia, H.-H. Park, M. A. Khayer, K. Alam, G. Klimeck, and R. K. Lake, "Material selection for minimizing direct tunneling in nanowire transistors," *IEEE*

Transactions on Electron Devices, vol. 59, no. 8, pp. 2064–2069, June 2012. DOI: 10.1109/TED.2012.2200688

[278] S. R. Mehrotra, S. Kim, T. Kubis, M. Povolotskyi, M. S. Lundstrom, and G. Klimeck, "Engineering nanowire n-MOSFETs at $L_g < 8$ nm," *IEEE Transactions on Electron Devices*, vol. 60, no. 7, pp. 2171–2177, June 2013. DOI: 10.1109/TED.2013.2263806

[279] Á. Szabó and M. Luisier, "Under-the-barrier model: An extension of the top-of-the-barrier model to efficiently and accurately simulate ultrascaled nanowire transistors," *IEEE Transactions on Electron Devices*, vol. 60, no. 7, pp. 2353–2360, June 2013. DOI: 10.1109/TED.2013.2263386

[280] Z. Jiang, J. Wang, H.-H. Park, A.-T. Pham, N. Xu, Y. Lu, S. Jin, W. Choi, M. A. Pourghaderi, J. Kim, and K.-H. Lee, "Comprehensive simulation study of direct source-to-drain tunneling in ultra-scaled Si, Ge, and III–V DG-FETs," *IEEE Transactions on Electron Devices*, vol. 64, no. 3, pp. 945–952, Feb. 2017. DOI: 10.1109/TED.2017.2656921

[281] W.-S. Cho and K. Roy, "The effects of direct source-to-drain tunneling and variation in the body thickness on (100) and (110) sub-10-nm Si double-gate transistors," *IEEE Electron Device Letters*, vol. 36, no. 5, pp. 427–429, Mar. 2015. DOI: 10.1109/LED.2015.2413785

[282] M. Salmani-Jelodar, S. R. Mehrotra, H. Ilatikhameneh, and G. Klimeck, "Design guidelines for sub-12 nm nanowire MOSFETs," *IEEE Transactions on Nanotechnology*, vol. 14, no. 2, pp. 210–213, Mar. 2015. DOI: 10.1109/TNANO.2015.2395441

[283] H. Wakabayashi, T. Ezaki, M. Hane, S. Yamagami, N. Ikarashi, K. Takeuchi, T. Yamamoto, T. Mogami, T. Ikezawa, T. Sakamoto, and H. Kawaura, "Transport properties of sub-10-nm planar-bulk-CMOS devices," in *International Electron Devices Meeting (IEDM) Technical Digest*, pp. 429–432, Dec. 2004. DOI: 10.1109/IEDM.2004.1419178

[284] N. Pandey, G. Pahwa, and Y. S. Chauhan, "Addressing source to drain tunneling in extremely scaled Si-transistors using negative capacitance," *Solid-State Electronics*, vol. 186, p. 108189, Dec. 2021. DOI: 10.1016/j.sse.2021.108189

[285] H. Ilatikhameneh, T. Ameen, B. Novakovic, Y. Tan, G. Klimeck, and R. Rahman, "Saving Moore's law down to 1 nm channels with anisotropic effective mass," *Scientific Reports*, vol. 6, p. 31501, Aug. 2016. DOI: 10.1038/srep31501

[286] S. B. Desai, S. R. Madhvapathy, A. B. Sachid, J. P. Llinas, Q. Wang, G. H. Ahn, G. Pitner, M. J. Kim, J. Bokor, C. Hu, H.-S. P. Wong, and A. Javey, "MoS_2 transistors with 1-nanometer gate lengths," *Science*, vol. 354, no. 6308, pp. 99–102, Oct. 2016. DOI: 10.1126/science.aah4698

[287] R. H. Fowler and L. Nordheim, "Electron emission in intense electric fields," *Proceedings of the Royal Society of London A*, vol. 119, no. 781, pp. 173–181, May 1928. DOI: 10.1098/rspa.1928.0091

[288] J. Zou, Q. Xu, J. Luo, R. Wang, R. Huang, and Y. Wang, "Predictive 3-D modeling of parasitic gate capacitance in gate-all-around cylindrical silicon nanowire MOSFETs," *IEEE Transactions on Electron Devices*, vol. 58, no. 10, pp. 3379–3387, Aug. 2011. DOI: 10.1109/TED.2011.2162521

[289] K. Yim, Y. Yong, J. Lee, K. Lee, H.-H. Nahm, J. Yoo, C. Lee, C. S. Hwang, and S. Han, "Novel high-κ dielectrics for next-generation electronic devices screened by automated *ab initio* calculations," *NPG Asia Materials*, vol. 7, no. 6, p. e190, June 2015. DOI: 10.1038/am.2015.57

[290] "Moore's law at 40—Happy birthday," *The Economist*, Mar. 2005. URL: https://www.economist.com/business/2005/03/23/happy-birthday

[291] C. Convertino, C. B. Zota, H. Schmid, D. Caimi, L. Czornomaz, A. M. Ionescu, and K. E. Moselund, "A hybrid III–V tunnel FET and MOSFET technology platform integrated on silicon," *Nature Electronics*, vol. 4, pp. 162–170, Feb. 2021. DOI: 10.1038/s41928-020-00531-3

[292] D. D. L. Chung, "Review graphite," *Journal of Materials Science*, vol. 37, no. 8, pp. 1475–1489, Apr. 2002. DOI: 10.1023/A:1014915307738

[293] K. S. Novoselov, A. K. Geim, S. V. Morozov, D. Jiang, Y. Zhang, S. V. Dubonos, I. V. Grigorieva, and A. A. Firsov, "Electric field effect in atomically thin carbon films," *Science*, vol. 306, no. 5696, pp. 666–669, Oct. 2004. DOI: 10.1126/science.1102896

[294] P. R. Wallace, "The band theory of graphite," *Physical Review*, vol. 71, no. 9, pp. 622–634, May 1947. DOI: 10.1103/PhysRev.71.622

[295] A. K. Geim and K. S. Novoselov, "The rise of graphene," *Nature Materials*, vol. 6, pp. 183–191, Mar. 2007. DOI: 10.1038/nmat1849

[296] K. S. Novoselov, V. I. Fal'ko, L. Colombo, P. R. Gellert, M. G. Schwab, and K. Kim, "A roadmap for graphene," *Nature*, vol. 490, pp. 192–200, Oct. 2012. DOI: 10.1038/nature11458

[297] Y. Zhang, T.-T. Tang, C. Girit, Z. Hao, M. C. Martin, A. Zettl, M. F. Crommie, Y. R. Shen, and F. Wang, "Direct observation of a widely tunable bandgap in bilayer graphene," *Nature*, vol. 459, pp. 820–823, June 2009. DOI: 10.1038/nature08105

[298] F. Schwierz, "Graphene transistors," *Nature Nanotechnology*, vol. 5, pp. 487–496, May 2010. DOI: 10.1038/nnano.2010.89

[299] A. H. Castro Neto, F. Guinea, N. M. R. Peres, K. S. Novoselov, and A. K. Geim, "The electronic properties of graphene," *Reviews of Modern Physics*, vol. 81, no. 1, pp. 109–162, Jan. 2009. DOI: 10.1103/RevModPhys.81.109

[300] O. Klein, "Die reflexion von elektronen an einem potentialsprung nach der relativistischen dynamik von Dirac," *Zeitschrift für Physik*, vol. 53, pp. 157–165, Mar. 1929. DOI: 10.1007/BF01339716

[301] M. Katsnelson, K. Novoselov, and A. Geim, "Chiral tunnelling and the Klein paradox in graphene," *Nature Physics*, vol. 2, pp. 620–625, Aug. 2006. DOI: 10.1038/nphys384

[302] T. R. Robinson, "On Klein tunneling in graphene," *American Journal of Physics*, vol. 80, no. 2, pp. 141–147, Jan. 2012. DOI: 10.1119/1.3658629

[303] N. Stander, B. Huard, and D. Goldhaber-Gordon, "Evidence for Klein tunneling in graphene p–n junctions," *Physical Review Letters*, vol. 102, no. 2, p. 026807, Jan. 2009. DOI: 10.1103/PhysRevLett.102.026807

[304] L. Xu, C. Qiu, L.-M. Peng, and Z. Zhang, "Transconductance amplification in Dirac-source field-effect transistors enabled by graphene/nanotube hereojunctions," *Advanced Electronic Materials*, vol. 6, no. 5, p. 1901289, Apr. 2020. DOI: 10.1002/aelm.201901289

[305] M. Xiao, Y. Lin, L. Xu, B. Deng, H. Peng, L.-M. Peng, and Z. Zhang, “n-type Dirac-source field-effect transistors based on a graphene/carbon nanotube heterojunction,” *Advanced Electronic Materials*, vol. 6, no. 7, p. 2000258, June 2020. DOI: 10.1002/aelm.202000258

[306] M. Liu, H. N. Jaiswal, S. Shahi, S. Wei, Y. Fu, C. Chang, A. Chakravarty, F. Yao, and H. Li, “Monolayer MoS_2 steep-slope transistors with record-high sub-60-mV/decade current density using Dirac-source electron injection,” in *International Electron Devices Meeting (IEDM) Technical Digest*, p. 12.5, Dec. 2020. DOI: 10.1109/IEDM13553.2020.9371961

[307] Z. Tang, C. Liu, X. Huang, S. Zeng, L. Liu, J. Li, Y.-G. Jiang, D. W. Zhang, and P. Zhou, “A steep-slope MoS_2/graphene Dirac-source field-effect transistor with a large drive current,” *Nano Letters*, vol. 21, no. 4, pp. 1758–1764, Feb. 2021. DOI: 10.1021/acs.nanolett.0c04657

[308] M. Liu, H. N. Jaiswal, S. Shahi, S. Wei, Y. Fu, C. Chang, A. Chakravarty, X. Liu, C. Yang, Y. Liu, Y. H. Lee, V. Perebeinos, F. Yao, and H. Li, “Two-dimensional cold electron transport for steep-slope transistors,” *ACS Nano*, vol. 15, no. 3, pp. 5762–5772, Mar. 2021. DOI: 10.1021/acsnano.1c01503

[309] L. Johnson and J. E. Meany, “Mass-producing graphene,” *American Scientist*, vol. 106, no. 3, p. 176, May 2018. DOI: 10.1511/2018.106.3.176

[310] A. Akkerman and M. Murat, “Electron–phonon interactions in silicon: Mean free paths, related distributions and transport characteristics,” *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*, vol. 350, pp. 49–54, May 2015. DOI: 10.1016/j.nimb.2015.03.024

[311] J. Lyu, J. Pei, Y. Guo, J. Gong, and H. Li, “A new opportunity for 2D van der Waals heterostructures: Making steep-slope transistors,” *Advanced Materials*, vol. 32, no. 2, p. 1906000, Nov. 2019. DOI: 10.1002/adma.201906000

[312] D. Logoteta, J. Cao, M. Pala, P. Dollfus, Y. Lee, and G. Iannaccone, “Cold-source paradigm for steep-slope transistors based on van der Waals heterojunctions,” *Phys-*

ical Review Research, vol. 2, no. 4, p. 043286, Nov. 2020. DOI: 10.1103/PhysRevResearch.2.043286

[313] Q. Wang, P. Sang, X. Ma, F. Wang, W. Wei, W. Zhang, Y. Li, and J. Chen, "Cold source engineering towards sub-60mV/dec p-type field-effect-transistors (pFETs): Materials, structures, and doping optimizations," in *International Electron Devices Meeting (IEDM) Technical Digest*, p. 22.4, Dec. 2020. DOI: 10.1109/IEDM13553.2020.9371895

[314] L. Esaki and R. Tsu, "Superlattice and negative differential conductivity in semiconductors," *IBM Journal of Research and Development*, vol. 14, no. 1, pp. 61–65, Jan. 1970. DOI: 10.1147/rd.141.0061

[315] R. Tsu and L. Esaki, "Tunneling in a finite superlattice," *Applied Physics Letters*, vol. 22, no. 11, pp. 562–564, June 1973. DOI: 10.1063/1.1654509

[316] L. L. Chang, L. Esaki, and R. Tsu, "Resonant tunneling in semiconductor double barriers," *Applied Physics Letters*, vol. 24, no. 12, pp. 593–595, June 1974. DOI: 10.1063/1.1655067

[317] M. T. Björk, S. F. Karg, J. Knoch, H. E. Riel, W. H. Riess, and H. Schmid, "Metal–oxide–semiconductor device including a multiple-layer energy filter," U.S. Patent 8 129 763, Mar. 2012.

[318] E. Gnani, P. Maiorano, S. Reggiani, A. Gnudi, and G. Baccarani, "Performance limits of superlattice-based steep-slope nanowire FETs," in *International Electron Devices Meeting (IEDM) Technical Digest*, p. 5.1, Dec. 2011. DOI: 10.1109/IEDM.2011.6131491

[319] D. Logoteta, M. G. Pala, J. Choukroun, P. Dollfus, and G. Iannaccone, "A steep-slope MoS_2-nanoribbon MOSFET based on an intrinsic cold-contact effect," *IEEE Electron Device Letters*, vol. 40, no. 9, pp. 1550–1553, July 2019. DOI: 10.1109/LED.2019.2928131

[320] F. Liu, "Switching at less than 60 mv/decade with a 'cold' metal as the injection source," *Physical Review Applied*, vol. 13, no. 6, p. 064037, June 2020. DOI: 10.1103/PhysRevApplied.13.064037

[321] D. Marian, E. G. Marin, G. Iannaccone, and G. Fiori, "Sub-maxwellian source injection and negative differential transconductance in decorated graphene nanoribbons,"

Physical Review Applied, vol. 14, no. 6, p. 064019, Dec. 2020. DOI: 10.1103/PhysRevApplied.14.064019

[322] S. Ye, Z. Wang, H. Wang, Q. Huang, J. He, and S. Chang, "Steep-slope transistors based on chiral graphene nanoribbons with intrinsic cold source," *IEEE Transactions on Electron Devices*, vol. 68, no. 8, pp. 4123–4128, June 2021. DOI: 10.1109/TED.2021.3087459

[323] Q. Wang, P. Sang, F. Wang, W. Wei, Y. Li, and J. Chen, "Strain engineered C_{31} field-effect-transistors: A new strategy to break 60 mV/decade by using electron injection from intrinsic isolated states," *Applied Physics Express*, vol. 14, no. 7, p. 074003, June 2021. DOI: 10.35848/1882-0786/ac0c66

[324] H. Lu and S.-D. Li, "Two-dimensional carbon allotropes from graphene to graphyne," *Journal of Materials Chemistry C*, vol. 1, no. 23, pp. 3677–3680, Apr. 2013. DOI: 10.1039/C3TC30302K

[325] S. Selberherr, *Analysis and Simulation of Semiconductor Devices.* Springer Science & Business Media, 2012. DOI: 10.1007/978-3-7091-8752-4

[326] E.-H. Toh, G. H. Wang, L. Chan, G.-Q. Lo, G. Samudra, and Y.-C. Yeo, "Strain and materials engineering for the I-MOS transistor with an elevated impact–ionization region," *IEEE Transactions on Electron Devices*, vol. 54, no. 10, pp. 2778–2785, Sep. 2007. DOI: 10.1109/TED.2007.904988

[327] K. Gopalakrishnan, P. B. Griffin, and J. D. Plummer, "I-MOS: A novel semiconductor device with a subthreshold slope lower than kT/q," in *International Electron Devices Meeting (IEDM) Technical Digest*, pp. 289–292, Dec. 2002. DOI: 10.1109/IEDM.2002.1175835

[328] K. Gopalakrishnan, P. B. Griffin, and J. D. Plummer, "Impact ionization MOS (I-MOS)—Part I: Device and circuit simulations," *IEEE Transactions on Electron Devices*, vol. 52, no. 1, pp. 69–76, Dec. 2004. DOI: 10.1109/TED.2004.841344

[329] K. Gopalakrishnan, R. Woo, C. Jungemann, P. B. Griffin, and J. D. Plummer, "Impact ionization MOS (I-MOS)—Part II: Experimental results," *IEEE Transactions on Electron Devices*, vol. 52, no. 1, pp. 77–84, Dec. 2004. DOI: 10.1109/TED.2004.841345

[330] D. J. Frank, S. E. Laux, and M. V. Fischetti, "Monte Carlo simulation of a 30 nm dual-gate MOSFET: How short can Si go?" in *International Electron Devices Meeting (IEDM) Technical Digest*, pp. 553–556, Dec. 1992. DOI: 10.1109/IEDM.1992.307422

[331] D. A. Bandurin, A. V. Tyurnina, L. Y. Geliang, A. Mishchenko, V. Zólyomi, S. V. Morozov, R. K. Kumar, R. V. Gorbachev, Z. R. Kudrynskyi, S. Pezzini, Z. D. Kovalyuk, U. Zeitler, K. S. Novoselov, P. Amalia, L. Eaves, G. I. V., V. I. Fal'ko, A. K. Geim, and Y. Cao, "High electron mobility, quantum hall effect and anomalous optical response in atomically thin InSe," *Nature Nanotechnology*, vol. 12, pp. 223–227, Nov. 2016. DOI: 10.1038/nnano.2016.242

[332] J. Keane and C. H. Kim, "An odomoeter for CPUs," *IEEE Spectrum*, vol. 48, no. 5, pp. 28–33, Apr. 2011. DOI: 10.1109/MSPEC.2011.5753241

[333] D. Kahng and S. M. Sze, "A floating gate and its application to memory devices," *The Bell System Technical Journal*, vol. 46, no. 6, pp. 1288–1295, July 1967. DOI: 10.1002/j.1538-7305.1967.tb01738.x

[334] A. Padilla, C. W. Yeung, C. Shin, C. Hu, and T.-J. K. Liu, "Feedback FET: A novel transistor exhibiting steep switching behavior at low bias voltages," in *International Electron Devices Meeting (IEDM) Technical Digest*, Dec. 2008. DOI: 10.1109/IEDM.2008.4796643

[335] C. W. Yeung, A. Padilla, T.-J. K. Liu, and C. Hu, "Programming characteristics of the steep turn-on/off feedback FET (FBFET)," in *Symposium on VLSI Technology*, pp. 176–177, June 2009. URL: https://ieeexplore.ieee.org/document/5200678

[336] J. Wan, S. Cristoloveanu, C. Le Royer, and A. Zaslavsky, "A feedback silicon-on-insulator steep switching device with gate-controlled carrier injection," *Solid-State Electronics*, vol. 76, pp. 109–111, July 2012. DOI: 10.1016/j.sse.2012.05.061

[337] Y. Solaro, P. Fonteneau, C. A. Legrand, C. Fenouillet-Beranger, P. Ferrari, and S. Cristoloveanu, "A sharp-switching device with free surface and buried gates based on band modulation and feedback mechanisms," *Solid-State Electronics*, vol. 116, pp. 8–11, Nov. 2015. DOI: 10.1016/j.sse.2015.10.010

[338] J. Wan, C. Le Royer, A. Zaslavsky, and S. Cristoloveanu, "Z^2-FET used as 1-transistor high-speed DRAM," in *2012 Proceedings of the European Solid-State Device Research Conference (ESSDERC)*, pp. 197–200, Sep. 2012. DOI: 10.1109/ESSDERC.2012.6343367

[339] Y. Solaro, P. Fonteneau, C.-A. Legrand, D. Marin-Cudraz, J. Passieux, P. Guyader, L.-R. Clement, C. Fenouillet-Beranger, P. Ferrari, and S. Cristoloveanu, "Innovative ESD protections for UTBB FD-SOI technology," in *International Electron Devices Meeting (IEDM) Technical Digest*, p. 7.3, Dec. 2013. DOI: 10.1109/IEDM.2013.6724580

[340] Y. Solaro, J. Wan, P. Fonteneau, C. Fenouillet-Beranger, C. Le Royer, A. Zaslavsky, P. Ferrari, and S. Cristoloveanu, "Z^2-FET: A promising FDSOI device for ESD protection," *Solid-State Electronics*, vol. 97, pp. 23–29, May 2014. DOI: 10.1016/j.sse.2014.04.032

[341] C. Lee, J. Sung, and C. Shin, "Understanding of feedback field-effect transistor and its applications," *Applied Sciences*, vol. 10, no. 9, p. 3070, Apr. 2020. DOI: 10.3390/app10093070

[342] J. B. Angell, S. C. Terry, and P. W. Barth, "Silicon micromechanical devices," *Scientific American*, vol. 248, no. 4, pp. 44–55, Apr. 1983. DOI: 10.1038/scientificamerican0483-44

[343] R. P. Feynman, "There's plenty of room at the bottom," in *Annual meeting of the American Physical Society*, Dec. 1959. URL: https://resolver.caltech.edu/CaltechES:23.5.1960Bottom

[344] F. Ohnesorge and G. Binnig, "True atomic resolution by atomic force microscopy through repulsive and attractive forces," *Science*, vol. 260, no. 5113, pp. 1451–1456, June 1993. DOI: 10.1126/science.260.5113.1451

[345] A. M. Ionescu, V. Pott, R. Fritschi, K. Banerjee, M. J. Declercq, P. Renaud, C. Hibert, P. Fluckiger, and G. A. Racine, "Modeling and design of a low-voltage SOI suspended-gate MOSFET (SG-MOSFET) with a metal-over-gate architecture," in *Proceedings International Symposium on Quality Electronic Design*, pp. 496–501, Mar. 2002. DOI: 10.1109/ISQED.2002.996794

[346] N. Abelé, R. Fritschi, K. Boucart, F. Casset, P. Ancey, and A. M. Ionescu, "Suspended-gate MOSFET: bringing new MEMS functionality into solid-state MOS transistor," in

International Electron Devices Meeting (IEDM) Technical Digest, pp. 479–481, Dec. 2005. DOI: 10.1109/IEDM.2005.1609384

[347] H. Kam, D. T. Lee, R. T. Howe, and T.-J. King, "A new nano-electro-mechanical field effect transistor (NEMFET) design for low-power electronics," in *International Electron Devices Meeting (IEDM) Technical Digest*, pp. 463–466, Dec. 2005. DOI: 10.1109/IEDM.2005.1609380

[348] J.-H. Kim, Z. C. Chen, S. Kwon, and J. Xiang, "Three-terminal nanoelectromechanical field effect transistor with abrupt subthreshold slope," *Nano Letters*, vol. 14, no. 3, pp. 1687–1691, Feb. 2014. DOI: 10.1021/nl5006355

[349] J. Cao and A. M. Ionescu, "Study on dual-lateral-gate suspended-body single-walled carbon nanotube field-effect transistors," *Solid-State Electronics*, vol. 74, pp. 121–125, Aug. 2012. DOI: 10.1016/j.sse.2012.04.022

[350] T. van Hemert and R. J. E. Hueting, "Active strain modulation in field effect devices," in *Proceedings of the European Solid-State Device Research Conference (ESSDERC)*, pp. 125–128, Sep. 2012. DOI: 10.1109/ESSDERC.2012.6343349

[351] T. van Hemert and R. J. E. Hueting, "Piezoelectric strain modulation in FETs," *IEEE Transactions on Electron Devices*, vol. 60, no. 10, pp. 3265–3270, Aug. 2013. DOI: 10.1109/TED.2013.2274817

[352] R. K. Jana, G. L. Snider, and D. Jena, "On the possibility of sub 60 mV/decade subthreshold switching in piezoelectric gate barrier transistors," *Physica Status Solidi C*, vol. 10, no. 11, pp. 1469–1472, Oct. 2013. DOI: 10.1002/pssc.201300280

[353] R. K. Jana, A. Ajoy, G. Snider, and D. Jena, "Sub-60 mV/decade steep transistors with compliant piezoelectric gate barriers," in *International Electron Devices Meeting (IEDM) Technical Digest*, p. 13.6, Dec. 2014. DOI: 10.1109/IEDM.2014.7047047

[354] B. Kaleli, R. J. E. Hueting, M. D. Nguyen, and R. A. M. Wolters, "Integration of a piezoelectric layer on Si FinFETs for tunable strained device applications," *IEEE Transactions on Electron Devices*, vol. 61, no. 6, pp. 1929–1935, May 2014. DOI: 10.1109/TED.2014.2316164

[355] H. Wang, X. Jiang, N. Xu, G. Han, Y. Hao, S.-S. Li, and D. Esseni, "Revised analysis of design options and minimum subthreshold swing in piezoelectric FinFETs," *IEEE Electron Device Letters*, vol. 39, no. 3, pp. 444–447, Jan. 2018. DOI: 10.1109/LED.2018.2791987

[356] A. M. Ionescu, L. Lattanzio, G. A. Salvatore, L. De Michielis, K. Boucart, and D. Bouvet, "The hysteretic ferroelectric tunnel FET," *IEEE Transactions on Electron Devices*, vol. 57, no. 12, pp. 3518–3524, Oct. 2010. DOI: 10.1109/TED.2010.2079531

[357] M. Kobayashi, K. Jang, N. Ueyama, and T. Hiramoto, "Negative capacitance for boosting tunnel FET performance," *IEEE Transactions on Nanotechnology*, vol. 16, no. 2, pp. 253–258, Jan. 2017. DOI: 10.1109/TNANO.2017.2658688

[358] W. Li, Q. Jia, Y. Pan, X. Chen, Y. Yin, Y. Wu, Y. Wang, Y. Wen, C. Wang, and S. Wang, "A T-shaped gate tunneling field effect transistor with negative capacitance, super-steep subthreshold swing," *Nanotechnology*, vol. 32, no. 39, p. 395202, July 2021. DOI: 10.1088/1361-6528/ac0d20

[359] N. Thoti and Y. Li, "Promised design of energy-efficient negative-capacitance vertical tunneling FET," *ECS Journal of Solid State Science and Technology*, vol. 10, no. 7, p. 075002, July 2021. DOI: 10.1149/2162-8777/ac10b4

[360] A. Saeidi, F. Jazaeri, F. Bellando, I. Stolichnov, G. V. Luong, Q.-T. Zhao, S. Mantl, C. C. Enz, and A. M. Ionescu, "Negative capacitance as performance booster for tunnel FETs and MOSFETs: An experimental study," *IEEE Electron Device Letters*, vol. 38, no. 10, pp. 1485–1488, Aug. 2017. DOI: 10.1109/LED.2017.2734943

[361] A. Saeidi, F. Jazaeri, I. Stolichnov, G. V. Luong, Q.-T. Zhao, S. Mantl, and A. M. Ionescu, "Effect of hysteretic and non-hysteretic negative capacitance on tunnel FETs DC performance," *Nanotechnology*, vol. 29, no. 9, p. 095202, Jan. 2018. DOI: 10.1088/1361-6528/aaa590

[362] Y. Long, J. Z. Huang, Q. Huang, N. Xu, X. Jiang, Z.-C. Niu, R. Huang, and S.-S. Li, "Piezoelectric tunnel FET with a steep slope," *IEEE Electron Device Letters*, vol. 41, no. 6, pp. 948–951, Apr. 2020. DOI: 10.1109/LED.2020.2988412