DISSERTATION

Simulation of Tunneling in Semiconductor Devices

ausgeführt zum Zwecke der Erlangung des akademischen Grades
eines Doktors der technischen Wissenschaften

eingereicht an der Technischen Universität Wien
Fakultät für Elektrotechnik und Informationstechnik
von

ANDREAS GEHRING

Messerschmidtgasse 2/2/7
A-1180 Wien, Österreich

Matr. Nr. 9525275
geboren am 5. Februar 1975 in Mistelbach

Wien, im November 2003

The road to wisdom? Well it's plain
and simple to express:
Err
and err
and err again
but less
and less
and less

Piet Hein

Kurzfassung

DIE MIKROELEKTRONIK hat einen Punkt erreicht, an dem quantenmechanische Effekte einen wesentlichen Einfluss auf die elektrischen Eigenschaften von Halbleiterbauelementen haben. Einer der wichtigsten dieser Effekte ist das quantenmechanische Tunneln von Ladungsträgern durch Schichten dünner Dielektrika. Einerseits führt dies zu einem erhöhten Leistungsverbrauch von Halbleiterbauelementen und limitiert dadurch die Dicke des Gatedielektrikums. Andererseits werden Tunneleffekte in nichtflüchtigen Speicherbauelementen verwendet um Ladung auf einen isolierten Speicherknoten zu transferieren.

Der Tunneleffekt basiert auf dem Übergang von Ladungsträgern von einer Elektrode durch eine klassisch isolierende Region auf eine andere Elektrode. Dieser Prozess wird durch drei Faktoren beeinflusst: Der energetischen Verteilung der Ladungsträger in beiden Elektroden, dem quantenmechanischen Transmissionskoeffizienten der Energiebarriere, und vorhandenen Störstellen im Dielektrikum die den Tunnelprozess beeinflussen.

Die energetische Verteilung der Ladungsträger in den Elektroden ist von fundamentaler Bedeutung für den Tunnelstrom. Üblicherweise wird eine FERMI-DIRAC oder MAXWELL-BOLTZMANN Verteilung angenommen. Diese Verteilungsfunktionen sind jedoch nur nahe des Gleichgewichtszustands gültig und scheitern bei der Beschreibung des Verhaltens heisser Ladungsträger. In dieser Arbeit wird eine neue Verteilungsfunktion verwendet, die auf der Konzentration, Temperatur, und Kurtosis der Ladungsträger basiert. Diese Verteilungsfunktion zeigt gute Übereinstimmung mit den Ergebnissen von Monte Carlo Simulationen und reproduziert die Verteilung der hochenergetischen Ladungsträger mit hoher Genauigkeit. Die heisse MAXWELL Verteilung, die nur auf der Konzentration und Temperatur der Ladungsträger basiert, kann die Verteilung der hochenergetischen Ladungsträger nicht reproduzieren und führt zu einer stark überhöhten Tunnelstromdichte.

Der quantenmechanische Transmissionskoeffizient wird durch Lösung der SCHRÖDINGER-Gleichung bestimmt und hängt von der Form der Energiebarriere im Dielektrikum ab. Dielektrika die aus einer einzigen Schicht bestehen zeigen eine lineare Potentialvariation in der Barriere die zu einem entweder dreieckigen oder trapezförmigen Banddiagramm führt. Für diesen Fall können analytische Modelle zur Berechnung des Transmissionskoeffizienten hergeleitet werden die auf der WENTZEL-KRAMERS-BRILLOUIN-Näherung oder der GUNDLACH-Formel beruhen.

Der stetige Miniaturisierungsprozess elektronischer Bauelemente führt jedoch zu einer entsprechenden Reduzierung der Dicke der Gatedielektrika in MOS Bauelementen, was für das fast ausschliesslich verwendete Material SiO_2 zu unzulässig hohen Leckströmen führt. Als Abhilfe wurden geschichtete Dielektrika aus Materialien mit höheren Dielektrizitätskonstanten vorgeschlagen.

In derartigen geschichteten Dielektrika hat das Banddiagramm einen nichtlinearen Verlauf und Modelle, die auf einer dreieckigen oder trapezförmigen Energiebarriere basieren, sind nicht mehr gültig. Stattdessen muss die SCHRÖDINGER-Gleichung mit Hilfe der Transfer-Matrix oder der Quantum Transmitting Boundary Methode gelöst werden. Diese Methoden wurden untersucht, wobei sich die Quantum Transmitting Boundary Methode auf Grund der höheren numerischen Stabilität und der Eignung für mehrdimensionale Probleme als vorteilhaft herausgestellt hat.

Nichtflüchtige Speicherbauelemente müssen bis zu 10^5 Schreib- und Löschvorgänge bei Spannungen in der Höhe von 8–12 V fehlerfrei ausführen. Diese wiederholte Belastung des Dielektrikums führt zur Bildung von Defekten, die störstellenunterstütztes Tunneln bei niedrigen Feldstärken ermöglichen. Diese Generation von Störstellen wird als einer der Hauptgründe für die Verschlechterung der Isolationseigenschaften des Dielektrikums angesehen. Störstellenunterstütztes Tunneln wird in dieser Arbeit als zweistufiger Prozess modelliert, bei dem Energie durch die Emission von Phononen frei wird. Die Besetzungsdichte der Defekte wird durch eine Ratengleichung beschrieben. Um diese Geichung zu lösen wird ein iteratives Modell verwendet.

Die beschriebenen Modelle wurden in den Bauelementsimulator MINIMOS-NT implementiert. Zahlreiche Anwendungen wurden untersucht, wobei eine Unterscheidung zwischen MOS Transistoren und nichtflüchtigen Speicherbauelementen gemacht wurde. Die Anwendbarkeit alternativer Dielektrika wurde untersucht und an Hand eines MOS Kondensators mit ZrO_2 Dielektrikum mit Messungen verglichen. Weiters wurden nichtflüchtige Speicherbauelemente wie EEPROMs und alternative Strukturen untersucht. Mit Hilfe der implementierten Modelle kann MINIMOS-NT für die Modellierung des Tunnelstroms in beliebigen Halbleiterbauelementen verwendet werden.

Abstract

MICROELECTRONICS has reached a point where quantum effects have a major impact on the electrical characteristics of semiconductor devices. One of the most important effects in this regime is the quantum-mechanical tunneling of carriers through thin dielectric layers. On the one hand, this leads to increased power consumption and thus limits the thickness of the gate dielectric. On the other hand, tunneling effects are used in non-volatile memory devices to transfer charge to an isolated floating gate.

The tunneling current is caused by the transition of carriers from one electrode through a classically isolating region to another electrode. Three major factors influence this process: the carrier energy distribution at both electrodes, the quantum-mechanical transmission coefficient of the energy barrier between the electrodes, and the presence of traps in the insulating layer which may assist in the tunneling process.

The carrier energy distribution is of major importance for the tunneling process. The FERMI-DIRAC or MAXWELL-BOLTZMANN distribution is frequently used to approximate this distribution. These expressions are, however, only valid near equilibrium and fail to describe the distribution of hot carriers. In this work an alternative expression for the distribution function, which is based on the carrier concentration, temperature, and kurtosis, was applied. This distribution shows good agreement with results from Monte Carlo simulations and accurately reproduces the high-energy tail of the distribution. The heated MAXWELLian distribution, which only accounts for the electron concentration and temperature, completely fails to reproduce the high-energy tail and highly overestimates the tunneling current density.

The quantum-mechanical transmission coefficient is calculated by solving the stationary SCHRÖDINGER equation in the region considered for tunneling. The coefficient depends on the shape of the energy barrier in the dielectric layer. Dielectrics which consist of a single layer give rise to a linear potential variation in the barrier, yielding either a trapezoidal or a triangular band diagram. Analytical models can be derived to approximately calculate the transmission coefficient in these cases, based on the WENTZEL-KRAMERS-BRILLOUIN approximation or on GUNDLACH's formula.

With reduced device dimensions, however, the gate dielectric in MOS devices must be scaled accordingly which, for the commonly used material SiO_2, leads to an intolerably high gate current density. To overcome this problem, gate dielectric stacks including high-κ dielectrics have been proposed.

In such dielectric stacks the band profile has a non-linear shape, and models based on triangular or trapezoidal barriers are no more valid. Instead, SCHRÖDINGER's equation must be solved using the transfer-matrix or the quantum transmitting boundary method. These methods have been studied and the quantum transmitting boundary method was found superior due to its better numerical stability and the possibility to apply it to two- and three-dimensional problems.

Non-volatile memory devices need to endure up to 10^5 write and erase cycles at a voltage of 8-12 V. This repeated high-field stress introduces defects in the tunneling dielectric, which give rise to trap-assisted tunneling current at low electric fields. That trap generation is considered a major reason for device degradation. In this work trap-assisted tunneling is modeled as a two-step process during which energy relaxation by phonon emission takes place. The trap occupancy in the dielectric is described by a time-dependent rate equation. To solve this equation, an iterative procedure is applied.

Models to describe the outlined processes have been implemented into the general-purpose device simulator MINIMOS-NT. Several applications are investigated, where a distinction between MOS transistors and non-volatile memory devices is made. The applicability of alternative dielectric materials is investigated and, as an example, a MOS capacitor with a ZrO_2 dielectric is simulated and compared with measurements. Non-volatile memory devices such as conventional EEPROM devices, trap-rich dielectric devices, multi-barrier tunneling devices, and devices which use layered tunnel barriers to improve the retention time are investigated. With the implemented models, MINIMOS-NT can be used for the evaluation of tunneling currents in device structures of arbitrary complexity.

Acknowledgment

FIRST AND FOREMOST I want to thank Prof. SIEGFRIED SELBERHERR for giving me the opportunity to join his research group, for providing the excellent infrastructure at the Institute for Microelectronics, and for the strong industrial network which allows his students to gain international experience.

I thank Prof. EMMERICH BERTAGNOLLI, who actually was one of the first to ignite my interest in microelectronics, that he was willing to serve on my examination committee. Furthermore, I am strongly indebted to Prof. ERASMUS LANGER, the head of the Institute for Microelectronics, who was a strict but very cooperative and helpful boss.

I enjoyed the luck to work closely together with two great advisors: Prof. HANS KOSINA and Prof. TIBOR GRASSER. Prof. KOSINA impressed me from the very first encounter with his deep knowledge on semiconductor device modeling and physics in general. His clear and understandable way to write down even the most abstract and complicated topics, sometimes consuming only the confined space of a napkin, provided the background of my work.
Prof. GRASSER, the head of the MINIMOS-NT development crew, convinced me very soon that programming is not simply a craft, but an elaborate art. His ingenious coding style and his sound knowledge on software architecture inspired me to improve my knowledge in these directions.

ROBERT KLIMA bootstrapped me at the institute, and he was always willing to discuss programming topics with me. Proofreading his texts on programming in C allowed me to share his deep understanding of programming.
I am also indebted to STEPHAN WAGNER, a real C++ expert, for explaining me numerous aspects of object-oriented programming, helping me with compilation problems, and regularly improving my code.

I am grateful to MARKUS GRITSCH who was a very pleasant and intelligent room mate. His comprehensive knowledge about the typesetting system LaTeX helped me on countless occasions. SERGEY SMIRNOV became my second room mate, and immediately impressed me with his strong knowledge in semiconductor physics and his incredible eagerness to work. Recently, STEPHAN HOLZER joined our office where he soon became very popular due to the chocolate sweets he used to distribute.

KLAUS DRAGOSITS was a never ending source of funny stories about all kind of topics, and I thank him for frequent support and discussions on CV simulations and quantum-mechanical modeling. His room mate VASSIL PALANKOVSKI impressed me with his long working hours and publication list, and with his comprehensive knowledge about bipolar devices. I also remember PETER FLEISCHMANN who provided me with some in-depth knowledge about Japanese culture and railroad timetables on our trip to Japan. Furthermore, I am grateful to JONG-MUN PARK who shared his strong experience with commercial TCAD simulation packages, and TESFAYE AYALEW for numerous discussions on high-power devices. RAINER SABELKA, JOHANNES CERVENKA, CHRISTIAN HARLANDER, ENZO UNGERSBÖCK, and ROBERT ENTNER took care of the network and computer infrastructure for which I want to thank them, too.

EWALD HASLINGER, MANFRED KATTERBAUER, and RENATE WINKLER provided the background work at the Institute, which much too often is completely undervalued. All other members of the Institute for Microelectronics deserve gratitude for assistance and for the stimulating working atmosphere they create.

BYOUNGHO CHEONG from the Samsung Advanced Institute for Technology was a very kind and patient project leader with whom I cooperated for years on several topics. He also introduced me into the Korean culture, habits, and cuisine, for which I am very grateful.
FRANCISCO JIMÉNEZ-MOLINOS spent two very productive months at our institute and I thank him for the good cooperation regarding the trap-assisted tunneling model.
STEFAN HARASEK provided me with some information about the real world of high-κ dielectric materials and I thank him for measurements and discussions on trap-assisted tunneling mechanisms.
HELMUT PUCHNER enabled an internship at Cypress Semiconductor at very short notice. He impressed me with his comprehensive knowledge about CMOS device and process technology and his high level of professionalism.

More important than any other support, my wife ELISABETH provided me with love and understanding, giving me a feeling for the really important things in life. Finally, none of my studies would have been possible without the continuous support of my parents.

Contents

Kurzfassung ii

Abstract iv

Acknowledgment vi

Contents viii

List of Abbreviations and Acronyms xiii

List of Symbols xv

Notation xv

Physical Quantities xvi

Constants xviii

1 Introduction 1

2 Fundamentals of CMOS Devices 3

2.1 Historical Overview 4

2.2 Obstacles to Device Miniaturization 6

2.2.1 Channel 7

2.2.2 Gate Stack 9

2.2.3 Source and Drain 10

2.2.4 Gate Dielectric 10

2.3 Novel Device Concepts . . . 11
2.3.1 Strained-Silicon Devices . . . 11
2.3.2 Depleted-Substrate Devices . . . 11
2.3.3 Vertical Transistors . . . 11
2.3.4 Carbon Nanotube FET . . . 12
2.4 Semiconductor Device Simulation . . . 13
2.4.1 Hierarchy of Semiconductor Device Simulation Models . . . 14
2.4.2 Classical Device Simulation . . . 14
2.4.2.1 The Drift-Diffusion Model . . . 15
2.4.2.2 The Energy-Transport Model . . . 15
2.4.2.3 Monte Carlo Device Simulation . . . 15
2.4.3 Quantum Device Simulation . . . 16
2.4.3.1 SCHRÖDINGER Equation and SCHRÖDINGER-POISSON Solvers . . 16
2.4.3.2 WIGNER Equation and the Density-Gradient Model . . . 17
2.4.3.3 Quantum Monte Carlo Device Simulation . . . 17
2.4.3.4 Non-Equilibrium GREEN's Function Device Simulation . . . 18

3 Tunneling in Semiconductors 19
3.1 Tunneling Mechanisms . . . 20
3.2 The TSU-ESAKI Model . . . 21
3.3 Supply Function Modeling . . . 24
3.3.1 FERMI-DIRAC Distribution . . . 24
3.3.2 MAXWELL-BOLTZMANN Distribution . . . 25
3.3.3 Non-MAXWELLian Distributions . . . 26
3.3.4 Normalization . . . 31
3.4 The Energy Barrier . . . 32
3.4.1 The Metal-Oxide-Semiconductor Capacitor . . . 32
3.4.2 Image Force Correction . . . 34
3.5 Transmission Coefficient Modeling . . . 36
3.5.1 The WENTZEL-KRAMERS-BRILLOUIN Approximation . . . 37
3.5.2 The GUNDLACH Method . . . 38
3.5.3 Transfer-Matrix Method . . . 40

3.5.3.1 Piecewise-Constant Potential 40
3.5.3.2 Piecewise-Linear Potential 41
3.5.4 Quantum Transmitting Boundary Method 43
3.5.5 Comparison 44
3.6 Bound and Quasi-Bound States 45
3.6.1 Eigenvalues of a Triangular Energy Well 46
3.6.2 Eigenvalues of Arbitrary Energy Wells 47
3.6.3 The Life Time of Quasi-Bound States 47
3.6.3.1 The Reflection Coefficient Resonances 48
3.6.3.2 The Quasi-Classical Formula 50
3.6.3.3 The Eigenvalues of the Non-HERMITian HAMILTONian 50
3.7 Compact Tunneling Models 52
3.8 Trap-Assisted Tunneling 53
3.8.1 Model Overview 53
3.8.1.1 CHANG's Model 54
3.8.1.2 IELMINI's Model 54
3.8.1.3 Compact Trap-Assisted Tunneling Models 55
3.8.2 The Model of JIMÉNEZ *et al.* 56
3.8.2.1 Capture and Emission Probabilities 56
3.8.2.2 Capture and Emission Times 59
3.8.2.3 Steady-State Current 61
3.8.2.4 Transient Current 61
3.9 Model Comparison 63

4 Implementation **64**
4.1 The Device Simulator MINIMOS-NT 64
4.2 The Tunneling Model 65
4.2.1 Single Segment Tunneling 67
4.2.2 Stacked Segment Tunneling 68
4.2.3 Trap-Assisted Tunneling 69
4.3 The SCHRÖDINGER Solver 71
4.3.1 Open and Closed Boundary conditions 71

4.3.2 System HAMILTONian 72
4.3.3 The Eigenvalue Solver 74

5 Applications **76**

5.1 Tunneling in MOS Transistors 77
5.1.1 Tunneling Paths in MOS Transistors 77
5.1.2 Channel Tunneling 78
5.1.2.1 Effect of the Polysilicon Gate Doping on the Channel Tunneling 79
5.1.2.2 Effect of the Substrate Doping on the Channel Tunneling 80
5.1.2.3 Effect of the Dielectric Thickness on the Channel Tunneling . . 81
5.1.2.4 Effect of the Barrier Height on the Channel Tunneling 82
5.1.2.5 Effect of the Carrier Mass on the Channel Tunneling 83
5.1.2.6 Effect of the Dielectric Permittivity on the Channel Tunneling . 84
5.1.2.7 Effect of the Lattice Temperature on the Channel Tunneling . . 85
5.1.2.8 Comparison to Measurements 86
5.1.2.9 Validity of Compact Models 87
5.1.3 Source and Drain Extension Tunneling 87
5.1.3.1 Effect of the Polysilicon Gate Doping on the Source and Drain Extension Tunneling 88
5.1.3.2 Effect of the Substrate Doping on the Source and Drain Extension Tunneling 89
5.1.3.3 Effect of the Dielectric Permittivity on the Source and Drain Extension Tunneling 90
5.1.3.4 Effect of the Lattice Temperature on the Source and Drain Extension Tunneling 91
5.1.4 Hot-Carrier Tunneling in MOS Transistors 92
5.1.5 Alternative Dielectrics for MOS Transistors 94
5.1.6 Trap-Assisted Tunneling in ZrO_2 Dielectrics 99
5.2 Tunneling in Non-Volatile Memory Devices 101
5.2.1 Conventional EEPROM Devices 101
5.2.1.1 Static SILC in EEPROMs 102
5.2.1.2 Transient SILC in EEPROMs 103
5.2.2 Alternative Non-Volatile Memory Devices 105

5.2.2.1 Non-Volatile Memory Devices Based on Trap-Rich Dielectrics . . 107
5.2.2.2 Multi-Barrier Tunneling Devices 109
5.2.2.3 Non-Volatile Memory Devices Based on Crested Barriers 112

6 Summary and Conclusions **114**

A The FOWLER-NORDHEIM Formula **116**
A.1 Original FOWLER-NORDHEIM Formula 117
A.2 Correction for Direct Tunneling 119

B The WKB Approximation **122**

C Wave Function Normalization for a Triangular Potential **124**

D User Interface **126**
D.1 Direct Tunneling 126
D.1.1 The Model `FNPure` 127
D.1.2 The Model `FNLenzlingerSnow` 129
D.1.3 The Model `DTSchuegraf` 129
D.1.4 The Model `FrenkelPoole` 129
D.1.5 The Model `TsuEsaki` 130
D.2 Stacked Segments 131
D.3 Oxide Traps 133
D.4 Trap-Assisted Tunneling 134

Bibliography **135**

Own Publications **156**

Curriculum Vitae **159**

List of Abbreviations and Acronyms

nMOS	...	n-type MOS
pMOS	...	p-type MOS
BTE	...	BOLTZMANN's transport equation
CMOS	...	Complementary MOS
CNT	...	Carbon nanotube
CSB	...	Central shutter barrier
CV	...	Capacitance-voltage
DIBL	...	Drain-induced barrier lowering
DRAM	...	Dynamical RAM
ECB	...	Electrons from the conduction band
EED	...	Electron energy distribution
EEPROM	...	Electrically erasable programmable read-only memory
EOT	...	Effective oxide thickness
EVB	...	Electrons from the valence band
FET	...	Field-effect transistor
FN	...	FOWLER-NORDHEIM
FWHM	...	Full-width half-maximum
HED	...	Hole energy distribution
HVB	...	Holes from the valence band
ITRS	...	International Technology Roadmap for Semiconductors
IV	...	Current-voltage
LDD	...	Lightly doped drain
MOCVD	...	Metal-organic chemical vapor deposition
MOS	...	Metal-oxide-semiconductor
MPU	...	Microprocessor unit
MOSFET	...	MOS field-effect transistor
NEGF	...	Non-equilibrium GREEN's function

NTRS	...	National Technology Roadmap for Semiconductors
NVM	...	Non-volatile memory
PIF	...	PROFILE INTERCHANGE FORMAT
PLEDM	...	Planar localized-electron device memory
PLEDTR	...	Planar localized-electron device transistor
QTBM	...	Quantum transmitting boundary method
QBS	...	Quasi-bound state
RAM	...	Random-access memory
RTA	...	Relaxation-time approximation
SIA	...	Semiconductor Industry Association
SILC	...	Stress-induced leakage current
SIMS	...	Secondary ion mass spectroscopy
SOI	...	Silicon on insulator
SONOS	...	Silicon-oxide-nitride-oxide-silicon
SRAM	...	Static random-access memory
STI	...	Shallow trench isolation
TAT	...	Trap-assisted tunneling
TCAD	...	Technology computer-aided design
TEM	...	Transmission electron microscopy
TM	...	Transfer matrix
WKB	...	WENTZEL-KRAMERS-BRILLOUIN
WSS	...	WAFER-STATE SERVER

List of Symbols

Notation

x	...	Scalar
x^*	...	Complex conjugate of x
$\mathbf{x}$	...	Vector
$\underline{A}$	...	Matrix
A_{ij}	...	Elements of the matrix $\underline{A}$
$\underline{A}^+$	...	Conjugate transposed matrix: $A_{ij} = A^*_{ji}$
$\mathbf{e}_x$	...	Unity vector in direction x
$\mathbf{x} \cdot \mathbf{y}$	...	Scalar (in) product
$\partial_t(\cdot)$	...	Partial derivative with respect to t
∇	...	Nabla operator
$\nabla\mathbf{x}$	...	Gradient of $\mathbf{x}$
$\nabla \cdot \mathbf{x}$	...	Divergence of $\mathbf{x}$
$\nabla \cdot \nabla = \nabla^2$	...	LAPLACE operator
$\Gamma(\cdot)$	...	Gamma function
$\Gamma_i(\cdot, \cdot)$	...	Incomplete gamma function
$\langle \cdot \rangle$	...	Statistical average
$f(\mathbf{r}, \mathbf{k}, t)$	...	Distribution function
$\mathcal{Q}(f)$	...	Collision operator
$\underline{H}$	...	HAMILTONian operator
$\underline{G}$	...	GREEN's function
$\underline{I}$	...	Unity matrix
$\underline{T}$	...	Transfer matrix
$\det(\cdot)$	...	Determinant of a matrix
$\mathcal{F}_i$	...	FERMI integral of order i
$\mathrm{Ai}(\cdot)$, $\mathrm{Bi}(\cdot)$	...	AIRY functions
$\mathrm{Ai}'(\cdot)$, $\mathrm{Bi}'(\cdot)$	...	Derivative of the AIRY functions

Physical Quantities

Symbol	Unit	Description
β_n	1	Electron kurtosis
β_{Bulk}	1	Electron kurtosis in the bulk
$q\chi_S$	eV	Electron affinity in the semiconductor
D_n	m^2s^{-1}	Electron diffusion coefficient
D_p	m^2s^{-1}	Hole diffusion coefficient
E	Vm^{-1}	Electric field
E_{diel}	Vm^{-1}	Electric field in the dielectric
$\mathcal{E}$	eV	Energy
$\mathcal{E}_f$	eV	FERMI energy
$\mathcal{E}_c$	eV	Conduction band edge energy
$\mathcal{E}_v$	eV	Valence band edge energy
$\mathcal{E}_{c,0}$	eV	Conduction band edge energy in the flat-band case
$\mathcal{E}_{v,0}$	eV	Valence band edge energy in the flat-band case
$\mathcal{E}_g$	eV	Band gap energy
$\mathcal{E}_i$	eV	Intrinsic energy
$\mathcal{E}_{\text{image}}$	eV	Image force correction energy
$\mathcal{E}_x$	eV	Energy component in the tunneling direction
$\mathcal{E}_\rho$	eV	Energy component perpendicular to the tunneling direction
$\mathcal{E}_i$	eV	Energy eigenvalue
$\mathcal{E}_{\text{im}}$	eV	Imaginary part of the energy eigenvalue
$\mathcal{E}_{\text{re}}$	eV	Real part of the energy eigenvalue
$\mathcal{E}_T$	eV	Trap energy level below the dielectric conduction band
$\mathcal{E}'$	eV	Trap energy
ϕ	V	Electrostatic potential
ϕ_{surf}	V	Surface potential
Φ_f	V	FERMI potential
$q\Phi_B$	eV	Barrier height
$q\Phi_W$	eV	Work function
$q\Phi_S$	eV	Work function of the semiconductor
$q\Phi_M$	eV	Work function of the metal
$q\Phi_{MS}$	eV	Work function difference between metal and semiconductor
$q\Phi$	eV	Upper edge of a triangular energy barrier
$q\Phi_0$	eV	Lower edge of a triangular energy barrier
$q\Phi_e$	eV	Electron energy barrier
$q\Phi_h$	eV	Hole energy barrier
f_P	1	Distribution of phonons in energy
f_T	1	Trap occupancy
g	$m^{-3}eV^{-1}$	Density of states
$\hbar\omega$	eV	Phonon energy
$\mathbf{J}$	Am^{-2}	Current density

Symbol	Unit	Description
$\mathbf{J_n}$	Am^{-2}	Electron current density
$\mathbf{J_p}$	Am^{-2}	Hole current density
k	m^{-1}	Wave number
$\mathbf{k}$	m^{-1}	Wave number vector
k_x	m^{-1}	Wave number component in the tunneling direction
k_ρ	m^{-1}	Wave number component perpendicular to the tunneling direction
k_f	m^{-1}	Radius of the FERMI sphere
κ	$\mathrm{AsV}^{-1}\mathrm{m}^{-1}$	Dielectric permittivity
κ_diel	$\mathrm{AsV}^{-1}\mathrm{m}^{-1}$	Dielectric permittivity of the dielectric layer
κ_si	$\mathrm{AsV}^{-1}\mathrm{m}^{-1}$	Dielectric permittivity in silicon
κ_sio2	$\mathrm{AsV}^{-1}\mathrm{m}^{-1}$	Dielectric permittivity in silicon dioxide
$\kappa_{\mathrm{high}-\kappa}$	$\mathrm{AsV}^{-1}\mathrm{m}^{-1}$	Dielectric permittivity in a high-κ dielectric
μ_n	$\mathrm{m}^2\mathrm{V}^{-1}\mathrm{s}^{-1}$	Electron mobility
μ_p	$\mathrm{m}^2\mathrm{V}^{-1}\mathrm{s}^{-1}$	Hole mobility
μ_s	$\mathrm{m}^2\mathrm{V}^{-1}\mathrm{s}^{-1}$	Energy flux mobility
m	kg	Mass
m_diel	kg	Carrier mass in the dielectric
m_eff	kg	Carrier effective mass in the semiconductor
n	m^{-3}	Electron concentration
n_i	m^{-3}	Intrinsic concentration
N	eV	Supply function
N_D	m^{-3}	Concentration of donors
N_A	m^{-3}	Concentration of acceptors
N_poly	m^{-3}	Concentration of dopants in the polysilicon
N_c	m^{-3}	Effective density of states of the conduction band
N_T	m^{-3}	Trap concentration
p	m^{-3}	Hole concentration
P	1	Number of phonons
Ψ	$\mathrm{m}^{-1/2}$	Wave function
Q_T	As	Trap charge state
$\mathbf{r}$	m	Space vector
ρ	1	Probability density
R	$\mathrm{s}^{-1}\mathrm{m}^{-3}$	Net recombination rate
RC	1	Reflection coefficient
R_tun	$\mathrm{m}^{-3}\mathrm{s}^{-1}$	Additional recombination term due to the tunneling current
S_n	$\mathrm{J\ m}^{-2}\ \mathrm{s}^{-1}$	Electron energy flux density
S	1	HUANG-RHYS factor
t	s	Time
$\tau_\mathcal{E}$	s	Energy relaxation time
τ_m	s	Momentum relaxation time
τ_s	s	Energy flux relaxation time
τ_β	s	Kurtosis relaxation time

Symbol	Unit	Description
τ_{q}	s	Life time of a quasi-bound state
τ_{c}	s	Capture time
τ_{e}	s	Emission time
τ_{ca}	s	Capture time to the anode
τ_{cc}	s	Capture time to the cathode
τ_{ea}	s	Emission time to the anode
τ_{ec}	s	Emission time to the cathode
t_{diel}	m	Thickness of a dielectric
$t_{\mathrm{sio_2}}$	m	Thickness of a SiO_2 dielectric
$t_{\mathrm{high}-\kappa}$	m	Thickness of a high-κ dielectric
T	K	Temperature
T_{L}	K	Lattice temperature
T_n	K	Electron temperature
TC	1	Transmission coefficient
v	ms^{-1}	Velocity
$\mathbf{v}$	ms^{-1}	Velocity vector
v_x	ms^{-1}	Velocity component in the tunneling direction
v_ρ	ms^{-1}	Velocity component perpendicular to the tunneling direction
V_{poly}	V	Voltage drop in the polysilicon
V_{GS}	V	Gate-source voltage
V_{DS}	V	Drain-source voltage
V_{CG}	V	Control gate voltage
V_{FG}	V	Floating gate voltage
V_{diel}	V	Voltage drop in the dielectric
V_{e}	Jm	Overlap integral
W	eV	Potential energy
W_{c}	$\mathrm{m}^2\mathrm{s}^{-1}$	Capture rate
W_{e}	$\mathrm{m}^2\mathrm{s}^{-1}$	Emission rate
x_{T}	m	Trap cube side length

Constants

h	...	PLANCK's constant	$6.6260755 \times 10^{-34}\,\mathrm{Js}$
$\hbar$	...	Reduced PLANCK's constant	$h/(2\,\pi)$
$\mathrm{k_B}$	...	BOLTZMANN's constant	$1.380662 \times 10^{-23}\,\mathrm{JK}^{-1}$
q	...	Elementary charge	$1.6021892 \times 10^{-19}\,\mathrm{C}$
$\mathrm{m_0}$	...	Electron rest mass	$9.1093897 \times 10^{-31}\,\mathrm{kg}$
κ_0	...	Dielectric constant	$8.8541878 \times 10^{-12}\,\mathrm{AsV}^{-1}\mathrm{m}^{-1}$
$\imath$	...	$\sqrt{-1}$	

'... many high barriers exist in this world: Barriers between nations, races, and creeds. Unfortunately, some barriers are thick and strong. But I hope, with determination, we will find a way to tunnel through these barriers easily and freely, to bring the world together so that everyone can share in the legacy of Alfred Nobel.'

Leo Esaki

Chapter 1

Introduction

THE INCREASING demand for higher computing power, smaller dimensions, and lower power consumption of electronic devices leads to a pressing need to downscale semiconductor components. This process has already led to length scales where the electrical device characteristics is dominated by quantum-mechanical effects. One of the most interesting of these effects is the quantum-mechanical tunneling of charge carriers through classically forbidden regions.

This effect is important for many aspects of microelectronic technology. On the one hand, tunneling currents are exploited in non-volatile memory cells such as EEPROM (electrically erasable programmable read-only memory) or Flash devices to transfer charge to an isolated floating gate by applying high voltages at a capacitively coupled contact. On the other hand, parasitic tunneling currents through the ultra-thin gate dielectric cause increased power consumption of deep-submicron MOS (metal-oxide-semiconductor) transistors. DRAM (dynamical random-access memory) and quasi-nonvolatile SRAM (static random-access memory) cells face reduced retention times due to leakage through the memory node isolation. Resonant tunneling diodes are based on the tunneling mechanism to achieve a negative differential resistance, resulting in extraordinarily high operating frequencies.

It is therefore necessary to account for tunneling effects in the design of semiconductor devices. This can be achieved using numerical simulation. In the field of microelectronics the term TCAD (technology computer-aided design) is used to describe the numerical simulation of the semiconductor manufacturing process and the prediction of the electrical characteristics of the resulting devices. **Chapter 2** describes the fundamentals of contemporary CMOS (complementary MOS) technology, gives a brief overview about the crucial topics encountered in device scaling, and outlines the hierarchy of TCAD simulation approaches.

Several models of varying complexity and accuracy can be derived to describe the tunneling current density in semiconductor devices. The models depend on two central quantities, namely the supply function, which describes the supply of available electrons, and the transmission coefficient, which describes the probability that an electron can tunnel through the barrier. The supply function is determined by the energy distribution of the electrons. In equilibrium, this distribution can be approximated by a MAXWELLian distribution.

However, the electric field in miniaturized devices is so high that non-MAXWELLian models have to be be considered to accurately describe the shape of the distribution function and especially the shape of the high-energy tail of the distribution.

To calculate the transmission coefficient of a dielectric layer, SCHRÖDINGER's equation must be solved. One of the most frequently used methods is the WENTZEL-KRAMERS-BRILLOUIN (WKB) approximation which, however, does not reproduce transmission coefficient oscillations as observed in thin gate dielectrics. To accurately describe tunneling through dielectric stacks, it is necessary to resolve the effects of wave function interference. This can be achieved using the transfer-matrix method with either constant or linear potential segments. However, this method is numerically stable only for layer thicknesses up to a few nanometers. It is therefore hardly applicable to the simulation of high-κ dielectric stacks, which may have thicknesses of up to 10 nm. A more promising approach is the quantum transmitting boundary method which allows a stable and reliable evaluation of the transmission coefficient.

Unlike assumed in idealized models, dielectric layers are not ideal insulators. Caused by electric stress or processing conditions, defects arise in the dielectric which give rise to trap-assisted tunneling. This results in increased tunneling current at low bias, which is referred to as SILC (stress-induced leakage current). The trap-assisted tunneling process is caused by inelastic transitions of carriers supported by the emission of phonons. As this is a transient process it is necessary to account for the creation and annihilation of traps in the dielectric based on the rate equation of the traps.

All these effects are discussed in **Chapter 3** which treats the theory of tunneling in semiconductors. This comprises modeling of the supply function, the transmission coefficient, and trap-assisted tunneling.

Modern device simulators are complex software packages and the integration of interfaces to allow tunneling of charge carriers between arbitrary places in a device is not a straightforward task. **Chapter 4** provides a short description of the device simulator MINIMOS-NT and summarizes the implementation of the tunneling models. Furthermore, the SCHRÖDINGER solver which is used for the calculation of the transmission coefficient is briefly sketched.

In **Chapter 5** several applications are presented. MINIMOS-NT is used for the simulation of gate leakage currents in MOS capacitors and MOSFETs (MOS field-effect transistors). Emphasis is put on the modeling of the different tunneling paths in MOS transistors and on the evaluation of alternative high-κ dielectric materials. Furthermore, several NVM (non-volatile memory) devices such as EEPROM devices, trap-rich dielectric, or multi-barrier tunneling based devices are investigated.

Finally, **Chapter 6** briefly summarizes the thesis with some conclusions.

'There is plenty of room at the bottom.'
Richard P. Feynman

Chapter 2

Fundamentals of CMOS Devices

SWITCHES are the main building blocks of any hardware logic implementation. Computers in today's meaning[1] have been realized using mechanical and later electromechanical switches. The main shortcomings of such components are their low speed and their high power consumption. Vacuum tubes, which are switches without moving parts, have been used as replacements, but suffered from poor reliability. The invention of semiconductor switches gave a fast and reliable alternative. Bipolar transistors allow a high switching speed and a large amplification, however, current flow into the base contact must be maintained to keep the switch open. In metal-oxide-semiconductor field-effect transistors the current flow is controlled by a voltage. Ideally, no power is needed to control the on- and off-state. Complementary MOS technology is based on complementary-type transistors where current flows only during the switching process. These devices allow hardware logic implementations with extremely low standby power, high speed, and small footprint. Fig. 2.1 shows a schematic and a simplified layout of a CMOS inverter, the workhorse of all modern computers. An n-type MOS (nMOS) and a p-type MOS (pMOS) device are fabricated on the same p-doped wafer, with the pMOS device embedded in an n-doped well. The footprint of these structures is very small and allows high integration densities.

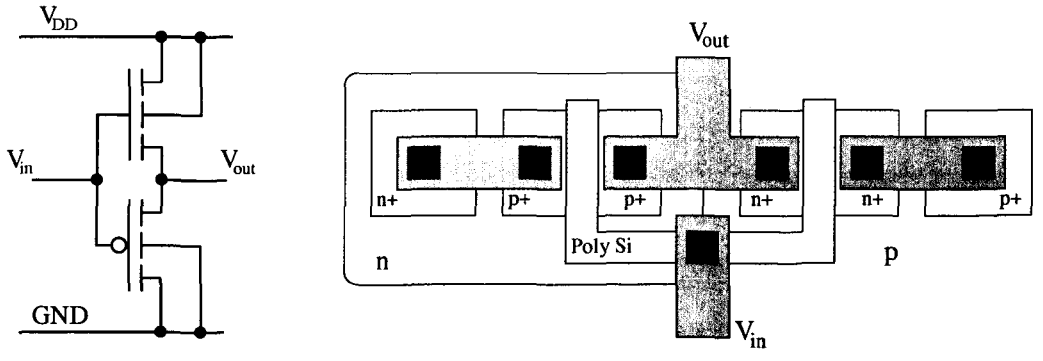


Figure 2.1: Schematics of a CMOS inverter and its layout.

[1] In early times, the term 'computer' denoted a person who does computations.

2.1 Historical Overview

The field-effect transistor (FET) was first proposed by LILIENFELD in 1926 and patented in 1930 [1]. However, the practical implementation was impossible due to material-related problems. On December 23rd 1947, BARDEEN and BRATTAIN, scientists at AT&T Bell Labs who worked in the group of SHOCKLEY, discovered the transistor effect [2–4], for which they received the NOBEL prize in 1956. The first integrated circuit was demonstrated by KILBY at Texas Instruments in 1959. In the same year NOYCE and MOORE, who had been working with SHOCKLEY, founded the company Fairchild Semiconductor, where they introduced the first commercially used semiconductor transistors[2]. The first field-effect transistor based on MOS technology was developed by KAHNG and ATALLA in 1960 [5]. MOORE, NOYCE, and GROVE left Fairchild Semiconductor and founded the company Intel in 1968. Soon, this company became the leading manufacturer of microprocessors. In 1965, MOORE reckoned that the number of transistors per integrated circuit approximately doubles every year, and he contributed this to three main effects: improvements in lithography, increased chip size, and gain from circuit and design innovation [6]. In 1975 he updated his statement and predicted that the number of transistors doubles every eighteen months to two years [7]. This statement has become widely known as MOORE's law, and it became the main paradigm of the microelectronics industry in the following decades.

The steady reduction of MOSFET device dimensions and integration densities found a theoretical basis in 1974 when DENNARD presented the constant-field scaling law [8] according to which the device dimensions can be reduced without altering the electrical characteristics if all dimensions, voltages, and doping concentrations are scaled in such a way that the electric field in the device remains constant. Hence, lengths and voltages are reduced by a factor s, while doping concentrations are increased by the same factor. This is shown schematically in Fig. 2.2 for a scaling factor $s = 2$ [9]. BACCARANI *et al.* presented a generalized scaling law [10] which takes into account that voltages cannot be reduced by the same factor as lengths. Instead, if voltages are scaled with a factor s_2 and lengths with a factor s_1, doping concentrations must be scaled by s_1^2/s_2.

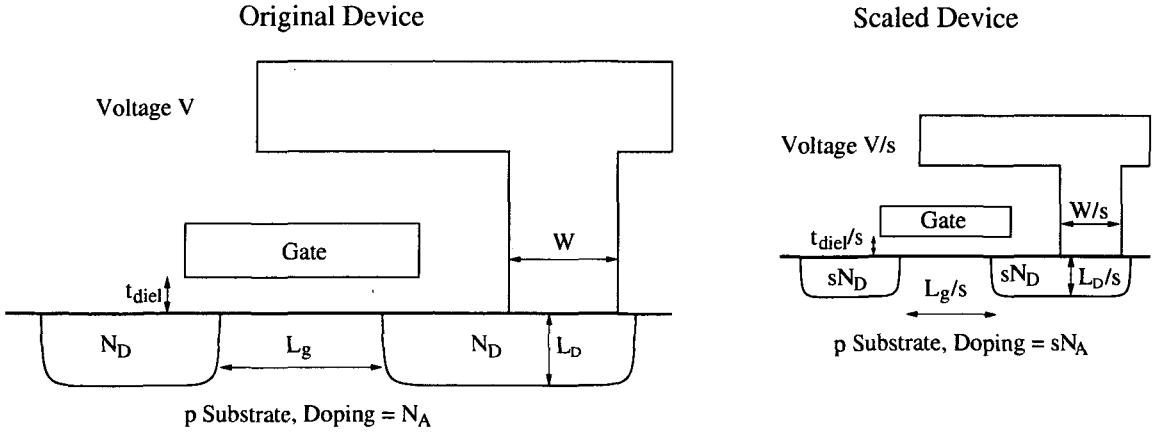


Figure 2.2: Constant-field scaling of MOS devices.

[2]The word *transistor* stems from *transconductance varistor*.

In 1992, the Semiconductor Industry Association (SIA) published the National Technology Roadmap for Semiconductors (NTRS) which was later replaced by the International Technology Roadmap for Semiconductors (ITRS). This document represents a collaborative effort to identify critical topics in semiconductor development. Every two years, comprehensive forecasts of the main technological parameters of semiconductor technology are published.
Two of the most important parameters to quantify device scaling are the DRAM (dynamical random-access memory) half pitch and the MPU (microprocessor unit) half pitch, defined as half the spacing of two connecting metal lines. Another important parameter is the gate length of MOSFETs L_g, where a distinction between printed and physical gate length must be made. Table 2.1 shows the predictions of the 2001 edition of the ITRS [11] compared with the values of the 1999 and 1997 edition.

	DRAM 1/2 pitch			MPU 1/2 pitch		MPU Printed L_g			MPU Physical L_g
	2001	1999	1997	2001	1999	2001	1999	1997	2001
2001	130	150	150	150	180	90	100	120	65
2002	115	130		130	160	75	85		53
2003	100	120	130	107	145	65	80	100	45
2004	90	110		90	130	53	70		37
2005	80	100		80	115	45	65		32
2006	70		100	70		40		70	28
2007	65			65		35			25
2008		70			80		45		
2009			70					50	
2010	45			45		25			18
2011		50			55		30		
2012			50					35	
2013	32			32		18			13
2014		35			40		20		
2016	22			22		13			9

Table 2.1: Predictions of the ITRS 2001 compared with the predictions of 1997 and 1999. Values are in nm. In the ITRS 1999 and 1997 no predictions for the physical gate length, and in 1997, no MPU half pitch is given.

It can be seen that the predictions of each roadmap exceed the ones of the predecessor, an observation which has been called *roadmap acceleration*: While in 1997 the 70 nm DRAM half-pitch was predicted for the year 2009, it was predicted for 2008 in 1999, and the 2001 roadmap sees it in the year 2006.

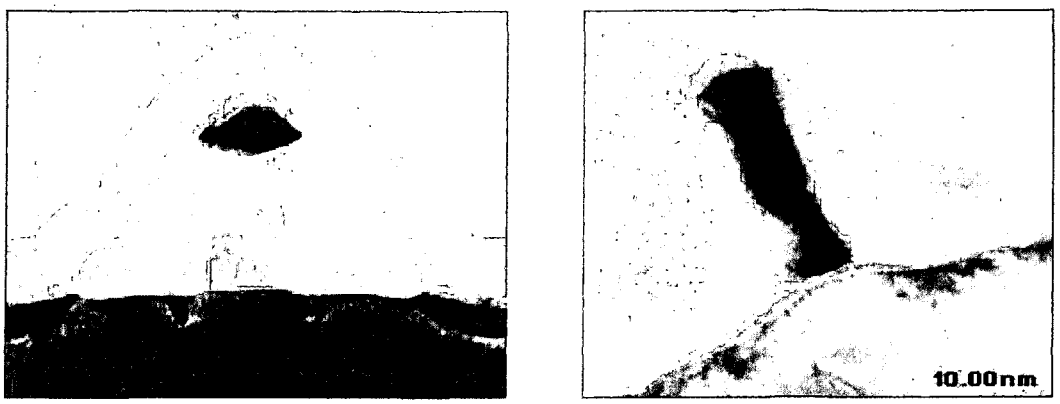


Figure 2.3: MOSFETs with 60 nm (left) and 10 nm (right) gate length [12, 13]. The gate dielectric thicknesses are 1.5 nm and 0.8 nm, respectively.

This continuous scaling has led to the development of transistors with gate lengths as small as 60 nm or even 10 nm in experimental devices, as shown in Fig. 2.3 [12, 13]. However, major obstacles arise when devices are scaled to such small dimensions.

2.2 Obstacles to Device Miniaturization

Several topics can be identified which represent severe handicaps to a further scaling of CMOS devices. Fig. 2.4 shows a cut through a typical CMOS inverter which consists of an nMOS and a pMOS device separated by shallow trench isolation (STI) [13, 14]. Crucial topics which must be taken into account to allow further device shrinkage are highlighted [15]. They will be briefly discussed in the following sections.

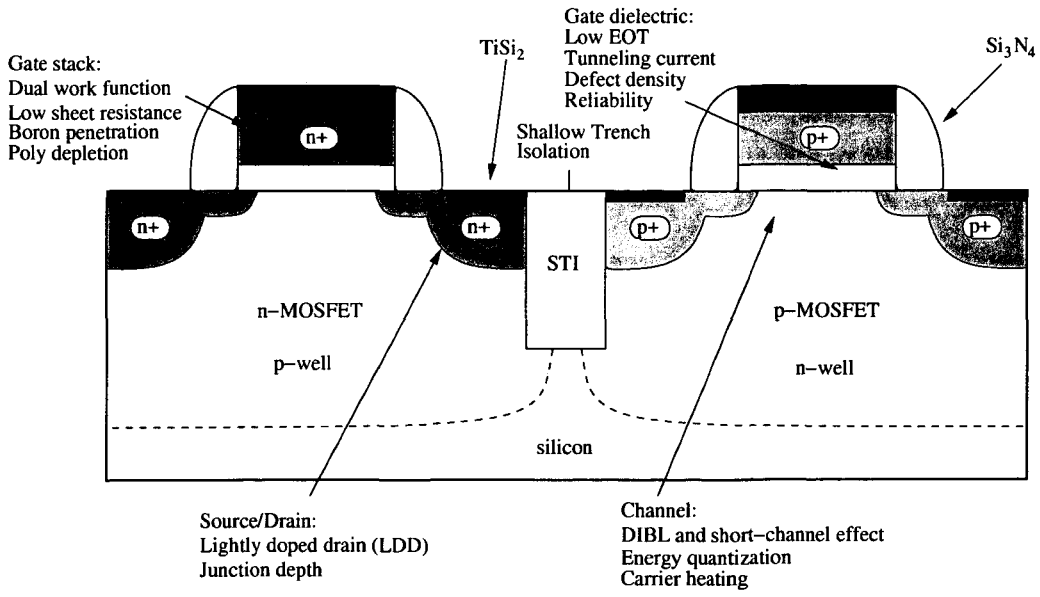


Figure 2.4: Important topics for further miniaturization of CMOS devices [15].

2.2.1 Channel

In the inversion layer of a MOSFET the strong band bending perpendicular to the channel leads to **energy quantization**. While the band edge energy along the channel varies only slightly, there is a strong gradient perpendicular to the channel. The inversion carriers are confined to a narrow quantum well beneath the gate dielectric which is called the two-dimensional electron gas. This is depicted in Fig. 2.5, where the carrier concentration in the channel is shown for a classical simulation with and without quantum correction.

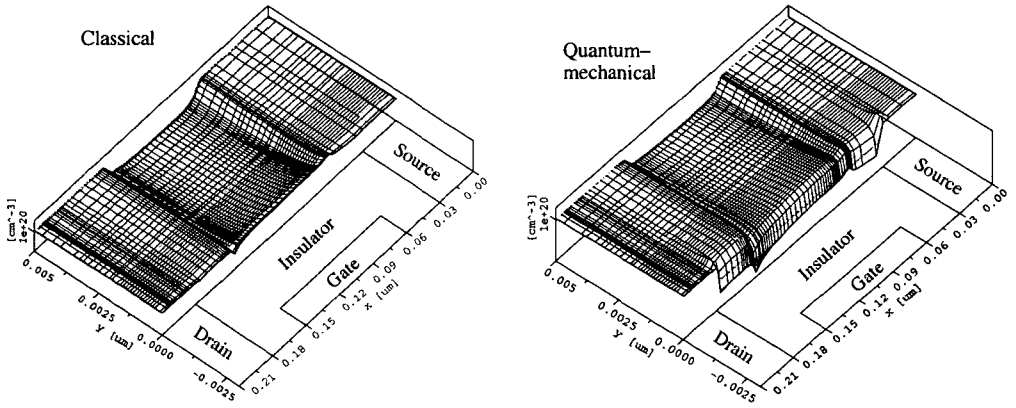


Figure 2.5: Carrier concentration without (left) and with (right) quantum correction. In the classical case the concentration peaks at the interface.

If it is assumed that the carrier wave function is blocked at the gate dielectric — that is, wave function penetration is neglected — discrete energy levels are formed [16]. The maximum of the electron concentration, the charge centroid, is not located at the interface to the gate dielectric, but forms inside the channel as shown in the left part of Fig. 2.6. This effect manifests as a reduced output current, as shown in the right part of Fig. 2.6, and can be modeled to some extent as a threshold voltage shift. Furthermore, the gate capacitance is reduced by this effect.

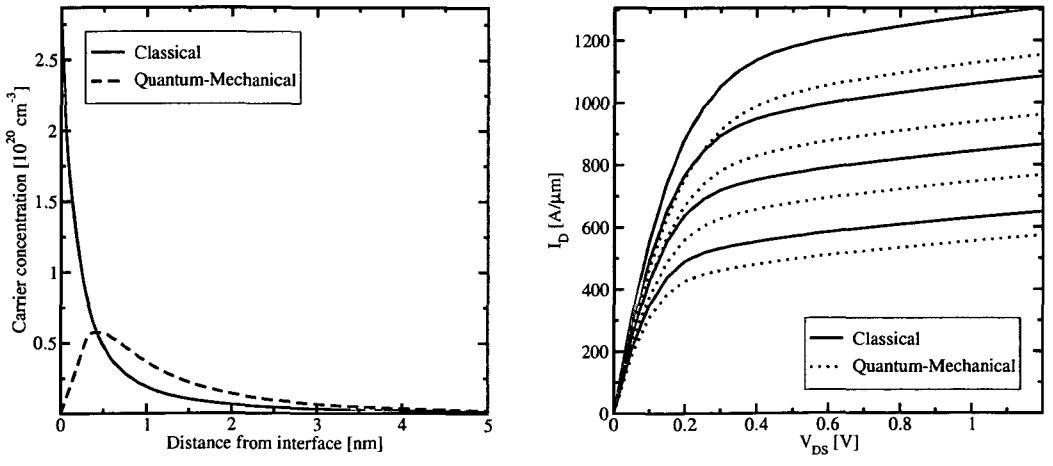


Figure 2.6: Carrier concentration in the channel (left) and output characteristics of a MOSFET (right) calculated with and without quantum correction.

Additional problems of device scaling are related to **hot-carrier effects**: When carriers in a turned-on MOSFET move from the source to the drain, they gain velocity and energy. Near the drain they have a high temperature which causes increased band-to-band tunneling, gate dielectric tunneling, and impact ionization (the phenomenon of hot-carrier tunneling will be reissued in Section 5.1.4.) The additional carriers created by these processes add to the substrate current, and thus to the leakage of the device. Furthermore, the hot-electron tunneling current leads to a degradation of the reliability of the gate dielectric.

Punchthrough poses a severe problem for miniaturized devices. It happens when a spurious path between source and drain of a turned-off MOSFET forms in the bulk region where the gate has no control over the charge. This results in a strongly increased leakage current. Fig. 2.7 shows the current density in a 90 nm turned-off MOSFET at V_{GS}=0.0 V, V_{DS}=1.2 V with a retrograde well (left) and without (right). Due to punchthrough, the current density in the right device is very high. It can be seen that the current does not flow through the channel but deeply in the substrate. Measures taken to reduce this effect are retrograde wells, halo implants, or pocket implants [17].

For devices with very short channels, an additional effect occurs which leads to increased leakage current. Due to the short distance between source and drain, the potential at the drain contact reduces the peak value of the energy barrier in the channel. This is shown in the left part of Fig. 2.8 for gate lengths of 250 nm down to 50 nm. It can be seen that the peak of the energy barrier near the source contact is strongly reduced, an effect which is called **drain-induced barrier lowering (DIBL)**. It leads to a decrease of the threshold voltage with reduced channel length. The resulting values of the threshold voltage for decreasing channel lengths, as shown in the right part of Fig. 2.8, give the so called 'roll-off' curve.

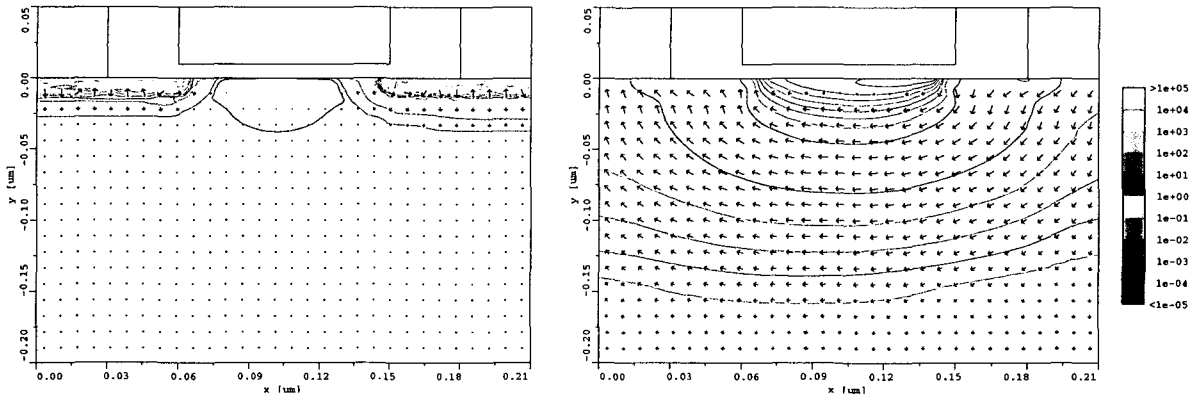


Figure 2.7: Current density in a 90 nm turned-off MOSFET without (left) and with a retrograde well implant (right). In the right device punchthrough leads to a high leakage current which mainly flows in the bulk region.

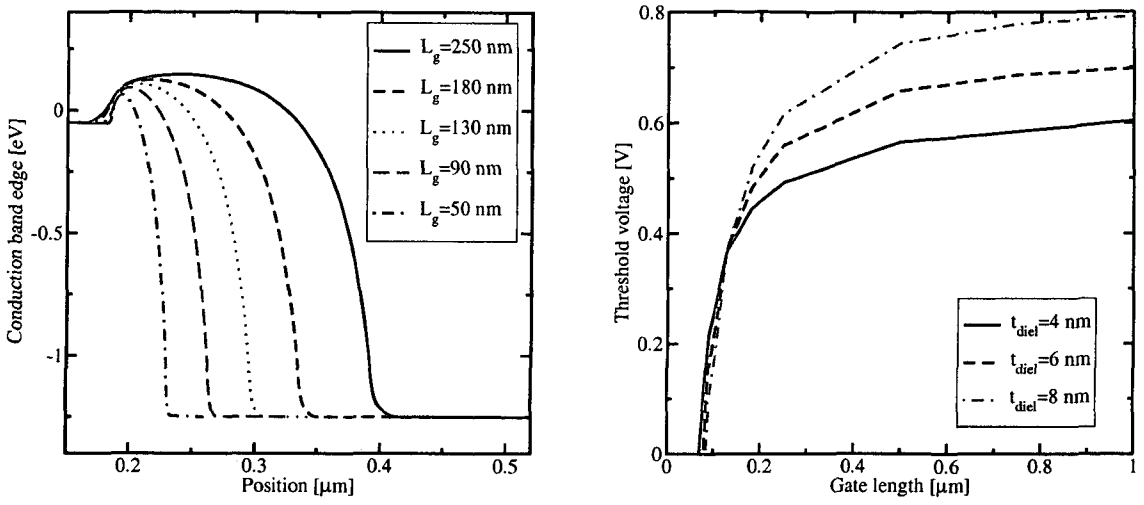


Figure 2.8: DIBL (left) and roll-off curve (right) for MOSFET devices with decreasing gate lengths and dielectric thicknesses at a drain bias of 1.2 V.

2.2.2 Gate Stack

For the realization of CMOS circuits it is necessary to integrate nMOS and pMOS devices closely together. Polysilicon gates allow an adjustment of the work function by doping and are thus ideally suited for large-scale integration [18], in contrast to metals where it is difficult to find materials with complementary work functions. However, if a voltage is applied on the polysilicon gate, a depletion layer forms at the interface to the gate dielectric. Within this layer a voltage drop occurs which is approximately given by [19]

$$V_{\text{poly}} \approx \frac{\kappa_{\text{diel}}^2 \cdot E_{\text{diel}}^2}{2q\kappa_{\text{si}} N_{\text{poly}}} ,$$

where κ_{diel} and κ_{si} denote the dielectric permittivity of the gate dielectric and the substrate, E_{diel} is the electric field in the dielectric, and N_{poly} the doping of the polysilicon. This effect is called **polysilicon depletion.** It leads to a reduced electron concentration at the interface and causes an effective increase of the dielectric thickness and an increase of the threshold voltage. The polysilicon depletion effect can be avoided by the use of metal gates such as nitrogen-doped molybdenum [20], which, however, is demanding from a process point of view.

Furthermore, polysilicon gates must be doped, and the material Boron is used as dopant for pMOS devices. However, during further process steps, the Boron tends to diffuse through the polysilicon gate and penetrate the dielectric layer and even the channel (**Boron penetration**) [21]. This causes a number of problems not only with the quality and reliability of the dielectric but especially with the device operation: Boron penetration increases the threshold voltage of MOS devices and degrades the MOSFET transconductance and its subthreshold slope.

2.2.3 Source and Drain

The doping profile of the source and drain region has an important impact on the device characteristics. On the one hand, it is desirable to have shallow junctions to reduce the influence of the drain on the channel and to improve the gate control over the inversion charge. On the other hand, a deep and heavily doped source and drain region reduces the series resistance. One possibility to achieve both is to introduce **lightly doped drain (LDD)** regions, where a deep implant is used at the contact and connected via a shallow implant to the channel. Another approach is to use **raised source/drain** contacts which are formed at a higher elevation than the channel [22].

2.2.4 Gate Dielectric

According to the scaling theory outlined in Section 2.1, the gate dielectric thickness must shrink with every new device generation, reaching values of 2.2 nm, 1.9 nm, and 1.4 nm for 180 nm, 150 nm, and 100 nm gate length devices [23]. However, the quantum-mechanical **tunneling** effect comes into play if the energy barrier between gate and semiconductor becomes to small. One remedy against this effect is to use dielectric materials which have a higher dielectric permittivity. These materials allow to achieve a high physical thickness together with a small effective oxide thickness (EOT). The EOT is defined as the thickness of a SiO_2 layer with equal capacitance. For a layer of SiO_2 and a high-κ dielectric, the EOT is

$$\mathrm{EOT} = t_{\mathrm{sio_2}} + t_{\mathrm{high}-\kappa} \cdot \frac{\kappa_{\mathrm{sio2}}}{\kappa_{\mathrm{high}-\kappa}} ,$$

where $t_{\mathrm{sio_2}}$ and $t_{\mathrm{high}-\kappa}$ denote the thickness of the SiO_2 and high-κ layer, and κ_{sio2} and $\kappa_{\mathrm{high}-\kappa}$ are the respective permittivities. With **high-κ dielectrics** it is possible to retain good control over the inversion charge even with physically thick dielectrics to block tunneling currents. This topic will be investigated in more detail in Section 5.1.5.

However, the **gate dielectric reliability** is a crucial issue, in particular with new materials. The parasitic tunneling current which flows through the dielectric gives rise to wear-out which means that the blocking capability of the dielectric is reduced, and even dielectric breakdown which is a sudden conductance increase. It is commonly assumed that this breakdown is caused by the gradual buildup of defects in the dielectric layer which may be caused by anode hole injection or the release of hydrogen from the Si-SiO_2 interface [24]. This is especially critical for high-κ dielectrics which do not form a native layer on silicon.

2.3 Novel Device Concepts

In addition to the ongoing scaling process, novel design concepts have arisen to enable a further increase in the integration density. These concepts span from strained-silicon MOS devices where the silicon channel is replaced by strained silicon to improve the mobility, to depleted-substrate devices such as single-gate or double-gate silicon on insulator (SOI) devices, FinFETs, vertical transistors, and even carbon nanotubes (CNTs) which represent a completely new device structure.

2.3.1 Strained-Silicon Devices

The mobility of carriers in silicon is enhanced if biaxial tensile strain is applied [25], because under tensile strain in (001) silicon, the fourfold-degenerate conduction band ellipsoids with the higher effective mass are lifted. Thus, more carriers remain in the two-fold degenerate ellipsoids with lower effective mass. Additionally, inter-valley scattering is reduced. Strained silicon channels can be realized by growing a thin layer of silicon on a material with a slightly larger lattice constant, such as silicon-germanium. The silicon layer must be thin enough to prevent relaxation and strain relief.

2.3.2 Depleted-Substrate Devices

As outlined above, punchthrough is one of the main problems in MOS devices. A straightforward countermeasure is to use devices where the substrate is partially or fully depleted [13]. Since there are no free carriers except in the channel, punchthrough cannot happen. Depleted-substrate devices can be realized using silicon on insulator substrates. The structure and conduction band edge of a fully-depleted **single-gate SOI** device is shown in the left part of Fig. 2.9. It consists of a standard MOSFET with a substrate that is insulated from the wafer by a layer of SiO_2. The gate can have an even better control over the inversion charge if a **double-gate SOI** transistor is considered, as shown in the right part of Fig. 2.9. Double- or even triple-gate MOSFETs can be achieved using a **FinFET**. This is a device where a small silicon channel — the fin — is surrounded at two or three sides by the gate electrode [25–27]. Fin thicknesses down to 6.5 nm have been reported [28] which means that the channel between source and drain consists of only about 15 atomic layers of silicon.

2.3.3 Vertical Transistors

MOSFETs which are used as access transistors in DRAM cells need a particularly small footprint to allow high integration densities [29]. The DRAM capacitor, which requires a capacitance of approximately 50 fF to allow practicable retention times, is usually built as a trench capacitor. One approach by which the footprint is reduced drastically is to turn the access transistor into the vertical direction directly above the trench capacitor [30–35].

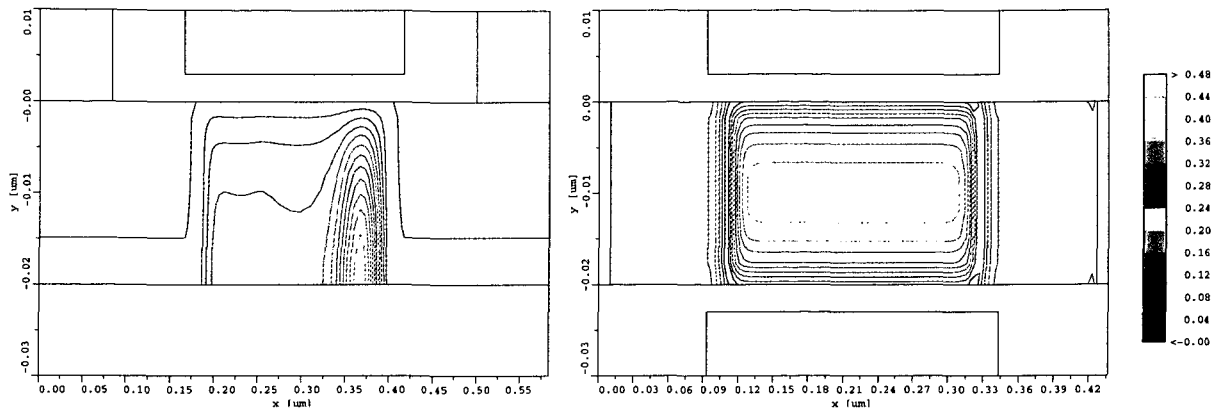


Figure 2.9: Conduction band edge in a fully-depleted turned-on single-gate (left) and double-gate SOI transistor (right).

2.3.4 Carbon Nanotube FET

Carbon nanotubes are cylindrical sheets of one ore more concentric layers of carbon atoms. Experiments have shown that the tubes can either have metallic or semiconducting properties. Their band structure depends on the position of the carbon atoms forming the tube. Particularly single-wall carbon nanotubes show superior electrical properties and are considered promising candidates for future nanoelectronic applications, either as interconnects or active devices. Semiconducting nanotubes can be used as active elements in field-effect transistor (FET) designs. Two possible applications of carbon nanotubes as transistor devices are shown in Fig. 2.10 [36, 37]. Single-wall carbon nanotubes are ballistic conductors, so the current is governed by LANDAUER's equation. This, however, implies that the minimum resistance of a metallic nanotube is $h/4q^2 \approx 6.5\,\mathrm{k}\Omega$. It is now generally accepted that the transport in the tubes is dominated by SCHOTTKY barriers at the metal contacts [38].

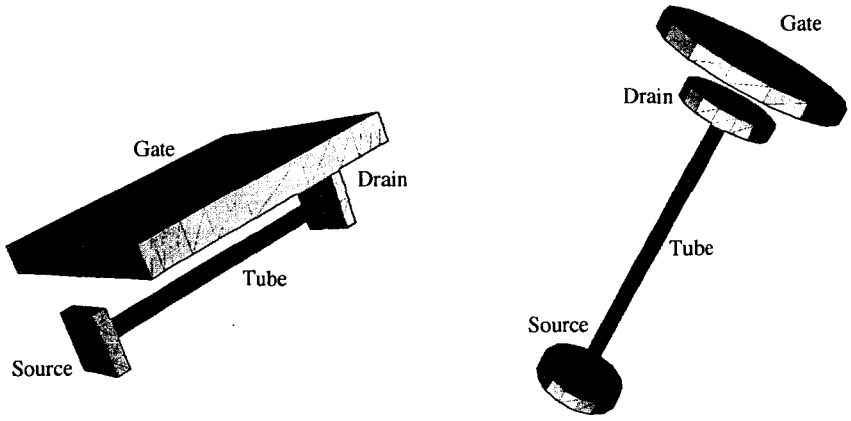


Figure 2.10: A lateral (left) and an axial (right) carbon nanotube FET.

2.4 Semiconductor Device Simulation

With the development of large scale integration in the late 1970's it became evident that the optimization of semiconductor manufacturing processes on a mere experimental basis is questionable. The numerical simulation of the fabrication process and the electrical characteristics of semiconductor devices offers a fast and inexpensive way to check device designs and processes. The tools for numerical simulation efforts can be separated into three categories (see Fig. 2.11): process simulation, device simulation, and circuit simulation. Process simulation is based on measurements such as doping profiles provided by SIMS (secondary ion mass spectroscopy), topography provided by TEM (transmission electron microscopy), the process recipe, and the lithography masks. Processes such as diffusion, oxidation, etching, lithography, and ion implantation are simulated. Device simulation uses the resulting device geometry and doping profile to reproduce and predict electrical data such as current-voltage (IV) curves, capacitance-voltage (CV) curves, or transfer frequencies. The output of device simulators can serve to calibrate compact models of circuit simulation programs. Integrated simulation packages can be used to perform these steps automatically. The abbreviation TCAD (technology computer-aided design) has been established to refer to process and device simulation approaches.

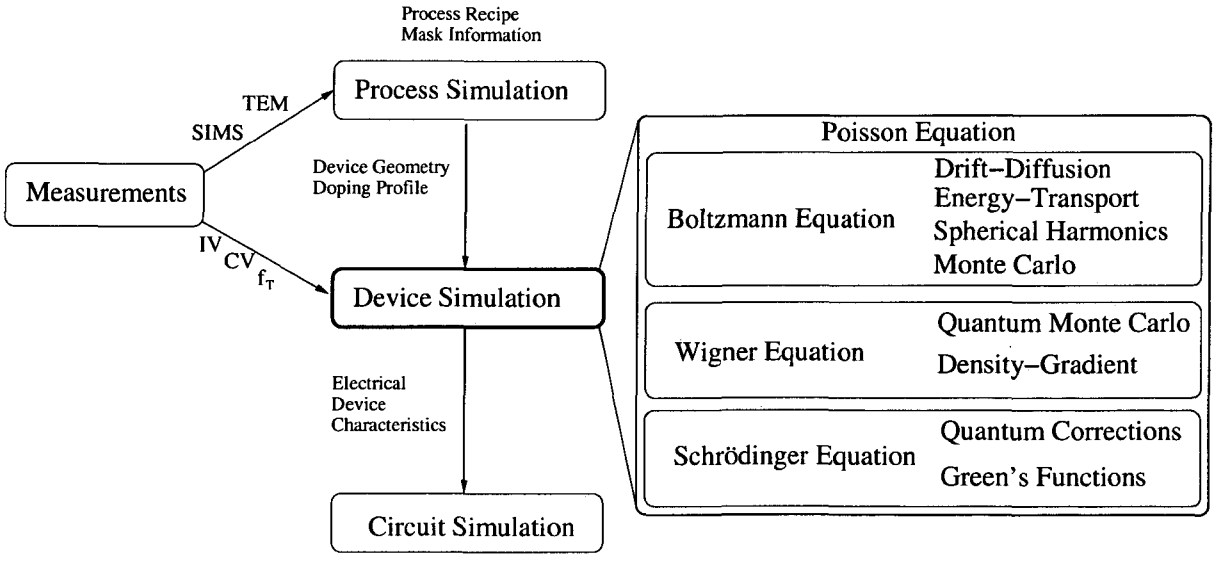


Figure 2.11: Hierarchy of process, device, and circuit simulation.

The simulation of semiconductor devices is either based on semi-classical or quantum-mechanical formulations. Based on fundamental equations — the POISSON[3], BOLTZMANN[4], WIGNER[5], or SCHRÖDINGER[6] equation — several models can be derived. They will be briefly described in the next sections.

[3] SIMÉON DENIS POISSON, French mathematician, 1781–1840.
[4] LUDWIG BOLTZMANN, Austrian physicist, 1844–1906.
[5] EUGENE PAUL WIGNER, Hungarian physicist, 1902–1995.
[6] ERWIN RUDOLF JOSEF ALEXANDER SCHRÖDINGER, Austrian physicist, 1887–1961.

2.4.1 Hierarchy of Semiconductor Device Simulation Models

Models of increasing sophistication can be derived for the simulation of charge transport in semiconductor devices, as shown in Fig. 2.11. The most important equation, which all models have in common, is POISSON's equation to determine the electrostatic potential

$$\nabla \cdot (\kappa \nabla \phi) = \mathrm{q}(n - p - C) \ , \tag{2.1}$$

where ϕ denotes the electrostatic potential, κ the dielectric permittivity, n and p the electron and hole concentration, and $C = N_\mathrm{D} - N_\mathrm{A}$ the net concentration of impurities. The transport of carriers is described by the BOLTZMANN transport equation (BTE) which is a semi-classical formulation of charge transport.

Quantum-mechanical effects are described by the SCHRÖDINGER equation. To incorporate quantum-mechanical effects into classical device simulation, BOLTZMANN's transport equation can be coupled to the SCHRÖDINGER equation, or the WIGNER equation can be applied [39–42]. Transport models based on solutions of the BOLTZMANN transport equation can be derived using the method of moments [43–45] which yields the drift-diffusion model [46], the energy-transport or hydrodynamic model [47], or higher-order transport models [48]. Furthermore, an approximate solution can be obtained by expressing the distribution function as a series expansion which leads to the spherical harmonics approach [49–53].

2.4.2 Classical Device Simulation

If the quantum-mechanical nature of electrons is neglected, carrier transport in a device can be described by BOLTZMANN's transport equation which is a seven-dimensional integro-differential equation in the phase space [46]. For electrons it reads

$$\frac{\partial f}{\partial t} + \mathbf{v} \cdot \nabla_r f - \frac{\mathrm{q}\mathbf{E}}{\hbar} \cdot \nabla_\mathbf{k} f = \mathcal{Q}(f) \ . \tag{2.2}$$

Here, $f(\mathbf{r}, \mathbf{k}, t)$ is the distribution of carriers in space ($\mathbf{r}$), momentum ($\hbar\mathbf{k}$), and time. On the right-hand side of this partial differential equation stands the collision operator $\mathcal{Q}(f)$ which describes scattering of particles due to phonons, impurities, interfaces, or other scattering sources. However, the direct solution of this equation is computationally prohibitive[7]. It is rather solved by approximate means applying the method of moments or using Monte Carlo methods. In the method of moments each term of (2.2) is multiplied with a weight function and integrated over $\mathbf{k}$-space. This yields a set of differential equations in the $(\mathbf{r}, t)$-space. The moments of the distribution function are defined by [54]

$$\langle \Phi \rangle = \frac{1}{4\pi^3} \int \Phi \, f(\mathbf{r}, \mathbf{k}, t) \, \mathrm{d}^3 k \ . \tag{2.3}$$

[7]Suppose we are only interested in the static case, we still have 6 solution variables. Considering a coarse mesh of only 100 grid points in each direction, this would require to solve an equation system with 10^{12} unknowns, which is beyond any computational feasibility.

2.4.2.1 The Drift-Diffusion Model

By multiplying (2.2) with the first two moments of the distribution function $\Phi_0 = 1$ and $\Phi_1 = \hbar\mathbf{k}$, integration over $\mathbf{k}$ space, using a parabolic dispersion relation, and applying the macroscopic relaxation-time approximation (RTA) for the integral of the collision operator, the following equation system can be derived

$$\nabla \cdot \mathbf{J_n} = \mathrm{q}R + \mathrm{q}\frac{\partial n}{\partial t} \,, \tag{2.4}$$

$$\nabla \cdot \mathbf{J_p} = -\mathrm{q}R - \mathrm{q}\frac{\partial p}{\partial t} \,, \tag{2.5}$$

$$\mathbf{J_n} = \mathrm{q}n\mu_n\mathbf{E} + \mathrm{q}D_n\nabla n \,, \tag{2.6}$$

$$\mathbf{J_p} = \mathrm{q}p\mu_p\mathbf{E} - \mathrm{q}D_p\nabla p \,. \tag{2.7}$$

In these equations $\mathbf{J}$ denotes the current density, R the net recombination rate, μ the mobility, $\mathbf{E}$ the electric field, and D the diffusion coefficient. Together with (2.1), these basic semiconductor equations form the drift-diffusion model which, due to its simplicity, is widely used for the simulation of semiconductor devices.

2.4.2.2 The Energy-Transport Model

Taking the first four moments of (2.2) into account yields the hydrodynamic model which, however, incorporates convective terms difficult to handle in a numerical simulator. If they are neglected, the following equation system can be derived (the expressions for holes are analogous and have been omitted)

$$\mathbf{J}_n = \mu_n \mathrm{k_B}\left(\nabla\left(nT_n\right) + \frac{\mathrm{q}}{\mathrm{k_B}}\mathbf{E}n\right) \,, \tag{2.8}$$

$$\mathbf{S}_n = -\frac{\tau_S}{\tau_m}\left(\frac{5\mathrm{k_B}^2}{2\mathrm{q}}\mu_n n T_n \nabla T_n + \frac{5\mathrm{k_B}^2}{2\mathrm{q}}T_n\mathbf{J}_n\right) \,, \tag{2.9}$$

$$\nabla \cdot \mathbf{J}_n = \mathrm{q}\left(R + \partial_t n\right) \,, \tag{2.10}$$

$$\nabla \cdot \mathbf{S}_n = -\frac{3}{2}\mathrm{k_B}\partial_t(nT_n) + \mathbf{E}\cdot\mathbf{J}_n - \frac{3}{2}\mathrm{k_B}n\frac{\mathrm{T}_n - T_L}{\tau_\mathcal{E}} + G_{\mathcal{E}n} \,. \tag{2.11}$$

This equation system is commonly known as energy-transport model. Here, $\mathbf{S}$ denotes the energy flux density, T_n the electron temperature, $\tau_\mathcal{E}$, τ_S, and τ_m the energy, energy flux, and momentum relaxation time, and $G_{\mathcal{E}n}$ the net energy generation rate.

2.4.2.3 Monte Carlo Device Simulation

In contrast to moment-based transport equations, the Monte Carlo method solves BOLTZMANN's transport equation by statistical means. It has been used extensively for the simulation of semiconductor devices [55–57]. Full-band Monte Carlo, which takes the correct shape of the band structure into account, is considered as the most rigorous method for the solution of BOLTZMANN's transport equation [58–62].

2.4.3 Quantum Device Simulation

The approaches described so far solve BOLTZMANN's transport equation, but do not take quantum effects into account. These effects can be incorporated by several methods: Coupling a SCHRÖDINGER-POISSON solver to BOLTZMANN's transport equation, solving WIGNER's equation, accounting for quantum effects in Monte Carlo simulations, or applying the non-equilibrium GREEN's function[8] formalism.

2.4.3.1 SCHRÖDINGER Equation and SCHRÖDINGER-POISSON Solvers

At the heart of quantum device simulation stands SCHRÖDINGER's equation [63]

$$-\frac{\hbar}{\imath}\frac{\partial\Psi(\mathbf{r},t)}{\partial t} = \underline{H}\Psi(\mathbf{r},t) \;, \tag{2.12}$$

where $\underline{H}$ denotes the HAMILTONian[9] of the system. For the stationary case, the SCHRÖDINGER equation can be written as [64]

$$\left(-\frac{\hbar^2}{2m}\nabla^2 + W(\mathbf{r})\right)\Psi(\mathbf{r}) = \mathcal{E}\Psi(\mathbf{r}) \;, \tag{2.13}$$

where $W(\mathbf{r})$ is an external potential energy. The central quantity is the wave function $\Psi(\mathbf{r})$. It is related to the probability $P_\mathcal{V}$ of finding an electron within a volume $\mathcal{V}$ by

$$P_\mathcal{V} = \int_\mathcal{V} \Psi(\mathbf{r})\Psi(\mathbf{r})^* \,\mathrm{d}\mathbf{r} = \int_\mathcal{V} |\Psi(\mathbf{r})^2| \,\mathrm{d}\mathbf{r} = \int_\mathcal{V} \rho(\mathbf{r}) \,\mathrm{d}\mathbf{r} \;, \tag{2.14}$$

where $\rho(\mathbf{r})$ denotes the probability density. The probability to find the electron *somewhere* must be unity, so

$$\int_{-\infty}^{\infty} \Psi(\mathbf{r})\Psi(\mathbf{r})^* \,\mathrm{d}\mathbf{r} = 1 \;. \tag{2.15}$$

From the wave function, the current density can be calculated via

$$\mathbf{J}(\mathbf{r}) = \frac{\imath\hbar\mathrm{q}}{2m}\left(\Psi\nabla\Psi^* - \Psi^*\nabla\Psi\right) \;, \tag{2.16}$$

which obeys the continuity equation

$$\nabla\cdot\mathbf{J}(\mathbf{r}) = -\mathrm{q}\frac{\partial\rho(\mathbf{r})}{\partial t} \;. \tag{2.17}$$

Common approaches to couple SCHRÖDINGER's equation to BOLTZMANN's transport equation perform a SCHRÖDINGER-POISSON self-consistent loop: The carrier concentration is calculated quantum-mechanically and used in POISSON's equation to obtain the electrostatic potential which is again used in SCHRÖDINGER's equation until convergence is reached. The resulting quantum-mechanical carrier concentration is used to derive correction factors for the solution of BOLTZMANN's transport equation [65, 66].

[8] GEORGE GREEN, British mathematician, 1793–1841.
[9] WILLIAM ROWAN HAMILTON, Irish Mathematician, 1805–1865.

2.4.3.2 WIGNER Equation and the Density-Gradient Model

The WIGNER function is defined as the FOURIER[10] transform of the product of wave functions at two points in space [67–69]

$$f(\mathbf{r},\mathbf{k},t) = \frac{1}{\pi^3}\int \Psi(\mathbf{r}-\mathbf{r}',t)\Psi^*(\mathbf{r}+\mathbf{r}',t)\exp(\imath 2\mathbf{r}'\mathbf{k})\,\mathrm{d}\mathbf{r}' \ . \tag{2.18}$$

Based on the WIGNER function, a transport equation — the BOLTZMANN-WIGNER equation — can be derived

$$\frac{\partial f}{\partial t} + \mathbf{v}\cdot\nabla_r f - \frac{\mathrm{q}}{\hbar}\sum_{\alpha=0}^{\infty}\frac{(-1)^{2\alpha}}{4^\alpha(2n+1)!}\nabla_r^{2n+1}V(\mathbf{r})\cdot\nabla_k^{2n+1}f = \left(\frac{\partial f}{\partial t}\right)_C \ , \tag{2.19}$$

where V denotes an external potential. Considering only the $\alpha = 0$ term yields the BOLTZMANN transport equation (2.2). If the $\alpha = 1$ term is also considered and a parabolic dispersion relation is assumed, the following transport equation, which is frequently referred to as the density-gradient model [70–76], is found

$$\frac{\partial f}{\partial t} + \frac{\hbar\cdot\mathbf{k}}{m}\nabla_r f - \frac{1}{\hbar}\nabla_r\left(V(\mathbf{r}) - \frac{\hbar^2}{12m}\nabla_r^2\ln(n)\right)\nabla_k f = \left(\frac{\partial f}{\partial t}\right)_C \ . \tag{2.20}$$

From this equation the quantum drift-diffusion or quantum hydrodynamic models can be derived by the method of moments. The quantum drift-diffusion model, for example, reads [77]

$$n = N_\mathrm{c}\exp\left(\frac{\mathcal{E}_\mathrm{f} - \mathcal{E}_\mathrm{c} - \Lambda}{\mathrm{k_B}T}\right) \ , \tag{2.21}$$

$$\mathbf{J}_n = -\mu_n \mathrm{k_B}T\nabla n - \mu_n n\nabla\left(\mathcal{E}_\mathrm{c} - \mathrm{k_B}T\ln N_\mathrm{c} + \Lambda\right) \ , \tag{2.22}$$

$$\Lambda = -\frac{\gamma\hbar^2}{12m_\mathrm{eff}}\left(\nabla^2\ln n + \frac{1}{2}(\nabla\ln n)^2\right) = -\frac{\gamma\hbar^2}{6m_\mathrm{eff}}\frac{\nabla^2\sqrt{n}}{\sqrt{n}} \ , \tag{2.23}$$

where the correction factors γ and Λ are used. Thus, the density-gradient model allows a local representation of quantum effects. It is therefore more suitable for the implementation in device simulators than a SCHRÖDINGER-POISSON solver which depends on non-local quantities, for example the thickness of a dielectric layer. The density-gradient method has been used by numerous authors [78–86]. However, it was reported that, while the carrier concentration in the inversion layer of a MOSFET can be modeled correctly, the method fails to reproduce tunneling currents as predicted by even more rigorous approaches [77].

2.4.3.3 Quantum Monte Carlo Device Simulation

Recently, strong efforts have been undertaken to couple the most accurate classical device simulation approach, the Monte Carlo technique, with quantum-mechanical formulations. These approaches are termed quantum Monte Carlo techniques [87–89].

One possibility is to use an effective potential instead of the solution of POISSON's equation [90, 91] in the Monte Carlo simulation. That can be achieved by convoluting the electrostatic potential with a GAUSS[11] function which leads to a smoothing of the original potential.

[10] JEAN BAPTISTE JOSEPH FOURIER, French mathematician, 1768–1830.
[11] JOHANN CARL FRIEDRICH GAUSS, German mathematician, 1777–1855.

A more rigorous approach is to solve the WIGNER transport equation (2.20) by means of Monte Carlo techniques. Unlike classical distribution functions, however, the WIGNER function (2.18) permits positive and negative values. Therefore, it cannot be interpreted as a probability distribution function, what is known as the *negative sign problem*. Instead, the WIGNER function can be modeled as the difference of two positive functions which describe in-scattering and out-scattering of particles [89]. This approach has the advantage that it allows for a seamless transition between classical and quantum-mechanical regions in a device.

2.4.3.4 Non-Equilibrium GREEN's Function Device Simulation

The non-equilibrium GREEN's function formalism (NEGF) provides a powerful means to handle open quantum systems. These are systems which are not confined but connected to reservoirs and have non-vanishing boundary conditions for the wave functions in SCHRÖDINGER's equation (2.13). The HAMILTONian of such a reservoir-coupled device can be written as

$$\begin{pmatrix} \underline{H} & \underline{C} \\ \underline{C}^{+} & \underline{H}_R \end{pmatrix} ,$$

where $\underline{H}$ and $\underline{H}_R$ denote the HAMILTONian of the device and the reservoir and $\underline{C}$ represents a coupling matrix. In real systems, the dimension of $\underline{H}_R$ is usually much larger than the dimension of $\underline{H}$. Note that $\underline{H}$ is not HERMITian[12], like in a closed system, and it therefore admits complex eigenvalues. The corresponding single-particle GREEN's function reads

$$\begin{pmatrix} \underline{G} & \underline{G}_{DR} \\ \underline{G}_{RD} & \underline{G}_R \end{pmatrix} = \begin{pmatrix} \mathcal{E}\underline{I} - \underline{H} & -\underline{C} \\ -\underline{C}^{+} & \mathcal{E}\underline{I} + \underline{H}_R \end{pmatrix}^{-1} , \tag{2.24}$$

where $\underline{G}_{DR}$ and $\underline{G}_{RD}$ refer to the coupling of the device to the reservoir, and $\underline{G}_R$ describes the reservoir itself. It can be shown that $\underline{G}$, the retarded GREEN's function, becomes

$$\underline{G} = (\mathcal{E}\underline{I} - \underline{H} - \underline{\Sigma})^{-1} , \tag{2.25}$$

where $\underline{\Sigma}$ denotes the self energy matrix which describes the interaction of the reservoir with the device [92–95]. This has the advantage that the reservoir, which may be of much larger dimensions than the device, only enters the problem via the self energy matrix which has the same dimension as the device HAMILTONian. From the retarded GREEN's function, the spectral function $\underline{A}$ can be derived

$$\underline{A}(\mathcal{E}) = \imath(\underline{G}(\mathcal{E}) - \underline{G}^{+}(\mathcal{E})) , \tag{2.26}$$

from which the carrier concentration in the device is calculated by

$$\underline{D} = \frac{m k_B T}{2\pi^2\hbar^2} \int \underline{A}(\mathcal{E}) \ln\left(1 + \exp\left(\frac{\mathcal{E}_f - \mathcal{E}}{k_B T}\right)\right) \mathrm{d}\mathcal{E} . \tag{2.27}$$

[12] CHARLES HERMITE, French mathematician, 1822–1901.

'It is quite wrong to try founding a theory on observable magnitudes alone... It is the theory which decides what we can observe.'

Albert Einstein

Chapter 3

Tunneling in Semiconductors

THIS CHAPTER outlines the theory of quantum-mechanical tunneling in semiconductor devices. Different tunneling mechanisms, such as direct-, FOWLER-NORDHEIM, and trap-assisted tunneling are covered. As a first step, the TSU-ESAKI model is derived. This model allows to distinguish between the supply function, which describes the supply of carriers for tunneling, and the transmission coefficient, which characterizes the penetrability of the considered energy barrier. The supply function depends on the energetic distribution of the carriers, an important quantity in semiconductor device modeling. Models which describe the shape of this distribution function are reviewed, namely the MAXWELLian[1], heated MAXWELLian, and non-MAXWELLian model.

The transmission coefficient can be found by a solution of SCHRÖDINGER's equation in the considered region. The WENTZEL-KRAMERS-BRILLOUIN- and GUNDLACH-methods, which are frequently encountered in the modeling of tunneling current, are shortly reviewed. However, for the proper simulation of transmission through arbitrary barriers, advanced models must be considered. Emphasis is put on the description of linear- and constant-potential transfer-matrix methods as well as on the quantum transmitting boundary method (QTBM).

The TSU-ESAKI tunneling formula finds the tunneling current density by an integration in the energy domain. In the channel of an inverted MOSFET, however, the strong electric field leads to the creation of bound and quasi-bound states. While bound states do not contribute to the tunneling process, tunneling from quasi-bound states can be understood using the concept of finite life times. Different numerical methods to calculate the life time of a quasi-bound state are reviewed.

The chapter continues with the description of trap-assisted tunneling and discusses some of the most frequently used models. Emphasis is put on the adaption of an inelastic trap-assisted tunneling model which incorporates energy loss by phonon emission and does not rely on the common assumptions of constant capture cross-sections.

Finally, a short summary and a comparison of the described methods is given.

[1] JAMES CLERK MAXWELL, British physicist, 1831–1879.

3.1 Tunneling Mechanisms

In the silicon-dielectric-silicon structure sketched in Fig. 3.1 a variety of tunneling processes can be identified. Considering the shape of the energy barrier alone, FOWLER-NORDHEIM (FN) tunneling and direct tunneling can be distinguished. However, a more rigorous classification distinguishes between ECB (electrons from the conduction band), EVB (electrons from the valence band), HVB (holes from the valence band), and TAT (trap-assisted tunneling) processes. The EVB process is caused by electrons tunneling from the valence band to the conduction band. It thus creates free carriers at both sides of the dielectric, which, for MOS transistors, gives rise to increased substrate current. The TAT process can either be elastic, which means that the energy of the carrier is conserved, or inelastic, where the carrier loses energy due to the emission of phonons. Furthermore, in dielectrics with a very high defect density, hopping conduction via multiple defects may occur.

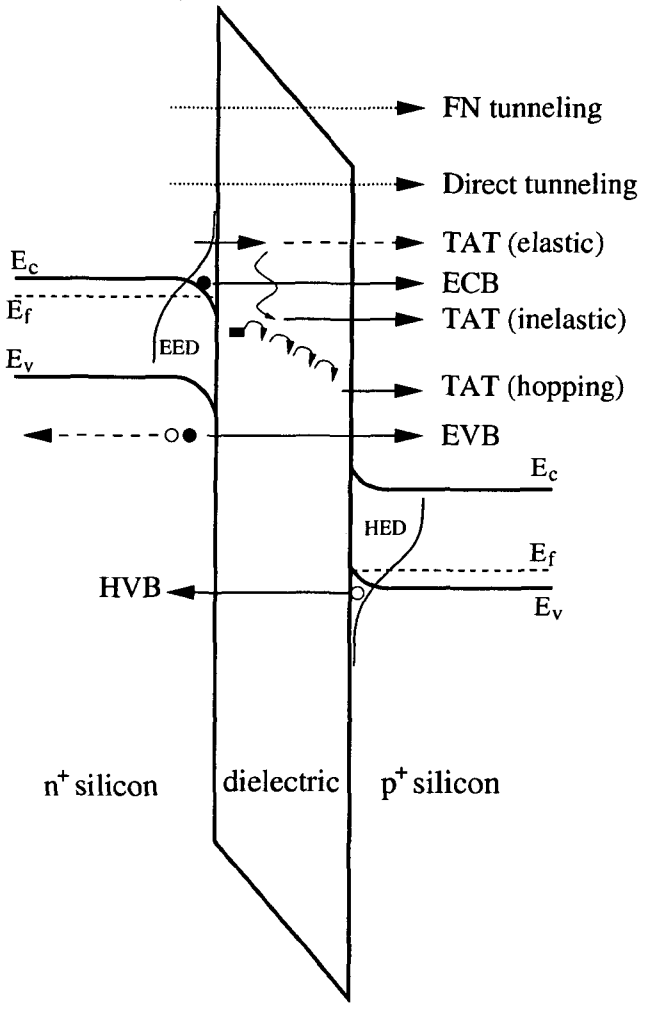


Figure 3.1: Schematic of the tunneling processes in a silicon-dielectric-silicon structure. The different tunneling processes are indicated by arrows and described in the text. The abbreviations EED and HED denote the electron and hole energy distribution function.

3.2 The TSU-ESAKI Model

The processes ECB and HVB shown in Fig. 3.1 can be investigated considering an energy barrier as shown in Fig. 3.2. Two semiconductor or metal regions are separated by an energy barrier with barrier height $q\Phi_B$, measured from the FERMI energy to the conduction band edge of the insulating layer. Electrons tunnel from Electrode 1 to Electrode 2. The distribution functions at both sides of the barrier are indicated in the figure.

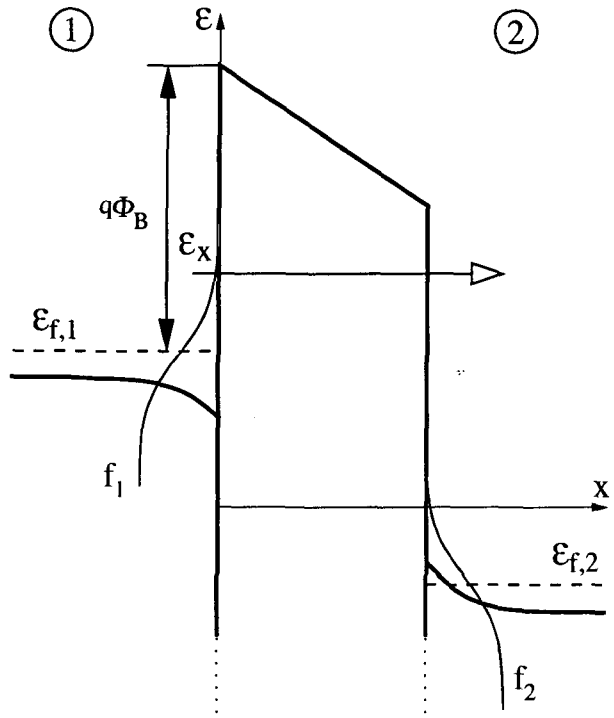


Figure 3.2: Energy barrier with two electrodes which can be used to describe the ECB or HVB processes.

In the following derivation some assumptions will be made which are necessary to allow an easy incorporation of the model in a device simulator. These are:

- Effective-mass approximation: The different masses corresponding to the band structure of the considered material are lumped into a single value for the effective mass. This is denoted by m_{eff} in the electrodes and m_{diel} in the dielectric layer.
- Parabolic bands: The dispersion relation in semiconductors is approximated by

$$\mathcal{E} = \frac{\hbar^2 \mathbf{k}^2}{2m_{\text{eff}}} = \frac{\hbar^2(k_x^2 + k_y^2 + k_z^2)}{2m_{\text{eff}}}, \tag{3.1}$$

  with the wave vector $\mathbf{k} = k_x\mathbf{e_x} + k_y\mathbf{e_y} + k_z\mathbf{e_z}$.
- Conservation of parallel momentum: Only transitions in the x-direction are considered, the parallel wave vector $\mathbf{k}_\rho = (k_y\mathbf{e_y} + k_z\mathbf{e_z})$ is not altered by the tunneling process.

The net tunneling current density from Electrode 1 to Electrode 2 can be written as the net difference between current flowing from Side 1 to Side 2 and *vice versa* [96, 97]

$$J = J_{1\to 2} - J_{2\to 1} \ . \tag{3.2}$$

The current density through the two interfaces depends on the perpendicular component of the wave vector k_x, the transmission coefficient TC, the perpendicular velocity v_x, the density of states g, and the distribution function at both sides of the barrier:

$$\begin{aligned} \mathrm{d}J_{1\to 2} &= \mathrm{q}TC(k_x)v_x g_1(k_x) f_1(\mathcal{E})(1 - f_2(\mathcal{E}))\,\mathrm{d}k_x \ , \\ \mathrm{d}J_{2\to 1} &= \mathrm{q}TC(k_x)v_x g_2(k_x) f_2(\mathcal{E})(1 - f_1(\mathcal{E}))\,\mathrm{d}k_x \ . \end{aligned} \tag{3.3}$$

In this expression it is assumed that the transmission coefficient only depends on the momentum perpendicular to the interface. The density of k_x states $g(k_x)$ is

$$g(k_x) = \int_0^\infty \int_0^\infty g(k_x, k_y, k_z)\,\mathrm{d}k_y\,\mathrm{d}k_z \ , \tag{3.4}$$

where $g(k_x, k_y, k_z)$ denotes the three-dimensional density of states in the momentum space. Considering the quantized wave vector components within a cube of side length L

$$\Delta k_x = \frac{2\pi}{L} \ , \quad \Delta k_y = \frac{2\pi}{L} \ , \quad \Delta k_z = \frac{2\pi}{L} \ , \tag{3.5}$$

yields for the density of states within the cube

$$g(k_x, k_y, k_z) = 2\frac{1}{\Delta k_x \Delta k_y \Delta k_z}\frac{1}{L^3} = \frac{1}{4\pi^3} \ , \tag{3.6}$$

where the factor 2 stems from spin degeneracy. For the parabolic dispersion relation (3.1) the velocity and energy components in tunneling direction obey

$$v_x = \frac{1}{\hbar}\frac{\partial \mathcal{E}}{\partial k_x} = \frac{\hbar k_x}{m_{\mathrm{eff}}} \ , \quad \mathcal{E}_x = \frac{\hbar^2 k_x^2}{2m_{\mathrm{eff}}} \ , \quad v_x \mathrm{d}k_x = \frac{1}{\hbar}\mathrm{d}\mathcal{E}_x \ . \tag{3.7}$$

Hence, expressions (3.3) become

$$\begin{aligned} \mathrm{d}J_{1\to 2} &= \frac{\mathrm{q}}{4\pi^3\hbar} TC(\mathcal{E}_x)\,\mathrm{d}\mathcal{E}_x \int_0^\infty \int_0^\infty f_1(\mathcal{E})\,(1 - f_2(\mathcal{E}))\ \mathrm{d}k_y\,\mathrm{d}k_z \ , \\ \mathrm{d}J_{2\to 1} &= \frac{\mathrm{q}}{4\pi^3\hbar} TC(\mathcal{E}_x)\,\mathrm{d}\mathcal{E}_x \int_0^\infty \int_0^\infty f_2(\mathcal{E})\,(1 - f_1(\mathcal{E}))\ \mathrm{d}k_y\,\mathrm{d}k_z \ . \end{aligned} \tag{3.8}$$

Using polar coordinates for the parallel wave vector components

$$\begin{aligned} k_\rho &= \sqrt{k_y^2 + k_z^2} \ , \qquad k_y = k_\rho \cos(\gamma) \ , \\ \gamma &= \arctan\left(\frac{k_z}{k_y}\right) \ , \quad k_z = k_\rho \sin(\gamma) \ , \end{aligned} \tag{3.9}$$

the current density evaluates to

$$J_{1\to 2} = \frac{4\pi m_{\text{eff}} q}{h^3} \int_{\mathcal{E}_{\min}}^{\mathcal{E}_{\max}} TC(\mathcal{E}_x)\,\mathrm{d}\mathcal{E}_x \int_0^{\infty} f_1(\mathcal{E})\,(1 - f_2(\mathcal{E}))\,\mathrm{d}\mathcal{E}_\rho\ ,$$
$$J_{2\to 1} = \frac{4\pi m_{\text{eff}} q}{h^3} \int_{\mathcal{E}_{\min}}^{\mathcal{E}_{\max}} TC(\mathcal{E}_x)\,\mathrm{d}\mathcal{E}_x \int_0^{\infty} f_2(\mathcal{E})\,(1 - f_1(\mathcal{E}))\,\mathrm{d}\mathcal{E}_\rho\ . \tag{3.10}$$

In these expressions the total energy $\mathcal{E}$ has been split into a longitudinal part $\mathcal{E}_\rho$ and a transversal part $\mathcal{E}_x$

$$\mathcal{E}_\rho = \frac{\hbar^2(k_y^2 + k_z^2)}{2m_{\text{eff}}} = \frac{\hbar^2 k_\rho^2}{2m_{\text{eff}}}\ ,\quad \mathcal{E}_x = \frac{\hbar^2 k_x^2}{2m_{\text{eff}}}\ . \tag{3.11}$$

Evaluating the difference $J = J_{1\to 2} - J_{2\to 1}$, the net current through the interface equals

$$J = \frac{4\pi m_{\text{eff}} q}{h^3} \int_{\mathcal{E}_{\min}}^{\mathcal{E}_{\max}} TC(\mathcal{E}_x)\,\mathrm{d}\mathcal{E}_x \int_0^{\infty} (f_1(\mathcal{E}) - f_2(\mathcal{E}))\,\mathrm{d}\mathcal{E}_\rho\ . \tag{3.12}$$

This expression is usually written as an integral over the product of two independent parts which only depend on the energy perpendicular to the interface: the transmission coefficient $TC(\mathcal{E}_x)$ and the supply function $N(\mathcal{E}_x)$:

$$J = \frac{4\pi m_{\text{eff}} q}{h^3} \int_{\mathcal{E}_{\min}}^{\mathcal{E}_{\max}} TC(\mathcal{E}_x) N(\mathcal{E}_x)\,\mathrm{d}\mathcal{E}_x\ , \tag{3.13}$$

which is the expression known as TSU-ESAKI formula. This model has been proposed by DUKE [98] and was used by TSU and ESAKI for the modeling of tunneling current in resonant tunneling devices [99]. The values of $\mathcal{E}_{\min}$ and $\mathcal{E}_{\max}$ depend on the considered tunneling process:

- Electrons tunneling from the conduction band (ECB): $\mathcal{E}_{\min}$ is the highest conduction band edge of the two electrodes, $\mathcal{E}_{\max}$ is the highest conduction band edge of the dielectric.
- Holes tunneling from the valence band (HVB): $\mathcal{E}_{\min}$ is the absolute value of the lowest valence band edge of the electrodes, $\mathcal{E}_{\max}$ is the absolute value of the lowest valence band edge of the dielectric. The sign of the integration must be changed.
- Electrons tunneling from the valence band (EVB): $\mathcal{E}_{\min}$ is the lowest conduction band edge of the two electrodes, $\mathcal{E}_{\max}$ the highest valence band edge of the two electrodes. It must be checked if $\mathcal{E}_{\min} < \mathcal{E}_{\max}$.

The next sections concentrate on the calculation of the supply function and the transmission coefficient.

3.3 Supply Function Modeling

The supply function describes the difference in the supply of carriers at the interfaces of the dielectric layer. Following (3.12), it is given as

$$N(\mathcal{E}_x) = \int_0^\infty (f_1(\mathcal{E}) - f_2(\mathcal{E}))\,\mathrm{d}\mathcal{E}_\rho\,, \tag{3.14}$$

where f_1 and f_2 denote the energy distribution functions near the interfaces. Since the exact shape of these distributions is usually not known, approximative shapes are commonly used. Furthermore it is assumed that the distributions are isotropic.

3.3.1 Fermi-Dirac Distribution

In equilibrium the energy distribution function of electrons or holes is given by the Fermi[2]-Dirac[3] statistics

$$f(\mathcal{E}) = \frac{1}{1+\exp\left(\dfrac{\mathcal{E}-\mathcal{E}_\mathrm{f}}{\mathrm{k_B}T}\right)}\,, \tag{3.15}$$

which can be derived from statistical thermodynamics [100]. Separating the longitudinal and transversal energy components $\mathcal{E} = \mathcal{E}_x + \mathcal{E}_\rho$ and splitting the integral in (3.14) $N(\mathcal{E}_x) = \xi_1(\mathcal{E}_x) - \xi_2(\mathcal{E}_x)$ the values of ξ_1 and ξ_2 become

$$\xi_i = \int_0^\infty f_i(\mathcal{E})\,\mathrm{d}\mathcal{E}_\rho = \int_0^\infty \frac{1}{1+\exp\left(\dfrac{\mathcal{E}_x+\mathcal{E}_\rho-\mathcal{E}_{\mathrm{f},i}}{\mathrm{k_B}T}\right)}\,\mathrm{d}\mathcal{E}_\rho \qquad i=1,2\,. \tag{3.16}$$

This expression can be integrated analytically using

$$\int \frac{\mathrm{d}x}{1+\exp(x)} = \ln\left(\frac{1}{1+\exp(-x)}\right) + C\,, \tag{3.17}$$

so expression (3.16) evaluates to

$$\xi_i = \mathrm{k_B}T\ln\left(1+\exp\left(-\frac{\mathcal{E}_x-\mathcal{E}_{\mathrm{f},i}}{\mathrm{k_B}T}\right)\right) \qquad i=1,2 \tag{3.18}$$

and the total supply function (3.14) becomes

$$N(\mathcal{E}_x) = \mathrm{k_B}T\ln\left(\frac{1+\exp\left(-\dfrac{\mathcal{E}_x-\mathcal{E}_{\mathrm{f},1}}{\mathrm{k_B}T}\right)}{1+\exp\left(-\dfrac{\mathcal{E}_x-\mathcal{E}_{\mathrm{f},2}}{\mathrm{k_B}T}\right)}\right)\,. \tag{3.19}$$

[2] Enrico Fermi, Italian physicist, 1901–1954.
[3] Paul Adrien Maurice Dirac, British physicist, 1902–1984.

3.3.2 MAXWELL-BOLTZMANN Distribution

For non-degenerate semiconductors the FERMI energy is located below the conduction band edge. Therefore, $\mathcal{E}_{\text{min}} - \mathcal{E}_{\text{f}} \gg \text{k}_\text{B}T$ holds in expression (3.13) and the FERMI-DIRAC distribution (3.15) can be approximated by a MAXWELL-BOLTZMANN (or MAXWELLian) distribution

$$f(\mathcal{E}) = \exp\left(\frac{\mathcal{E}_\text{f} - \mathcal{E}}{\text{k}_\text{B}T}\right) . \tag{3.20}$$

This expression is compared to the FERMI-DIRAC distribution in Fig. 3.3. It can be seen that only for energies well above the FERMI energy the expressions deliver equal results.

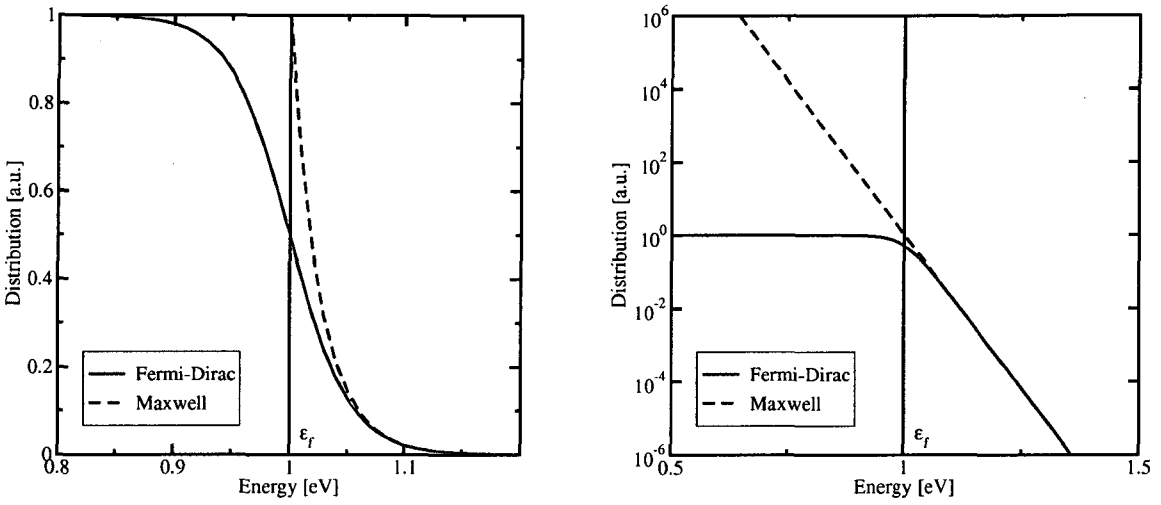


Figure 3.3: Comparison of the FERMI-DIRAC and the MAXWELL-BOLTZMANN distribution on a linear scale (left) and on a logarithmic scale (right). At energies above the FERMI energy the expressions yield similar results.

Using this expression, ξ in (3.14) becomes

$$\xi_i = \int_0^\infty f_i(\mathcal{E})\,\text{d}\mathcal{E}_\rho = \int_0^\infty \exp\left(-\frac{\mathcal{E}_x + \mathcal{E}_\rho - \mathcal{E}_{\text{f},i}}{\text{k}_\text{B}T}\right)\text{d}\mathcal{E}_\rho \qquad i = 1,2 \tag{3.21}$$

which evaluates to

$$\xi_i = \text{k}_\text{B}T \exp\left(-\frac{\mathcal{E}_x - \mathcal{E}_{\text{f},i}}{\text{k}_\text{B}T}\right) \qquad i = 1,2 \tag{3.22}$$

and yields a supply function of

$$N(\mathcal{E}_x) = \text{k}_\text{B}T\left(\exp\left(-\frac{\mathcal{E}_x - \mathcal{E}_{\text{f},1}}{\text{k}_\text{B}T}\right) - \exp\left(-\frac{\mathcal{E}_x - \mathcal{E}_{\text{f},2}}{\text{k}_\text{B}T}\right)\right) . \tag{3.23}$$

3.3.3 Non-MAXWELLian Distributions

The FERMI-DIRAC or MAXWELL-BOLTZMANN distribution functions are frequently used to describe the distribution of carriers in equilibrium since they are the solution of BOLTZMANN's transport equation for the case of vanishing applied electric field. In the channel region of a MOSFET, however, the energy distribution deviates from the ideal shape implied by expressions (3.15) or (3.20). Carriers gain energy by the electric field in the channel, and they experience scattering events. Models to describe the distribution function of such hot carriers have been studied by numerous authors [101–103]. One possibility to describe the distribution of hot carriers is to use a heated MAXWELLian distribution function

$$f(\mathcal{E}) = A \exp\left(-\frac{\mathcal{E}}{k_B T_n}\right) , \tag{3.24}$$

where T_n denotes the electron temperature and A is a normalization constant. The validity of this approach, however, is limited. Fig. 3.4 shows in the left part the contour lines of the heated MAXWELLian distribution function at the Si-SiO_2 interface in comparison to Monte Carlo results[4] for a MOSFET with a gate length of $L_g = 180\,nm$ and a thickness of the gate dielectric of 1.8 nm at a bias of $V_{DS} = V_{GS} = 1\,V$. It is evident that the heated MAXWELLian distribution (full lines) yields only poor agreement with the Monte Carlo results (dashed lines). The distribution function at two points near the middle of the channel (point A) and near the drain contact (point B) are shown in the right part of this figure. Particularly the high-energy tail in the middle of the channel is heavily overestimated by the heated MAXWELLian model. This is unsatisfactory since a correct description of the high energy tail is crucial for the evaluation of hot-carrier injection at the drain side used for programming and erasing of EEPROM devices.

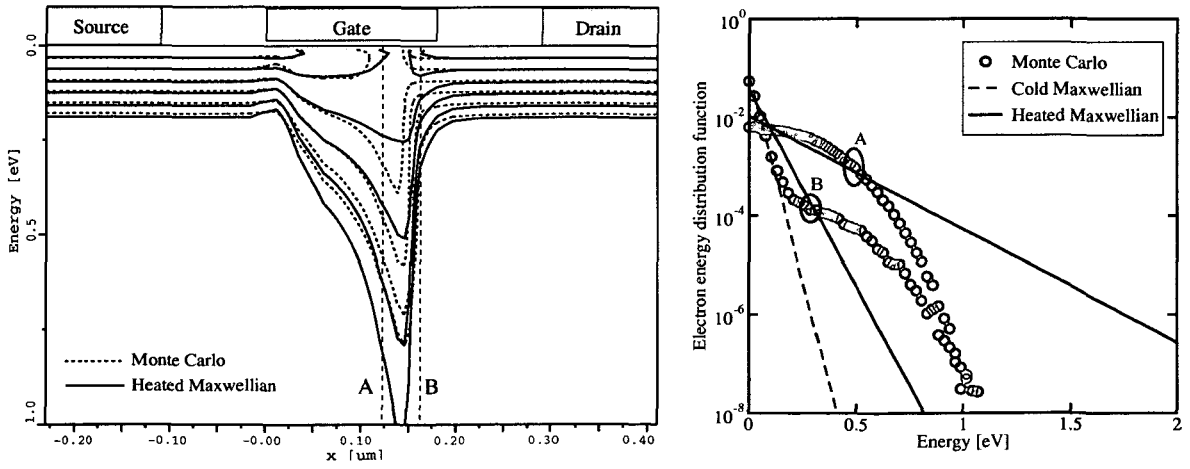


Figure 3.4: Comparison of the heated MAXWELLian distribution (full lines) with the results from a Monte Carlo simulation (dotted lines) in a turned-on 180 nm MOSFET. Neighboring lines differ by a factor of 10. The distributions at point A and B are compared with a cold MAXWELLian distribution in the right figure.

[4] A Monte Carlo simulator employing analytical non-parabolic bands was used for this simulation.

To obtain a better prediction of hot-carrier effects, CASSI and RICCÓ presented an expression to account for the non-MAXWELLian shape of the electron energy distribution function [101]

$$f(\mathcal{E}) = A \exp\left(-\frac{\chi \mathcal{E}^3}{E^{1.5}}\right) , \tag{3.25}$$

with χ as fitting parameter and E being the local electric field in the channel. This local-field dependence was soon questioned by other authors such as FIEGNA *et al.* [104] who replaced the electric field with an effective field calculated from the average electron energy to model the EEPROM writing process. HASNAT *et al.* used a similar form for the distribution function [105]

$$f(\mathcal{E}) = A \exp\left(-\frac{\mathcal{E}^\xi}{\eta(\mathrm{k_B} T_n)^\nu}\right) . \tag{3.26}$$

They obtained values of $\xi = 1.3$, $\eta = 0.265$, and $\nu = 0.75$ by fitting simulation results to measured gate currents. However, these values fail to describe the shape of the distribution function along the channel when compared to Monte Carlo results [106]. A quite generalized approach for the EED has been proposed by GRASSER *et al.*

$$f(\mathcal{E}) = A \exp\left(-\left(\frac{\mathcal{E}}{\mathcal{E}_{\mathrm{ref}}}\right)^b\right) . \tag{3.27}$$

In this expression the values of $\mathcal{E}_{\mathrm{ref}}$ and b are mapped to the solution variables T_n and β_n of a six moments transport model [107]. Expression (3.27) has been shown to appropriately reproduce Monte Carlo results in the source and the middle region of the channel of a turned-on MOSFET. However, this model is still not able to reproduce the high energy tail of the distribution function near the drain side of the channel because it does not account for the population of cold carriers coming from the drain. This was already visible in the right part Fig. 3.4 near the drain side of the channel: The distribution consists of a cold MAXWELLian, a high-energy tail, and a second cold MAXWELLian at higher energies. Expression (3.27) cannot reproduce the low-energy MAXWELLian. A distribution function accounting for the cold carrier population near the drain contact was proposed by SONODA *et al.* [103], and an improved model has been suggested by GRASSER *et al.* [106]:

$$f(\mathcal{E}) = A\left(\exp\left(-\left(\frac{\mathcal{E}}{\mathcal{E}_{\mathrm{ref}}}\right)^b\right) + c \exp\left(-\frac{\mathcal{E}}{\mathrm{k_B} T_{\mathrm{L}}}\right)\right) . \tag{3.28}$$

Here the pool of cold carriers in the drain region is correctly modeled by an additional cold MAXWELLian subpopulation. The values of $\mathcal{E}_{\mathrm{ref}}$, b, and c are again derived from the solution variables of a six moments transport model [106]. Fig. 3.5 shows again the results from Monte Carlo simulations in comparison to the analytical model. A good match between this non-MAXWELLian distribution and the Monte Carlo results can be seen.

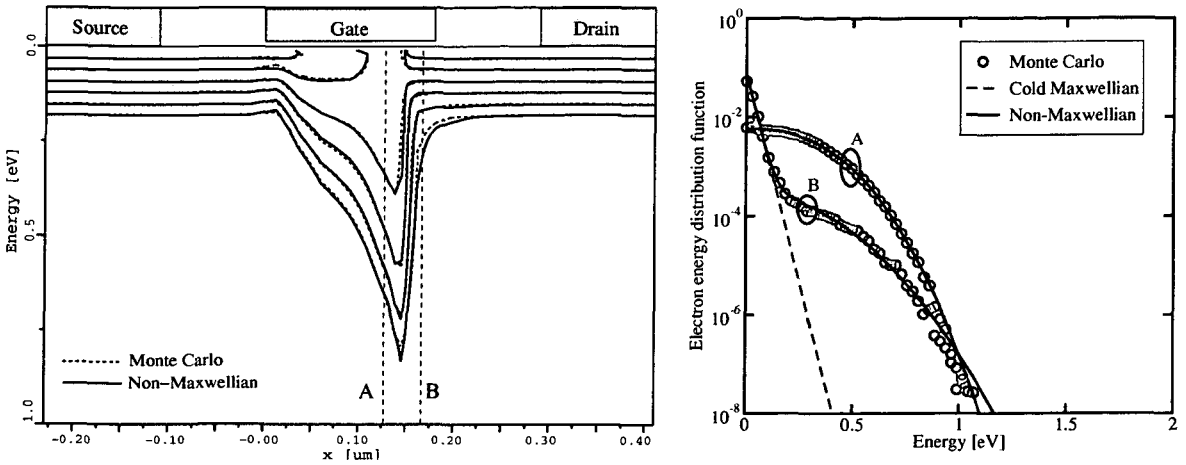


Figure 3.5: Comparison of the non-MAXWELLian distribution (full lines) with the results from a Monte Carlo simulation (dotted lines) in a turned-on 180 nm MOSFET. Neighboring lines differ by a factor of 10. The distributions at point A and B are compared with a cold MAXWELLian distribution in the right figure.

This model for the distribution function, however, requires to calculate the third even moment of the distribution function, the kurtosis β_n. As an approximation β_n can be calculated by an expression obtained for a bulk semiconductor where a fixed relationship between β_n, T_n, and the lattice temperature T_L exists:

$$\beta_\mathrm{Bulk}(T_n) = \frac{T_\mathrm{L}^2}{T_n^2} + 2\frac{\tau_\beta}{\tau_\mathcal{E}}\frac{\mu_S}{\mu_n}\left(1 - \frac{T_\mathrm{L}}{T_n}\right) . \tag{3.29}$$

In this expression $\tau_\mathcal{E}$, τ_β, μ_n, and μ_S are the energy relaxation time, the kurtosis relaxation time, the electron mobility, and the energy flux mobility, respectively. The value of $\tau_\beta\mu_S/\tau_\mathcal{E}\mu_n$ can be approximated by a fit to Monte Carlo data [106]. Estimating the kurtosis from (3.29), the distribution (3.27) can be used within the energy-transport or hydrodynamic model. For a parabolic band structure, the expressions

$$T_n = \frac{2}{3}\frac{\Gamma\left(\frac{5}{2b}\right)}{\Gamma\left(\frac{3}{2b}\right)}\frac{\mathcal{E}_\mathrm{ref}}{\mathrm{k_B}} , \tag{3.30}$$

$$\beta_n = \frac{3}{5}\frac{\Gamma\left(\frac{3}{2b}\right)\Gamma\left(\frac{7}{2b}\right)}{\Gamma\left(\frac{5}{2b}\right)^2} \tag{3.31}$$

are found [107], where $\Gamma(x)$ denotes the Gamma function

$$\Gamma(x) = \int_0^\infty \exp(-\alpha)\alpha^{x-1}\,\mathrm{d}\alpha . \tag{3.32}$$

While (3.30) can easily be inverted to obtain $\mathcal{E}_{\mathrm{ref}}(T_n)$, the inversion of (3.31) to find $b(T_n)$ at $\beta_n(b) = \beta_{\mathrm{Bulk}}(T_n)$ cannot be given in a closed form. Instead, a fit expression

$$b(T_n) = 1 + b_0 \left(1 - \frac{T_\mathrm{L}}{T_n}\right)^{b_1} + b_2 \left(1 - \frac{T_\mathrm{L}}{T_n}\right)^{b_3} \tag{3.33}$$

with the parameters b_0=38.82, b_1=101.11, b_2=3.40, and b_3=12.93 can be used. Using $\mathcal{E}_{\mathrm{ref}}(T_n)$ and $b(T_n)$ the Monte Carlo distribution can be approximated without knowledge of β_n. Fig. 3.6 shows simulation results for a 500 nm MOSFET using the heated MAXWELLian distribution (3.24), the non-MAXWELLian distribution (3.28), and the non-MAXWELLian distribution (3.27) using (3.30) and (3.33) to calculate the values of $\mathcal{E}_{\mathrm{ref}}$ and b. It can be seen that the fit to the results from Monte Carlo simulations is good. However, the emerging population of cold carriers near the drain end of the channel leads to a significant error in the shape of the distribution at low energy. This is important for certain processes, while in the case of tunneling the high-energy tail is more crucial.

With expression (3.27) for the distribution function and the assumption of a FERMI-DIRAC distribution in the polysilicon gate, the supply function (3.14) becomes

$$N(\mathcal{E}) = A_1 \frac{\mathcal{E}_{\mathrm{ref}}}{b} \Gamma_\mathrm{i}\left(\frac{1}{b}, \left(\frac{\mathcal{E}}{\mathcal{E}_{\mathrm{ref}}}\right)^b\right) - A_2 \mathrm{k_B} T_\mathrm{L} \ln\left(1 + \exp\left(-\frac{\mathcal{E} + \Delta\mathcal{E}_\mathrm{c}}{\mathrm{k_B} T_\mathrm{L}}\right)\right) , \tag{3.34}$$

where $\Gamma_\mathrm{i}(\alpha, \beta)$ denotes the incomplete gamma function

$$\Gamma_\mathrm{i}(x, y) = \int_y^\infty \exp(-\alpha)\alpha^{x-1}\,\mathrm{d}\alpha \; .$$

In (3.34) the explicit value of the FERMI energy was replaced by the shift of the two conduction band edges $\Delta\mathcal{E}_\mathrm{c}$. Assuming a MAXWELLian distribution in the polysilicon gate, the supply function can be further simplified to

$$N(\mathcal{E}) = A_1 \frac{\mathcal{E}_{\mathrm{ref}}}{b} \Gamma_\mathrm{i}\left(\frac{1}{b}, \left(\frac{\mathcal{E}}{\mathcal{E}_{\mathrm{ref}}}\right)^b\right) - A_2 \mathrm{k_B} T_\mathrm{L} \exp\left(-\frac{\mathcal{E} + \Delta\mathcal{E}_\mathrm{c}}{\mathrm{k_B} T_\mathrm{L}}\right) \; . \tag{3.35}$$

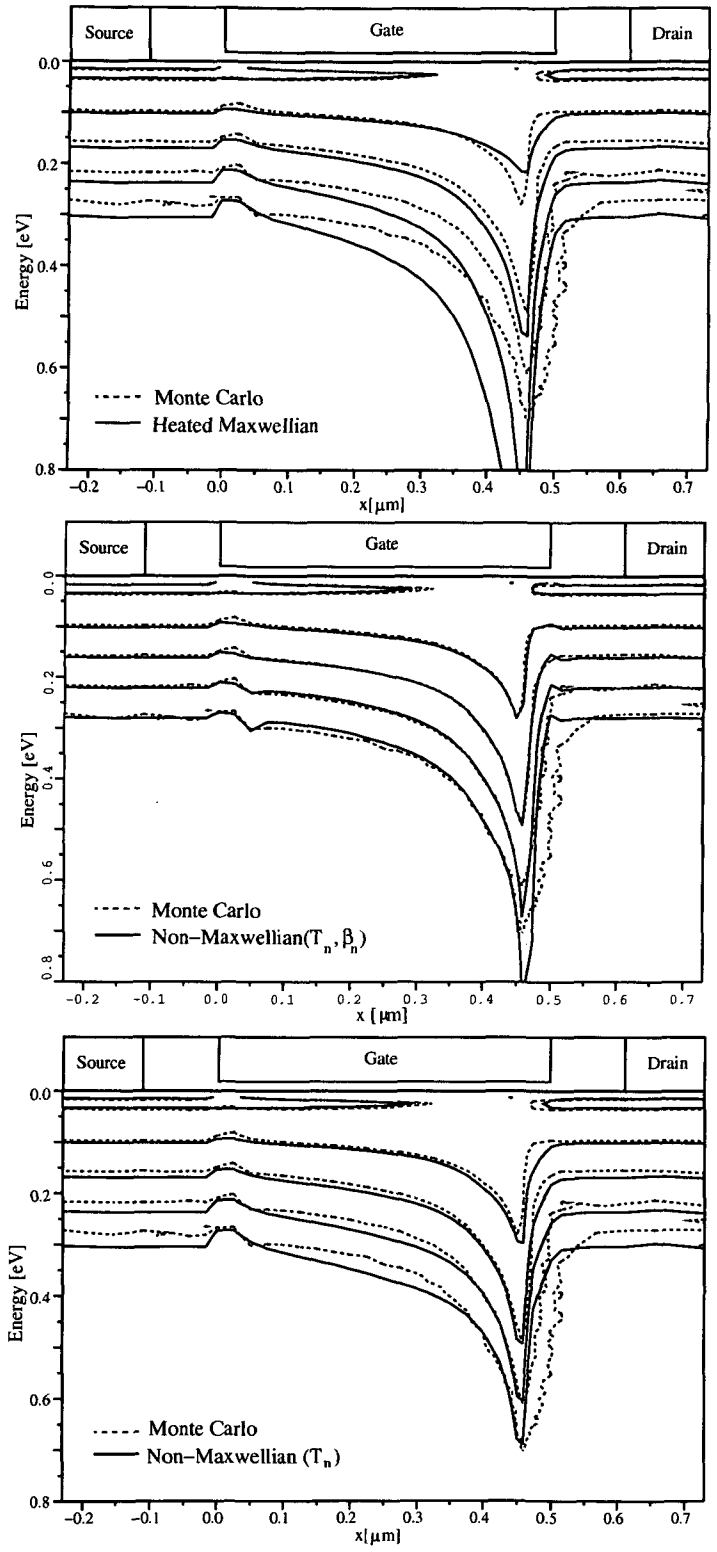
Using the accurate shape of the distribution (3.28), the expressions for the supply function become

$$\begin{aligned} N(\mathcal{E}) = {} & A_1 \frac{\mathcal{E}_{\mathrm{ref}}}{b} \Gamma_\mathrm{i}\left(\frac{1}{b}, \left(\frac{\mathcal{E}}{\mathcal{E}_{\mathrm{ref}}}\right)^b\right) + A_1 c \mathrm{k_B} T_2 \exp\left(-\frac{\mathcal{E}}{\mathrm{k_B} T_\mathrm{L}}\right) \\ & - A_2 \mathrm{k_B} T_\mathrm{L} \ln\left(1 + \exp\left(-\frac{\mathcal{E} + \Delta\mathcal{E}_\mathrm{c}}{\mathrm{k_B} T_\mathrm{L}}\right)\right) \end{aligned} \tag{3.36}$$

for a FERMI-DIRAC distribution, and

$$N(\mathcal{E}) = A_1 \frac{\mathcal{E}_{\mathrm{ref}}}{b} \Gamma_\mathrm{i}\left(\frac{1}{b}, \left(\frac{\mathcal{E}}{\mathcal{E}_{\mathrm{ref}}}\right)^b\right) + A_1 c \mathrm{k_B} T_2 \exp\left(-\frac{\mathcal{E}}{\mathrm{k_B} T_\mathrm{L}}\right) - A_2 \mathrm{k_B} T_\mathrm{L} \exp\left(-\frac{\mathcal{E} + \Delta\mathcal{E}_\mathrm{c}}{\mathrm{k_B} T_\mathrm{L}}\right) \tag{3.37}$$

assuming a MAXWELLian distribution in the polysilicon gate.



$$f(\mathcal{E}) = A \exp\left(-\frac{\mathcal{E}}{k_B T_n}\right)$$

$$f(\mathcal{E}) = A\left(\exp\left(-\left(\frac{\mathcal{E}}{\mathcal{E}_{ref}}\right)^b\right) + c\exp\left(-\frac{\mathcal{E}}{k_B T_L}\right)\right)$$

$\mathcal{E}_{ref}$, b, and c derived from n, T_n, and β_n.

$$f(\mathcal{E}) = A \exp\left(-\left(\frac{\mathcal{E}}{\mathcal{E}_{ref}}\right)^b\right)$$

$\mathcal{E}_{ref}$ and b derived from n and T_n.

Figure 3.6: Different expressions for the energy distribution function in a 500 nm MOSFET compared with Monte Carlo results.

3.3.4 Normalization

When implementing the analytical expressions for the distribution function and the supply function into a device simulator it is necessary to assure consistency: the carrier concentration defined by the analytical distribution function must match the carrier concentration from the transport model used. Therefore, the normalization prefactor A has to be evaluated from

$$n = \langle 1 \rangle = \frac{1}{4\pi^3} \int f(\mathbf{k})\, \mathrm{d}^3k \; . \tag{3.38}$$

This equation can be transformed to spherical coordinates using $k = (k_x^2 + k_y^2 + k_z^2)^{1/2}$

$$n = \frac{1}{4\pi^3} \int_{-\pi}^{\pi} \mathrm{d}\alpha \int_{0}^{\pi} \sin\theta \, \mathrm{d}\theta \int_{0}^{\infty} f(k) k^2 \, \mathrm{d}k \; . \tag{3.39}$$

For a parabolic dispersion relation we have $\mathrm{d}k = m_{\mathrm{eff}}/k\hbar^2 \mathrm{d}\mathcal{E}$ which finally leads to

$$n = \int_{0}^{\infty} f(\mathcal{E}) \frac{4\pi\sqrt{2m_{\mathrm{eff}}^3}}{h^3} \sqrt{\mathcal{E}} \, \mathrm{d}\mathcal{E} \; , \tag{3.40}$$

where the integration is performed from the conduction band edge $\mathcal{E}_{\mathrm{c}} = 0$. For a MAXWELLian or heated MAXWELLian distribution (expressions (3.20) or (3.24)), the normalization constant evaluates to

$$A = \frac{nh^3}{4\pi(\mathrm{k_B}T_\nu)^{3/2}\,\Gamma\left(\frac{3}{2}\right)\sqrt{2m_{\mathrm{eff}}^3}} \tag{3.41}$$

where T_ν is either the lattice temperature (for the assumption of a MAXWELLian distribution) or the carrier temperature (for the assumption of a heated MAXWELLian distribution). Using the non-MAXWELLian distribution (3.27) the normalization constant evaluates to

$$A = \frac{nh^3 b}{4\pi {\mathcal{E}_{\mathrm{ref}}}^{3/2}\,\Gamma\left(\frac{3}{2b}\right)\sqrt{2m_{\mathrm{eff}}^3}} \; , \tag{3.42}$$

while for expression (3.28) it is

$$A = \frac{nh^3}{4\pi\left(\frac{{\mathcal{E}_{\mathrm{ref}}}^{1/2}}{b}\Gamma\left(\frac{3}{2b}\right) + c(\mathrm{k_B}T_{\mathrm{L}})^{3/2}\Gamma\left(\frac{3}{2}\right)\right)\sqrt{2m_{\mathrm{eff}}^3}} \; . \tag{3.43}$$

3.4 The Energy Barrier

For the calculation of the transmission coefficient it is necessary to take the shape of the energy barrier into account. Electrons tunnel from a semiconductor or metal segment through a dielectric layer to another semiconductor or metal segment. Thus, the band diagram of a metal-oxide-semiconductor (MOS) capacitor has to be investigated. Furthermore, the image force, which leads to a reduction of both the electron and hole energy barrier for thin dielectrics, will be described in this section.

3.4.1 The Metal-Oxide-Semiconductor Capacitor

Fig. 3.7 shows the band diagram and the electrostatic potential in a metal-oxide-semiconductor structure for different voltages at the metal contact [108–110]. A central quantity is the work function which is defined as the energy required to extract an electron from the FERMI energy to the vacuum level. The work function of the semiconductor is

$$q\Phi_S = q\chi_S + \mathcal{E}_g - \mathcal{E}_i + \mathcal{E}_v + q\Phi_f \ , \tag{3.44}$$

where χ_s denotes the electron affinity of the semiconductor. The work function difference between the work function in the metal $q\Phi_M$ and the work function in the semiconductor $q\Phi_S$ is

$$q\Phi_{MS} = q\Phi_M - q\Phi_S \ . \tag{3.45}$$

The values of Φ_M and χ_S depend on the material, as shown in Table 3.1 [100, 111, 112]. However, the actual value of the work function of a metal deposited on SiO_2 is not exactly the same as that of the metal in vacuum [112].

As long as BOLTZMANN statistics can be applied, the FERMI potential Φ_f depends on the doping concentration of the semiconductor in the following way:

$$\mathrm{p-type:}\quad \Phi_f = \frac{k_B T}{q} \ln\left(\frac{N_A}{n_i}\right) > 0 \ , \tag{3.46}$$

$$\mathrm{n-type:}\quad \Phi_f = -\frac{k_B T}{q} \ln\left(\frac{N_D}{n_i}\right) < 0 \ . \tag{3.47}$$

The concentration-independent part of (3.45) is labeled Φ'_{MS}:

$$q\Phi'_{MS} = q\Phi_M - q\chi_S - \mathcal{E}_g + \mathcal{E}_i - \mathcal{E}_v \ . \tag{3.48}$$

The voltage which has to be applied to achieve flat bands is denoted the flatband voltage. If we deviate from this voltage, a space charge region forms near the interface between the dielectric and the semiconductor. The total potential drop across this space charge region is the surface potential ϕ_{surf}. Due to this potential all energy levels in the conduction and valence bands are shifted by a constant amount, therefore

$$\begin{aligned} \mathcal{E}_c(x) &= \mathcal{E}_{c,0} - q\phi(x) \ , \\ \mathcal{E}_v(x) &= \mathcal{E}_{v,0} - q\phi(x) \ , \end{aligned} \tag{3.49}$$

where $\mathcal{E}_{c,0}$ and $\mathcal{E}_{v,0}$ are the conduction and valence bands in the flatband case. Note that in the flatband case $\phi(x) = 0$ in the whole structure.

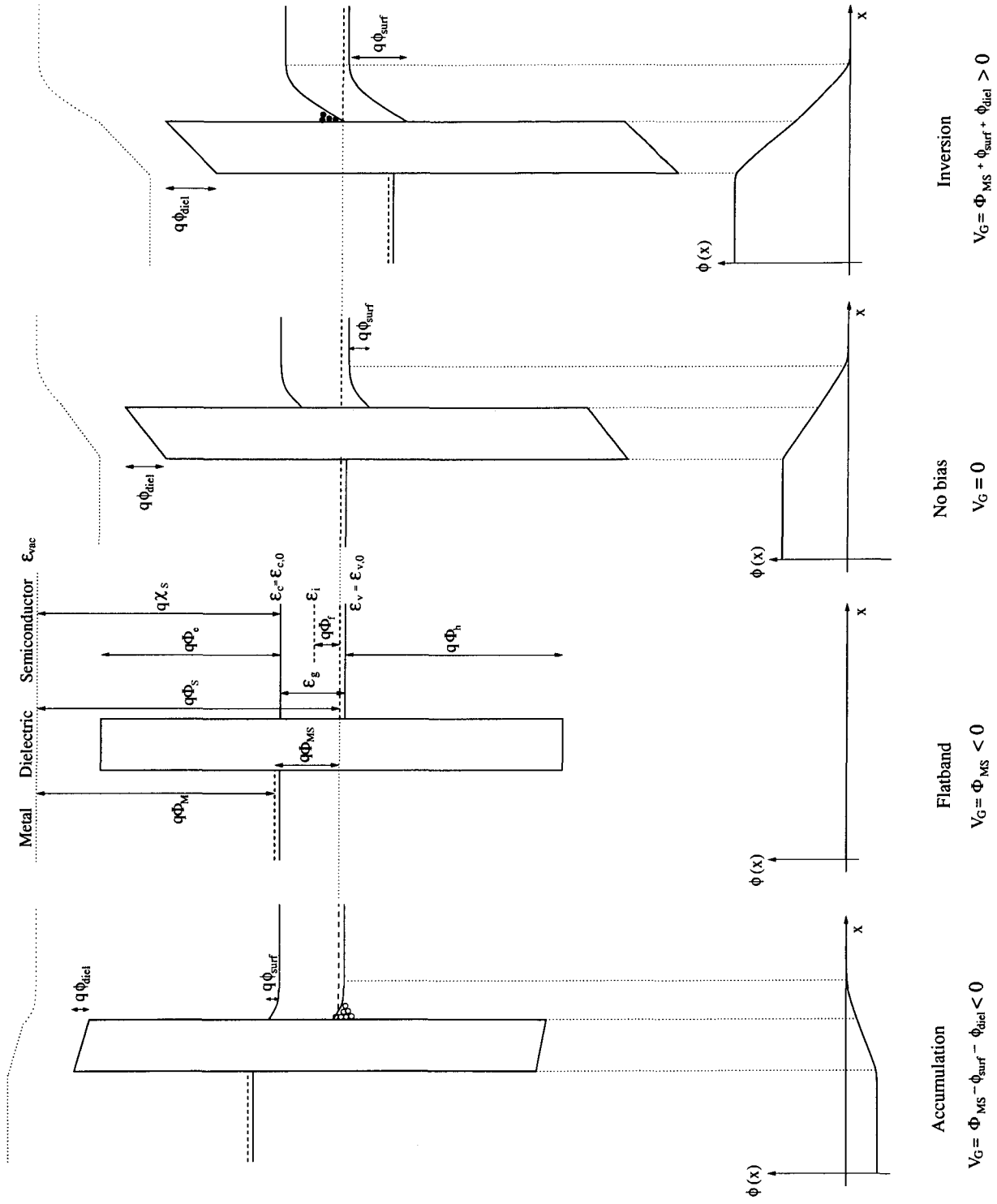


Figure 3.7: Band diagram and electrostatic potential in an nMOS structure (negative work function difference) in accumulation, under flatband condition, without bias, and under inversion condition.

Semiconductor	χ_S [V]	Metal	$q\Phi_M$ [eV]	k_f [nm^{-1}]
Si	4.05	Al	4.28	17.52
Ge	4.00	Pt	5.65	
GaAs	4.07	W	4.63	
GaP	3.80	Mg	3.66	13.74
GaSb	4.06	Ag	4.30	12.04
InAs	4.90	Au	4.80	12.06
InP	4.38	Cu	4.25	13.61
InSb	4.59	Cr	4.50	

Table 3.1: Electron affinity of various semiconductors (left), work function and the radius of the FERMI sphere of various metals (right) [113, 114].

In metals the FERMI energy is located at a higher energy level than the conduction band. The difference between the conduction band edge in the metal and the FERMI energy in the metal can be calculated considering the free-electron theory of metals which assumes that the metal electrons are unaffected by their metallic ions. The sphere of radius k_f (the FERMI wave vector) contains all occupied levels and determines the electron concentration

$$k_f = \sqrt[3]{3\pi^2 n} \; . \tag{3.50}$$

The values of the metal work function and k_f for various metals are summarized in the right part of Table 3.1 [114]. The value of $\mathcal{E}_f - \mathcal{E}_c$ can then directly be calculated from the carrier concentration assuming a parabolic dispersion relation and a MAXWELLian distribution function.

At the semiconductor side the height of the energy barrier is given by $q\Phi_e$ for electrons and $q\Phi_h$ for holes. Note that in the derivation of the TSU-ESAKI formula the barrier height $q\Phi_B$, which denotes the energetic difference between the FERMI energy and the band edge in the dielectric, is used. Depending on the considered tunneling process, $q\Phi_B$ must be calculated from $q\Phi_e$ or $q\Phi_h$.

3.4.2 Image Force Correction

When an electron approaches a dielectric layer, it induces a positive charge on the interface which acts like an image charge within the layer. This effect leads to a reduction of the barrier height for both electrons and holes [115–117]: The conduction band bends downward and the valence band bends upward, respectively. To account for this effect, the band edge energies (3.49) must be modified

$$\begin{aligned} \mathcal{E}_c(x) &= \mathcal{E}_{c,0} - q\phi(x) + \mathcal{E}_{image}(x) \; , \\ \mathcal{E}_v(x) &= \mathcal{E}_{v,0} - q\phi(x) + \mathcal{E}_{image}(x) \; , \end{aligned} \tag{3.51}$$

where the image force correction in the dielectric with thickness t_{diel} is calculated as [118]

$$\mathcal{E}_{image}(x) = -\frac{q^2}{16\pi\kappa_{diel}} \sum_j^\infty (k_1 k_2)^j \left(\frac{k_1}{|x| + jt_{diel}} + \frac{k_2}{(j+1)t_{diel} - |x|} + \frac{2k_1 k_2}{(j+1)t_{diel}} \right) \; , \tag{3.52}$$

where $x = 0$ is at the interface to the dielectric. The symbols k_1 and k_2 are calculated from the dielectric permittivities in the neighboring materials

$$k_1 = \frac{\kappa_{\text{diel}} - \kappa_{\text{si}}}{\kappa_{\text{diel}} + \kappa_{\text{si}}} , \quad k_2 = \frac{\kappa_{\text{diel}} - \kappa_{\text{metal}}}{\kappa_{\text{diel}} + \kappa_{\text{metal}}} = -1 . \tag{3.53}$$

Here, k_2 accounts for the interface between the insulator and the metal and evaluates to -1.

In the semiconductor the band edge energies are also altered

$$\mathcal{E}_{\text{image}}(x) = -\frac{\mathrm{q}^2}{16\pi\kappa_{\text{si}}} \sum_j^{\infty} (k_1 k_2)^j \left(\frac{-k_1}{|x| + jt_{\text{diel}}} + \frac{k_2}{(j+1)t_{\text{diel}} + |x|} \right) . \tag{3.54}$$

In practice it is sufficient to evaluate the sums in (3.52) and (3.54) up to $j = 11$ [119]. Fig. 3.8 shows the band edge energies in an MOS structure for a dielectric layer with a thickness of 2 nm and different dielectric permittivities for an applied bias of 0 V (left) and 2 V (right). A lower dielectric permittivity leads to a stronger band bending due to the image force and therefore strongly influences the transmission coefficient.

However, there is still some uncertainty if the image force has to be considered for tunneling calculations. While it is used in some works [119–122], others neglect it or report only minor influence on the results [123–127]. For rigorous investigations, however, its necessary to include it in the simulations. This, however, raises the need for a high spatial resolution along the dielectric. Simple models like the analytical WKB formula or the GUNDLACH formula are not valid for this case, as described in the following sections. It may therefore be justified to account for the image force barrier lowering by correction factors.

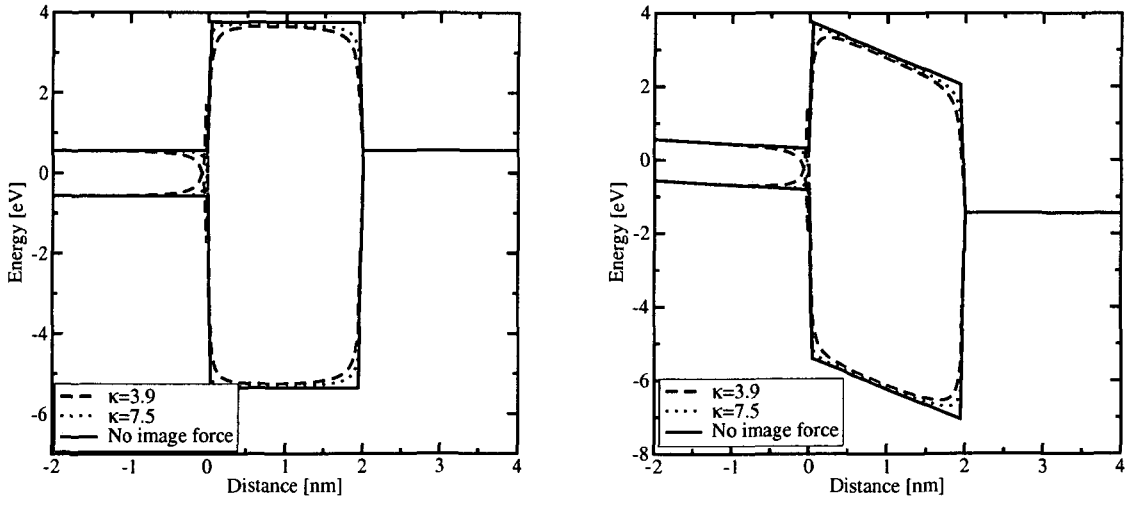


Figure 3.8: Effect of the image force in an nMOS device with a dielectric thickness of 2 nm at a gate bias of 0 V (left) and 2 V (right).

3.5 Transmission Coefficient Modeling

Now that the shape of the energy barrier has been treated, the calculation of the quantum-mechanical transmission coefficient of such a barrier can be investigated. The transmission coefficient TC is defined as the ratio of the quantum-mechanical current density (2.16) due to an incident wave in Region 1 and a transmitted wave in Region N, see Fig. 3.9. The assumption of plane waves in both regions[5]

$$\begin{aligned} \Psi_1(x) &= A_1 \exp(\imath k_1 x) \ , \\ \Psi_N(x) &= A_N \exp(\imath k_N x) \ , \end{aligned} \tag{3.55}$$

leads to the transmission coefficient

$$TC = \frac{J_N}{J_1} = \frac{k_1 m_1}{k_N m_N} \frac{|A_N|^2}{|A_1|^2} \ . \tag{3.56}$$

The wave function amplitudes A_1 and A_N can be found by solving the stationary SCHRÖDINGER equation (2.13) in the barrier region. This can be achieved by various methods. The WENTZEL-KRAMERS-BRILLOUIN approximation can be applied either analytically for a linear barrier, or numerically for arbitrary barriers. GUNDLACH's method can be used for a single linear energy barrier, while the transfer-matrix and quantum transmitting boundary methods are applicable for arbitrary-shaped barriers. The transfer-matrix method can be applied using either constant or linear potential segments as shown in Fig. 3.9. The different methods will be described in this section and a brief comparison at the end summarizes their advantages and shortcomings.

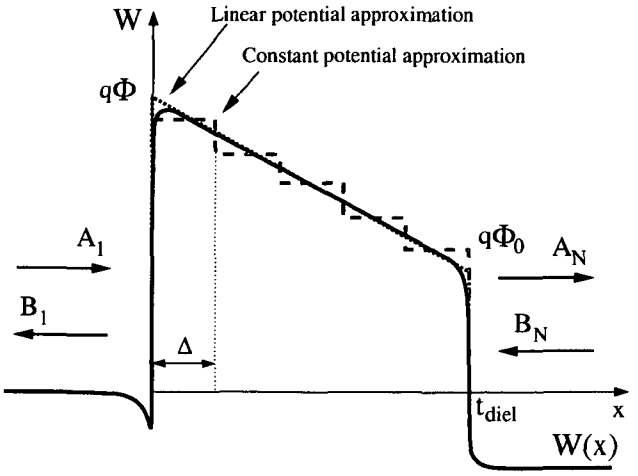


Figure 3.9: The energy barrier of a single-layer dielectric. The potential energy $W(x)$ may either be the conduction band or the valence band energy, depending on the tunneling process. The linear and constant potential approximations refer to the transfer-matrix method described in Section 3.5.3.

[5]In the stationary case, the quantum-mechanical current density (2.16) is, of course, equal in Region 1 and Region N, since the right hand side of (2.17) is zero. Considering only the incident wave in Region 1 and the transmitted wave in Region N allows to define a transmission coefficient $TC \leq 1$.

3.5.1 The WENTZEL-KRAMERS-BRILLOUIN Approximation

The WENTZEL-KRAMERS-BRILLOUIN[6] (WKB) approximation is one of the most frequently encountered assumptions for the quantum-mechanical wave function. It is often used for tunneling simulations and has been implemented in device simulators [96, 128, 129]. Within the WKB approximation, the transmission coefficient can be written as (for a detailed derivation see Appendix B) [130, 131]

$$TC(\mathcal{E}) = \exp\left(-\frac{2}{\hbar}\int_{x_1}^{x_2}\sqrt{2m_{\text{diel}}\left(W(x)-\mathcal{E}\right)}\,\mathrm{d}x\right) . \tag{3.57}$$

In this expression the integration is performed only within the classical turning points x_1 and x_2, defined by the region where $\mathcal{E} \leq W(x)$ and the integrand in (3.57) is real. Thus, only the decaying part of the wave function is considered. For a linear energy barrier the numerical calculation of the integral in (3.57) can be avoided. Still, it is necessary to distinguish between regions where direct or FOWLER-NORDHEIM tunneling takes place. For the direct tunneling regime $\mathcal{E} < \mathrm{q}\Phi_0$ holds (see Fig. 3.9). Therefore, the transmission coefficient

$$TC(\mathcal{E}) = \exp\left(-\frac{2}{\hbar}\int_{0}^{t_{\text{diel}}}\sqrt{2m_{\text{diel}}\left(\mathrm{q}\Phi-\mathrm{q}E_{\text{diel}}x-\mathcal{E}\right)}\,\mathrm{d}x\right) \tag{3.58}$$

evaluates to

$$TC(\mathcal{E}) = \exp\left(-4\frac{\sqrt{2t_{\text{diel}}}}{3\hbar\mathrm{q}E_{\text{diel}}}\left((\mathrm{q}\Phi-\mathcal{E})^{3/2}-(\mathrm{q}\Phi_0-\mathcal{E})^{3/2}\right)\right) , \tag{3.59}$$

with E_{diel} being the electric field defined as $V_{\text{diel}}/t_{\text{diel}}$ and m_{diel} the electron mass in the dielectric. The symbols Φ and Φ_0 denote the upper and lower barrier heights, as shown in Fig. 3.9. The value of Φ_0 is calculated assuming a linear potential in the barrier

$$\Phi_0 = \Phi - E_{\text{diel}}t_{\text{diel}} . \tag{3.60}$$

For the FOWLER-NORDHEIM tunneling regime it holds $\mathcal{E} > \mathrm{q}\Phi_0$ and therefore with x_1 defined by $\mathrm{q}\Phi - \mathrm{q}E_{\text{diel}}x_1 = \mathcal{E}$ the transmission coefficient

$$TC(\mathcal{E}) = \exp\left(-\frac{2}{\hbar}\int_{0}^{x_1}\sqrt{2m_{\text{diel}}\left(\mathrm{q}\Phi-\mathrm{q}E_{\text{diel}}x-\mathcal{E}\right)}\,\mathrm{d}x\right) , \tag{3.61}$$

evaluates to

$$TC(\mathcal{E}) = \exp\left(-4\frac{\sqrt{2m_{\text{diel}}}}{3\hbar\mathrm{q}E_{\text{diel}}}(\mathrm{q}\Phi-\mathcal{E})^{3/2}\right) . \tag{3.62}$$

The WKB tunneling coefficient is frequently multiplied by an oscillating prefactor to reproduce FOWLER-NORDHEIM-induced oscillations [132–136]. However, since no wave function interference is taken into account, the general validity of this method is questionable.

[6] MARCEL LOUIS BRILLOUIN, French physicist, 1854–1948.

3.5.2 The GUNDLACH Method

The GUNDLACH method [137] provides an analytical solution of SCHRÖDINGER's equation for a linear energy barrier. The one-dimensional time-independent SCHRÖDINGER equation in this case reads

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2}\Psi(x) + \frac{2m}{\hbar^2}\left(\mathcal{E} - W(x)\right)\Psi(x) = 0 \,, \tag{3.63}$$

with the linear potential energy $W(x)$ between the points x_0 and x_1, $W_0 = W(x_0)$, and $W_1 = W(x_1)$,

$$W(x) = W_0 + (x - x_0)\frac{W_1 - W_0}{x_1 - x_0} \tag{3.64}$$

for $x_0 < x < x_1$. Using the abbreviations

$$\begin{aligned} l &= -\left(\frac{\hbar^2}{2m}\frac{x_1 - x_0}{W_1 - W_0}\right)^{1/3} \,, \\ \lambda &= -\left(\frac{2m}{\hbar^2}\right)^{1/3}\left(\frac{x_1 - x_0}{W_1 - W_0}\right)^{2/3}\left(\mathcal{E} - W_0 + x_0\frac{W_1 - W_0}{x_1 - x_0}\right) \,, \end{aligned} \tag{3.65}$$

and $u(x) = \lambda - x/l$, expression (3.63) turns into

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2}\Psi(x) - \frac{1}{l^2}u(x)\Psi(x) = 0 \,. \tag{3.66}$$

With

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2}\Psi(x) = \frac{\mathrm{d}}{\mathrm{d}u}\frac{\mathrm{d}u}{\mathrm{d}x}\left(\frac{\mathrm{d}}{\mathrm{d}u}\frac{\mathrm{d}u}{\mathrm{d}x}\Psi\left(u(x)\right)\right) = \frac{1}{l^2}\frac{\mathrm{d}^2}{\mathrm{d}u^2}\Psi\left(u(x)\right) \tag{3.67}$$

SCHRÖDINGER's equation evolves into the AIRY[7] differential equation

$$\frac{\mathrm{d}^2}{\mathrm{d}u^2}\Psi\left(u(x)\right) - u(x)\Psi\left(u(x)\right) = 0 \,. \tag{3.68}$$

The solutions of this differential equation are the AIRY functions $\mathrm{Ai}\left(u(x)\right)$ and $\mathrm{Bi}\left(u(x)\right)$ [138], which are depicted in Fig. 3.10 together with their derivatives. The wave functions consist of linear superpositions of these AIRY functions

$$\Psi(x) = A\mathrm{Ai}\left(u(x)\right) + B\mathrm{Bi}\left(u(x)\right) \,, \tag{3.69}$$

where the function $u(x)$ is given as

$$u(x) = -\left(\frac{2m}{\hbar^2}\right)^{1/3}\left(\frac{x_1 - x_0}{W_1 - W_0}\right)^{2/3}\left(\mathcal{E} - W(x)\right) \,. \tag{3.70}$$

Assuming a constant electron mass in the dielectric, GUNDLACH derives an expression for the transmission coefficient

$$TC = \frac{k_n}{k_1}\frac{4}{\pi^2}\left(\left(\frac{z'}{k_1}A + \frac{k_n}{z'}B\right)^2 + \left(\frac{k_n}{k_1}C + D\right)^2\right)^{-1} \,, \tag{3.71}$$

[7] GEORGE BIDDELL AIRY, British mathematician, 1801–1892.

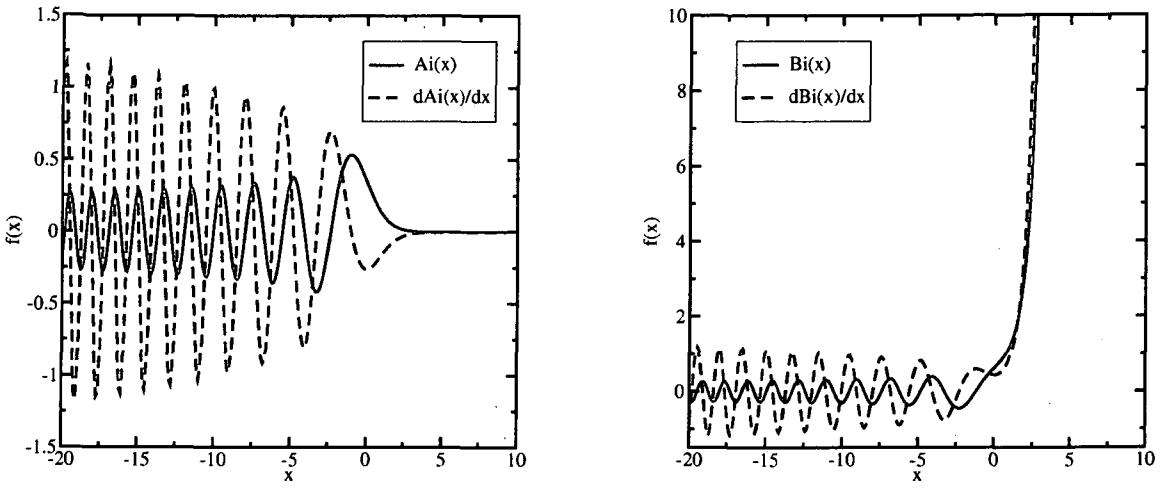


Figure 3.10: The AIRY functions Ai and Bi and their derivatives.

where the abbreviations

$$A = \mathrm{Ai}'(z_0)\mathrm{Bi}'(z_s) - \mathrm{Ai}'(z_s)\mathrm{Bi}'(z_0)\ , \tag{3.72}$$

$$B = \mathrm{Ai}(z_0)\mathrm{Bi}(z_s) - \mathrm{Ai}(z_s)\mathrm{Bi}(z_0)\ , \tag{3.73}$$

$$C = \mathrm{Ai}(z_s)\mathrm{Bi}'(z_0) - \mathrm{Ai}'(z_0)\mathrm{Bi}(z_s)\ , \tag{3.74}$$

$$D = \mathrm{Ai}(z_0)\mathrm{Bi}'(z_s) - \mathrm{Ai}'(z_s)\mathrm{Bi}(z_0)\ , \tag{3.75}$$

have been used and the symbols z_0, z_s, and z' are given by

$$z_0 = (\mathrm{q}\Phi_0 - \mathcal{E})\left(\frac{at_{\mathrm{diel}}}{2\mathrm{q}(\Phi - \Phi_0)}\right)^{2/3}\ ,\quad z_s = (\mathrm{q}\Phi - \mathcal{E})\left(\frac{at_{\mathrm{diel}}}{2\mathrm{q}(\Phi - \Phi_0)}\right)^{2/3}\ , \tag{3.76}$$

and

$$z' = -\left(\frac{a^2}{4}\frac{\mathrm{q}\Phi - \mathrm{q}\Phi_0}{t_{\mathrm{diel}}}\right)^{1/3}\ ,\quad a = \frac{2}{\hbar}\sqrt{2m_{\mathrm{diel}}}\ . \tag{3.77}$$

The symbols $\mathrm{q}\Phi$ and $\mathrm{q}\Phi_0$ denote the two edges of the energy barrier as shown in Fig. 3.9. The GUNDLACH method is frequently used in the literature [121, 139] and implemented in device simulators. Numerical problems may occur for flat barriers ($\Phi \approx \Phi_0$) due to the exponential increase of the AIRY functions Bi and Bi′ for positive arguments. In practical implementations the values of z_0 and z_s have been bounded to values below ≈ 200 to avoid floating point overflow.

3.5.3 Transfer-Matrix Method

The use of the transfer-matrix (TM) method for the calculation of the transmission coefficient of energy barriers is based on the work of TSU and ESAKI on electron tunneling through one-dimensional super lattices [99]. It has been used by numerous authors to describe tunneling processes in semiconductor devices [140–144]. The basic principle of the transfer-matrix method is the approximation of an arbitrary-shaped energy barrier by a series of piece-wise constant or piece-wise linear functions. Since the wave function in such barriers can easily be calculated, the total transfer matrix can be derived by a number of subsequent matrix computations. From the transfer matrix, the transmission coefficient can easily be derived.

3.5.3.1 Piecewise-Constant Potential

If an arbitrary potential barrier is segmented into N regions with constant potentials (see Fig. 3.9) the wave function in each region can be written as the sum of an incident and a reflected wave [93] $\Psi_j(x) = A_j \exp(\imath k_j x) + B_j \exp(-\imath k_j x)$ with the wave number $k_j = \sqrt{2m_j(\mathcal{E} - W_j)}/\hbar$. The wave amplitudes A_j, B_j, the carrier mass m_j, and the potential energy W_j are assumed constant for each region j. With the interface conditions for energy and momentum conservation

$$\Psi_j(x^-) = \Psi_{j+1}(x^+) \;, \tag{3.78}$$

$$\frac{1}{m_j}\frac{\mathrm{d}\Psi_j(x^-)}{\mathrm{d}x} = \frac{1}{m_{j+1}}\frac{\mathrm{d}\Psi_{j+1}(x^+)}{\mathrm{d}x} \;, \tag{3.79}$$

the outgoing wave of a layer relates to the incident wave by a complex transfer matrix:

$$\begin{pmatrix} A_j \\ B_j \end{pmatrix} = \underline{T}_j \begin{pmatrix} A_{j-1} \\ B_{j-1} \end{pmatrix} \qquad 2 \le j \le \mathrm{N} \;. \tag{3.80}$$

The transfer matrices are of the form

$$\underline{T}_j = \frac{1}{2}\begin{pmatrix} \left(1 + \frac{k_{j-1}}{k_j}\right)\gamma^{-k_j} & \left(1 - \frac{k_{j-1}}{k_j}\right)\gamma^{-k_j} \\ \left(1 - \frac{k_{j-1}}{k_j}\right)\gamma^{k_j} & \left(1 + \frac{k_{j-1}}{k_j}\right)\gamma^{k_j} \end{pmatrix} \begin{pmatrix} \gamma^{k_{j-1}} & 0 \\ 0 & \gamma^{-k_{j-1}} \end{pmatrix} \qquad 2 \le j \le \mathrm{N} \;, \tag{3.81}$$

with the phase factor $\gamma = \exp(\imath\Delta(j-2))$. The transmitted wave in Region N can then be calculated from the incident wave by subsequent multiplication of transfer matrices:

$$\begin{pmatrix} A_\mathrm{N} \\ B_\mathrm{N} \end{pmatrix} = \prod_{j=2..\mathrm{N}} \underline{T}_j \begin{pmatrix} A_1 \\ B_1 \end{pmatrix} \;. \tag{3.82}$$

If it is assumed that there is no reflected wave in Region N and the amplitude of the incident wave is unity, (3.82) simplifies to

$$\begin{pmatrix} A_\mathrm{N} \\ 0 \end{pmatrix} = \begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix} \begin{pmatrix} 1 \\ B_1 \end{pmatrix} \;, \tag{3.83}$$

and the transmission coefficient can be calculated from (3.56). The transfer-matrix method based on constant potential segments has the obvious shortcoming that, for practical barriers, the accuracy of the resulting matrix strongly depends on the chosen resolution. A more rigorous approach is to use linear potential segments.

3.5.3.2 Piecewise-Linear Potential

A general barrier may consist of several segments with linear potential sandwiched between contact segments where the potential is constant, as depicted in Fig. 3.11. The wave functions within these four regions can be written as (confer (3.69) and (3.70) for a linear potential)

$$\Psi_1(x) = A_1 \exp(\imath k_1 x) + B_1 \exp(-\imath k_1 x) \;, \tag{3.84}$$
$$\Psi_2(x) = A_2 \mathrm{Ai}(u_2(x)) + B_2 \mathrm{Bi}(u_2(x)) \;, \tag{3.85}$$
$$\Psi_3(x) = A_3 \mathrm{Ai}(u_3(x)) + B_3 \mathrm{Bi}(u_3(x)) \;, \tag{3.86}$$
$$\Psi_4(x) = A_4 \exp(\imath k_4 x) + B_4 \exp(-\imath k_4 x) \;, \tag{3.87}$$

with $u(x)$ from (3.70) and the x-independent derivative

$$u' = \frac{\mathrm{d}u(x)}{\mathrm{d}x} = -\left(\frac{2m}{\hbar^2}\right)^{1/3} \left(\frac{W_2 - W_1}{x_2 - x_1}\right)^{1/3} \;. \tag{3.88}$$

The conditions for continuity of the wave functions and their derivatives yield the following equation system, where abbreviations for the left and right value of $u(x)$ in a layer $\overleftarrow{u}_j = u_j(l_{j-2})$, $\overrightarrow{u}_j = u_j(l_{j-1})$, and their derivatives u'_j for $2 \le j \le \mathrm{N}-1$ have been used.

$$\begin{array}{lcllcl}
A_1 \exp(\imath k_1 l_0) & + & B_1 \exp(-\imath k_1 l_0) & = & A_2 \mathrm{Ai}(\overleftarrow{u}_2) & + & B_2 \mathrm{Bi}(\overleftarrow{u}_2) \;, \\
A_1 \imath k_1 \exp(i k_1 l_0) & - & B_1 \imath k_1 \exp(-i k_1 l_0) & = & A_2 \mathrm{Ai}'(\overleftarrow{u}_2) u'_2 & + & B_2 \mathrm{Bi}'(\overleftarrow{u}_2) u'_2 \;, \\
A_2 \mathrm{Ai}(\overrightarrow{u}_2) & + & B_2 \mathrm{Bi}(\overrightarrow{u}_2) & = & A_3 \mathrm{Ai}(\overleftarrow{u}_3) & + & B_3 \mathrm{Bi}(\overleftarrow{u}_3) \;, \\
A_2 \mathrm{Ai}'(\overrightarrow{u}_2) u'_2 & + & B_2 \mathrm{Bi}'(\overrightarrow{u}_2) u'_2 & = & A_3 \mathrm{Ai}'(\overleftarrow{u}_3) u'_3 & + & B_3 \mathrm{Bi}'(\overleftarrow{u}_3) u'_3 \;, \\
A_3 \mathrm{Ai}(\overrightarrow{u}_3) & + & B_3 \mathrm{Bi}(\overrightarrow{u}_3) & = & A_4 \exp(\imath l_2 k_4) & + & B_4 \exp(-\imath l_2 k_4) \;, \\
A_3 \mathrm{Ai}'(\overrightarrow{u}_3) u'_3 & + & B_3 \mathrm{Bi}'(\overrightarrow{u}_3) u'_3 & = & A_4 \imath k_4 \exp(\imath l_2 k_4) & - & B_4 \imath k_4 \exp(-\imath l_2 k_4) \;,
\end{array} \tag{3.89}$$

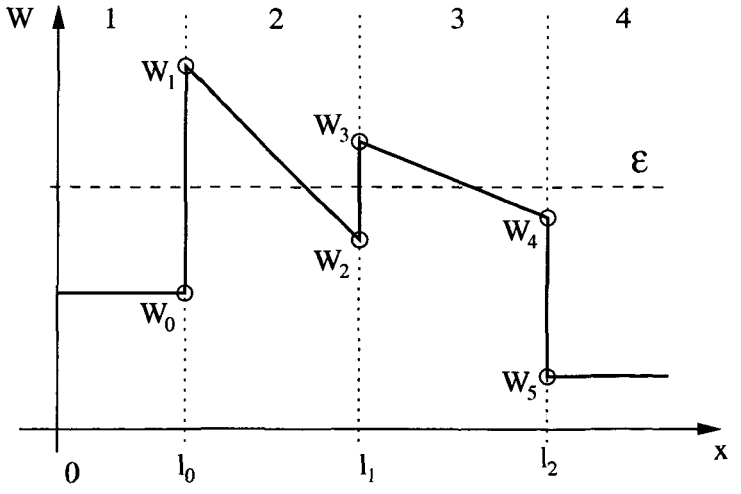


Figure 3.11: An energy barrier consisting of constant and linear potential segments.

The transfer matrices between adjacent layers are again calculated from (3.80). Using the first two equations of (3.89) and the WRONSKIan[8] [138]

$$\mathrm{Wr}\{\mathrm{Ai}(z),\mathrm{Bi}(z)\} = \mathrm{Ai}(z)\mathrm{Bi}'(z) - \mathrm{Ai}'(z)\mathrm{Bi}(z) = \pi^{-1}\,, \tag{3.90}$$

the matrix $\underline{T}_1$ can be simplified to

$$\underline{T}_1 = \pi \begin{pmatrix} \exp(\imath k_1 l_0)\left(\mathrm{Bi}'(\overleftarrow{u}_2) - \mathrm{Bi}(\overleftarrow{u}_2)\frac{\imath k_1}{u_2'}\right) & \exp(-\imath k_1 l_0)\left(\mathrm{Bi}'(\overleftarrow{u}_2) + \mathrm{Bi}(\overleftarrow{u}_2)\frac{\imath k_1}{u_2'}\right) \\ \exp(\imath k_1 l_0)\left(-\mathrm{Ai}'(\overleftarrow{u}_2) + \mathrm{Ai}(\overleftarrow{u}_2)\frac{\imath k_1}{u_2'}\right) & \exp(-\imath k_1 l_0)\left(-\mathrm{Ai}'(\overleftarrow{u}_2) - \mathrm{Ai}(\overleftarrow{u}_2)\frac{\imath k_1}{u_2'}\right) \end{pmatrix}.$$

Using the next two lines of (3.89) yields

$$\underline{T}_2 = \pi \begin{pmatrix} \mathrm{Ai}(\overrightarrow{u}_2)\mathrm{Bi}'(\overleftarrow{u}_3) - \frac{u_2'}{u_3'}\mathrm{Bi}(\overleftarrow{u}_3)\mathrm{Ai}'(\overrightarrow{u}_2) & \mathrm{Bi}(\overrightarrow{u}_2)\mathrm{Bi}'(\overleftarrow{u}_3) - \frac{u_2'}{u_3'}\mathrm{Bi}(\overleftarrow{u}_3)\mathrm{Bi}'(\overrightarrow{u}_2) \\ \frac{u_2'}{u_3'}\mathrm{Ai}(\overleftarrow{u}_3)\mathrm{Ai}'(\overrightarrow{u}_2) - \mathrm{Ai}(\overrightarrow{u}_2)\mathrm{Ai}'(\overleftarrow{u}_3) & \frac{u_2'}{u_3'}\mathrm{Ai}(\overleftarrow{u}_3)\mathrm{Bi}'(\overrightarrow{u}_2) - \mathrm{Bi}(\overrightarrow{u}_2)\mathrm{Ai}'(\overleftarrow{u}_3) \end{pmatrix},$$

and the last two equations yield with the phase factor $\gamma = \exp(\imath l_2 k_4)$

$$\underline{T}_3 = \frac{1}{2}\begin{pmatrix} \mathrm{Ai}(\overrightarrow{u}_3)\gamma^{-1} + \frac{u_3'}{\imath k_4}\mathrm{Ai}'(\overrightarrow{u}_3)\gamma^{-1} & \mathrm{Bi}(\overrightarrow{u}_3)\gamma^{-1} + \frac{u_3'}{\imath k_4}\mathrm{Bi}'(\overrightarrow{u}_3)\gamma^{-1} \\ \mathrm{Ai}(\overrightarrow{u}_3)\gamma - \frac{u_3'}{\imath k_4}\mathrm{Ai}'(\overrightarrow{u}_3)\gamma & \mathrm{Bi}(\overrightarrow{u}_3)\gamma - \frac{u_3'}{\imath k_4}\mathrm{Bi}'(\overrightarrow{u}_3)\gamma \end{pmatrix}. \tag{3.91}$$

While being more accurate than the constant potential approach this method is computationally more expensive. This drawback, however, is offset by the fact that a lower resolution and thus fewer matrix multiplications are necessary to resolve an energy barrier consisting of linear potential segments.

Simulations using the transfer-matrix method have been reported by several authors [145–148]. Others compared the constant and linear potential approaches and found the constant potential method more feasible for device simulation [149]. The main advantage of the linear-potential transfer-matrix method is, that for linear potential segments the accuracy does not depend on the resolution as it does for the constant-potential transfer-matrix method. However, the evaluation of the AIRY functions must be carefully implemented to avoid overflow.

Although the transfer-matrix method for constant or linear potential segments is intuitively easy to understand and implement, the main shortcoming of the method is that it becomes numerically instable for thick barriers. This has been observed by several authors [149–153]. The reason for the numerical problems is that during the matrix multiplications exponentially growing and decaying states have to be multiplied, leading to rounding errors which eventually exceed the amplitude of the wave function itself for thick barriers.

These problems have been overcome by a further segmentation of the barrier into slices with more accurate transfer matrices [150], the use of scattering matrices instead of transfer matrices [151], iterative methods [152], or by simply setting the transfer matrix entries to zero if the decay factor $\sum k_j x_j$ exceeds a certain value of about 20 [149]. In the next section a method will be presented which avoids this problem and allows a fast and reliable transmission coefficient estimation.

[8] JOSEF HOËNÉ DE WRONSKI, Polish mathematician, 1778–1853.

3.5.4 Quantum Transmitting Boundary Method

An alternative method to solve the SCHRÖDINGER equation has been proposed by FRENSLEY and EINSPRUCH [154] which is based on the tight-binding quantum transmitting boundary method (QTBM) introduced by LENT [155]. It has been used to simulate electron transport in resonant tunneling diodes [153]. The method is based on the finite-difference approximation of the stationary one-dimensional SCHRÖDINGER equation (3.63) on an equidistant grid with an effective mass m_j and a grid spacing Δ

$$\underline{H}\Psi_j = -s_{j-1}\Psi_{j-1} + d_j\Psi_j - s_{j+1}\Psi_{j+1} = \mathcal{E}\Psi_j \ , \tag{3.92}$$

where $s_j = \hbar^2/(2m_j\Delta^2)$ and $d_j = \hbar^2/(m_j\Delta^2) + W_j$. For the evaluation of the transmission coefficient it is necessary to assume open boundary conditions. They are introduced by writing the wave functions at the boundaries of the simulation domain as

$$\Psi_1 = a_1 + b_1 \tag{3.93}$$
$$\Psi_N = a_N + b_N \tag{3.94}$$

and relate them to the wave functions outside of the simulation domain by

$$\Psi_0 = a_1 \exp(-\imath k_1\Delta) + b_1 \exp(\imath k_1\Delta) \ , \tag{3.95}$$
$$\Psi_{N+1} = a_N \exp(-\imath k_N\Delta) + b_N \exp(\imath k_N\Delta) \ . \tag{3.96}$$

This introduces four unknowns and two equations into the system. Setting

$$a_1 = \zeta_1\Psi_0 + \xi_1\Psi_1 \tag{3.97}$$
$$a_N = \zeta_N\Psi_{N+1} + \xi_N\Psi_N \tag{3.98}$$

eliminates the unknown values of b_1 and b_N and gives a linear system for the $N+2$ complex values Ψ_j

$$\begin{pmatrix} \zeta_1 & \xi_1 & & & & & \\ -s_1 & d_1-\mathcal{E} & -s_2 & & & & \\ & -s_2 & d_1-\mathcal{E} & -s_3 & & & \\ & & & \cdots & & & \\ & & & & -s_N & d_N-\mathcal{E} & -s_N \\ & & & & & \xi_N & \zeta_N \end{pmatrix} \begin{pmatrix} \Psi_0 \\ \Psi_1 \\ \Psi_2 \\ \cdots \\ \Psi_N \\ \Psi_{N+1} \end{pmatrix} = \begin{pmatrix} a_1 \\ 0 \\ 0 \\ \cdots \\ 0 \\ a_N \end{pmatrix} \ . \tag{3.99}$$

Setting $a_1 = 1$ and $a_N = 0$ yields the values of the wave function in the whole simulation domain for an incident wave from the left side like in the transfer-matrix method. The method is easy to implement, fast, and more robust than the transfer-matrix method. A further advantage of this method is its suitability for two- and three-dimensional problems. It thus represents a much more powerful method than the transfer-matrix based methods which are limited to one-dimensional problems only. Note that the QTBM is closely linked with the non-equilibrium GREEN's function formalism (NEGF, see Section 2.4.3.4): The matrix in expression (3.99) is the inverse of the retarded GREEN's function (2.25) for an open system without scattering. However, the values of ζ and ξ are complex, so the matrix admits complex eigenvalues and complex solving routines are necessary.

3.5.5 Comparison

Fig. 3.12 shows the transmission coefficient for the described methods for a triangular energy barrier (left) and a two-step non-linear energy barrier (right). The inset shows the energy barrier and the values of $|\Psi|^2$ for an energy of 2.8 eV on a logarithmic scale. The dotted lines refer to the constant-potential transfer-matrix method. In the left figure the numerical instability of the transfer-matrix method leads to an increasing transmission coefficient for energies below 1 eV. These numerical problems occur for both the constant-potential and the linear-potential approaches.

The GUNDLACH and analytical WKB methods deliver similar results for the triangular barrier. For the stacked dielectric shown in the right figure, the analytical WKB and GUNDLACH methods cannot be used. The numerical WKB, transfer-matrix, and QTB methods deliver similar results, however, the WKB method does not resolve oscillations in the transmission coefficient.

It can be concluded that for a single-layer dielectric, the analytical WKB method yields reasonable accuracy as compared to the other, computationally more expensive methods. For stacked dielectrics, however, only the numerical WKB, transfer-matrix, or QTB methods can be used in the first place. Since transfer-matrix based methods exhibit problems regarding numerical stability, only the QTBM and the numerical WKB methods remain. Since the numerical WKB method also needs a numerical integration, its advantage in terms of computational effort is not high enough to rule out the QTBM. Furthermore, if resonance effects — such as in dielectrics with quantum wells, see Section 5.2.2.3 — have to be taken into account, the QTBM remains as the method of choice for a reliable transmission coefficient estimation.

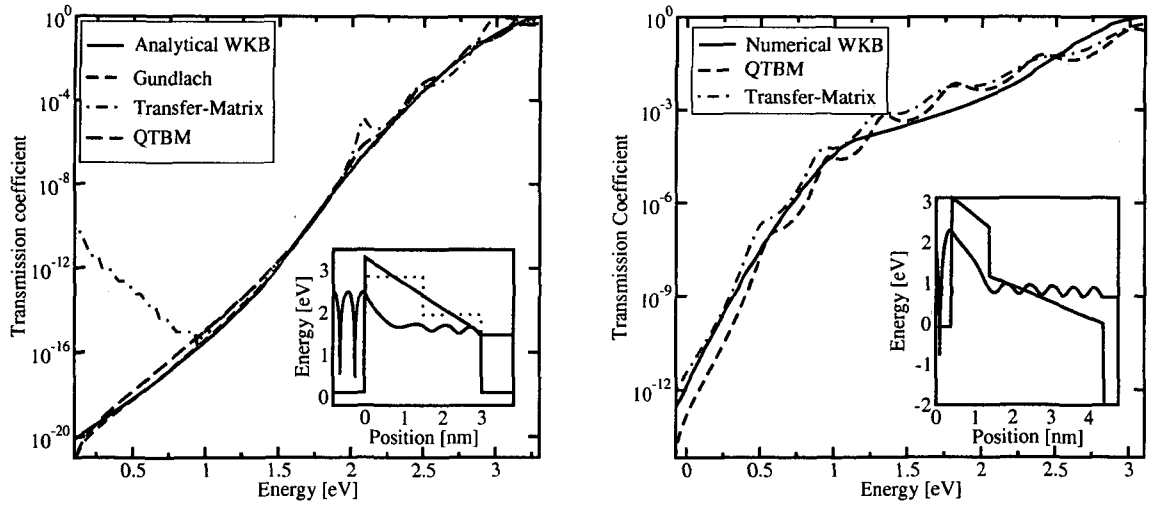


Figure 3.12: The transmission coefficient using different methods for a dielectric consisting of a single layer (left) and for a dielectric consisting of two layers (right). The shape of the energy barrier and the wave function at 2.8 eV is shown in the inset.

3.6 Bound and Quasi-Bound States

Up to now it has been assumed that all energetic states in the substrate contribute to the tunneling current. However, the high doping and the high electric field in the channel leads to a quantum-mechanical quantization of carriers as described in Section 2.2.1 [156, 157]. If it is assumed that the wave function does not penetrate into the gate, discrete energy levels can be identified. However, it cannot be assumed that electrons tunnel from these energies, since for the derivation of the levels it was assumed that there is no wave function penetration into the dielectric. This leads to the *paradox* which was addressed by MAGNUS and SCHOENMAKER [158]: How can a bound state, which has vanishing current density, lead to tunneling current?

The answer is that it cannot. Taking a closer look at the conduction band edge of a MOSFET in inversion reveals that, depending on the boundary conditions, different types of quantized energy levels must be distinguished [159], see Fig. 3.13: Bound states are formed at energies for which the wave function decays to zero at both sides. Quasi-bound states (QBS) have closed boundary conditions at one side and open boundary conditions at the other side. Free states, finally, are states which do not decay at any side. The total tunnel current density therefore consists of current from the QBS and from the free states:

$$J = \mathrm{q}\sum_i \frac{n_\nu(\mathcal{E}_i)}{\tau_\mathrm{q}(\mathcal{E}_i)} + \frac{4\pi m_\mathrm{eff}\mathrm{q}}{h^3}\int_{\mathcal{E}_\mathrm{min}}^{\mathcal{E}_\mathrm{max}} TC(\mathcal{E})N(\mathcal{E})\,\mathrm{d}\mathcal{E}\ , \tag{3.100}$$

where the symbol $n_\nu(\mathcal{E}_i)$ denotes the two-dimensional carrier concentration [160]

$$n_\nu = g_\nu \frac{m\mathrm{k_B}T}{\pi\hbar^2}\ln\left(1+\exp\left(\frac{\mathcal{E}_\mathrm{f}-\mathcal{E}_i}{\mathrm{k_B}T}\right)\right)\ , \tag{3.101}$$

the symbol g_ν is the valley degeneracy, and τ_q is the life time of the quasi-bound state $\mathcal{E}_i$. The life time is based on GAMOW's theory of nuclear decay [40] and denotes the time constant with which an electron leaks through the energy barrier. Since bound and quasi-bound states are closely related, the computation of bound states will be described first.

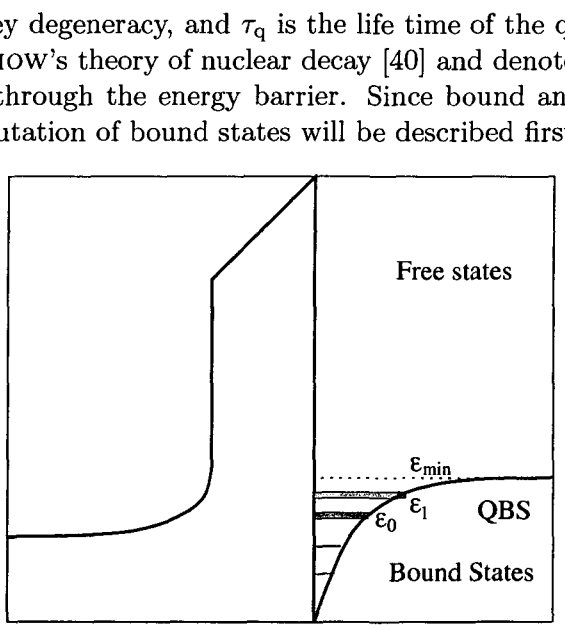


Figure 3.13: Free, bound, and quasi-bound states in a typical MOS inversion layer.

3.6.1 Eigenvalues of a Triangular Energy Well

To first order the conduction band edge in a MOSFET inversion layer can be approximated by a linear potential (this is actually done by various authors, see [161–164]). The solution of SCHRÖDINGER's equation for a linear potential has been derived in Section 3.5.2 and consists of a linear superposition of AIRY's functions. If the triangular energy well is defined as

$$W(x) = W_0 + \frac{W_1 - W_0}{x_1 - x_0} x \tag{3.102}$$

and no wave function penetration for $x <= x_0$ is taken into account, the wave function for $x > 0$ can be written as [156]

$$\Psi(x) = A\mathrm{Ai}(u(x)) \ , \tag{3.103}$$

$$\Psi(x_0) = A\mathrm{Ai}(u(x_0)) = 0 \ . \tag{3.104}$$

Therefore, $u(x_0)$ must equal one of the zeros of the AIRY function z_i:

$$u(x_0) = z_i < 0 \ . \tag{3.105}$$

With $u(x)$ from expression (3.70) the energy eigenvalues are found as

$$\mathcal{E}_i = W_0 - z_i \left(\frac{\hbar^2}{2m}\right)^{1/3} \left(\frac{W_1 - W_0}{x_1 - x_0}\right)^{2/3} \ . \tag{3.106}$$

The first five zeros of the AIRY function are -2.34, -4.09, -5.52, -6.79, and -7.94. These values are often used to approximate the quantized carrier concentration in the channel of MOS devices. The value of the normalizing constant A becomes (the derivation is shown in Appendix C)

$$A = \left(\frac{\left(\frac{2m\mathrm{q}E}{\hbar^2}\right)^{1/3}}{\mathrm{Ai}'^2(\lambda_0) - \lambda_0 \mathrm{Ai}^2(\lambda_0)} \right)^{1/2} \ , \tag{3.107}$$

where E is the constant electric field in the energy well, and the value of λ_0 depends on the energy eigenvalue $\mathcal{E}_i$ via

$$\lambda_0 = -\frac{\mathcal{E}_i}{\mathrm{q}E} \left(\frac{2m\mathrm{q}E}{\hbar^2}\right)^{1/3} \ . \tag{3.108}$$

This method can be used to get an estimate of the first few eigenvalues of the system, or to find initial values for the calculation of the eigenvalues described in the next section.

3.6.2 Eigenvalues of Arbitrary Energy Wells

To calculate the eigenvalues of an arbitrary energy well it is necessary to solve SCHRÖDINGER's equation. This can be done using the method of finite differences. It is based on a discretization of the HAMILTONian on a spatial grid and given by (3.92) which is repeated here for convenience

$$\underline{H}\Psi_j = -s_j\Psi_{j-1} + d_j\Psi_j - s_{j+1}\Psi_{j+1} = \mathcal{E}\Psi_j \ .$$

While in Section 3.5.4, a constant value of the electron mass in the simulated region was used, a discretization which allows for a position-dependent carrier mass reads

$$d_j = \frac{\hbar^2}{4\Delta^2}\left(\frac{1}{m_{j-1}} + \frac{2}{m_j} + \frac{1}{m_{j+1}}\right) + W_j \tag{3.109}$$

and

$$s_j = \frac{\hbar^2}{4\Delta^2}\left(\frac{1}{m_{j-1}} + \frac{1}{m_j}\right) \ . \tag{3.110}$$

The system HAMILTONian is tridiagonal and, for a six-point example, can be written similar to (3.99) but without the entries for ζ and ξ:

$$\begin{pmatrix} d_1 & -s_2 & & \\ -s_2 & d_2 & -s_3 & \\ & -s_3 & d_3 & -s_4 \\ & & -s_4 & d_4 \end{pmatrix} \begin{pmatrix} \Psi_1 \\ \Psi_2 \\ \Psi_3 \\ \Psi_4 \end{pmatrix} = \mathcal{E} \begin{pmatrix} \Psi_1 \\ \Psi_2 \\ \Psi_3 \\ \Psi_4 \end{pmatrix} \ . \tag{3.111}$$

The values Ψ_0 and Ψ_5 must be 0 in this case, that is closed boundary conditions are assumed. The system HAMILTONian is real and symmetric, therefore all eigenvalues are real. While this matrix equation looks similar to (3.99), there are important differences. Here it is necessary to solve the eigenvalue equation to get a value for $\mathcal{E}_i$ and Ψ_i. In (3.99), any value of $\mathcal{E}$ leads to a valid solution for Ψ_i, and the solution is obtained by solving a complex equation system.

3.6.3 The Life Time of Quasi-Bound States

The tunneling current from quasi-bound states in (3.100) depends on their quantum-mechanical life time τ_q: In contrast to electrons in bound states, which have an infinite life time, electrons in quasi-bound states have a non-zero probability to tunnel through the energy barrier, thus their life time is finite [165–167]. This can be seen if the time time evolution of the states is considered [168]

$$\Psi(t) = \Psi_0 \exp\left(-\imath\frac{\mathcal{E}_\mathrm{i}}{\hbar}t\right) \ , \tag{3.112}$$

where Ψ_0 is the initial wave function and the complex eigenenergy is

$$\mathcal{E}_\mathrm{i} = \mathcal{E}_\mathrm{re} - \imath\mathcal{E}_\mathrm{im} \ . \tag{3.113}$$

The time-dependent probability becomes

$$P(t) = \Psi^*(t)\Psi(t) = \Psi_0^2 \exp\left(-\frac{2\mathcal{E}_{\text{im}}}{\hbar}t\right) = \Psi_0^2 \exp\left(-\frac{t}{\tau_{\text{q}}}\right) . \tag{3.114}$$

Thus, the imaginary component of the eigenenergy $\mathcal{E}$ is related to the decay time constant by

$$\tau_{\text{q}} = \frac{\hbar}{2\mathcal{E}_{\text{im}}} . \tag{3.115}$$

The QBS are frequently used for tunneling current calculations [169–174]. Three methods are established to compute the life time of a quasi-bound state in MOS inversion layers: Computing the full-width half-maximum (FWHM) of the reflection coefficient resonances, using the quasi-classical formula based on the WENTZEL-KRAMERS-BRILLOUIN-method, or from the complex eigenvalues of the non-HERMITian HAMILTONian. These methods will be described in the following.

3.6.3.1 The Reflection Coefficient Resonances

A quasi-bound state forms if one of the system boundary conditions is open ($\neq 0$) and the other one is closed ($= 0$). The carrier wave function is reflected at the interface, there is no transmitted wave. Using the transfer-matrix method described in Section 3.5.3, the system can be described by

$$\begin{pmatrix} A_{\text{N}} \\ B_{\text{N}} \end{pmatrix} = \begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix} \begin{pmatrix} A_1 \\ B_1 \end{pmatrix} , \tag{3.116}$$

where the wave functions are plane waves

$$\Psi_j(x) = A_j \exp(\imath k_j x) + B_j \exp(-\imath k_j x) . \tag{3.117}$$

However, no transmission coefficient can be defined for a quasi-bound state: The transmitted wave amplitude A_{N} must vanish to fulfill the assumption of closed boundary conditions. Instead, a reflection coefficient can be defined which is

$$RC(\mathcal{E}) = \frac{B_1}{A_1} = -\frac{T_{21}}{T_{22}} . \tag{3.118}$$

It is shown in [165] that for a quasi-bound state, the transfer matrix is not HERMITian[9] and its elements obey

$$T_{11} = T_{12}^* ,$$
$$T_{21} = T_{22}^* .$$

[9] For free states, which is the kind of application investigated in Section 3.5.3, the transfer matrix is HERMITian:

$$T_{11} = T_{22}^*$$
$$T_{12} = T_{21}^*$$

Therefore, the reflection coefficient $RC(\mathcal{E})$ can be written as

$$RC(\mathcal{E}) = \exp\left(\imath\Theta(\mathcal{E})\right) \ . \tag{3.119}$$

The phase $\Theta(\mathcal{E})$ varies only weakly at energies away from the resonance energy of the QBS, while near the QBS the phase changes strongly. Near the complex energy levels $\mathcal{E}_\mathrm{i}$ the derivative of the phase factor $\Theta(\mathcal{E})$ follows a LORENTZian[10] distribution

$$\frac{\mathrm{d}\Theta}{\mathrm{d}\mathcal{E}} = \frac{2\mathcal{E}_\mathrm{i}}{(\mathcal{E}-\mathcal{E}_\mathrm{re})^2 + \mathcal{E}_\mathrm{im}^2} \ , \tag{3.120}$$

where $2\mathcal{E}_\mathrm{im}$ is the full-width half-maximum (FWHM) value of $\mathrm{d}\Theta/\mathrm{d}\mathcal{E}$. Thus, by calculating the phase of the reflection coefficient as a function of energy, the life times can be determined. This method has been studied intensely by CASSAN *et al.* [160, 175]. They reported numerical difficulties in the calculation of the value of $\mathrm{d}\Theta/\mathrm{d}\mathcal{E}$ which is prone to numerical noise. Similar problems have been reported by other groups [176].

An alternative approach has been presented by CLERC *et al.* who noted that the life times can also be extracted directly from the transfer matrix [144]. For a free state, $B_\mathrm{N} = 0$ in (3.116) and the transmission coefficient becomes

$$TC = \left|\frac{A_\mathrm{N}}{A_1}\right|^2 = \frac{1}{|T_{11}|^2} \ . \tag{3.121}$$

For a quasi-bound state, $A_\mathrm{N} = 0$. Therefore,

$$A_1 = T_{12}B_\mathrm{N} \ , \tag{3.122}$$

but, since $T_{11} = T_{12}^*$, the value of $|T_{11}|^{-2}$ may be evaluated as well — even if it cannot be interpreted as a transmission coefficient. The life time of the QBS is again found from the resonance peak of the LORENTZian around the real component of the eigenenergy $\mathcal{E}_\mathrm{re}$

$$\frac{1}{|T_{11}|^2} = \frac{1}{(\mathcal{E}-\mathcal{E}_\mathrm{re})^2 + \dfrac{\hbar^2}{4\tau_\mathrm{q}^2}} \ , \tag{3.123}$$

but no derivative must be calculated this time. As an example of this method the left part of Fig. 3.14 shows the shape of the conduction band edge of a MOS structure in the substrate, dielectric, and polysilicon gate. In the substrate a triangular quantum well forms. Considering closed boundaries, eigenvalues and wave functions can be calculated. The corresponding wave functions are shown in the figure, where closed boundary conditions have been used at the boundaries of the simulation domain. Note the wave function penetration into the classically forbidden region of the dielectric layer. The eigenvalues of the quasi-bound states are located at 0.27, 0.47, 0.63, 0.76, 0.86, and 0.95 eV. The same information can be found when the value of $|T_{11}|^{-2}$ is investigated, as shown in right part of Fig. 3.14: Every quasi-bound state in the inversion layer manifests as a peak in the value of $|T_{11}|^{-2}$. The width of each peak is directly related to its life time.

[10] Hendrik Antoon Lorentz, Dutch physicist, 1853–1928.

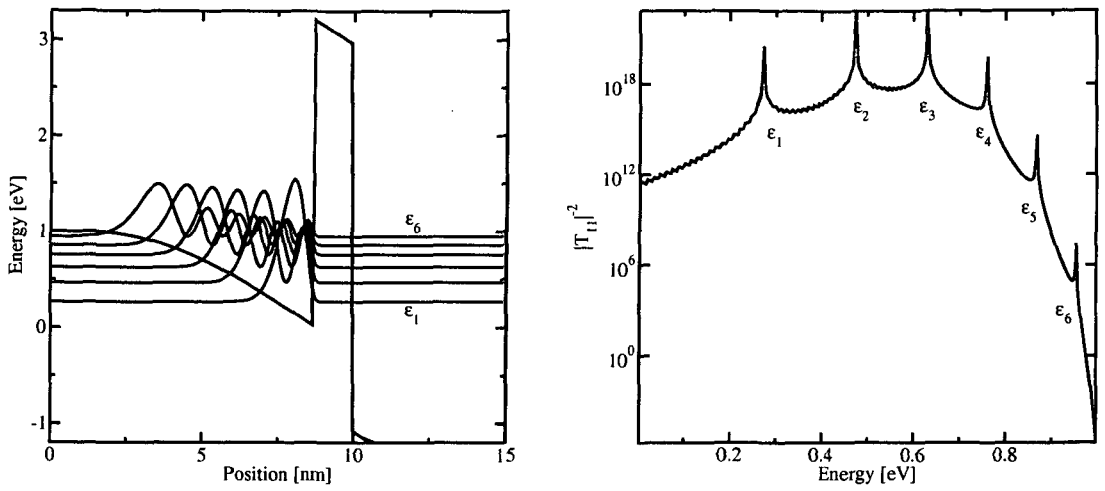


Figure 3.14: Wave function of quasi-bound states. Note the wave function penetration into classically forbidden regions (left). The respective value of $|T_{11}|^{-2}$ as a function of energy is shown in the right plot. The energy broadening around the poles is clearly visible.

3.6.3.2 The Quasi-Classical Formula

The calculation of the life times using the approaches shown so far is cumbersome and error-prone, since a precise value for the FWHM in regions where different QBS overlap is difficult to obtain. As an approximation the life time of a QBS can be computed from the quasi-classical formula [176]

$$\tau_{\mathrm{q}} = \frac{1}{TC(\mathcal{E}_{\mathrm{i}})} \int_0^{x_i} \sqrt{\frac{2m_i}{\mathcal{E}_{\mathrm{i}} - \mathcal{E}_{\mathrm{c}}(x)}}\, \mathrm{d}x \,, \tag{3.124}$$

where $\mathcal{E}_{\mathrm{i}}$ is the resonance energy of the respective bound state and x_i the classical turning point for this energy. The transmission coefficient $TC(\mathcal{E}_{\mathrm{i}})$ can be calculated by the transfer-matrix method or any other method that solves SCHRÖDINGER's equation.

3.6.3.3 The Eigenvalues of the Non-HERMITian HAMILTONian

For open-boundary conditions, the system is described by a HAMILTONian which is not HERMITian and admits complex eigenvalues. The most straightforward way to calculate the life times is to directly find the complex eigenvalues of the system HAMILTONian. This, however, is not easily possible because the eigenvalue problem is nonlinear: The matrix elements depend on the eigenvalue [177]. The numerical implementation of this method will be described in Section 4.3.3.

The complex eigenvalues have been used to calculate the life times of the structure shown in the left part of Fig. 3.14. The complex energies and life times found are shown in Table 3.2. The

values perfectly agree with the values found using the method based on the evaluation of the reflection-coefficient.

The life times of the first and second QBS have been evaluated as a function of the gate bias and the thickness of the dielectric layer as shown in the left part of Fig. 3.15. The life time decreases with increasing gate bias which is due to the higher penetrability of the energy barrier. The results of the gate current density (3.100) is shown in the right part of Fig. 3.15, where the TSU-ESAKI tunneling current was not considered.

This method, however, is by far the most computationally demanding one and it has not been implemented in MINIMOS-NT since problems regarding the stability of the underlying algorithms have been observed.

$\mathcal{E}_i$	$\mathcal{E}_{re}$ [eV]	$\mathcal{E}_{im}$ [eV]	τ_q [s]
1	0.2695	1.503×10^{-20}	4.376×10^{4}
2	0.4695	1.830×10^{-19}	3.594×10^{3}
3	0.6256	5.285×10^{-15}	1.244×10^{-1}
4	0.7549	2.794×10^{-11}	2.354×10^{-4}
5	0.8629	4.231×10^{-8}	1.555×10^{-8}
6	0.9503	2.005×10^{-5}	3.281×10^{-11}

Table 3.2: Eigenvalues found by using a resonance-finding algorithm based on the determinant of the open-boundary HAMILTONian.

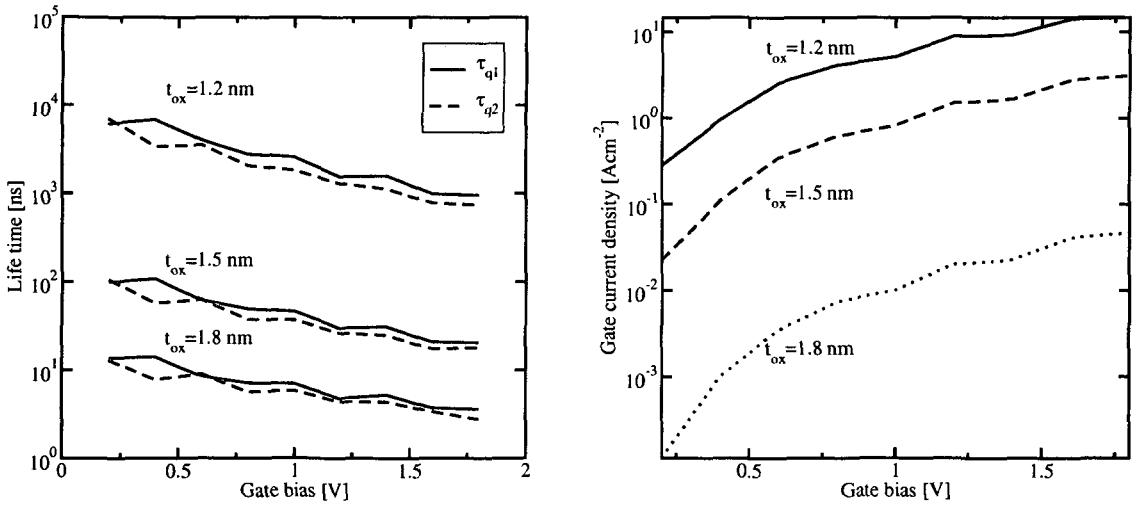


Figure 3.15: The life time of the first and second QBS for different gate dielectric thicknesses and gate voltages (left) and the resulting gate current density considering the first three quasi-bound states (right).

3.7 Compact Tunneling Models

The above presented models for the calculation of tunneling currents require a considerable computational effort. However, for practical device simulation, it is desirable to use compact models which do not require large computational resources. That may be necessary for a quick estimation of the dielectric thickness from IV data or to predict the impact of gate leakage on the performance of CMOS circuits [178–183]. The most frequently used model to describe tunneling is the FOWLER-NORDHEIM formula [184]

$$J = AE_{\mathrm{diel}}^2 \exp\left(-\frac{B}{E_{\mathrm{diel}}}\right) \tag{3.125}$$

which was originally used to describe tunneling between metals under intense electric fields. The parameters A and B have been refined by LENZLINGER and SNOW [185]:

$$J = \frac{\mathrm{q}^3 m_{\mathrm{eff}}}{8\pi m_{\mathrm{diel}} h \mathrm{q}\Phi_{\mathrm{B}}} E_{\mathrm{diel}}^2 \exp\left(-\frac{4\sqrt{2m_{\mathrm{diel}}(\mathrm{q}\Phi_{\mathrm{B}})^3}}{3\hbar \mathrm{q} E_{\mathrm{diel}}}\right) . \tag{3.126}$$

This expression can be derived from the TSU-ESAKI formula (3.13) by the assumption of zero temperature, a triangular energy barrier, and equal materials on both sides of the dielectric (the derivation is shown in Appendix A). Thus, it is not valid for direct tunneling where the barrier is of trapezoidal shape. Furthermore, $\mathrm{q}\Phi_{\mathrm{B}}$ denotes the difference between the FERMI energy in the electrode and the conduction band edge in the dielectric, and not the conduction band offset, as it is often found in the literature.

SCHUEGRAF and HU derived correction terms for this expression to make it applicable to the regime of direct tunneling [186]

$$J = \frac{\mathrm{q}^3 m_{\mathrm{eff}}}{8\pi m_{\mathrm{diel}} h \mathrm{q}\Phi_{\mathrm{B}} B_1} E_{\mathrm{diel}}^2 \exp\left(-\frac{4\sqrt{2m_{\mathrm{diel}}(\mathrm{q}\Phi_{\mathrm{B}})^3} B_2}{3\hbar \mathrm{q} E_{\mathrm{diel}}}\right) , \tag{3.127}$$

with the correction terms B_1 and B_2 given as (the derivation can also be found in Appendix A)

$$B_1 = \left(1 - \left(1 - \frac{\mathrm{q}E_{\mathrm{diel}} t_{\mathrm{diel}}}{\mathrm{q}\Phi_{\mathrm{B}}}\right)^{1/2}\right)^2 , \tag{3.128}$$

and

$$B_2 = \left(1 - \left(1 - \frac{\mathrm{q}E_{\mathrm{diel}} t_{\mathrm{diel}}}{\mathrm{q}\Phi_{\mathrm{B}}}\right)^{3/2}\right) . \tag{3.129}$$

For a triangular barrier the correction factors become $B_1 = B_2 = 1$ and the expression simplifies to (3.126). Note that using these equations, the minimum tunneling current occurs for $E_{\mathrm{diel}} = 0\,\mathrm{V/m}$ which, for a work function difference $\neq 0$, does not occur at the minimum applied bias.

3.8 Trap-Assisted Tunneling

Besides direct or FOWLER-NORDHEIM tunneling, which are one-step tunneling processes, defects in the dielectric layer give rise to tunneling processes based on two or more steps. This tunneling component is mainly observed after writing-erasing cycles in electrically erasable programmable read-only-memories (EEPROMs). It is therefore assumed that traps arise in the dielectric layer due to the repeated high voltage stress. The increased tunneling current at low bias is called stress-induced leakage current (SILC) and is mainly responsible for the degradation of the retention time of non-volatile memory devices [187]. It is now generally accepted that it is caused by inelastic trap-assisted tunnel transitions and that the traps are created by the electric high-field stress during the writing and erasing processes [187–192]. SILC has been widely studied and modeled in MOS capacitors [193–195] and EEPROM devices [196].

This section gives a brief overview of trap-assisted tunneling models, describes two frequently encountered models (CHANG's and IELMINI's model) and elaborates on one of the most sophisticated models which was originally proposed by JIMENEZ *et al.*. The adaption of this model to allow its inclusion in the device simulator MINIMOS-NT is described in some detail.

3.8.1 Model Overview

Numerous models have been presented to describe trap-assisted tunneling in the gate dielectric of MOS devices. These models usually share the equation for the current density which is given by an integration along the gate dielectric [197]:

$$J = \mathrm{q} \int_0^{t_{\mathrm{diel}}} \frac{N_{\mathrm{T}}(x)}{\tau_{\mathrm{c}}(x) + \tau_{\mathrm{e}}(x)} \, \mathrm{d}x \; . \tag{3.130}$$

In this expression N_{T} denotes the trap concentration, and τ_{c} and τ_{e} denote the capture and emission times of the considered trap. Since both processes – capture and emission – must happen in sequence, they both determine the current density. However, differences exist in how the capture and emission times are calculated. Some models use constant capture and emission cross sections to calculate the respective times. Another important point is the distribution in space, where the traps are usually assumed to follow a GAUSSian distribution. The distribution in energy is also crucial. Commonly it is either assumed that traps have a GAUSSian distribution in energy or that they are located at a certain energy level below the dielectric conduction band. The assumption of a discrete energy level for specific trap types is backed by spectroscopic analyses [198]. Additionally, the tunneling process can either be elastic, where the energy of the tunneling electron is conserved, or inelastic, where the energy of the tunneling electron changes. Recent studies and experiments have shown strong evidence for the tunneling process being inelastic [199–201].

3.8.1.1 CHANG's Model

A frequently used model is the generalized trap-assisted tunneling model presented by CHANG *et al.* [202, 203]. The current density reads

$$J = \mathrm{q} \int_0^{t_{\mathrm{diel}}} A N_{\mathrm{T}}(x) \frac{P_1(x) P_2(x)}{P_1(x) + P_2(x)} \, \mathrm{d}x \ , \tag{3.131}$$

where A denotes a fitting constant, $N_{\mathrm{T}}(x)$ the spatial trap concentration, and P_1 and P_2 the transmission coefficients of electrons captured and emitted by traps. Using $\tau_{\mathrm{c}} \sim P_1/P_2$ and $\tau_{\mathrm{e}} \sim P_2/P_1$, this expression reduces to (3.130). A similar model was used by GHETTI *et al.* [169]

$$J = \int_0^{t_{\mathrm{diel}}} C_{\mathrm{T}} N_{\mathrm{T}}(x) \frac{J_{\mathrm{in}} J_{\mathrm{out}}}{J_{\mathrm{in}} + J_{\mathrm{out}}} \, \mathrm{d}x \ , \tag{3.132}$$

who assumed a constant capture cross section C_{T} for the traps. The symbols J_{in} and J_{out} denote the capture and emission currents. Essentially the same formula was used by other authors as well [200, 204].

3.8.1.2 IELMINI's Model

Considerable research has been done by IELMINI *et al.* [205–208] who describe inelastic TAT and also take hopping conduction into account [209, 210]. They derive the trap-assisted current by an integration along the dielectric thickness and energy

$$J = \int_0^{t_{\mathrm{diel}}} \mathrm{d}x \int_{\mathcal{E}_{\mathrm{min}}}^{\mathcal{E}_{\mathrm{max}}} \tilde{J}(\mathcal{E}_{\mathrm{T}}, x) \, \mathrm{d}\mathcal{E} \ ,$$

where $\tilde{J}$ denotes the net current flowing through the dielectric, given as the difference between capture and emission currents through either side (left or right), as shown in Fig. 3.16

$$\tilde{J}(\mathcal{E}_{\mathrm{T}}, x) = J_{\mathrm{cl}} - J_{\mathrm{el}} = J_{\mathrm{er}} - J_{\mathrm{cr}} = \mathrm{q} N'_{\mathrm{T}} W_{\mathrm{c}} \left(1 - \frac{f_{\mathrm{T}}(\mathcal{E}_{\mathrm{T}}, x)}{f_{\mathrm{l}}(\mathcal{E}_{\mathrm{T}}, x)} \right) \ ,$$

where f_{T} is the trap occupancy, $\mathcal{E}_{\mathrm{T}}$ the trap energy, W_{c} the capture rate, and f_{l} the energy distribution function at the left interface. The symbol N'_{T} denotes the trap concentration in space and energy. IELMINI further develops the model to include transient effects and notes that in this case, the net difference between current from the left and right interfaces equals the change in the trap occupancy multiplied by the trap charge

$$(J_{\mathrm{cl}} - J_{\mathrm{el}}) + (J_{\mathrm{cr}} - J_{\mathrm{er}}) = \mathrm{q} N_{\mathrm{T}} \frac{\partial f_{\mathrm{T}}}{\partial t} \ , \tag{3.133}$$

an observation that will be revisited in Section 3.8.2.4. The main shortcoming of this model, despite its sophistication, is the assumption of a constant capture cross-section.

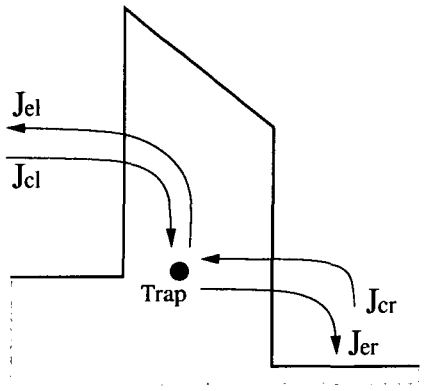


Figure 3.16: Schematic capture and emission currents through the left and right interfaces of the dielectric layer.

3.8.1.3 Compact Trap-Assisted Tunneling Models

For application in circuit simulators, or to catch a quick glimpse at the effects of trap-assisted tunneling, compact models are required. A frequently used expression is based on the work of RICCO *et al.* [193]. They describe the trapping- and detrapping processes by

$$J_{\mathrm{TAT}} = JC_{\mathrm{T}}TC_1(N_{\mathrm{T}} - n_{\mathrm{T}}) = \mathrm{q}\nu n_{\mathrm{T}}TC_2 \ , \tag{3.134}$$

where J is the supply current density at the interface, C_{T} the capture cross section, TC_1 and TC_2 the transmission coefficients from the left and right side of the dielectric to the trap, n_{T} the concentration of trapped electrons which is smaller or equal than the trap concentration N_{T}, and ν their escape frequency. The highest contribution comes from traps which have $TC_1 \approx TC_2$, therefore the trap-assisted tunnel current becomes

$$J_{\mathrm{TAT}} = \mathrm{q}\nu n_{\mathrm{T}}TC = \mathrm{q}\nu C_{\mathrm{T}}N_{\mathrm{T}}\frac{J}{JC_{\mathrm{T}} + \mathrm{q}\nu}TC \ . \tag{3.135}$$

A modified version of this expression was used by GHETTI *et al.* [195, 211]. Other more or less empirical trap-assisted tunneling models based on SILC measurements are presented in [212]. These comprise hopping conduction

$$J = C_1 E_{\mathrm{diel}} \exp\left(-\frac{\mathrm{q}\Phi_{\mathrm{a}}}{\mathrm{k_B}T}\right) \ , \tag{3.136}$$

where Φ_{a} is an activation potential, and the frequently applied POOLE-FRENKEL tunneling formula [212–218]. This model describes the emission of trapped electrons and reads

$$J = AE_{\mathrm{diel}} \exp\left(-\frac{\mathcal{E}_{\mathrm{T}}}{\mathrm{k_B}T}\right) \exp\left(\frac{\mathrm{q}}{\mathrm{k_B}T}\sqrt{\frac{\mathrm{q}E_{\mathrm{diel}}}{\pi\kappa_0 r^2}}\right) \ , \tag{3.137}$$

where r is the refractive index of the dielectric, $\mathcal{E}_{\mathrm{T}}$ is the difference between the conduction band in the dielectric and the trap energy, and the coefficient A depends on the trap concentration. The main motivation to use this expression is that the trap-assisted gate current density was found to be a linear function of the square root of the dielectric field, in contrast to the FOWLER-NORDHEIM tunneling current which is a linear function of the dielectric field. Note, however, that no trapping-detrapping considerations enter this equation.

3.8.2 The Model of JIMÉNEZ *et al.*

A model for trap-assisted inelastic tunneling has been developed by JIMÉNEZ *et al.* [219]. Their model is based on the theory of non-radiative capture and emission of electrons by multiphonon processes [220]. The main difference to the models described before is that it does not require constant capture cross sections as fitting parameters but calculates them for each trap based on the trap energy level and the shape of the energy barrier.

3.8.2.1 Capture and Emission Probabilities

The tunneling model is based on a two-step tunneling process via traps in the dielectric which incorporates energy loss by phonon emission [219]. Fig. 3.17 shows the basic two-step process of an electron tunneling from a region with higher FERMI energy (the cathode) to a region with lower FERMI energy (the anode). To avoid integration in energy, the initial electron energy is assumed to be located at the average kinetic energy, which, for the parabolic dispersion relation (3.1) and the MAXWELLian distribution (3.20), is

$$\frac{\langle \mathcal{E} \rangle}{\langle 1 \rangle} = \frac{\int_0^\infty \mathcal{E} f(\mathcal{E}) g(\mathcal{E}) \, \mathrm{d}\mathcal{E}}{\int_0^\infty f(\mathcal{E}) g(\mathcal{E}) \, \mathrm{d}\mathcal{E}} = \frac{\int_0^\infty \mathcal{E}^{3/2} \exp\left(-\frac{\mathcal{E}}{\mathrm{k_B}T}\right) \mathrm{d}\mathcal{E}}{\int_0^\infty \mathcal{E}^{1/2} \exp\left(-\frac{\mathcal{E}}{\mathrm{k_B}T}\right) \mathrm{d}\mathcal{E}} = \frac{3}{2}\mathrm{k_B}T \, . \tag{3.138}$$

During the capture process (W_c), the difference in total energy between the initial and final state is released by means of phonon emission ($\hbar\omega$). An electron captured by a trap can then be emitted into the anode (W_e).

The rate with which an electron with energy $\mathcal{E}$ is captured by a trap located at position x and energy $\mathcal{E}'$ is given by [221]

$$W_\mathrm{c}(x, \mathcal{E}', \mathcal{E}) = \frac{\pi}{\hbar^2 \omega} |V_\mathrm{e}|^2 \, S \left(1 - \frac{P}{S}\right)^2 I_P(\xi) \exp\left(-(2 f_\mathrm{P} + 1) \, S + \frac{\Delta\mathcal{E}}{2\mathrm{k_B}T}\right) \, . \tag{3.139}$$

Here, S is the HUANG-RHYS factor which characterizes the electron-phonon interaction [222], $\hbar\omega$ is the energy of the phonons involved in the transitions, $\Delta\mathcal{E} = \mathcal{E} - \mathcal{E}'$, and $P = \Delta\mathcal{E}/\hbar\omega$ is the number of phonons emitted due to this energy difference. In the simulations the value of $S\hbar\omega$ was used as fitting parameter.

The population of phonons f_P is given by the BOSE[11]-EINSTEIN[12] statistics

$$f_\mathrm{P} = \left(\exp\left(\frac{\hbar\omega}{k_B T}\right) - 1\right)^{-1} \, . \tag{3.140}$$

[11] SATYENDRA NATH BOSE, *Indian physicist, 1894–1974.*
[12] ALBERT EINSTEIN, German physicist, 1879–1955.

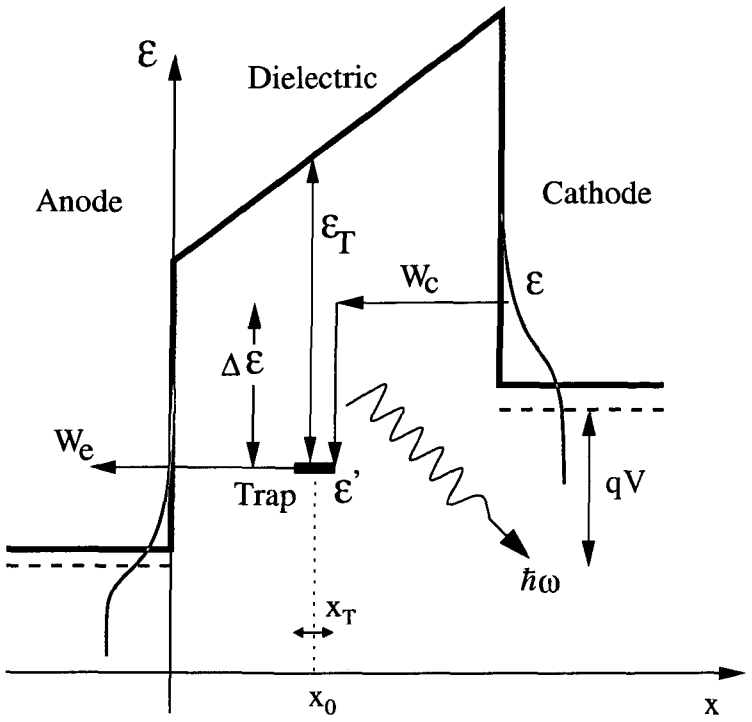


Figure 3.17: The trap-assisted tunneling process.

The function $I_P(\xi)$ is the modified BESSEL[13] function of order P, with

$$\xi = 2S\sqrt{f_P(f_P+1)} \; . \tag{3.141}$$

The term $|V_e|^2$ in (3.139) denotes the transition matrix element which is calculated by an integration over the trap cube [220]

$$|V_e|^2 = 5\pi S(\hbar\omega)^2 \frac{\hbar^2}{2m_{diel}\mathcal{E}_T} \int_{x_0-x_T/2}^{x_0+x_T/2} |\Psi(x)|^2 \, dx \; . \tag{3.142}$$

In this expression x_T denotes the side length of the trap cube, estimated as

$$x_T = \frac{\hbar}{\sqrt{2m_{diel}\mathcal{E}_T}} \left(\frac{4\pi}{3}\right)^{1/3} \; . \tag{3.143}$$

The symbol $\mathcal{E}_T$ denotes the energy difference between the trap energy and the barrier conduction band edge as shown in Fig. 3.17. For the emission of electrons from the trap to the anode, elastic tunneling is assumed. Hence, the probability of emission to the anode is equal to the probability of capture from the anode, which is calculated from (3.139).

[13] FRIEDRICH WILHELM BESSEL, German mathematician, 1784–1846.

The numerical evaluation of (3.142) requires the calculation of the wave functions in the dielectric layer, which, however, degrades the computational efficiency of a multi-purpose device simulator where simulation speed is crucial. To avoid this, the barriers have been transformed to take advantage of the well known solutions for constant potentials. Two cases must be distinguished, namely the case of a trapezoidal barrier and the case of a triangular barrier. The two cases are depicted in Fig. 3.18.

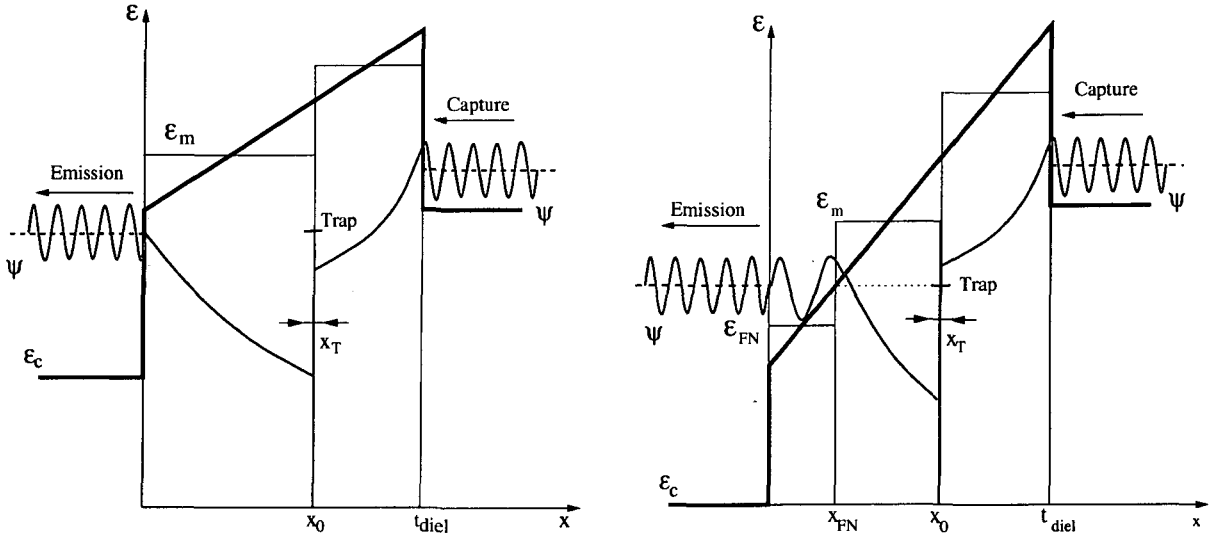


Figure 3.18: The approximate shape of the barrier in the direct (left) and FOWLER-NORDHEIM regime (right).

For capture processes and for emission processes where the electron faces a **trapezoidal barrier**, the barrier is transformed into a step function of height equal to the potential at the middle point between $x = 0$ and $x = x_0$ ($\mathcal{E}_\mathrm{m}$ in the left part of Fig. 3.18), x_0 being the position of the trap inside the dielectric. Assuming

$$\begin{aligned} \Psi(x \leq 0) &= A \sin(k_1 x + \alpha) \ , \\ \Psi(x > 0) &= B \exp(-k_2 x) \ , \end{aligned} \tag{3.144}$$

the wave function at the position of the trap becomes

$$\Psi(x) = A \sin\left(\arctan\left(\frac{m_\mathrm{diel}}{m_\mathrm{eff}} \frac{k_1}{k_2}\right)\right) \exp(-k_2 x) \ , \tag{3.145}$$

where m_diel and m_eff are the electron masses in the dielectric and the neighboring electrode, respectively. The wave numbers are given by

$$\begin{aligned} k_1 &= \frac{1}{\hbar}\sqrt{2 m_\mathrm{eff}(\mathcal{E} - \mathcal{E}_\mathrm{c})} \ , \\ k_2 &= \frac{1}{\hbar}\sqrt{2 m_\mathrm{diel}(\mathcal{E}_\mathrm{m} - \mathcal{E})} \ . \end{aligned} \tag{3.146}$$

For emission processes in which the barrier is **triangular** (the electron energy is above the dielectric conduction band at some point between the trap and the anode), two regions in the dielectric must be distinguished. The first one, between the interface at $x = 0$ and the point $x = x_{\mathrm{FN}}$ (see the right part of Fig. 3.18) has the height $\mathcal{E}_{\mathrm{FN}}$. The height of the approximated barrier in the other region is then the value of the barrier, $\mathcal{E}_{\mathrm{m}}$, in the middle point between $x = x_{\mathrm{FN}}$ and the position of the trap $x = x_0$. With this new barrier and the assumptions for the wave functions in the three regions

$$\Psi(x \leq 0) = A \sin(k_1 x + \alpha_1) \ , \tag{3.147}$$

$$\Psi(0 < x \leq x_{\mathrm{FN}}) = B \sin(k_2 x + \alpha_2) \ , \tag{3.148}$$

$$\Psi(x_{\mathrm{FN}} < x \leq x_0) = C \exp(-k_3(x - x_{\mathrm{FN}})) \ , \tag{3.149}$$

the wave function at the position of the trap becomes

$$\Psi(x) = A \frac{\sin \alpha_1}{\sin \alpha_2} \sin\left(k_2 x_{\mathrm{FN}} + \alpha_2\right) \exp\left(-k_3(x - x_{\mathrm{FN}})\right) \ , \tag{3.150}$$

with the symbols

$$\begin{aligned} \alpha_1 &= \arctan\left(\frac{k_1}{k_2} \tan \alpha_2\right) \ , \\ \alpha_2 &= \arctan\left(\frac{k_2}{k_3}\right) - k_2 x_{\mathrm{FN}} \ . \end{aligned} \tag{3.151}$$

The corresponding wave numbers are given as

$$\begin{aligned} k_1 &= \frac{1}{\hbar}\sqrt{2m_{\mathrm{eff}}(\mathcal{E} - \mathcal{E}_{\mathrm{c}})} \ , \\ k_2 &= \frac{1}{\hbar}\sqrt{2m_{\mathrm{diel}}(\mathcal{E} - \mathcal{E}_{\mathrm{FN}})} \ , \\ k_3 &= \frac{1}{\hbar}\sqrt{2m_{\mathrm{diel}}(\mathcal{E}_{\mathrm{m}} - \mathcal{E})} \ . \end{aligned} \tag{3.152}$$

Using expression (3.145) and (3.150), the integration in (3.142) can be performed analytically which allows the capture and emission probabilities to be calculated without the need for numerical integration.

3.8.2.2 Capture and Emission Times

Once the capture and emission probabilities have been obtained, the corresponding times can be calculated. The inverse of the capture time is given by [219, 223]

$$\tau_{\mathrm{c}}^{-1}(x) = \int_{\mathcal{E}'}^{\infty} W_{\mathrm{c}}(x, \mathcal{E}', \mathcal{E}) g_{\mathrm{c}}(\mathcal{E}) f_{\mathrm{c}}(\mathcal{E}) \, \mathrm{d}\mathcal{E} \ , \tag{3.153}$$

where $g_c(\mathcal{E})$ denotes the two-dimensional density of states and $f_c(\mathcal{E})$ the electron energy distribution function in the cathode. For the above stated assumption that all electrons are captured from the same energy level $\mathcal{E}_c + 3/2 k_B T$ in the cathode, this expression can be approximated by

$$\tau_{\mathrm{c}}^{-1}(x) \approx W_{\mathrm{c}}\left(x, \mathcal{E}', \mathcal{E}_{\mathrm{c}} + \frac{3}{2} k_B T\right) n_c \ , \tag{3.154}$$

where n_c is the sheet carrier concentration in the cathode, which is determined by the transport model used in the device simulator. The inverse of the emission time is [219]

$$\tau_e^{-1}(x) = \int_{-\infty}^{\mathcal{E}'} W_e(x, \mathcal{E}', \mathcal{E}) g_a(\mathcal{E}) \left(1 - f_a(\mathcal{E})\right) \mathrm{d}\mathcal{E} \ . \tag{3.155}$$

Assuming $f_a(\mathcal{E}) \approx 0$ in the anode and elastic tunneling for the emission process ($\mathcal{E} = \mathcal{E}'$), the emission time becomes

$$\tau_e^{-1}(x) \approx W_e(x, \mathcal{E}', \mathcal{E}') g_a(\mathcal{E}') \hbar\omega \ , \tag{3.156}$$

where the energy loss is restricted to values less than $\hbar\omega$. To check the validity of the approximations for the wave functions, the resulting capture and emission times have been compared to results using a SCHRÖDINGER-POISSON solver for a MOS capacitor with the parameters $\mathcal{E}_T$=2.8 eV, $S\hbar\omega$=1.6 eV, and a trap concentration of $N_T = 10^{19}\,\mathrm{cm}^{-3}$. As can be seen in Fig. 3.19, the analytical and the numerical results are very close. Electrons are captured from the right and emitted to the left in this figure. Thus, for traps near the right side of the barrier the capture time is very low and the emission time is very high. The oscillations in the emission time for high bias are due to the fact that in this regime, the energy barrier has a triangular shape which gives rise to an oscillating wave function, in contrast to the decaying wave function for a trapezoidal barrier.

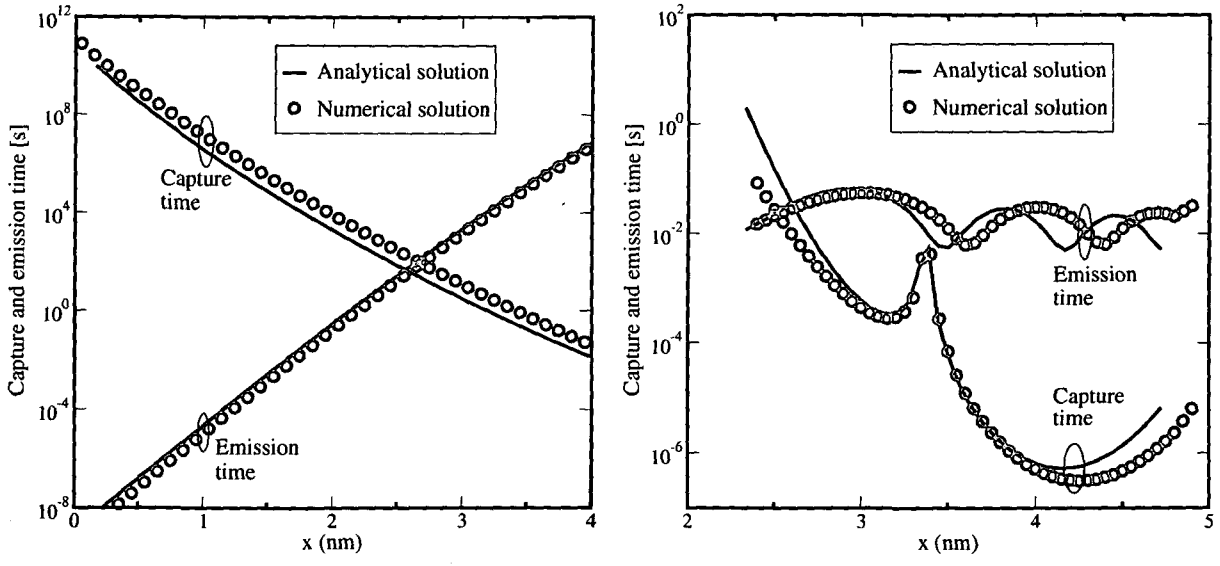


Figure 3.19: Comparison of the analytic solution with a numerical solution for the capture and emission times at a gate bias of 3 V (left) and 7 V (right).

3.8.2.3 Steady-State Current

The total steady-state tunneling current is derived as the sum of the trap-assisted tunneling current (3.130) and the direct tunneling current computed from the TSU-ESAKI formula (3.13)

$$J = J_{\mathrm{TAT}} + J_{\mathrm{Tsu-Esaki}} \,. \tag{3.157}$$

Fig. 3.20 shows the dependence of the gate current density on the model parameters $\mathcal{E}_{\mathrm{T}}$ (trap energy level) and $S\hbar\omega$ for a fixed phonon energy of $\hbar\omega$=10 meV in an MOS capacitor. For a low trap energy level traps are located near the conduction band edge in the dielectric, and direct tunneling prevails. With increasing trap energy level, the trap-assisted component becomes stronger and exceeds the direct tunneling current for low bias. The current density shows a peak at low bias which is due to the alignment of the trap energy level with the cathode conduction band edge. The HUANG-RHYS factor has only a minor influence on the results, as shown in the right part of Fig. 3.20.

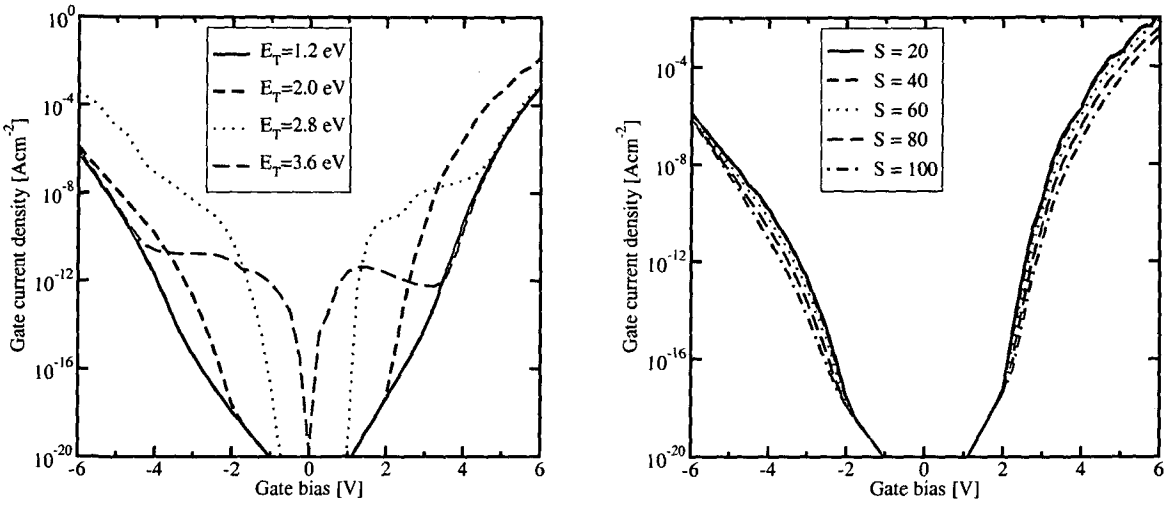


Figure 3.20: Dependence of the tunneling current on the trap energy level (left) and on the HUANG-RHYS factor for a fixed phonon energy of 10 meV (right).

3.8.2.4 Transient Current

Models of trap-assisted transitions are commonly employed to calculate steady-state SILC in MOS capacitors, while transient SILC has hardly been studied [194, 205]. However, transient tunneling current becomes important at high switching speed where the transients of the trap charging and discharging processes may degrade signal integrity. For the calculation of transient SILC it is necessary to calculate capture and emission times at each time step. Considering a spatial trap distribution $N_{\mathrm{T}}(x)$ across the dielectric layer, the rate equation for the concentration of occupied traps at position x reads

$$N_{\mathrm{T}}(x)\frac{\mathrm{d}f_{\mathrm{T}}(x,t)}{\mathrm{d}t} = N_{\mathrm{T}}(x)\left(1 - f_{\mathrm{T}}(x,t)\right)\tau_{\mathrm{c}}^{-1}(x,t) - N_{\mathrm{T}}(x)f_{\mathrm{T}}(x,t)\tau_{\mathrm{e}}^{-1}(x,t) \,, \tag{3.158}$$

where $f_T(x,t)$ is the trap occupancy function and $\tau_c(x,t)$ and $\tau_e(x,t)$ are the inverse capture and emission times of electrons by a trap placed at position x. In the static case capture and emission processes are in equilibrium and $df_T(x,t)/dt = 0$. In the transient case, however, capture and emission times include transitions from the cathode and the anode (compare Section 3.8.1.2 and Fig. 3.16)

$$\begin{aligned} \tau_c^{-1}(x,t) &= \tau_{ca}^{-1}(x,t) + \tau_{cc}^{-1}(x,t) \ , \\ \tau_e^{-1}(x,t) &= \tau_{ea}^{-1}(x,t) + \tau_{ec}^{-1}(x,t) \ , \end{aligned} \tag{3.159}$$

where τ_{ca} and τ_{cc} are the capture times to the anode and to the cathode, and τ_{ea} and τ_{ec} the corresponding emission times. To calculate the local trap occupancy, the differential equation (3.158) must be solved. If the capture and emission times τ_c^{-1} and τ_e^{-1} are constant over time, like in a discharging process with a constant potential distribution, the solution of (3.158) can be given in a closed form

$$f_T(x,t) = f_T(x,0)\exp\left(-\frac{t}{\tau_m(x,t)}\right) + \frac{\tau_m(x,t)}{\tau_c(x,t)}\left(1-\exp\left(-\frac{t}{\tau_m(x,t)}\right)\right) \ , \tag{3.160}$$

with $\tau_m^{-1} = \tau_c^{-1} + \tau_e^{-1}$.

A more general approach is to look at the change of the trap distribution at discrete time steps. Integration of (3.158) in time between t_i and t_{i+1} and changing to discrete time steps yields

$$f_T(x,t_i) - f_T(x,t_{i-1}) \approx \tau_c^{-1}(x,t_{i-1})\Delta t_i - \tau_m^{-1}(x,t_{i-1})\overline{f_i}\Delta t_i \ ,$$

where the abbreviations $\Delta t_i = t_i - t_{i-1}$ and $\overline{f_i} = \left(f_T(x,t_i) + f_T(x,t_{i-1})\right)/2$ have been used. Thus it is possible to write the trap distribution over time in the following recursive manner:

$$f_T(x,t_i) = A_i + B_i f_T(x,t_{i-1}) \ , \tag{3.161}$$

where the symbols A_i, B_i, and C_i are calculated from

$$\begin{aligned} A_i &= \frac{\tau_c^{-1}(x,t_i)\Delta t_i}{1+C_i} \ , \\ B_i &= \frac{1-C_i}{1+C_i} \ , \\ C_i &= \frac{\tau_m^{-1}(x,t_i)\Delta t_i}{2} \ . \end{aligned} \tag{3.162}$$

Once the time-dependent occupancy function in the dielectric is known, the tunnel current through each of the interfaces is

$$J_{TAT,Anode}(t) = q\int_0^{t_{diel}} N_T(x)\left(\tau_{ca}^{-1}(x,t) - f_T(x,t)\left(\tau_{ca}^{-1}(x,t) + \tau_{ea}^{-1}(x,t)\right)\right) dx \ , \tag{3.163}$$

$$J_{TAT,Cathode}(t) = q\int_0^{t_{diel}} N_T(x)\left(\tau_{cc}^{-1}(x,t) - f_T(x,t)\left(\tau_{cc}^{-1}(x,t) + \tau_{ec}^{-1}(x,t)\right)\right) dx \ . \tag{3.164}$$

3.9 Model Comparison

This chapter outlined a number of tunneling models useful for the simulation of tunneling in semiconductor devices. For practical device simulation, however, it is often not clear which model to select for the application at hand. Therefore, Table 3.3 summarizes the main model features and also gives the approximate computational effort. The following points can be concluded:

- Especially the FOWLER-NORDHEIM, SCHUEGRAF, and FRENKEL-POOLE models have a very low computational effort since they are compact models. However, they do not correctly reproduce the device physics and can only be used after careful calibration.
- The TSU-ESAKI formula with the analytical WKB or GUNDLACH method for the transmission coefficient combines moderate computational effort with reasonable accuracy. This approach can be used for the simulation of tunneling in devices with single-layer dielectrics.
- The inelastic TAT model allows simulation of all effects related with traps in the dielectric and, due to the analytical calculation of the overlap integral, poses only moderate computational effort. This model can be used for the simulation of leakage in EEPROMs or *trap-rich dielectric devices* (see Section 5.2.2.1).
- The TSU-ESAKI model with the numerical WKB, transfer-matrix, or QTB method to calculate the transmission coefficient represents the most accurate method usable for the simulation of tunneling through dielectric stacks, however, with high computational effort. The transfer-matrix method should not be used due to its poor numerical stability.

	FOWLER-NORDHEIM model	SCHUEGRAF model	TSU-ESAKI analytic WKB	TSU-ESAKI GUNDLACH	TSU-ESAKI numeric WKB	TSU-ESAKI transfer-matrix	TSU-ESAKI QTBM	Inelastic TAT (Section 3.8.2)	FRENKEL-POOLE
FN tunneling	✓	✓	✓	✓	✓	✓	✓		
Direct tunneling		✓	✓	✓	✓	✓	✓		
EVB tunneling process			✓	✓	✓	✓	✓		
QM current oscillations				✓		✓	✓		
Dielectric stacks					✓	✓	✓		
Numerical stability						–			
Trap-assisted tunneling								✓	✓
Trap occupancy modeling								✓	
Transient TAT								✓	
Computational effort	**low**	**low**			**high**	**high**	**high**		**low**

Table 3.3: A hierarchy of tunneling models and their properties.

'If programming in Pascal is like being put in a straight jacket, then programming in C is like playing with knives, and programming in C++ is like juggling chain saws.'

Anonymous

Chapter 4

Implementation

FOR A RIGOROUS STUDY of tunneling effects in modern semiconductor devices it is mandatory to consider arbitrary device geometries, which is only possible using a general-purpose device simulator. This chapter describes the implementation of the outlined tunneling models into the device simulator MINIMOS-NT. First, a brief description of MINIMOS-NT is given. Then, the discretization of the tunneling current density in dielectrics is described, covering tunneling through single segments and stacked segments and followed by the trap-assisted tunneling interface. Finally, the developed closed- and open-boundary SCHRÖDINGER solver is briefly explained.

4.1 The Device Simulator MINIMOS-NT

The general-purpose device simulator MINIMOS-NT [224] is the successor of the highly successful device simulator MINIMOS 6 [225]. The project was started in 1992 by FISCHER [226] and SIMLINGER [227]. In contrast to its predecessor, MINIMOS-NT is able to analyze arbitrary-shaped devices with the number of grid points being limited only by the available memory. It solves the drift-diffusion or energy-transport equations using the box-integration method and the direct, BiCGStab, or GMRES numerical solving routines. It was extended to allow the solution of the lattice heat-flow equation by KNAIPP in 1998 [228] and the simulation of mixed-mode circuits by GRASSER in 1999 [229]. It was further improved by a comprehensive material database including a wide range of alloy semiconductors by PALANKOVSKI in 2000 [230], AC small-signal analysis and complex solver routines by WAGNER in 2001 [231], and an advanced input deck language and the extension to three dimensional device simulation by KLIMA in 2002 [232]. For the solving process the user can choose among several iteration schemes to speed-up and improve the convergence of the simulation. MINIMOS-NT uses the PIF (PROFILE INTERCHANGE FORMAT) file format for device description which has been introduced by DUVALL in 1988 [233]. Several grid types such as ortho-product or triangular grids can be supplied. Ongoing work is concentrated on the coupling of MINIMOS-NT to the WAFER-STATE SERVER (WSS) [234, 235], the development of advanced gridding routines, and on the coupling of MINIMOS-NT to multi-dimensional SCHRÖDINGER and Monte Carlo modules.

4.2 The Tunneling Model

The main problem for the integration of tunneling current models in a device simulator such as MINIMOS-NT is that tunneling is a non-local effect. In contrast to the current density described by the drift-diffusion (2.4) or energy-transport model (2.8), the current density at a certain point does not only depend on quantities at the same point, but on geometrical properties such as the thickness of the segment considered for tunneling. Thus, the tunneling current contribution cannot be simply derived from local quantities alone. In MINIMOS-NT the tunneling current is calculated between two boundaries of insulator or semiconductor segments. The boundaries are either specified by the user (see Appendix D) or found automatically. In the latter case the tunneling boundaries are identified as the first two boundaries of the specified segment to neighboring non-insulating materials which have the smallest distance[1]. For each grid node at the specified boundary, the node on the other boundary with minimum distance is selected as partner node. It may happen that some nodes share their partner nodes, such as the nodes i, and $i+1$ in Fig. 4.1. Thus, this implementation is valid for non-orthogonal grids, too.

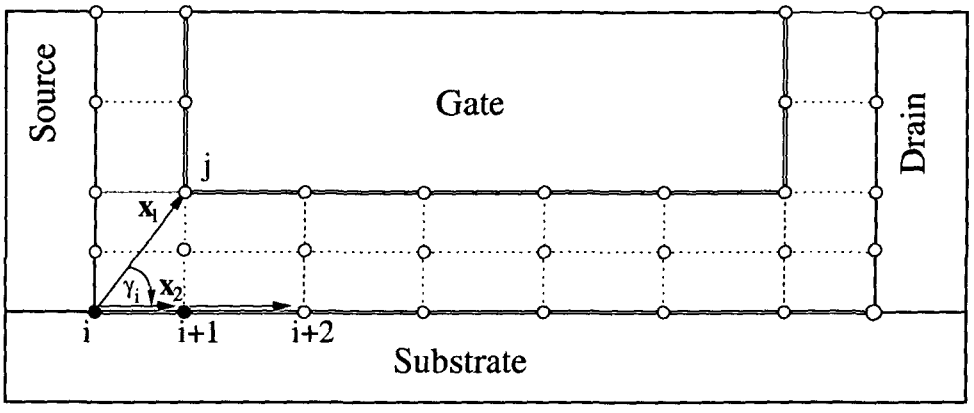


Figure 4.1: Boundary node - partner node pairs. The considered boundaries are indicated by bold lines.

The physical quantities at the neighboring segments, such as the carrier concentration, the electrostatic potential, and the carrier temperature, are passed to the tunneling model which is evaluated for each boundary grid point. Then, the tunneling current density is calculated by one of the models described in Section 3 and the total tunneling current is found by summation of the current density along the boundary and multiplication with the area of the grid element. A projection factor α_i is calculated for every node i to account for pair nodes which do not lie directly opposite to each other:

$$\alpha_i = \left| \frac{\mathbf{x_1} \cdot \mathbf{x_2}}{x_1 x_2} \right| = |\cos(\gamma_i)| \ , \tag{4.1}$$

[1]Note that, especially in the three-dimensional version of MINIMOS-NT, there may be several boundaries with equal minimum distance. Furthermore, the boundaries with the minimum distance are not necessarily the boundaries with the highest tunnel current density. A manual specification of the boundaries may be necessary to avoid ambiguities.

where $\mathbf{x_1}$ points from the boundary node to the partner node and $\mathbf{x_2}$ to the next node on the boundary. In Fig. 4.1, for example, the tunneling current is calculated for the boundary nodes i and $i+1$ with respect to the partner node j.

The total tunneling current is calculated by a summation along the boundary with length L which consists of N segments

$$I = w \int_0^L J(x)\,\mathrm{d}x \approx w \sum_{i=1..N} J_i \Delta x_i \cos(\gamma_i) \;, \tag{4.2}$$

where w is the gate width, J_i the local tunneling current density, and Δx_i the interface length associated with the node i. The local tunneling current density J_i is added self-consistently to the continuity equation of the neighboring segments by means of an additional recombination term $R_{\mathrm{tun}} = J_{\mathrm{tun}}/\mathrm{q}w$

$$\begin{aligned} \nabla \cdot \mathbf{J_n} &= \mathrm{q}R + \mathrm{q}\frac{\partial n}{\partial t} + \mathrm{q}R_{\mathrm{tun,n}} \;, \\ \nabla \cdot \mathbf{J_p} &= -\mathrm{q}R - \mathrm{q}\frac{\partial p}{\partial t} - \mathrm{q}R_{\mathrm{tun,p}} \;. \end{aligned} \tag{4.3}$$

In MINIMOS-NT the NEWTON[2] method is used to calculate the solution vector consisting of n, p, and ϕ at step $k+1$ from the matrix equation

$$\begin{pmatrix} \phi_{k+1} \\ n_{k+1} \\ p_{k+1} \end{pmatrix} = \begin{pmatrix} \phi_k \\ n_k \\ p_k \end{pmatrix} - \underbrace{\begin{pmatrix} \frac{\partial f_\phi}{\partial \phi} & \frac{\partial f_\phi}{\partial n} & \frac{\partial f_\phi}{\partial p} \\ \frac{\partial f_n}{\partial \phi} & \frac{\partial f_n}{\partial n} & \frac{\partial f_n}{\partial p} \\ \frac{\partial f_p}{\partial \phi} & \frac{\partial f_p}{\partial n} & \frac{\partial f_p}{\partial p} \end{pmatrix}^{-1}}_{\underline{J}^{-1}} \cdot \begin{pmatrix} f_\phi(\phi_k, n_k, p_k) \\ f_n(\phi_k, n_k, p_k) \\ f_p(\phi_k, n_k, p_k) \end{pmatrix}, \tag{4.4}$$

where f_ϕ, f_n, and f_p denote the control equations determining the electrostatic potential, the electron concentration, and the hole concentration. Since R_{tun} modifies all solution variables, the JACOBIan[3] $\underline{J}$ must be modified to achieve better convergence of the NEWTON solver. Therefore, the derivatives of the additional recombination term with respect to the potential, electron concentration, and hole concentration

$$\frac{\partial R_{\mathrm{tun}}}{\partial \phi}, \qquad \frac{\partial R_{\mathrm{tun}}}{\partial n}, \qquad \frac{\partial R_{\mathrm{tun}}}{\partial p}$$

have to be calculated. For the FOWLER-NORDHEIM, SCHUEGRAF, and FRENKEL-POOLE model, the derivatives are calculated analytically while for all other models they are calculated numerically.

[2] Sir ISAAC NEWTON, English mathematician and physicist, 1643–1727.
[3] CARL GUSTAV JACOB JACOBI, German mathematician, 1804–1851.

4.2.1 Single Segment Tunneling

In MINIMOS-NT the tunneling current density is calculated between two boundaries of a segment which has been specified by the user (see the description of the user interface in Appendix D). Unlike other models, however, the tunneling current density formulae outlined in Section 3 depend on physical quantities from neighboring segments. Therefore, the concept of **neighbor quantities** has been introduced: First, the segment where tunneling is calculated is — arbitrarily — assigned a reference and an opposite boundary, see Fig. 4.2. Interface models are called which transfer the necessary quantities of the reference and opposite segment to the tunneling segment. This is done by additional equations in the system matrix. The neighbor quantities are

- the electrostatic potential,
- the electron and hole concentration,
- the conduction and valence band edge,
- the lattice temperature,
- the electron and hole temperature (for energy-transport simulation),
- the electron and hole effective density of states, and
- the dielectric permittivity (for calculation of the image force correction energy).

In the tunneling model the tunneling current density is calculated by one of the models presented above for all points along a boundary node – partner node pair. The resulting current density is added as a generation or recombination term to the continuity equations of the reference and opposite segments as described above. For neighboring metal segments, the tunneling current is directly added to the contact current. Again, this step is achieved by means of additional matrix entries.

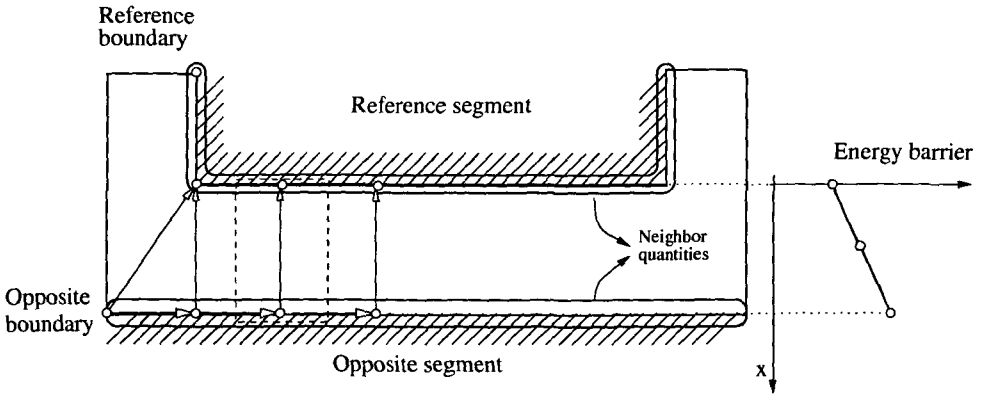


Figure 4.2: Tunneling through a single segment. After identifying the reference and opposite boundary, neighbor quantities are handed over to the tunneling model.

4.2.2 Stacked Segment Tunneling

As outlined in Section 2 advanced CMOS devices apply stacks of alternative dielectric materials and silicon dioxide to achieve a large physical, but small electrical thickness of gate dielectrics. Furthermore, non-volatile memories rely on stacked gate dielectrics to achieve asymmetry between the on- and off-state (see Section 5.2.2.3). Tunneling through such dielectric stacks requires models such as the numerical WKB method, the transfer-matrix method, or the QTBM, since the energy barrier has a non-linear shape.

MINIMOS-NT allows the definition of rectangular dielectric stacks consisting of an arbitrary number of independent segments, as shown in Fig. 4.3 (see also the user interface in Appendix D). The tunneling model, however, must only be evaluated once. Therefore, the segment with the highest index in the stack is chosen as master segment. As in the single-segment case, a reference and an opposite boundary is assigned to the stack. Only at these boundaries, the neighbor quantities are transferred to the master segment.

Further quantities which are necessary for tunneling, such as the conduction and valence band edge, the distance from the reference boundary, or the trap concentration, are transferred from each stack member segment to the master segment. A two-dimensional array is built up which describes the quantities in the whole stack region. In the master segment finally the chosen tunneling current model is evaluated and the calculated tunneling current is transferred back to the boundary node and partner node located at the reference and opposite boundary. There it is added to the continuity equation of the neighboring segments or to the contact current in case of metal contacts as described above.

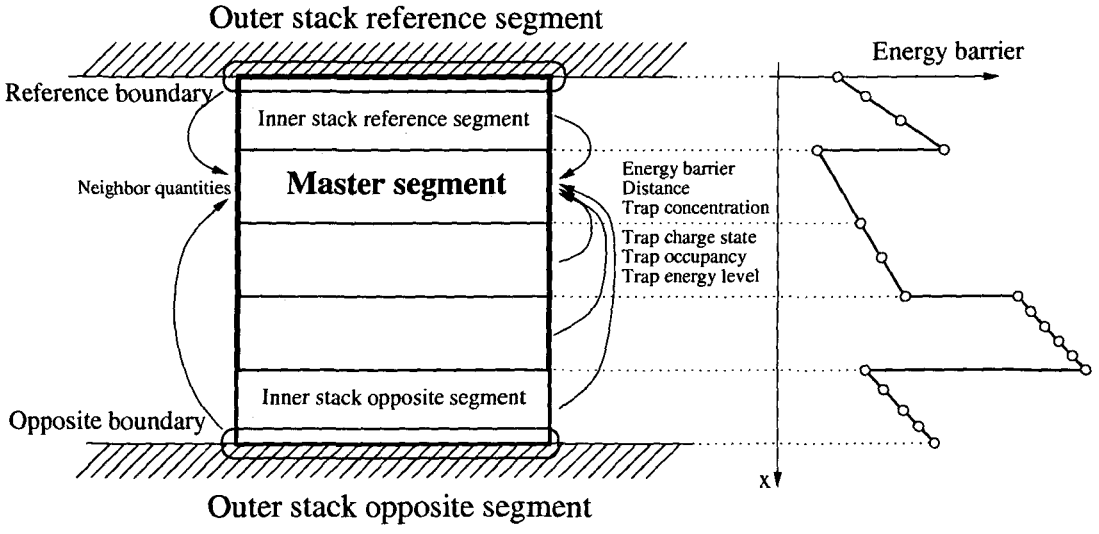


Figure 4.3: A stack consisting of five segments. The neighbor quantities are transferred from the outer stack reference and opposite segment to the master segment. Furthermore, the energy barrier, distance, trap concentration, trap charge state, and trap occupancy is transferred from each stack member segment to the master segment.

4.2.3 Trap-Assisted Tunneling

For the calculation of trap-assisted tunneling, several additional quantities are necessary. These are

- the trap charge state,
- the trap concentration,
- the trap energy level, and
- the trap occupancy.

The trap charge state is constant — either positive, neutral, or negative. The trap concentration and the trap energy level are also constant and can be specified by the user (see Appendix D). These quantities are initialized at startup and do not change.

The trap occupancy f_T is also initialized at startup. In each iteration the charge of occupied traps is included in the right hand side of the POISSON equation according to

$$\nabla(\kappa\nabla\phi) = q(n - p - C) + N_T f_T Q_T \ , \tag{4.5}$$

where N_T is the trap concentration, f_T the trap occupancy, and Q_T the trap charge state. If a trap-assisted tunneling model is evaluated in a transient simulation, the values of the trap occupancy change according to (3.158). For electron tunneling occupied neutral or positive traps become negative or neutral. For hole tunneling occupied neutral or negative traps become positive or neutral. This mechanism is shown in Fig. 4.4. However, a trap is only allowed to capture one carrier, so a negative trap cannot become positive and *vice versa*. For stacked segments, the trap occupancy and the trap charge state are transferred back to their segments after the evaluation of the tunneling model.

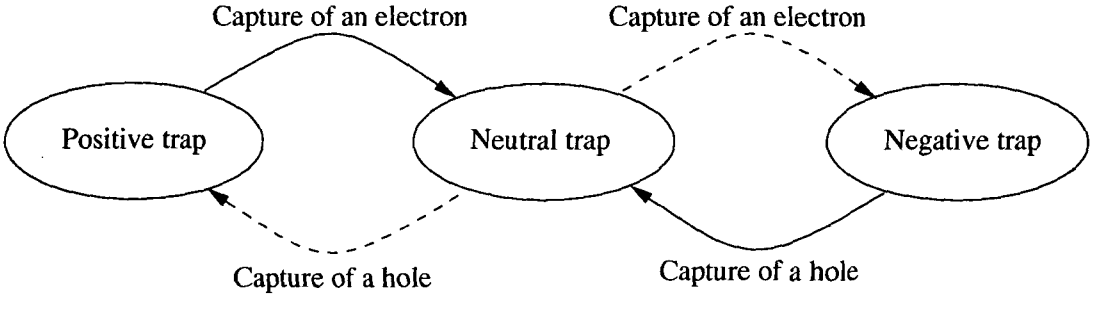


Figure 4.4: Positive, neutral, and negative trap charge states. Positive traps cannot become negative and *vice versa*.

A flow chart of the tunneling model in MINIMOS-NT is shown in Fig. 4.5. The functionality has been implemented in several steps. First, the tunneling segments, stacks, boundaries, and master segments are identified. Then, the neighbor quantities are transferred to the master segment, which is done by special interface models.

After this step the tunneling model is evaluated for all boundary node – partner node pairs. Interface routines transfer the calculated tunnel current density to the continuity equation of the neighboring segments, or directly add it to the contact current if the neighboring segment is a metal.

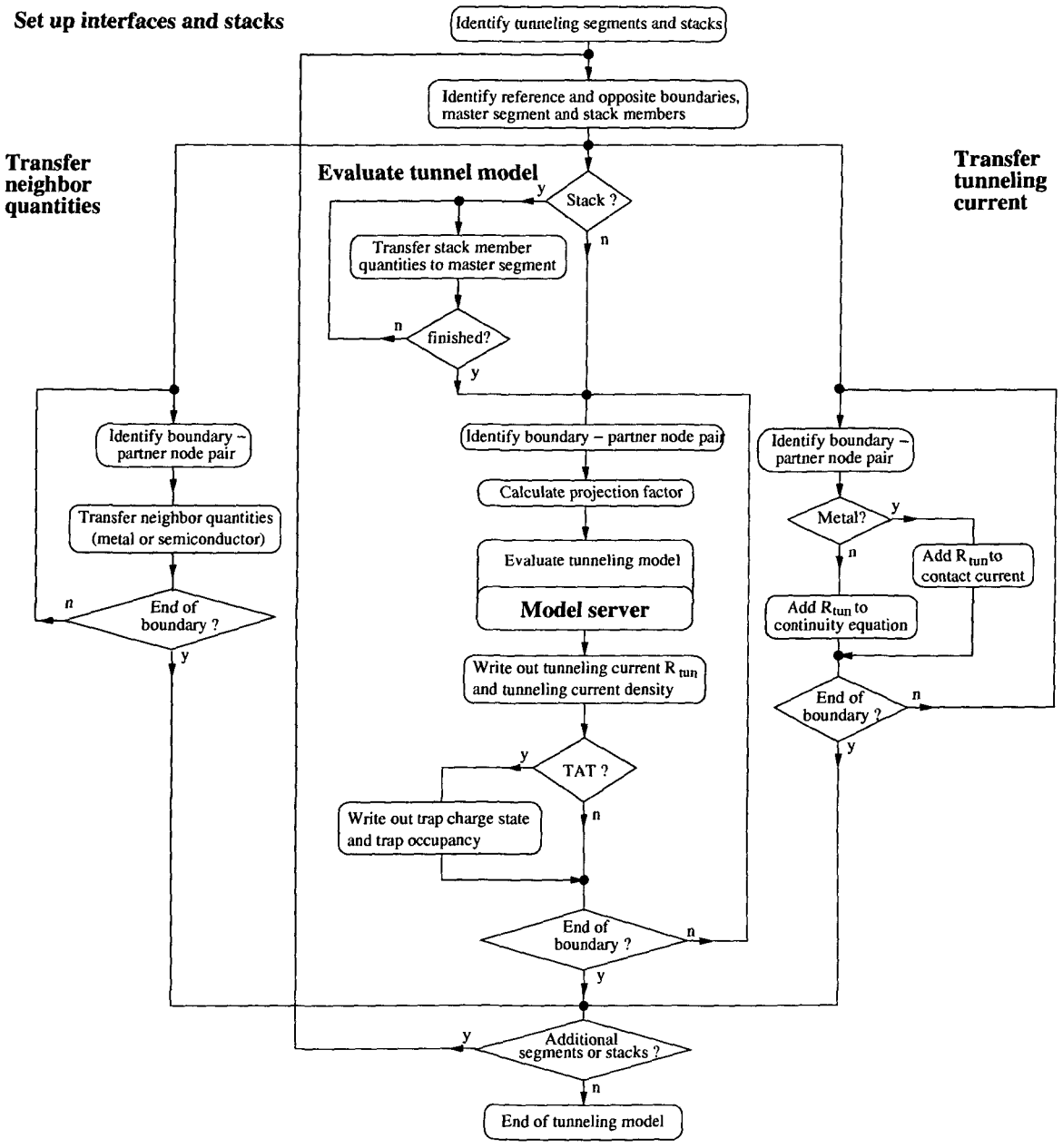


Figure 4.5: Flowchart of the tunneling model in MINIMOS-NT.

4.3 The SCHRÖDINGER Solver

For the calculation of the transmission coefficient of non-linear energy barriers, the stationary one-dimensional SCHRÖDINGER equation must be solved. To allow an easy extension to two- or three-dimensional problems, a multi-dimensional SCHRÖDINGER solver based on the finite-difference discretization has been developed. For the calculation of the transmission coefficient using the QTBM (see Section 3.5.4) it is necessary to solve SCHRÖDINGER's equation with open boundary conditions, while for the evaluation of tunneling from quasi-bound states as described in Section 3.6.3 it is necessary to solve the same problem with closed boundary conditions. Therefore, a solver which allows flexible treatment of both cases has been developed.

4.3.1 Open and Closed Boundary conditions

First, the closed-boundary matrix equation (3.111) is set up, as indicated by the one-dimensional energy barrier $W(x)$ in Fig. 4.6. There are closed boundary conditions at the points 0 and 9, respectively. If the system is coupled to a reservoir at so called connection points, injection points must be given which determine the values of W, $\mathcal{E}_\mathrm{f}$, and m at the reservoir. As described in Section 3.5.4, the coupling entries are calculated by expressions such as (3.95) and (3.96), where the values of the wave vector are

$$k_j = \frac{1}{\hbar}\sqrt{2m_j(\mathcal{E} - W_j)}\ . \tag{4.6}$$

Note that these values may be complex. Injection points are stored in a table which holds the information about the electrostatic potential, the electron mass, and the FERMI level at the injection point. If the transmission coefficient has to be calculated, the points which are considered for tunneling — the boundary nodes and their partner nodes — are used to set up these injection and connection points. The transmission coefficient is then calculated from the wave functions entering and leaving the simulation domain.

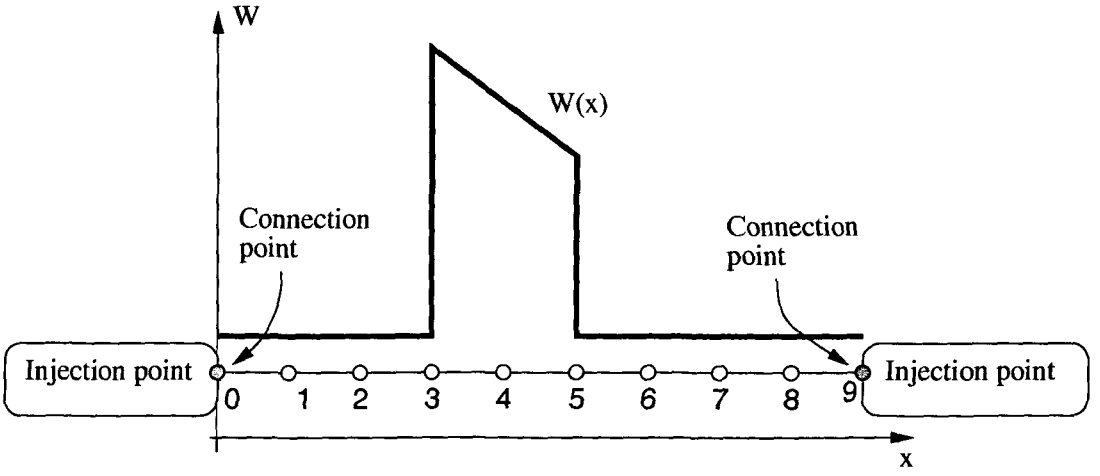


Figure 4.6: One-dimensional energy barrier: Injection points are coupled to connection points.

4.3.2 System HAMILTONian

The HAMILTONian for a system consisting of n grid points coupled to a reservoir at m of these points is described by the following matrix equation, where a simple one-dimensional finite-difference discretization was performed. The lines indicate non-zero elements.

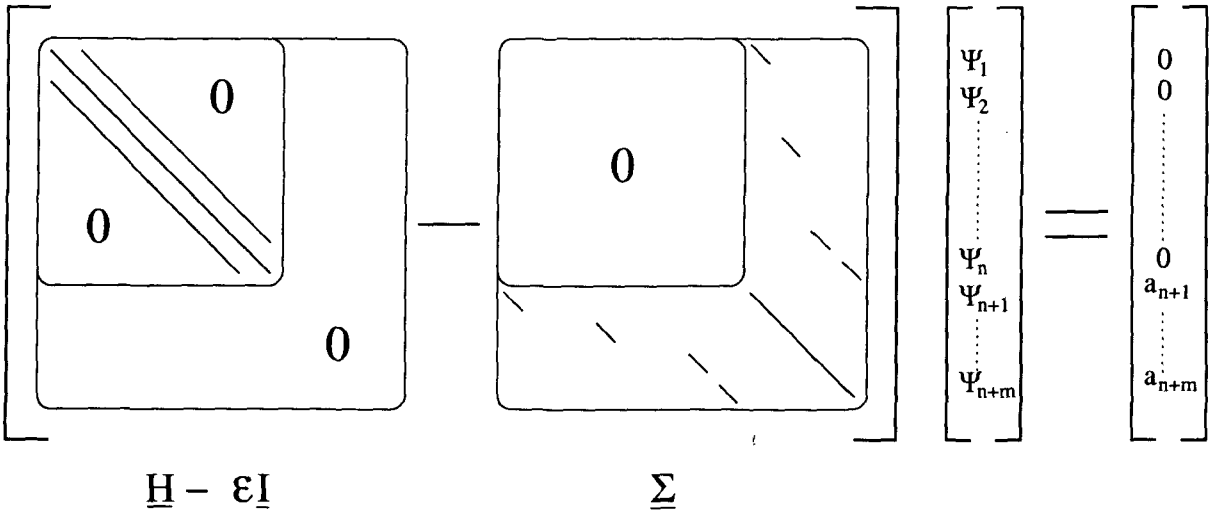


Figure 4.7: Matrix equation for the system HAMILTONian.

The applicability of this solver is therefore not limited to the calculation of transmission coefficients **or** the calculation of eigenvalues and life times, but can perform both operations based on the same data. If the right hand side of one of the points is set to a value $\neq 0$, the transmission coefficient can be calculated according to the value of the wave function at the corresponding node. Furthermore, the module allows to calculate the wave function and the carrier concentration for both open- and closed-boundary cases.

Fig. 4.8 shows a flowchart of a possible application of the Schrödinger solver module where optional modules are indicated by dotted boxes. At the beginning the constructor is invoked to initialize the variables and the memory for the barrier is allocated. In the next step the closed-boundary HAMILTONian is set up. This step has an interface to read the energy barrier from MINIMOS-NT, but it is also possible to specify the barrier manually. Optionally the values in the barrier can be checked and printed to a file. By means of the **open** flag the open- and closed-boundary solver is distinguished.

If injection points are added, the equation system is solved by means of a complex solver. Otherwise the eigenvalues of the closed system are found using an eigenvalue solver (see Section 4.3.3). Both solvers are part of the numerical library of MINIMOS-NT. In both cases the carrier concentration and the wave function can be calculated, while the transmission coefficient can only be calculated for the open system and is directly returned to the tunneling model in MINIMOS-NT.

The output of the program consists of eigenvalues, wave functions, the transmission coefficient, and the carrier concentration. It can either be written in CRV-format (one-dimensional, for use in the program XCRV [224]), in PIF-format (two-dimensional, for use in the program XPIF2D [224]), in DX-format (three-dimensional, for use in the program DATA EXPLORER [236]), or in WSS-format (three-dimensional, for use in the program SMARTVIEW [237]).

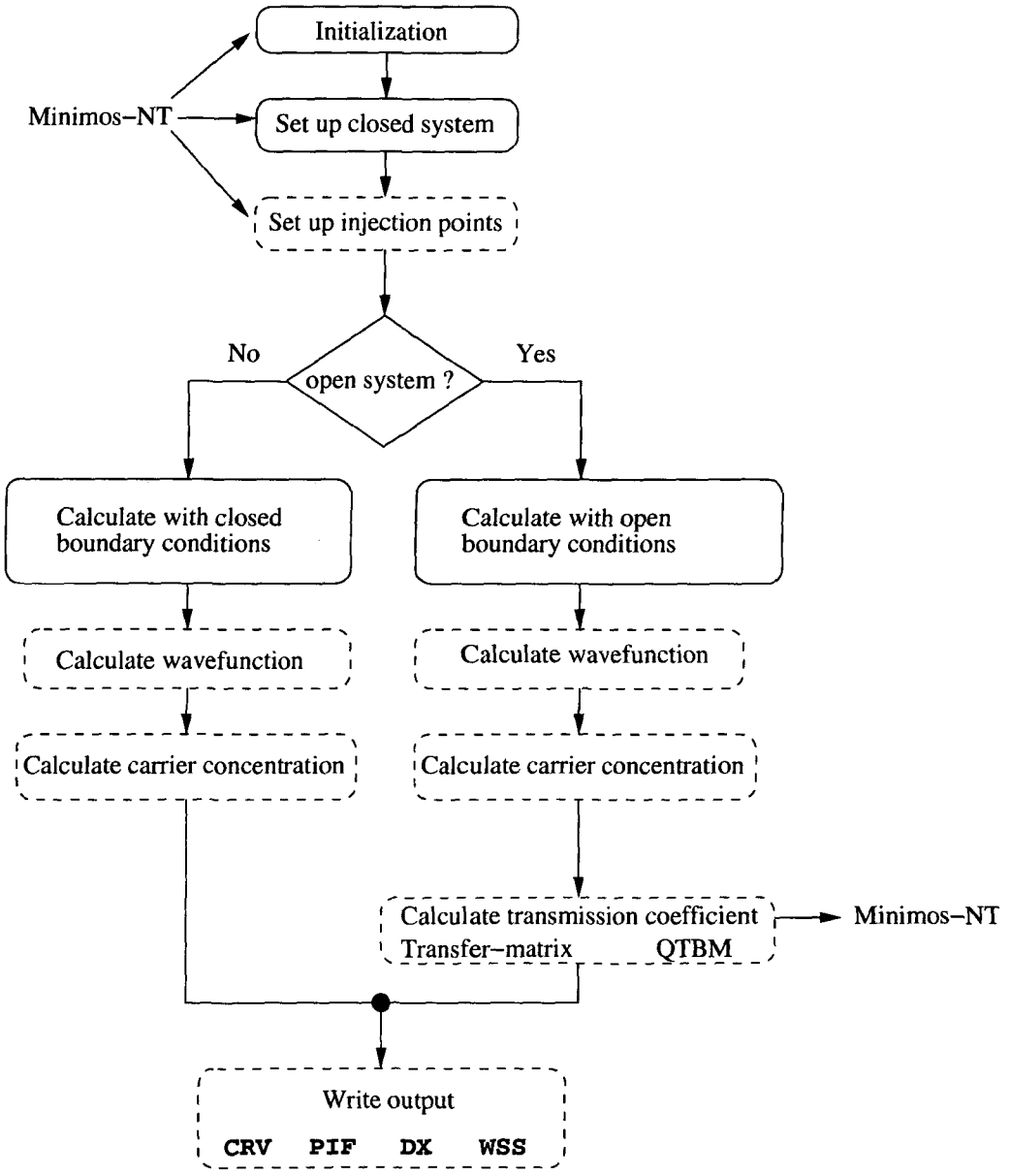


Figure 4.8: Flow chart of the SCHRÖDINGER solver. Optional modules are indicated by dotted boxes.

4.3.3 The Eigenvalue Solver

For closed boundary conditions (3.111), which represents an eigenvalue equation, must be solved. Such matrix eigenvalue problems arise in many applications of science and engineering. They are given by the matrix equation [177, 238]

$$\underline{A}\mathbf{x} = \lambda\mathbf{x} \ , \tag{4.7}$$

where $\underline{A}$ is a square $n \times n$ matrix, $\mathbf{x}$ a non-zero n by 1 vector, and λ a scalar. The polynomial

$$m(\lambda) = \det(\lambda\underline{I} - \underline{A}) \ , \tag{4.8}$$

where $\underline{I}$ is the unity matrix, is the characteristical polynomial of $\underline{A}$. The roots λ_i of the equation

$$m(\lambda) = 0 \tag{4.9}$$

are the eigenvalues of $\underline{A}$. Since the degree of $m(\lambda)$ is n, the characteristical polynomial has n roots, and so $\underline{A}$ has n eigenvalues. A vector $\mathbf{x}_i$ that satisfies

$$\underline{A}\mathbf{x}_i = \lambda_i\mathbf{x}_i \tag{4.10}$$

is called an eigenvector of $\underline{A}$. The matrix $\underline{A}$ is *positive definite*, if all eigenvalues are positive, *positive semidefinite*, if $\lambda_i \geq 0$, *negative definite*, if all eigenvalues are negative, and *negative semidefinite*, if $\lambda_i \leq 0$. If both positive and negative eigenvalues occur, the matrix is *indefinite*.

Based on the properties of the matrix $\underline{A}$, several cases can be distinguished. The matrix $\underline{A}$ can be HERMITian

$$\underline{A} = \underline{A}^{+} \quad : \quad A_{ij} = A_{ji}^{*} \tag{4.11}$$

or non-HERMITian. Furthermore, the matrix elements can be real or complex. A real HERMITian matrix is also denoted a symmetric matrix. A HERMITian matrix has only real eigenvalues, while a non-HERMITian matrix also permits complex eigenvalues. Based on the different cases, different numerical solvers have been used for the solution. Table 4.1 summarizes the different cases.

Matrix elements	Symmetry	Eigenvalues	Eigenvectors	Solver	Reference
real	HERMITian	real	real	CEPHES	[239]
real	non-HERMITian	complex	complex	EIGCOM	[240]
complex	HERMITian	real	complex	QRIHRM	[240]
complex	non-HERMITian	complex	complex	EIGCOM	[240]

Table 4.1: Eigenvalues and eigenvectors of matrices with different properties and the numerical solvers used.

As described in Section 3.6.3.3, calculation of the life times of quasi-bound states requires to find the eigenvalues of the inverse retarded GREEN's function $\underline{G}^{-1}$ (3.99). Since the coupling entries ζ and ξ are in general complex, the matrix is complex too. Furthermore, the matrix is not HERMITian. However, it is not possible to straightforwardly calculate the eigenvalues of $\underline{G}^{-1}$ because the eigenvalue problem is nonlinear [177]: The values of the matrix elements ζ and ξ depend on the eigenvalue $\mathcal{E}$.

Sophisticated methods have been developed to allow an easy solution of this matrix so that the life times can be calculated [241–244]. First, the closed-boundary HAMILTONian is constructed and the eigenvalues are calculated. In the one-dimensional case the matrix is tridiagonal. It is shown in [245] that in this case, the LU algorithm is advantageous for the calculation of eigenvalues compared to the commonly used QR algorithm which transforms the matrix into an upper HESSENBERG matrix [246]. This is also done by the CEPHES solver. However, since the solver will be used for two- and three-dimensional problems as well, where the LU algorithm shows no advantages, the QR algorithm was applied.

Then, the eigenvalues are filtered so that only the values remain which are located in the considered energy range. These values are then used as initial values for a NEWTON search around the closed-boundary eigenvalue [242, 244]. This is motivated by the fact that for $\mathcal{E}_i$ being an eigenvalue of $\underline{H}$, the determinant

$$m(\mathcal{E}_i) = \det(\underline{H} - \mathcal{E}_i \underline{I}) = 0 \tag{4.12}$$

must be zero. To find the roots of this equation, a NEWTON search around the closed-boundary eigenvalues $\mathcal{E}_i$ is used

$$\mathcal{E}_{i,j+1} = \mathcal{E}_{i,j} - \frac{m(\mathcal{E}_{i,j})}{m'(\mathcal{E}_{i,j})} , \tag{4.13}$$

where $m'(\mathcal{E})$ denotes the derivative of the determinant

$$m'(\mathcal{E}) = \frac{\mathrm{d}m(\mathcal{E})}{\mathrm{d}\mathcal{E}} . \tag{4.14}$$

For a tridiagonal matrix, it is possible to find an analytical expression for $m'(\mathcal{E})$ [247, 248]. For general situations, however, the derivative can only be found numerically by

$$m'(\mathcal{E}_i) \approx \frac{m(\mathcal{E}_i + \Delta\mathcal{E}/2) - m(\mathcal{E}_i - \Delta\mathcal{E}/2)}{\Delta\mathcal{E}} . \tag{4.15}$$

This has the advantage that it is not limited to one-dimensional problems but can be applied to any shape of the HAMILTONian.

'A crude model of the future is more valuable than an accurate model of the past.'

Michael Duane

Chapter 5

Applications

GATE LEAKAGE is one of the most important issues for contemporary CMOS devices. Based on the tunneling models outlined in Section 3 two different application areas will be investigated in this section. First, gate leakage in contemporary MOS transistors will be studied and compared to measurements. Emphasis is put on the distinction between the different sources of the tunneling current, namely the region below the gate and the region near the drain and source extensions.

Device engineers commonly rely on gate leakage measurements of turned-off devices to evaluate the power consumption of CMOS circuits. This may lead to erroneous results since for turned-on devices, hot-carrier tunneling prevails which may exceed the turned-off tunneling current. Models which are based on simplified assumptions of the carrier energy distribution function fail to predict gate leakage in such cases.

Advanced CMOS devices will use alternative dielectric materials as gate dielectrics. However, a pronounced trade-off between the height of the energy barrier and the dielectric permittivity exists. This makes the use of optimization necessary to find the optimum layer composition. Furthermore, alternative dielectrics are not ideal insulators but contain defects which give rise to trap-assisted tunneling. As a state-of-the-art example, tunneling in ZrO_2-based MOS capacitors will be studied and compared to measurements.

As a second important application area, non-volatile memories will be studied. Unlike MOS transistors, non-volatile memory devices represent an application where tunneling is not a spurious effect, but crucial for the device functionality. After a short review of non-volatile memory technology, the tunneling current of conventional EEPROMs and advanced structures will be studied. In contrast to these devices SONOS (silicon-oxide-nitride-oxide-silicon) EEPROM devices store the charge not on an isolated contact, but in a layer of trap-rich dielectric.

Recent efforts to reduce the charging time of non-volatile memory devices resulted in multi-barrier tunneling devices and EEPROMs with asymmetrically layered tunnel dielectrics. The operation of these devices will briefly be described at the end of this chapter.

5.1 Tunneling in MOS Transistors

The gate leakage current in contemporary MOS transistors poses a major problem for further device scaling. This section describes simulation results of MOS transistors, outlines the effect of various device parameters, shows how to account for hot-carrier tunneling in turned-on devices, and elaborates on the use of alternative dielectric materials to replace SiO_2 as a gate dielectric. First, however, the tunneling paths in MOS transistor structures will be reviewed.

5.1.1 Tunneling Paths in MOS Transistors

Tunneling in an MOS transistor, as shown in the left part of Fig. 5.1, basically can be separated into a path between the gate and the channel, and a path between the gate and the the source and drain extension areas [249]. Tunneling in the source and drain extension areas can exceed tunneling in the channel by orders of magnitude. This is related to two effects: First, instead of n-p or p-n tunneling, n-n or p-p tunneling prevails. Second, the potential difference and thus the bending of the energy barrier is high. This increased tunneling current in the source and drain extension areas can be a serious problem if measurements are performed on long-channel MOSFETs to characterize their short-channel pendants, because the edge tunneling currents exceed the channel tunneling current by orders of magnitude. Furthermore, there is a fundamental difference between tunneling in MOS transistors and MOS capacitors [96, 250]. In contrast to MOS transistors, MOS capacitors which are biased in strong inversion cannot supply the amount of carriers as predicted by the tunneling model. This effect is termed *substrate-limited* tunneling, because the tunneling current is limited by the generation rate in the substrate. In the channel of an inverted MOS transistor, on the other hand, carriers can always be supplied by the source and drain contacts. This effect is depicted in the right part of Fig. 5.1.

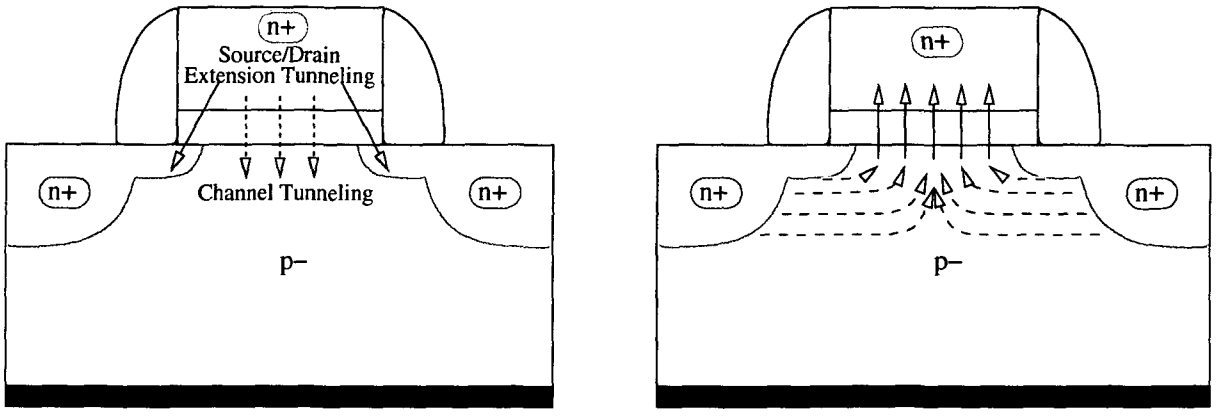


Figure 5.1: The different tunneling paths (channel tunneling, source and drain extension tunneling) in an MOS transistor (left). In an MOS transistor biased in inversion (right), tunneling electrons are supplied from the source and drain reservoirs, which is not possible in an MOS capacitor.

5.1.2 Channel Tunneling

In this section the effects of various device parameters on the gate leakage of MOS capacitors are studied. This is equivalent to tunneling in MOS transistors, if only channel tunneling (n-p or p-n) is considered and the source, drain, and bulk contacts are grounded. The parameters investigated are

- the doping of the polysilicon gate contact,
- the doping of the substrate,
- the thickness of the dielectric layer,
- the barrier height of the dielectric,
- the carrier mass in the dielectric,
- the dielectric permittivity, and
- the lattice temperature.

The typical shape of the gate current density in turned-off nMOS and pMOS devices is depicted in Fig. 5.2. A SiO_2 gate dielectric thickness of 2 nm and an acceptor or donor doping of $5 \times 10^{17}\,cm^{-3}$ and polysilicon gates was chosen. In the nMOS device the majority electron tunneling current always exceeds the hole tunneling current due to the lower electron mass and barrier height (3.2 eV instead of 4.65 eV for holes). In the pMOS capacitor, however, the majority hole tunneling exceeds electron tunneling only for negative and low positive bias. For positive bias the conduction band electron current again dominates due to its much lower barrier height [251].

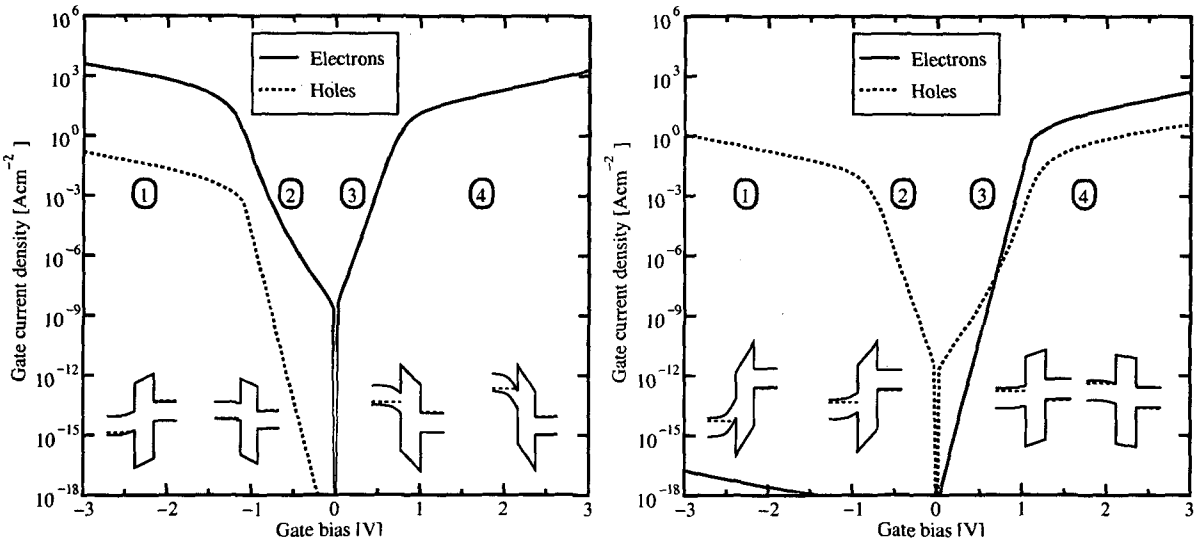


Figure 5.2: Channel tunneling regions in an nMOS (left) and a pMOS (right). The insets show the approximate shape of the band edge energies.

5.1.2.1 Effect of the Polysilicon Gate Doping on the Channel Tunneling

As outlined in Section 2.2.2, heavily doped polysilicon is used as material for the gate contact to allow adjustable work functions and realize CMOS circuits. Fig. 5.3 shows the electron and hole tunneling current density for different doping of the polysilicon gate contact. In the nMOS gate leakage generally increases with increasing doping of the polysilicon gate because tunneling current is dominated by electrons. In the pMOS a higher polysilicon doping leads to reduced electron tunneling current and increased hole tunneling current. The effect on the overall leakage depends on the doping and the gate bias.

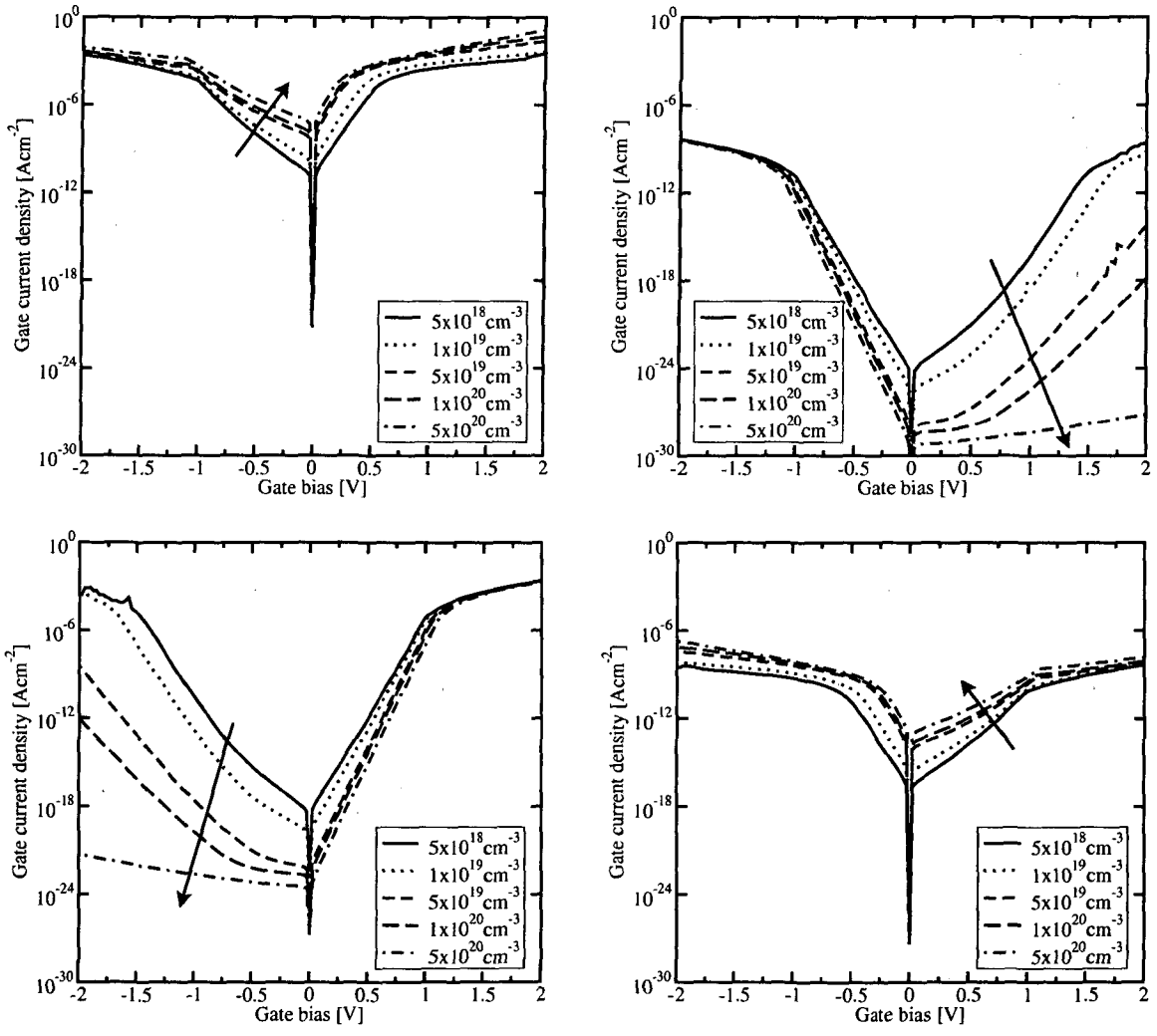


Figure 5.3: Electron (left) and hole (right) current density in an nMOS (top) and a pMOS (bottom) with different doping of the polysilicon gate. Substrate doping is $10^{18}\,\mathrm{cm}^{-3}$, dielectric thickness is 2 nm.

5.1.2.2 Effect of the Substrate Doping on the Channel Tunneling

Fig. 5.4 shows the electron and hole tunneling current density for different doping of the substrate. With increasing substrate doping, the majority tunneling component (electrons in the nMOS, holes in the pMOS) is reduced in both the nMOS and pMOS devices, while the minority component increases.

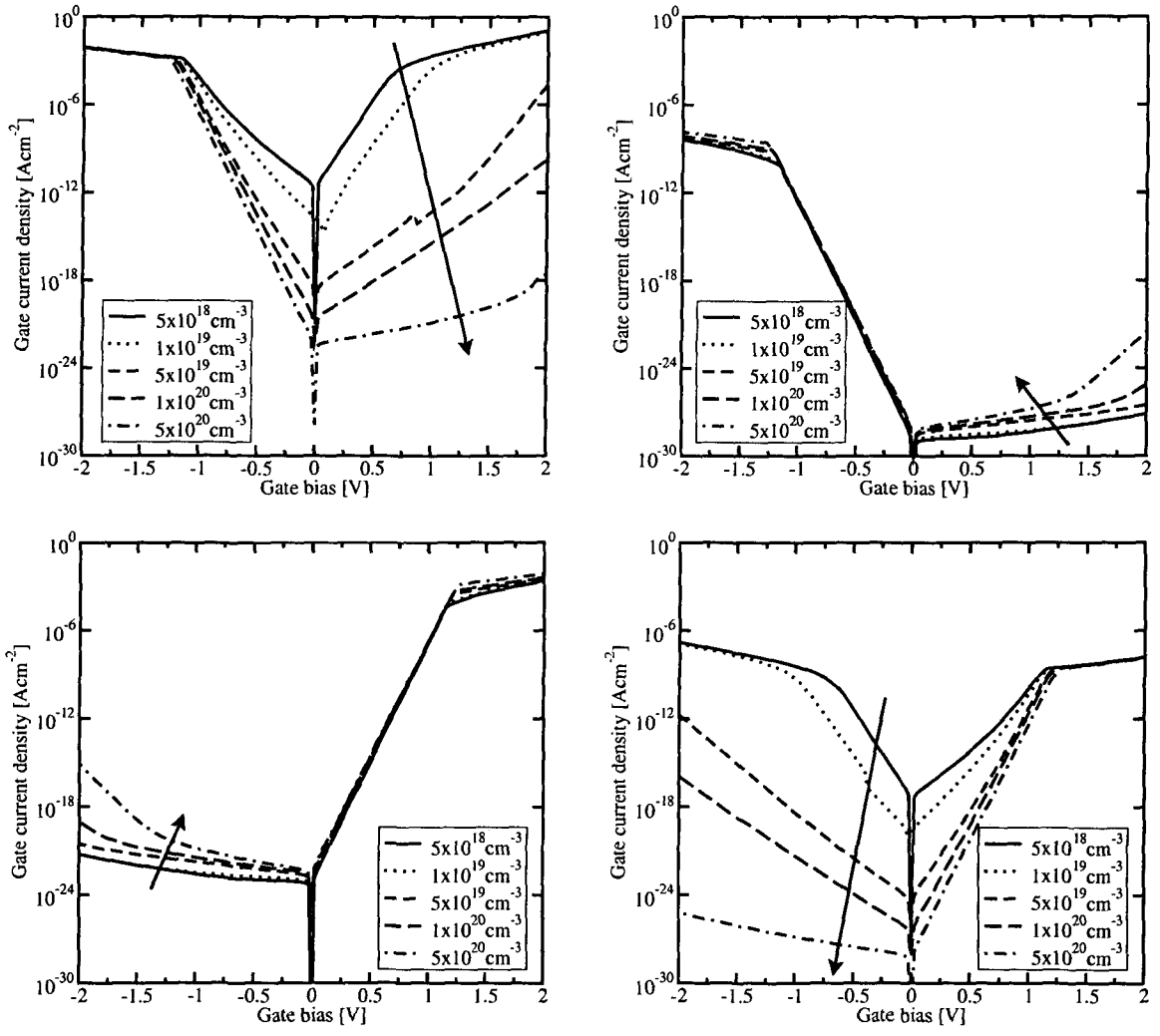


Figure 5.4: Electron (left) and hole (right) current density in an nMOS (top) and a pMOS (bottom) with different doping of the substrate. Gate polysilicon doping is $5 \times 10^{20}\,\mathrm{cm}^{-3}$, dielectric thickness is 2 nm.

5.1.2.3 Effect of the Dielectric Thickness on the Channel Tunneling

The physical thickness of the dielectric has the largest impact on the gate current density, as shown in Fig. 5.5. Increasing the gate dielectric thickness by 0.4 nm leads to a decrease of all tunneling current components by several orders of magnitude.

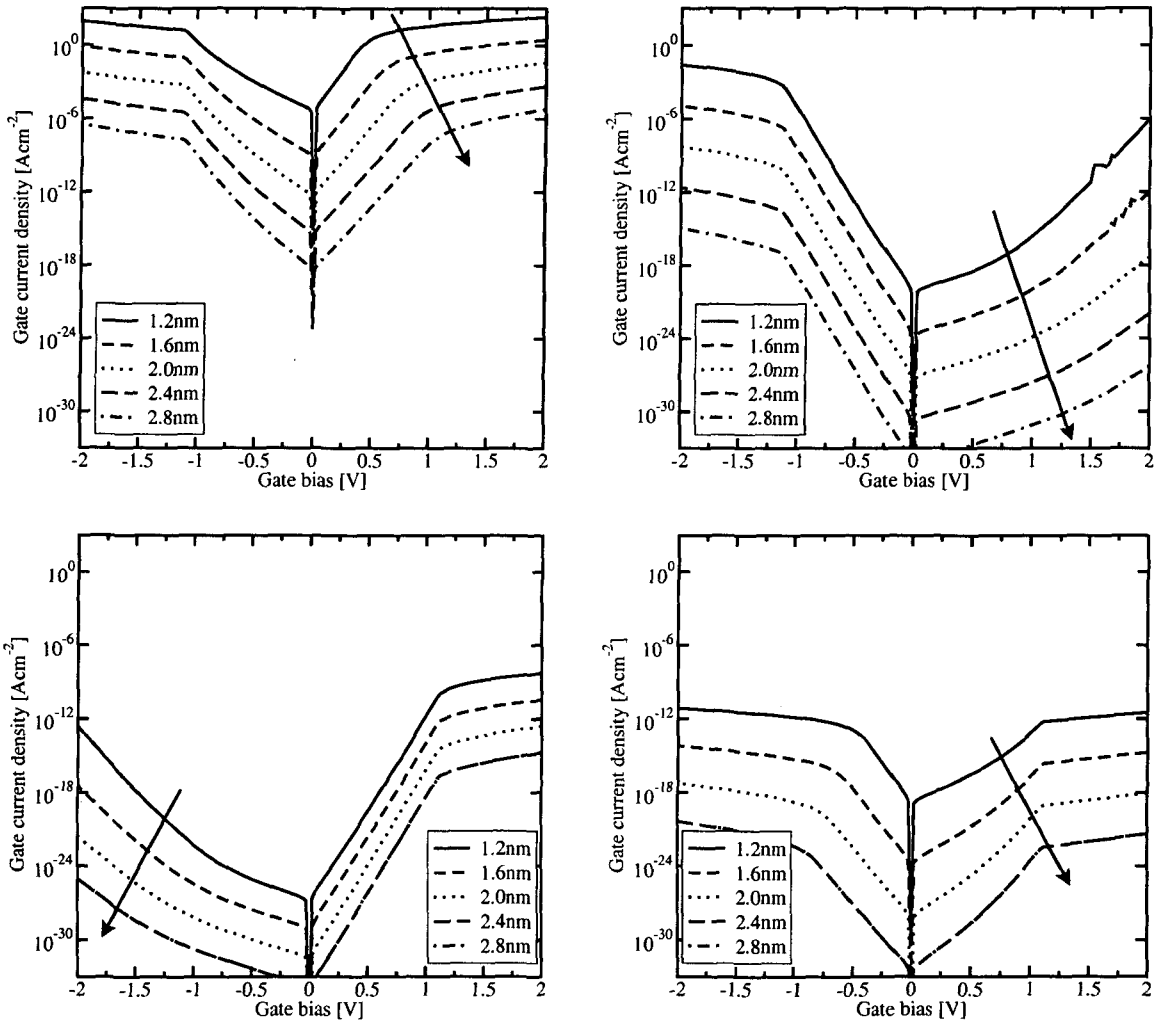


Figure 5.5: Electron (left) and hole (right) current density in an nMOS (top) and a pMOS (bottom) with different thickness of the dielectric layer. Gate polysilicon doping is $5 \times 10^{20}\,\mathrm{cm}^{-3}$, substrate doping is $5 \times 10^{18}\,\mathrm{cm}^{-3}$.

5.1.2.4 Effect of the Barrier Height on the Channel Tunneling

The main parameter, besides the thickness of the dielectric, influencing tunneling current is the height of the energy barrier. The influence of this parameter is depicted in Fig. 5.6. Different dielectric materials strongly differ in their work function difference to silicon. It must be distinguished between the barrier height for electrons and for holes. The most frequently used dielectric material SiO_2 has an electron barrier height of about 3.2 eV and a hole barrier height of approximately 4.6 eV. The measurement of these material parameters is difficult and values in the available literature vary widely (see Section 5.1.5).

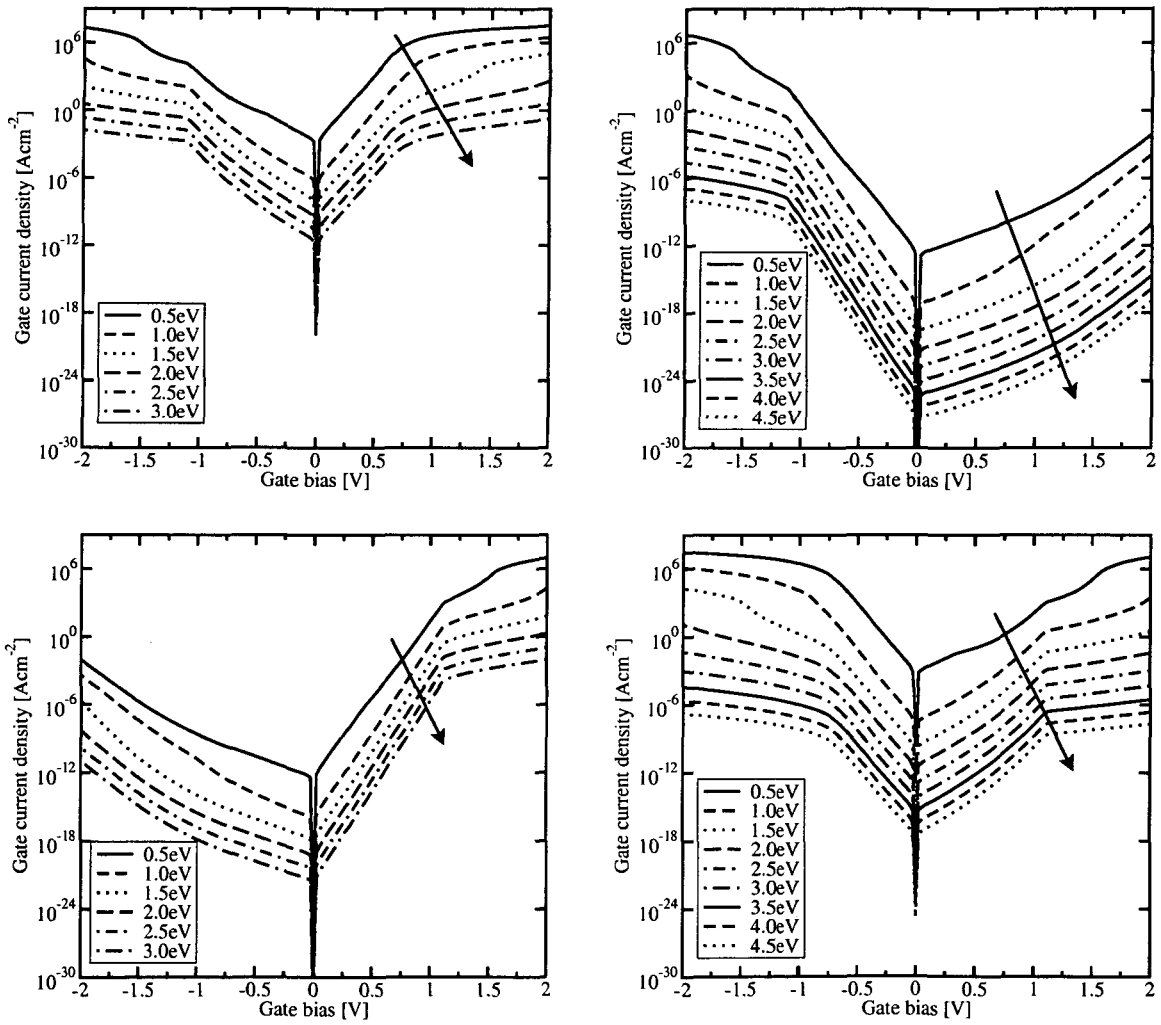


Figure 5.6: Effect of the electron and hole barrier height on electron tunneling current (left) and hole tunneling current (right) in an nMOS (top) and a pMOS (bottom) with 2 nm dielectric thickness, $10^{20}\,cm^{-3}$ polysilicon and $5 \times 10^{18}\,cm^{-3}$ substrate doping.

5.1.2.5 Effect of the Carrier Mass on the Channel Tunneling

Being the parameter with the highest uncertainty, the electron and hole mass in the dielectric is commonly used as a fitting parameter to reproduce measurements. Its influence on the gate current density is shown in Fig. 5.7. An increase in the carrier mass by $0.1m_0$ leads to a reduction in the gate current density by about a factor of 10. It must, of course, be held in mind that with the approaches described in Section 3, tunneling is described by a single value for the carrier mass. Its use as a fitting parameter may thus well be justified. Recent investigations, however, report an increase of the electron mass with reducing thickness of the dielectric layer, which is backed by measurements and tight-binding band structure calculations [252–254].

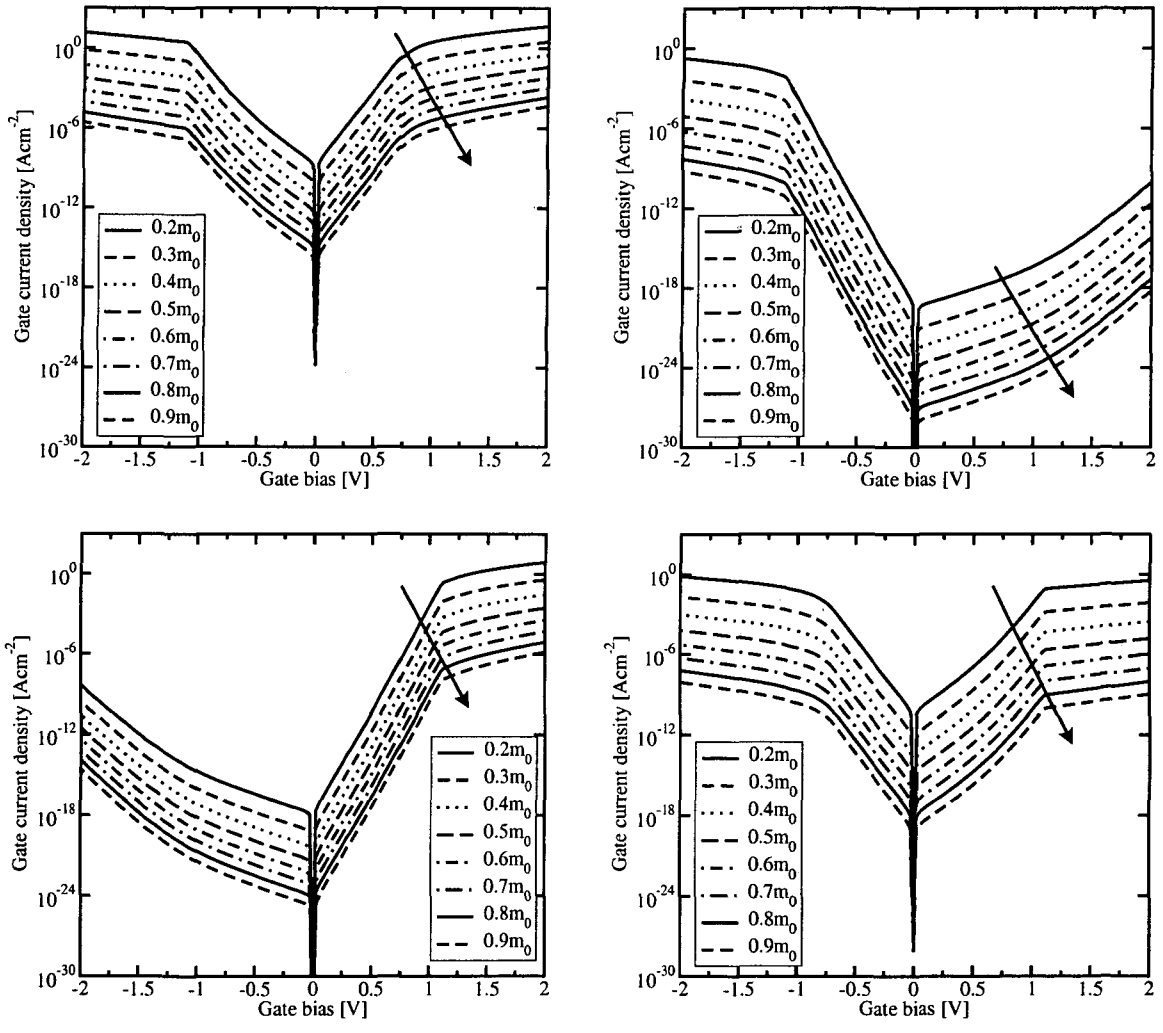


Figure 5.7: Effect of the carrier mass on electron tunneling current (left) and hole tunneling current (right) in an nMOS (top) and a pMOS (bottom) with 2 nm dielectric thickness, $10^{20}\,\mathrm{cm}^{-3}$ polysilicon and $5 \times 10^{18}\,\mathrm{cm}^{-3}$ substrate doping.

5.1.2.6 Effect of the Dielectric Permittivity on the Channel Tunneling

The permittivity of the dielectric layer influences the tunneling current density in two ways: First, the shape of the energy barrier — and thus the transmission coefficient — changes. Second, the inversion charge — and thus the band edge energy — in the channel is affected. The effect of varying dielectric permittivity is shown in Fig. 5.8. Especially in the low-bias regime, a higher permittivity strongly increases the gate current density.

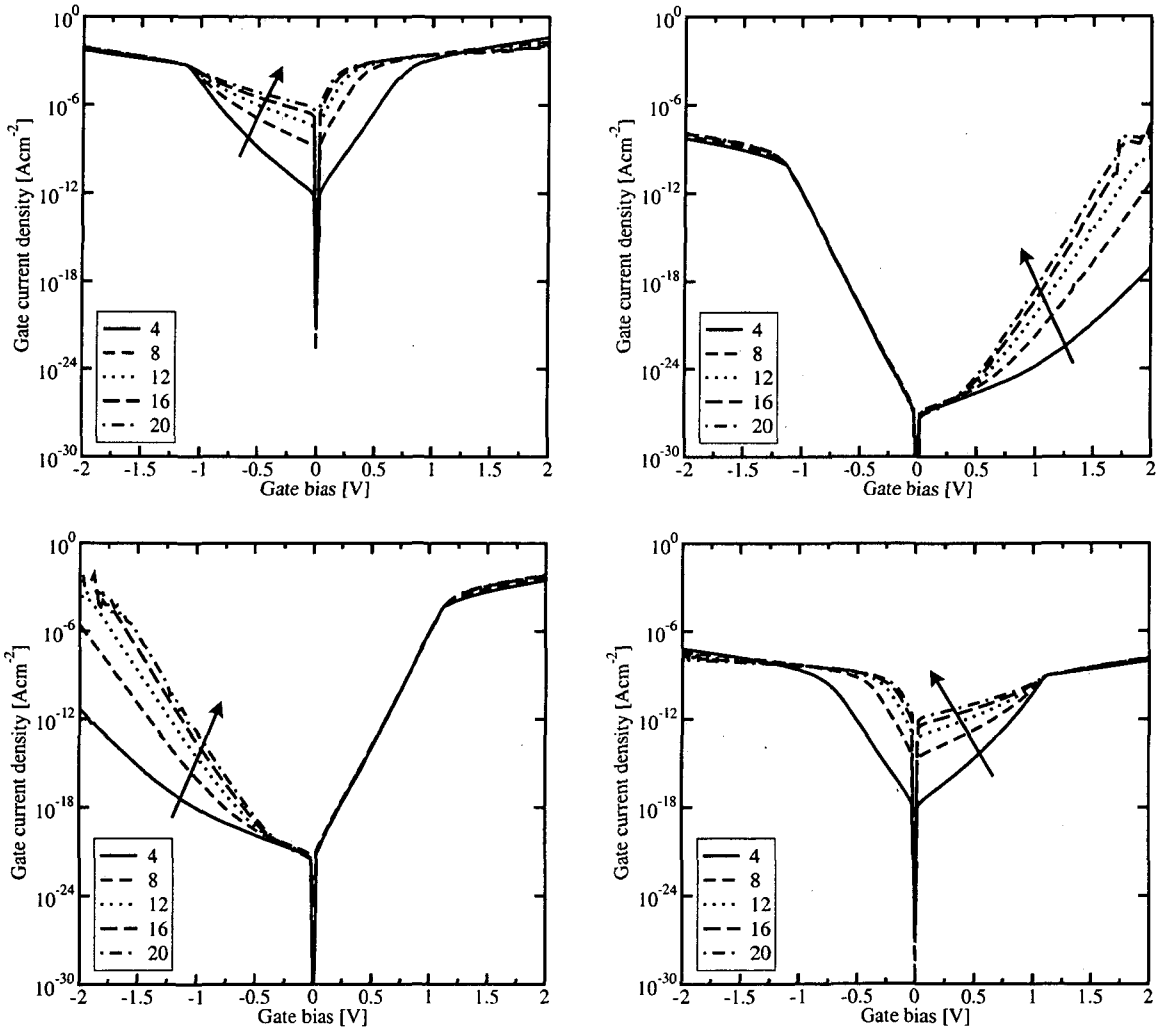


Figure 5.8: Effect of the dielectric permittivity κ/κ_0 on electron tunneling current (left) and hole tunneling current (right) in an nMOS (top) and a pMOS (bottom) with 2 nm dielectric thickness, $10^{20}\,\mathrm{cm}^{-3}$ polysilicon and $5 \times 10^{18}\,\mathrm{cm}^{-3}$ substrate doping.

5.1.2.7 Effect of the Lattice Temperature on the Channel Tunneling

The lattice temperature enters the gate tunneling current via the electron energy distribution functions in the polysilicon gate and in the channel. The transmission coefficient, being based on quantum-mechanical reasoning alone, is not affected by the lattice temperature. However, the supply function depends on the lattice temperature. The impact on the gate current density is shown in Fig. 5.9. Rising temperature increases the tunneling current density in all cases.

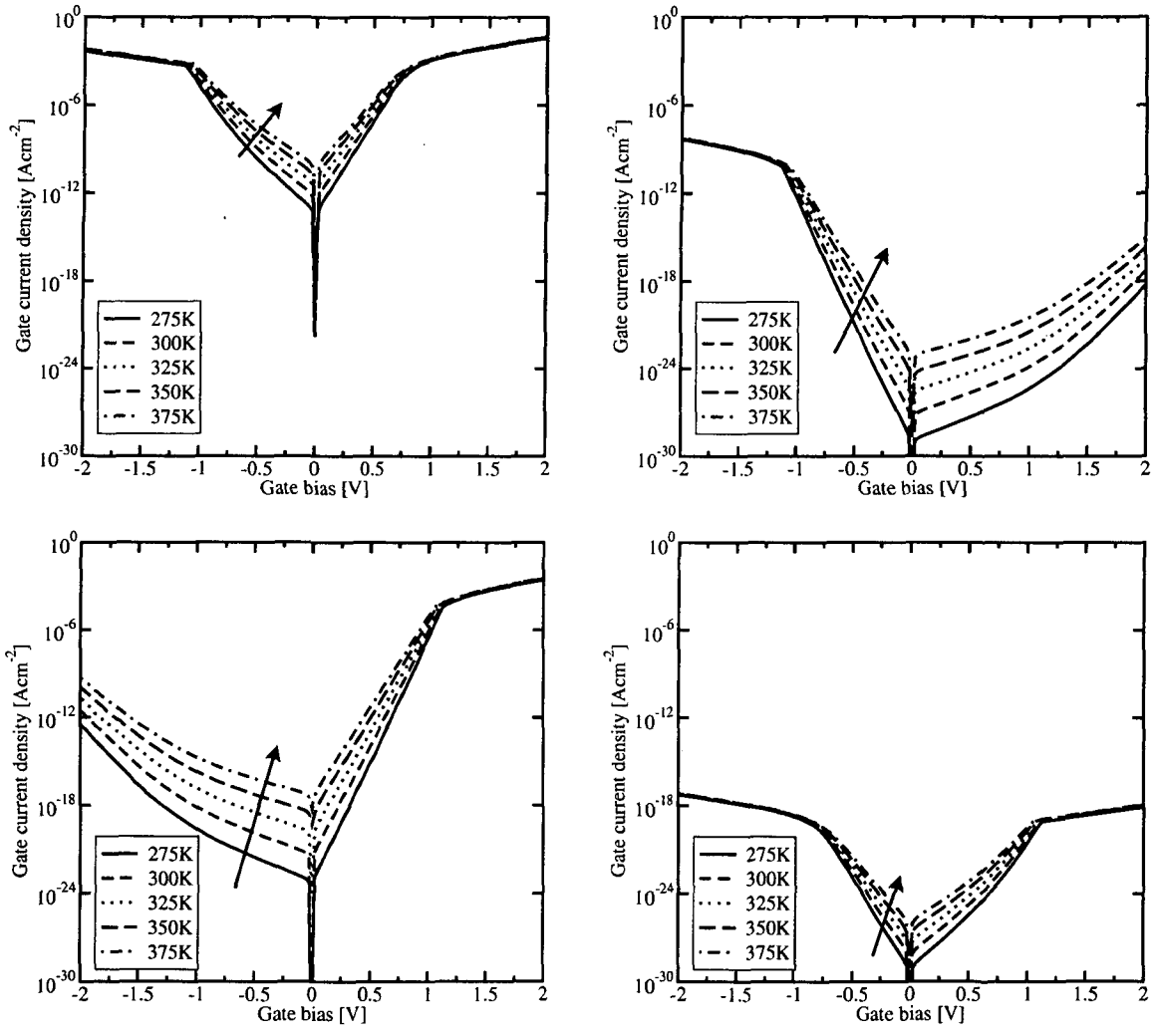


Figure 5.9: Effect of the lattice temperature on electron tunneling current (left) and hole tunneling current (right) in an nMOS (top) and a pMOS (bottom) with 2 nm dielectric thickness, $10^{20}\,\mathrm{cm}^{-3}$ polysilicon and $5 \times 10^{18}\,\mathrm{cm}^{-3}$ substrate doping.

5.1.2.8 Comparison to Measurements

Since almost all available measurements of gate leakage in MOS devices are performed on turned-off MOS transistors, a comparison with measurements will be given before turned-on devices are investigated in Section 5.1.4. The TSU-ESAKI model with an analytical WKB transmission coefficient is in good agreement with recently reported data for devices with different gate lengths and bulk doping [96, 249] as shown in Fig. 5.10 for nMOS (left) and pMOS devices (right) [255]. It can be seen that the gate current density can be reproduced over a wide range of dielectric thicknesses with a single set of physical parameters. Additional measurements have been performed on MOSFETs with a gate dielectric thickness of 1.5 nm (see the lower part of Fig. 5.10) and compared with the results of other simulators (UTQUANT [256] and MEDICI [257]). Under inversion condition the fit is not perfect while under accumulation the measurements can be reproduced well. Note that with UTQUANT, the low-bias tunneling current cannot be reproduced and MEDICI completely failed for the pMOS device.

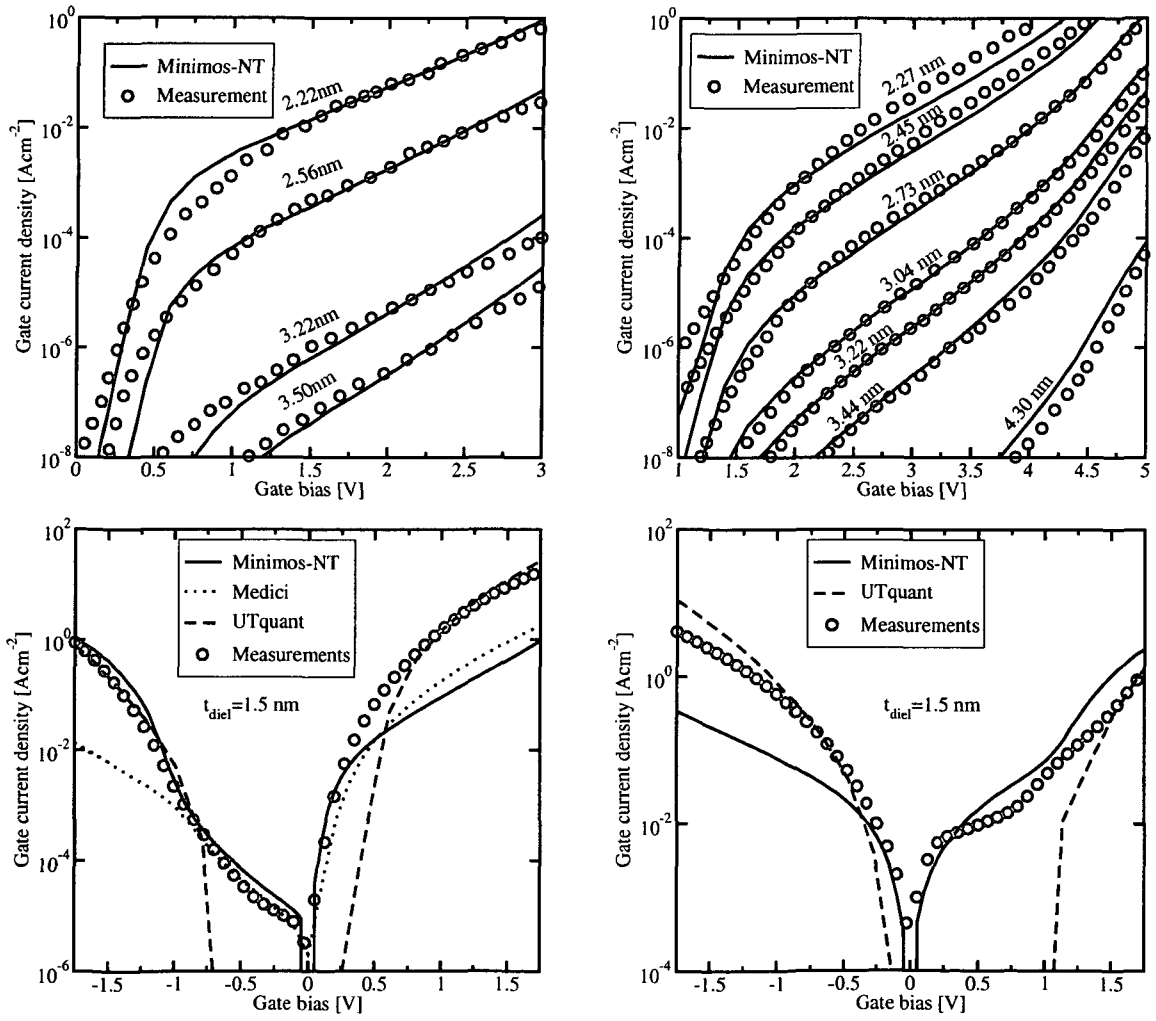


Figure 5.10: Comparison of simulations using different simulators with measurements of nMOS (left) and pMOS (right) devices [96, 249, 255].

5.1.2.9 Validity of Compact Models

Since the computational effort for the numerical integration in TSU-ESAKI's formula or the evaluation of the quasi-bound states is numerically expensive, it is reasonable to ask if compact models can describe tunneling, at least for single-layer dielectrics. The compact tunneling models outlined in Section 3.7 are compared in Fig. 5.11 for a symmetrical metal-dielectric-metal structure (left) and for an nMOS structure with 3 nm dielectric thickness (right). For the metal-dielectric-metal structure, SCHUEGRAF's model yields almost the same results as the computationally much more expensive TSU-ESAKI model. The FOWLER-NORDHEIM model delivers correct values only for high bias. It is thus only applicable to describe high-field transport through gate dielectrics, like program and erase cycles in EEPROM devices. For the MOS structure in the right part of Fig. 5.11, the SCHUEGRAF model fails to describe the tunneling current density at low bias. For high bias, however, it may be used to provide an estimation of the gate current. The FOWLER-NORDHEIM model totally fails for this application. Furthermore, the FOWLER-NORDHEIM model shows the minimum gate current at minimum electric field in the dielectric, and not for the minimum gate bias.

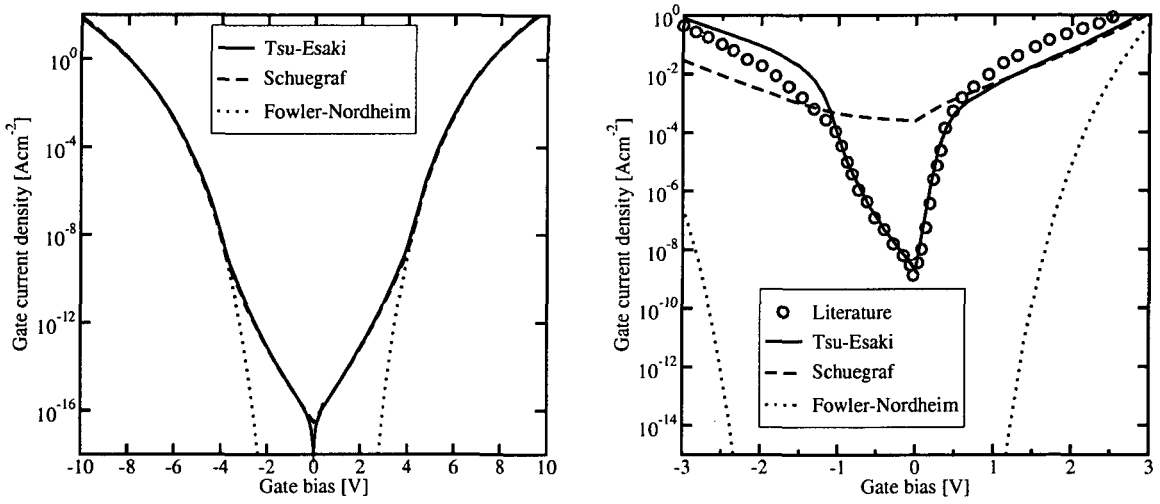


Figure 5.11: Compact models for a metal-dielectric-metal structure (left) and an nMOS structure (right, literature values from [249]).

5.1.3 Source and Drain Extension Tunneling

In the following examples the same devices as in Section 5.1.1 are investigated, but this time only the tunneling current in the source and drain extension areas (n-n or p-p) is taken into account. Since the barrier height, carrier mass, and dielectric thickness shows the same impact on the gate current density as for the case of channel tunneling, the corresponding figures are omitted.

5.1.3.1 Effect of the Polysilicon Gate Doping on the Source and Drain Extension Tunneling

Fig. 5.12 shows the effect of the doping concentration in the polysilicon gate on the extension region gate current density. Increasing the polysilicon doping leads to a slight increase of the main tunneling component and to a strong decrease of the minority tunneling component in both nMOS and pMOS devices.

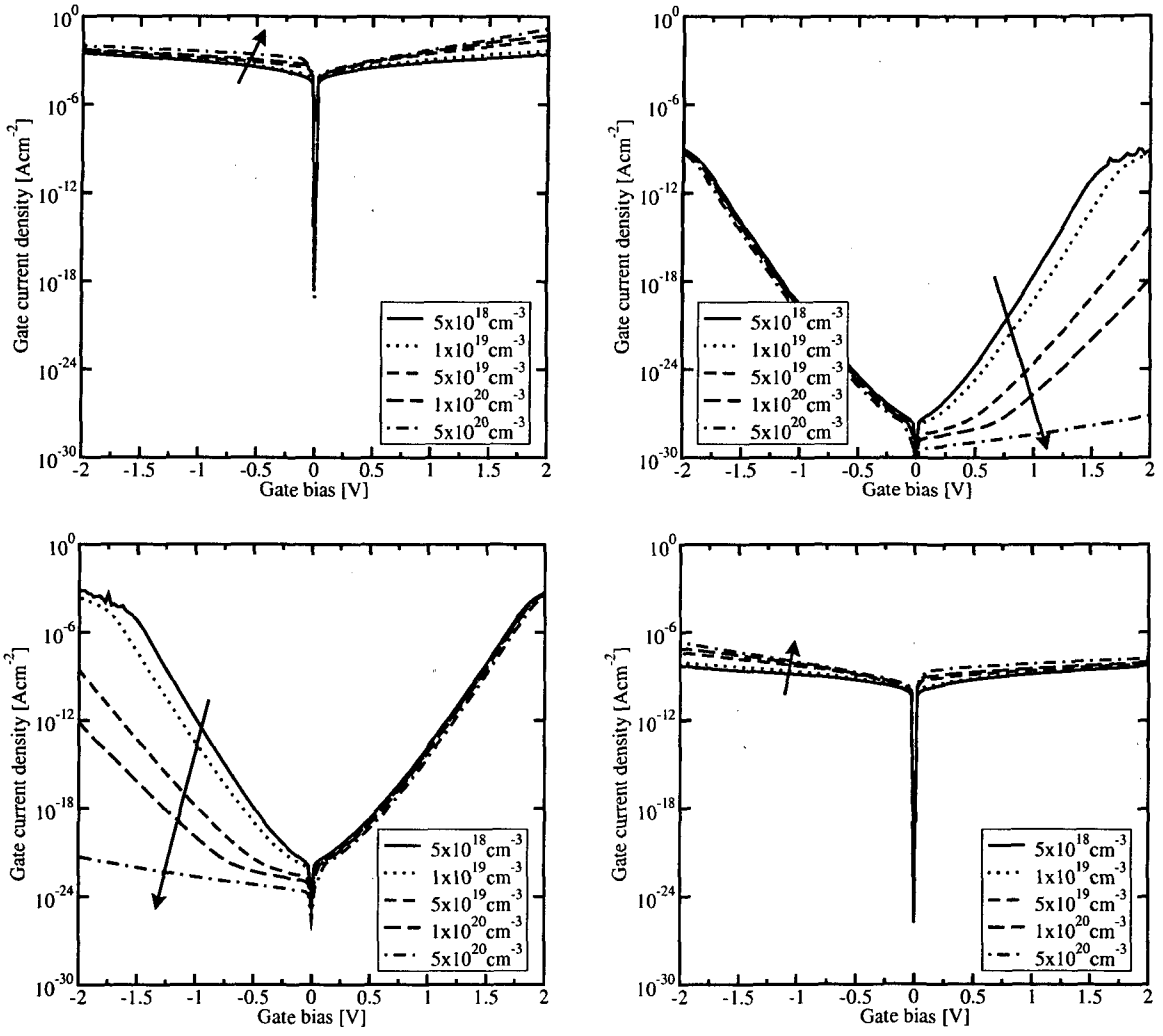


Figure 5.12: Effect of the polysilicon doping on the electron tunneling current (left) and the hole tunneling current (right) in the source and drain extension region of an nMOS (top) and a pMOS (bottom) with 2 nm dielectric thickness and 5×10^{18} cm^{-3} substrate doping.

5.1.3.2 Effect of the Substrate Doping on the Source and Drain Extension Tunneling

Fig. 5.13 shows the effect of the substrate doping concentration on the extension region gate current density. Similar to the polysilicon gate doping, a higher substrate doping leads to increased majority and decreased minority tunneling current.

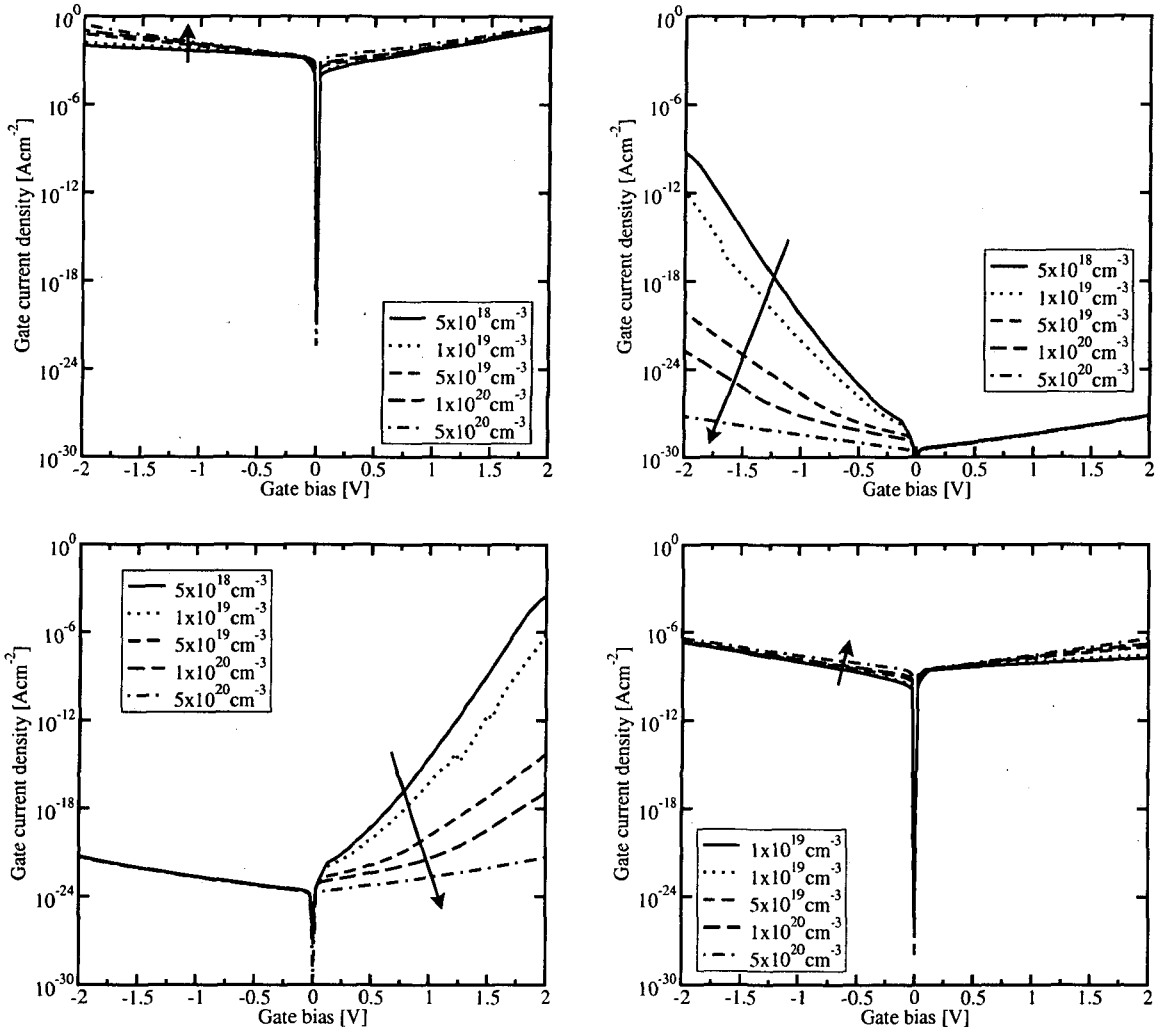


Figure 5.13: Effect of the substrate doping on the electron tunneling current (left) and the hole tunneling current (right) in the source and drain extension region of an nMOS (top) and a pMOS (bottom) with 2 nm dielectric thickness and $5 \times 10^{20}\,\mathrm{cm}^{-3}$ polysilicon doping.

5.1.3.3 Effect of the Dielectric Permittivity on the Source and Drain Extension Tunneling

Fig. 5.14 shows the effect of the dielectric permittivity on the extension region gate current density. In contrast to the channel-tunneling case, the low-bias regime is not influenced by the permittivity. Furthermore, the influence on the majority tunneling current component depends on the bias: The electron tunneling component in the nMOS decreases for negative bias and increases for positive bias. The hole tunneling component in the pMOS shows exactly the inverse trend.

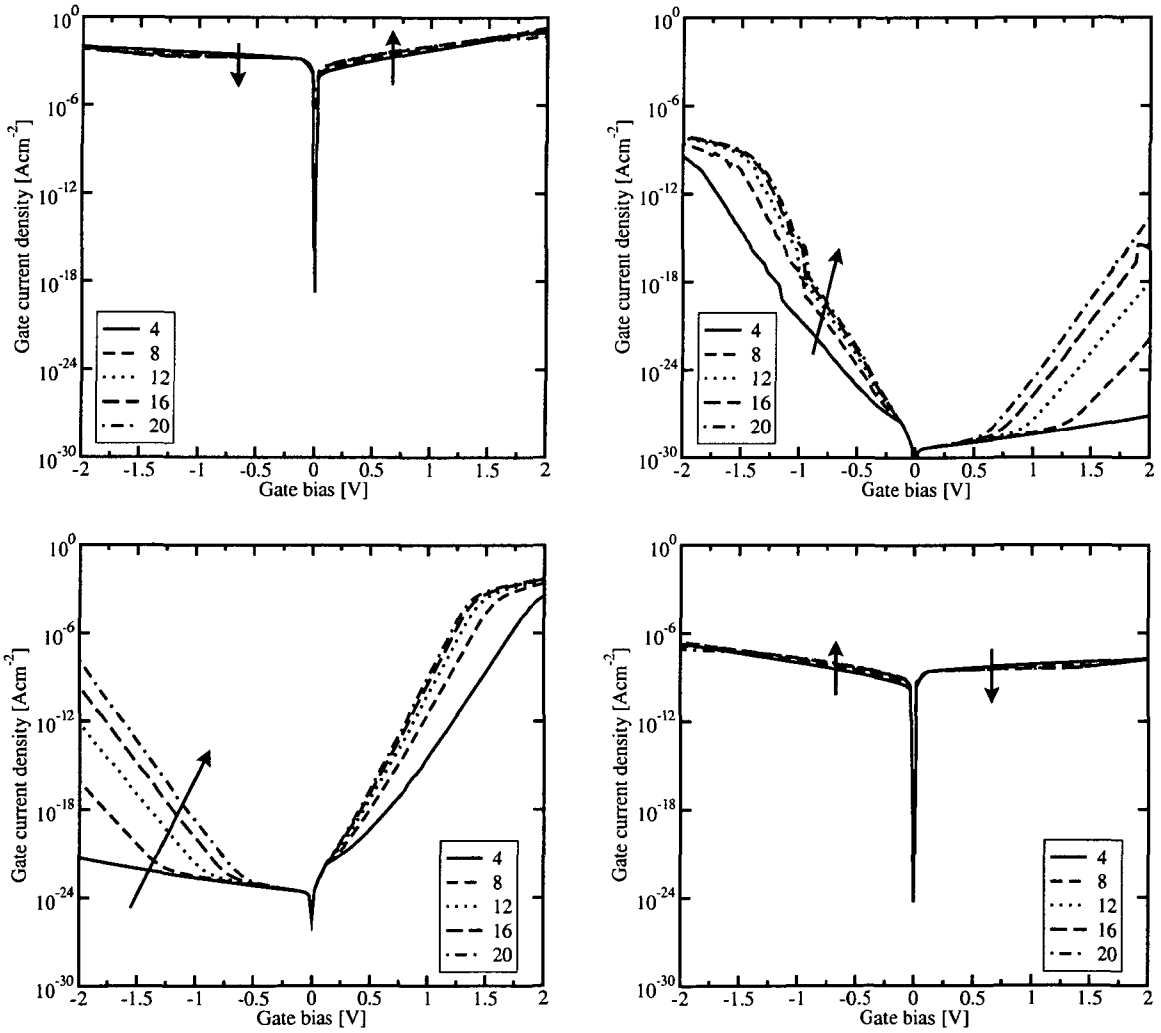


Figure 5.14: Effect of the dielectric permittivity κ/κ_0 on the electron tunneling current (left) and the hole tunneling current (right) in the source and drain extension region of an nMOS (top) and a pMOS (bottom) with 2 nm dielectric thickness and $5 \times 10^{18}\,\mathrm{cm}^{-3}$ substrate doping.

5.1.3.4 Effect of the Lattice Temperature on the Source and Drain Extension Tunneling

Fig. 5.15 shows the effect of the temperature on the extension region gate current density. Especially the minority carriers (holes in the nMOS, electrons in the pMOS) show strongly increased tunneling current with higher temperature. Unlike in the channel tunneling case, the majority tunneling component is hardly influenced by the temperature.

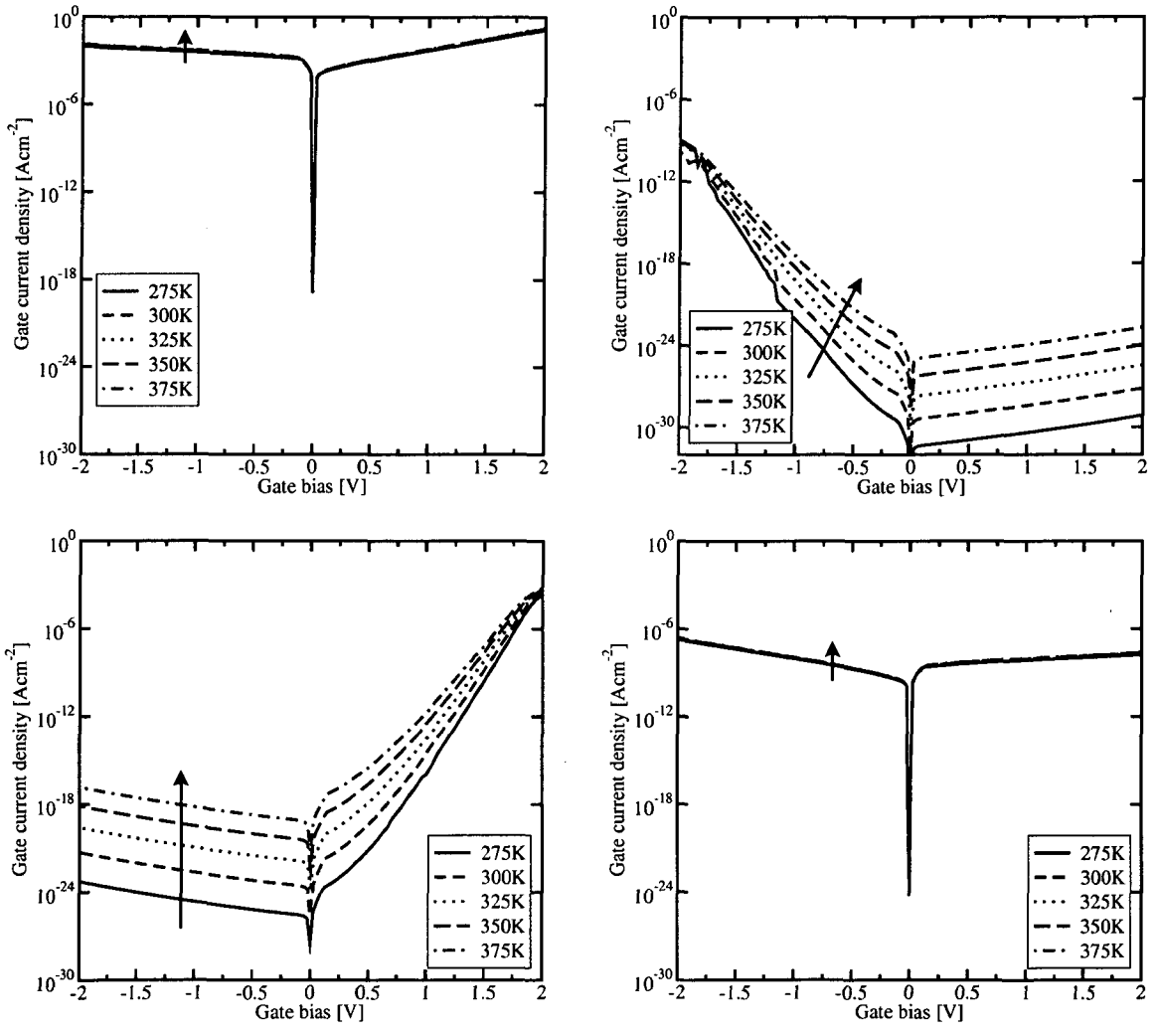


Figure 5.15: Effect of the lattice temperature on the electron tunneling current (left) and the hole tunneling current (right) in the source and drain extension region of an nMOS (top) and a pMOS (bottom) with 2 nm dielectric thickness and $5 \times 10^{18}\,\mathrm{cm}^{-3}$ substrate doping.

5.1.4 Hot-Carrier Tunneling in MOS Transistors

It has been shown in Section 3.3 that the distribution function in the channel of a turned-on MOS transistor heavily deviates from the shape implied by a FERMI-DIRAC or MAXWELLian distribution. A model for the non-MAXWELLian shape of the distribution function was presented which accurately reproduced the carrier energy distribution along the channel.

To check the impact of this wrong high-energy behavior, the integrand of the TSU-ESAKI formula, namely the expression $TC(\mathcal{E})N(\mathcal{E})$ has been evaluated for a standard device, as shown in the left part of Fig. 5.16, and compared to Monte Carlo results. The simulated device had a gate length of 100 nm and a gate dielectric thickness of 3 nm. While at low energies the difference between the non-MAXWELLian distribution function (3.28) and the heated MAXWELLian distribution (3.24) seems to be negligible, the amount of overestimation of the incremental gate current density for the heated MAXWELLian distribution reaches several orders of magnitude at 1 eV and peaks when the electron energy exceeds the barrier height. This spurious effect is clearly more pronounced for points at the drain end of the channel where the electron temperature is high. The non-MAXWELLian shape of the distribution function, indicated by the full line, reproduces the Monte Carlo results very well.

The region of high electron temperature is confined to only a small area near the drain contact, as shown in the right part of Fig. 5.16, where the gate current density along the channel is compared to Monte Carlo results. At the point of the peak electron temperature, which is located at approximately $x = 0.8L_g$, the heated MAXWELLian approximation overestimates the gate current density by a factor of almost 10^6. It will therefore have a large impact on the total gate current density. The cold MAXWELLian approximation underestimates the gate current density in this region, while the non-MAXWELLian distribution correctly reproduces the Monte Carlo results.

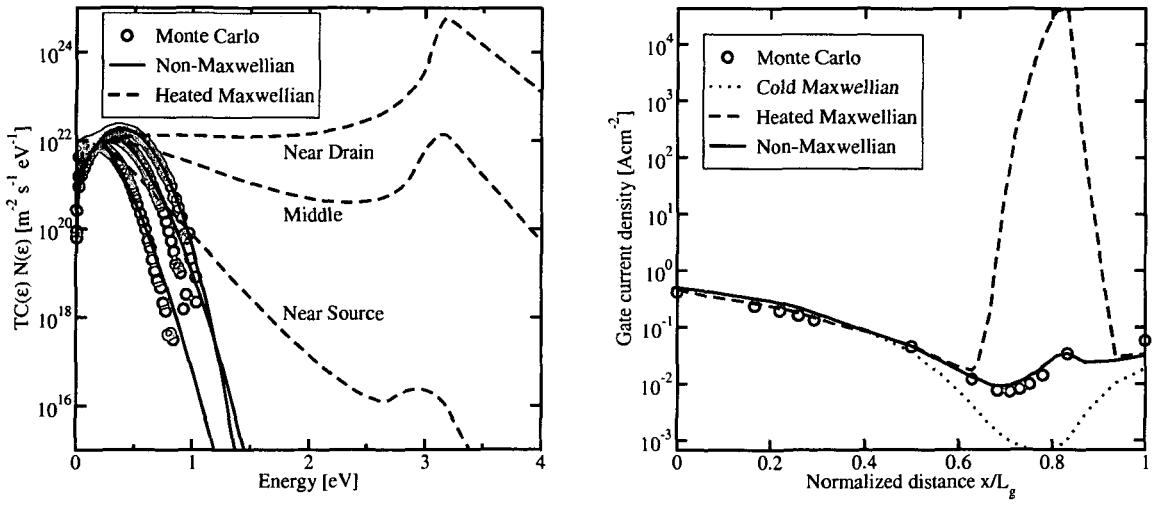


Figure 5.16: Integrand of TSU-ESAKI's equation (left) and gate current density along the channel (right) of a MOSFET with 100 nm gate length and 3 nm gate dielectric thickness.

The non-MAXWELLian shape yields excellent agreement, while the heated MAXWELLian approximation substantially overestimates the gate current density especially near the drain region. Instead of the heated MAXWELLian distribution it appears to be better to use a cold MAXWELLian distribution in that regime since it leads to a comparably low underestimation of the gate current density.

The effect of hot-carrier tunneling on the total gate current of the devices is shown in Fig. 5.17. In the left part of this figure the gate current density for a 0.5 μm turned-on MOSFET with a dielectric thickness of 4 nm is shown as a function of the gate bias. Results from Monte Carlo simulations are also shown in this figure. For low gate voltages ($V_{GS}<V_{DS}$) the peak electric field in the channel increases with increasing gate bias. The electron temperature is high and the heated MAXWELLian approximation massively overestimates the total gate current. If the gate bias exceeds the drain-source voltage, however, the peak electric field in the channel is reduced [258]. Therefore, for $V_{GS} > V_{DS}$ the electron temperature reduces with increasing gate bias and the heated MAXWELLian approximation delivers correct results. The non-MAXWELLian model (3.28) delivers correct results for all gate voltages.

The question remains if the hot-carrier tunneling current strongly depends on the gate length of the device. In the right part of Fig. 5.17 the gate current is given as a function of the gate length for different gate dielectric thicknesses (2.2 nm – 3.0 nm). Again, Monte Carlo simulation results are used as reference. It can be seen that the heated MAXWELLian distribution delivers correct results only for large gate lengths, while it totally fails for smaller devices. The use of a cold MAXWELLian distribution, on the other hand, underestimates the gate current only slightly and seems to be the better choice if accurate modeling of the device physics is not that important or only a quick estimation is asked for. The non-MAXWELLian model correctly reproduces the Monte Carlo results for all gate lengths and gate dielectric thicknesses.

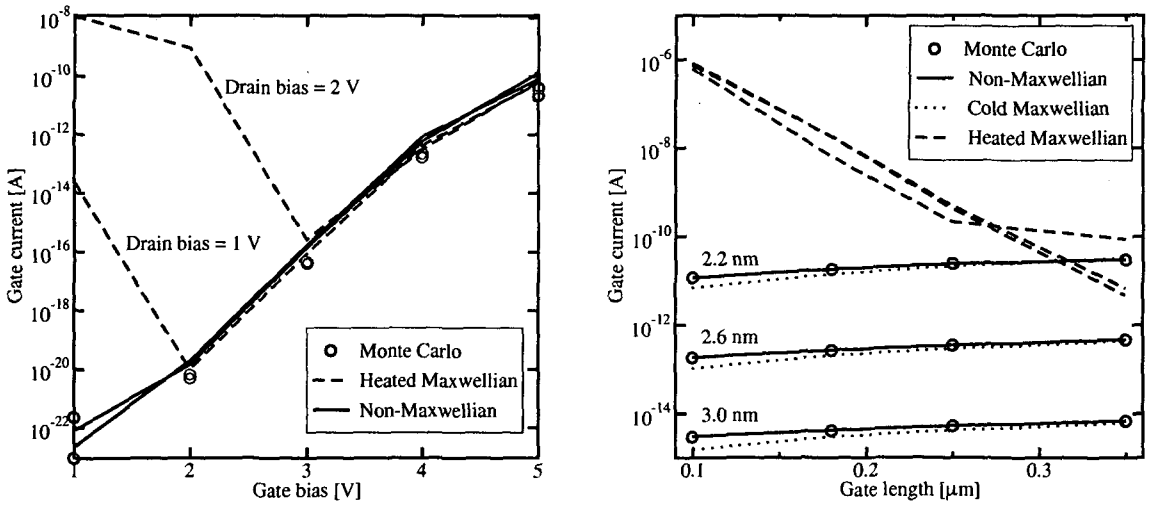


Figure 5.17: Gate current for different values of the gate bias (left). Dependency of the total gate current on the gate length (right).

5.1.5 Alternative Dielectrics for MOS Transistors

It has been outlined in Section 2.2.4 that the further reduction of device dimensions makes the introduction of alternative dielectric materials necessary. Since none of the possible materials forms a native oxide on silicon, a thin interfacial layer of SiO_2 can hardly be avoided. Thus, a two-layer band edge diagram is commonly assumed, as depicted in Fig. 5.18 [259]. A wide

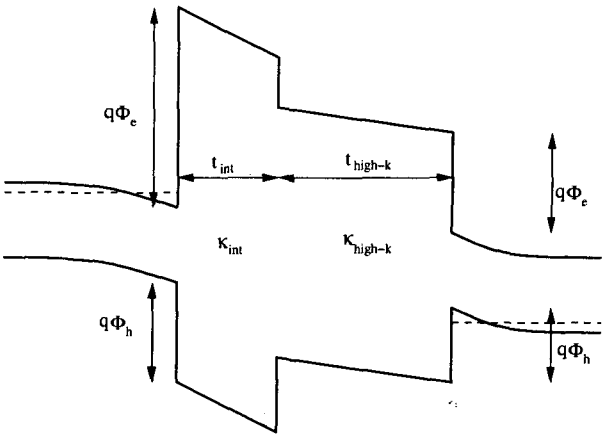


Figure 5.18: Band energy diagram of a stacked dielectric consisting of a thin underlying interface layer and a thick layer of a high-κ material with higher dielectric permittivity, but lower barrier height.

variety of high-κ materials can be considered as alternative dielectrics. However, several points must be considered when evaluating these materials:

1. The dielectric permittivity κ.
2. The barrier height for electrons $q\Phi_e$ and holes $q\Phi_h$ on silicon. These values are equivalent to the band edge offsets $\Delta\mathcal{E}_c$ and $\Delta\mathcal{E}_v$.
3. The thermodynamic stability of the dielectric material on silicon: The material must withstand all following processing steps.
4. The quality of the interfaces: High interface roughness may cause increased scattering which reduces the carrier mobility in the channel.
5. The trap concentration which leads to trap-assisted tunneling.
6. The feasibility and integrability of the deposition method in the fabrication process.

Only the permittivity, the trap concentration, and the barrier heights influence the tunneling current. When looking at the barrier height and permittivity of various dielectrics in Table 5.1, one notices a strong trade-off between the barrier height and the dielectric permittivity: dielectrics with a high energy barrier have a low permittivity and *vice versa*, see Fig. 5.19. Hence, optimization becomes necessary to find the optimum material.

	κ/κ_0	Band gap $\mathcal{E}_g$	Conduction band offset $\Delta\mathcal{E}_c$	Valence band offset $\Delta\mathcal{E}_v$	Reference
	[1]	[eV]	[eV]	[eV]	
SiO_2	**3.9**	**9.00**	**3.00**	**4.90**	[260]
	3.9	9.00	3.50	4.40	[261]
	3.9	9.00	3.15	4.75	[22]
	3.9	8.90	3.20	4.60	[262]
		9.00	3.50	4.40	[25, 263]
	3.9	9.00	3.00	4.90	[136]
Si_3N_4	**7.5**	**5.00**	**2.00**	**1.90**	[260]
	7.6	5.00 – 5.30	2.40	1.50 – 1.80	[261]
	7.9	5.30	2.40	1.80	[22]
	7.0	5.10	2.00	2.00	[262]
		5.30	2.40	1.80	[25, 263]
	7.5	5.00	2.00	1.90	[136]
Ta_2O_5	**25.0**	**4.40**	**1.40**	**1.90**	[136, 260]
	23.0 – 25.0	4.40	0.30	3.00	[261]
	25.0	4.40	0.36	2.94	[22, 25]
	26.0	4.50	1.00-1.50	1.90 – 2.40	[262]
		4.40	0.36	2.94	[263]
TiO_2	**40.0**	**3.50**	**1.10**	**1.30**	[136, 260]
	39.0–110.0	3.00–3.27	0.00	1.90 – 1.97	[261]
	80.0–170.0	3.05	0.00	1.95	[22]
	80.0	3.50	1.20	1.20	[262]
		3.05	0.00	1.95	[263]
Al_2O_3	**9.0**	**8.70**	**2.80**	**4.80**	[262]
	8.0 – 9.0	8.8–9.00	2.78–2.80	4.92 – 5.10	[261]
	9.5–12.0	8.8	2.80	4.90	[22]
		8.80	2.80	4.90	[263]
	10.0	8.80	2.80	4.90	[25]
ZrO_2	**23.0**	**5.80**	**1.40**	**3.30**	[25]
	25.0	7.80	1.40	5.30	[260, 262]
	22.0 – 25.0	5.00 – 5.80	1.40	2.50 – 3.30	[261]
	12.0–16.0	5.70–5.80	1.40–1.50	3.10 – 3.30	[22]
		5.80	2.50	2.20	[263]
HfO_2	**25.0**	**5.70**	**1.50**	**3.10**	[260, 262]
	22.0 – 40.0	6.00	1.50	3.50	[261]
	16.0–30.0	4.50–6.00	1.50	1.90 – 3.40	[22]
		6.00	1.50	3.40	[263]
	20.0	6.00	1.50	3.40	[25]
Y_2O_3	**15.0**	**5.60**	**2.30**	**2.20**	[262]
	11.3 – 18.0	5.50–6.00	1.30	3.10 – 3.60	[261]
	4.4	6.00	1.30	3.60	[263]
	15.0	6.00	2.30	2.60	[25]
$ZrSiO_4$	**12.6**	**6.00**	**1.50**	**3.40**	[261]
		4.50	0.70	2.70	[22]
	3.8	6.00	1.50	3.40	[263]
		6.00	1.50	3.40	[25]

Table 5.1: Band gap energy and conduction band offset of various dielectric materials.

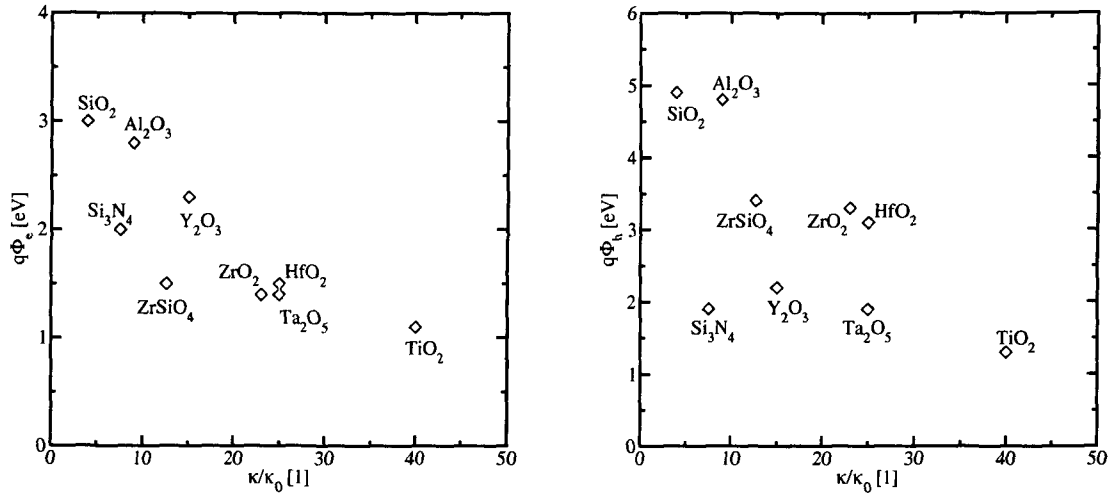


Figure 5.19: Trade-off between electron barrier height (left) or hole barrier height (right) and the permittivity of various dielectric materials.

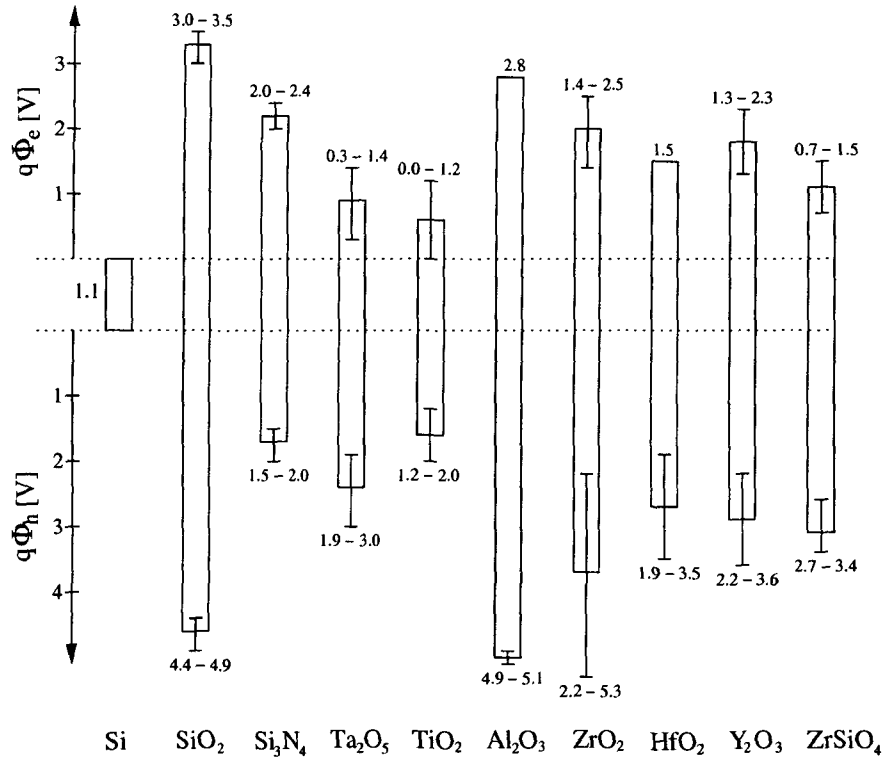


Figure 5.20: Conduction and valence band edges of various dielectric materials compared to silicon.

Choosing the highlighted material parameters from Table 5.1 the gate current density can be computed as a function of the gate bias. It is commonly assumed that an underlying layer of SiO_2 cannot be avoided — or is even deliberately introduced to achieve a lower trap density at the interface to silicon. Thus, an underlying SiO_2 layer with a thickness of 0.5 nm was assumed. The thickness of the high-κ layer was adjusted so that the effective oxide thickness (EOT) remains unchanged at 1 nm. The gate current density is shown in the left part of Fig. 5.21 as a function of the gate bias for different material combinations. The commonly assumed limit of 1Acm^{-2} gate leakage is also indicated. Both SiO_2 and Si_3N_4 show a much too high leakage, while Ta_2O_5, ZrO_2, and HfO_2 stay below $1\,\text{Acm}^{-2}$ at V_{GS}=1 V. Due to the low conduction band offset, TiO_2 shows an especially pronounced current increase for positive gate bias.

To assess the material parameters necessary to reach a specific maximum gate current density the gate current has been calculated as a function of the conduction band offset and dielectric permittivity as shown in the right part of Fig. 5.21. Since it is often not possible to vary the thickness of the underlying SiO_2 layer it was again fixed at 0.5 nm and the high-κ thickness was adjusted to reach an EOT of 1.5 nm. The gate current density was evaluated at a fixed bias point of V_{GS}=1.5 V and V_{DS}=0 V. The current density decreases strongly with increasing conduction band offset. Increasing the value of the dielectric permittivity κ also strongly reduces the leakage current due to the higher physical stack thickness. However, materials with a conduction band offset below 1 eV never reach acceptable gate current densities.

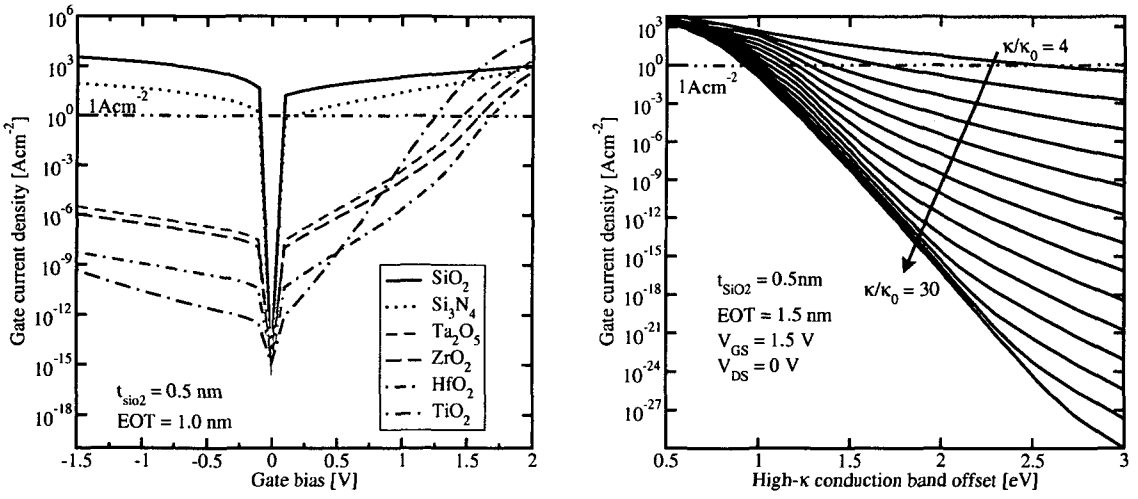


Figure 5.21: Gate current density as a function of the gate voltage for different materials. The dielectric stack consists of a 0.5 nm SiO_2 layer and a high-κ layer with a total EOT of 1.0 nm (left). Dependence of the gate current on the high-κ conduction band offset and dielectric permittivity of a stack with EOT=1.5 nm, a 0.5 nm SiO_2 interface layer at a gate bias of 1.5 V (right).

It may be asked which thickness of the high-κ layer is necessary to achieve a certain gate current density. In the left part of Fig. 5.22 the gate current density is shown for an effective oxide thickness ranging from 0.5 nm to 2.0 nm as a function of the high-κ layer thickness. Again, the stack consists of an underlying 0.5 nm layer of SiO_2 and the simulations are performed at a fixed bias point of V_{GS}=1.5 V and V_{DS}=0 V. In this plot the curves are only drawn for an EOT of 0.5 nm – 2.0 nm, and conduction band offsets of $q\Phi_e = 1$ eV to $q\Phi_e = 3$ eV have been considered. For a conduction band offset of 1 eV, large high-κ thicknesses are necessary to reduce the leakage. Such large stacks may pose problems due to fringing fields from the drain contact which reduce the threshold voltage of the device.

The tradeoff between the dielectric permittivity and the conduction band offset gives rise to further effects as shown in the right part of Fig. 5.22. If the EOT has to be held at a fixed value, an increase of the SiO_2 layer thickness causes a reduced thickness of the high-κ layer. This is shown for different values of the permittivity ($\kappa = 8.0$ – $\kappa = 24.0$). So, the total stack thickness may be larger than 8 nm for $\kappa = 24$, or as small as 1.5 nm if only SiO_2 is used. Such a reduction of the total stack thickness, however, has no clear effect on the leakage. It may cause the gate current density at a specific bias point to stay constant, increase, or even decrease depending on the material parameters. For example, the gate leakage for a material with $\kappa = 24$ and a conduction band offset of 1 eV shows the maximum leakage at a SiO_2 layer thickness of approximately 0.8 nm. Therefore, a clear statement about the optimum thickness of the interface layer obviously depends on the material parameters.

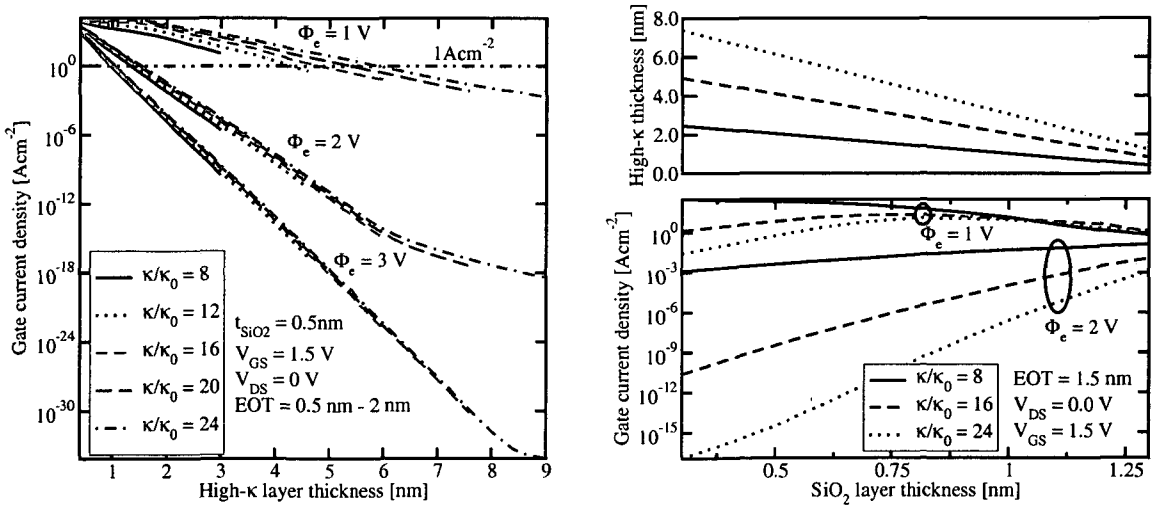


Figure 5.22: Dependence of the gate current on the high-κ layer thickness, conduction band offset, and permittivity of a stack with EOT=2.0 nm and a 0.5 nm SiO_2 interface layer at a gate bias of 1.5 V (left). Dependence of the gate current on the interface layer thickness, conduction band offset, and permittivity of a stack with EOT=1.5 nm at a gate bias of 1.5 V (right).

5.1.6 Trap-Assisted Tunneling in ZrO_2 Dielectrics

Since ZrO_2 offers good material parameters, it was further investigated by means of experiments and numerous results were published [264, 265]. ZrO_2 pMOS capacitors have been fabricated by MOCVD (metal-organic chemical vapor deposition) on p-type (100) silicon wafers with an acceptor doping of $1.5 \times 10^{18}\,\mathrm{cm}^{-3}$ and Al gate electrodes [265]. The overall thicknesses of the dielectric layers have been evaluated by spectroscopic ellipsometry. Employing a dielectric permittivity of the high-κ material of κ/κ_0=18, which has been found for thicker films, the comparison of optical measurements and the results of CV characterization implicates the presence of an interfacial layer with a permittivity in the range of 4 to 8. Table 5.2 summarizes the thicknesses of the high-κ films and interfacial layers. Also given is the effective oxide thickness EOT. The values t_{int} and $t_{\mathrm{high}-\kappa}$ denote the thicknesses of the interface and the high-κ layer.

Layer thickness	t_{int}	$t_{\mathrm{high}-\kappa}$	EOT
6.9	0.75 – 2.0	6.15 – 4.9	2.0
12.7	0.3 – 1.0	12.4 – 11.7	3.0

Table 5.2: Layer thicknesses and EOT of MOCVD-deposited ZrO_2 layers in nm, after HARASEK [265].

In the left part of Fig. 5.23 the measured gate current is shown for the two dielectric layers with the approximate shape of the energy barrier sketched in the insets. As reference the figure also shows the gate current for a 2 nm and a 3 nm SiO_2 layer (dotted lines). As expected, the measured current density is lower than for the SiO_2 counterparts. However, the TSU-ESAKI model cannot reproduce the measurements as it yields tunneling currents orders of magnitude lower than the measurements. This indicates the presence of strong trap-assisted tunneling due to a high trap concentration in the dielectric layer. By assuming a FRENKEL-POOLE like conduction through the dielectric layer the measurements could be reproduced (full lines). Note that in previous studies [264] tunneling through ZrO_2 layers fabricated by magnetron sputtering could be reproduced without considering trap-assisted tunneling. That indicates the presence of a high trap concentration due to the MOCVD process, in contrast to the sputtering process.

To clarify the trap energy level and concentration, the step response of the MOS capacitors has been measured as shown in the right part of Fig. 5.23 for the 12.7 nm ZrO_2 layer annealed in reducing conditions (forming gas) and the 6.9 nm layer annealed under oxidizing conditions. The gate voltage is turned off after being fixed at a value of 2.5 V and the resulting gate current is measured over time. The transient gate current exceeds the static gate current by orders of magnitude and decays very slowly. This behavior can be explained assuming defects in the dielectric layer [266]. Using the trap-assisted tunneling model outlined in Section 3.8.2, a trap energy level of 1.3 eV below the ZrO_2 conduction band edge, a trap concentration of $4.5 \times 10^{18}\,\mathrm{cm}^{-3}$ and an energy loss of 1.5 eV have been found. For the dielectric layer annealed under oxidizing conditions a trap concentration of $4 \times 10^{17}\,\mathrm{cm}^{-3}$ was found.

To predict the performance of devices based on ZrO_2 dielectrics a well-tempered MOSFET as described in [267] with an effective channel length of 50 nm has been simulated. EOT thicknesses of 2 nm and 3 nm SiO_2 and respective ZrO_2 layers have been considered. The left part of Fig. 5.24 depicts the conduction band edge in the channel for different gate-source voltages. It can be seen that the barrier is slightly lower for the ZrO_2 layer at V_{GS}=1.2 V, while it is strongly reduced at V_{GS}=0.1 V, which is due to the pronounced fringing fields from the drain contact.

An additional topic of interest for high-κ dielectrics is the influence of trapped charges in the high-κ layer on the threshold voltage of the device. The trap concentration in the ZrO_2 layer was increased from $10^{15}\,cm^{-3}$ to $10^{19}\,cm^{-3}$ with full trap occupancy in the dielectric layer. It can be seen in the right part of Fig. 5.24 that the threshold voltage strongly increases with rising trap concentration. This effect is therefore contrary to the decrease of the threshold voltage due to fringing fields described above.

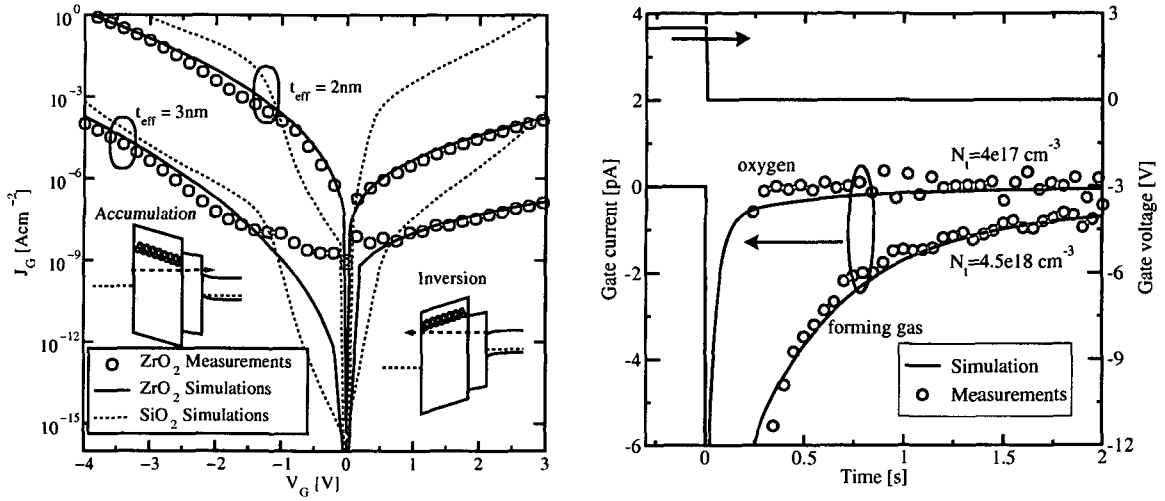


Figure 5.23: Stationary (left) and transient (right) gate current measurements of the ZrO_2 layers performed by HARASEK [265], compared with simulations.

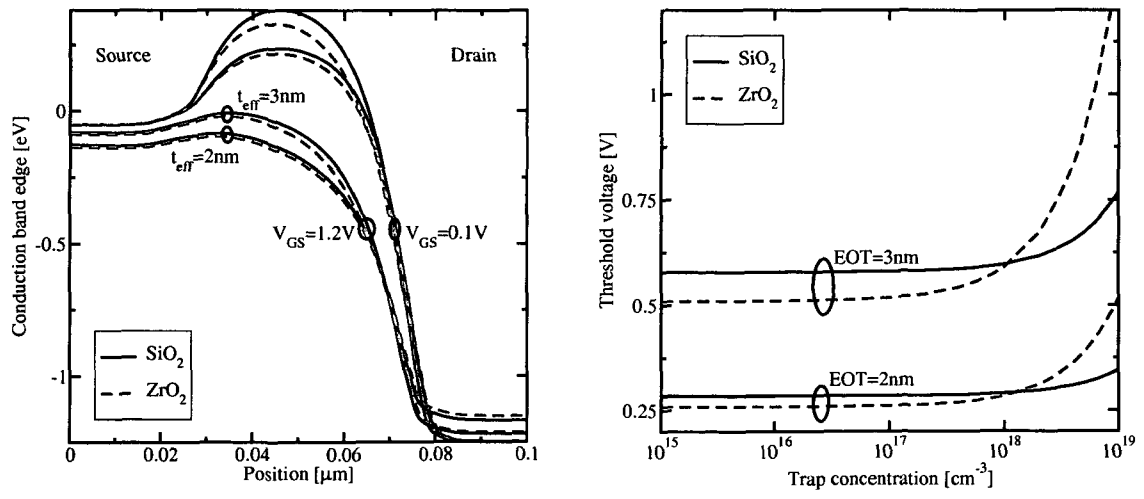


Figure 5.24: Well-tempered MOSFET conduction band edge along the channel for SiO_2 and ZrO_2 dielectrics (left). Influence of the dielectric trap concentration on the MOSFET threshold voltage (right).

5.2 Tunneling in Non-Volatile Memory Devices

Tunneling effects are crucial not only for MOS transistors but also for non-volatile semiconductor memory devices. In contrast to volatile memory devices they retain the stored information without external power supply. NVM devices can be read and programmed like random-access memory (RAM) devices, have a low power consumption, are mechanically robust, and offer the possibility of large-scale integration. They constitute about 10% of the total semiconductor memory market [268]. However, simulation of such devices is often carried out using simplified compact models [269–274]. For the case of stacked gate dielectrics or hot electron injection such models do not capture the device physics and can reproduce measured data only on a fit-formula level. In this section some examples of conventional EEPROM and alternative devices will be studied using the tunneling models described above.

5.2.1 Conventional EEPROM Devices

The basic operating principle of an EEPROM has been presented by KAHNG and SZE in 1967 at Bell Laboratories [275]. The device consists of a control gate and a floating gate on top of a conventional MOS transistor, see Fig. 5.25. A thin tunnel dielectric separates the floating gate from the channel. It must be thick enough to allow up to 10^5 writing and erasing cycles without breakdown — common thicknesses are 6-8 nm. Applying a high positive voltage (about 8-12 V) on the control gate raises the potential of the floating gate by capacitive coupling. The high electric field in the tunnel dielectric ($\approx 10^9$ V/m) leads to FOWLER-NORDHEIM tunneling of electrons from the substrate to the floating gate. The charge on the floating gate changes the threshold voltage of the underlying MOS transistor and is retained even if the control gate voltage is removed. A retention time of 10 years is required for consumer applications like memory cards. While EEPROM cells offer random access for writing and erasing of individual bits, Flash cells can be programmed selectively but erased only at once. This has the advantage of lower cell size. Due to the high electric field in the dielectric, degradation or even breakdown of the dielectric is a major concern. A comprehensive survey of NVM technology is given in [276] and [277].

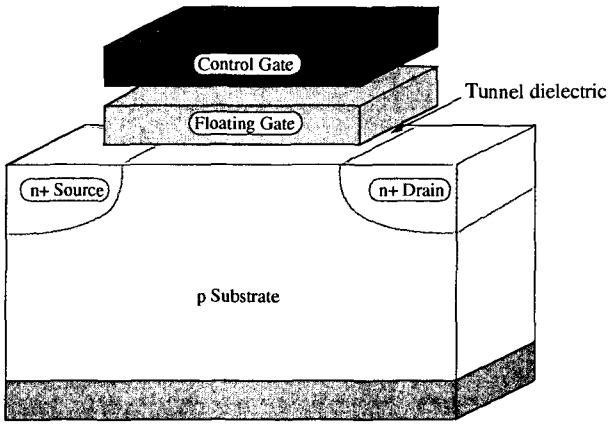


Figure 5.25: The standard EEPROM device.

5.2.1.1 Static SILC in EEPROMs

The speed of the programming and erasing process is one of the main figures of merit of an EEPROM cell. Therefore, strong electric fields are applied at the control gate to allow FOWLER-NORDHEIM tunneling of carriers during programming and erasing cycles. However, due to this repeated high-field stress, trap centers in the dielectric are formed which allow trap-assisted tunneling at low fields and thus reduce the retention time of the devices. This additional current at low bias is known as stress-induced leakage current (SILC) and represents one of the major reliability concerns in contemporary EEPROM devices [196, 219]. In the left part of Fig. 5.26 measured SILC after different stress times for a MOS capacitor with a dielectric thickness of 5.5 nm is shown [189]. The trap-assisted tunneling model outlined in Section 3.8.2 yields excellent agreement with the measured data if the trap concentration is used as a fitting parameter dependent on the stressing time (the model parameters are stated in the figure caption). The transition from the region of mainly trap-assisted tunneling for $V_{GS} < 5\,V$ to the region of FOWLER-NORDHEIM tunneling for $V_{GS} > 5\,V$ is clearly visible. The right part of Fig. 5.26 shows the trap occupancy f_T across the gate dielectric of a MOS capacitor using the gate voltage as parameter. The regions near the gate (right) and near the substrate (left) are only sparsely occupied. Near the gate, the emission time is much smaller than the capture time, and near the substrate, the trap energy lies above the electron energy in the cathode. Some of the trapped electrons face a triangular barrier for the emission process, giving rise to an additional peak in the trap occupancy near the gate side (the anode) of the dielectric. This is due to the wave function interference in the FOWLER-NORDHEIM region (the oscillations are also observed in the emission time of the traps shown in Fig. 3.19).

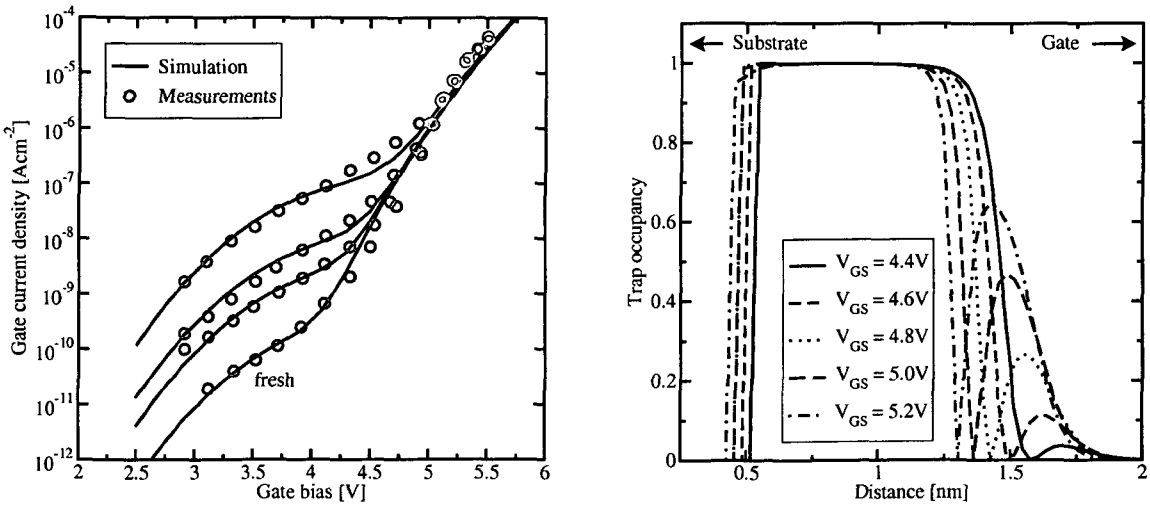


Figure 5.26: Comparison of simulations with measurements of an MOS capacitor with a dielectric thickness of 5.5 nm (left) [189]. The trap energy is 2.7 eV, the phonon energy 130 meV and the HUANG-RHYS factor 10. The trap concentration was set to $9 \times 10^{17}\,cm^{-3}$, $10^{17}\,cm^{-3}$, $3 \times 10^{16}\,cm^{-3}$, and $3 \times 10^{15}\,cm^{-3}$ to fit the measurements (from top to bottom). The trap occupancy across the gate dielectric at different gate voltages (right).

5.2.1.2 Transient SILC in EEPROMs

It has been shown that the transient trap-assisted tunneling current can be described by a rate equation which gives rise to an exponential behavior of the tunneling current over time, see Section 3.8.2.4. The left part of Fig. 5.27 shows measurements of the gate current density of MOS capacitors as a function of time with dielectric thicknesses of 8.5 nm and 13.0 nm, compared to simulations [188]. Initially, the traps are empty which can be achieved by applying flat band conditions. At $t = 0$ s, the gate voltage is turned on (-5.8 V and -8.3 V for the thinner and the thicker dielectric, respectively) and the traps are filled according to their specific capture and emission time constants. This charging current consists of an emission and a capture current, which may exceed the steady-state current by orders of magnitude. A good fit to the measured data can be achieved using the trap parameters indicated in the figure caption.

The right part of Fig. 5.27 shows the gate current of an MOS capacitor for an applied rectangular pulse with a frequency of 100 kHz assuming initial flat band conditions. It can be seen that the time constants of the trap filling and emptying processes are not equal but depend on the applied voltage, since different voltages lead to different capture and emission times. The spikes in this figure are due to the sudden voltage change while the trap concentration remains constant: In the transition from 3.0 V to 3.5 V the barrier shape changes suddenly, and traps are rapidly emptied. Traps near the cathode are filled and it takes several micro seconds until the new steady state is reached. Thus, dielectric materials which have such a high trap concentration may lead to considerable problems for high-frequency applications.

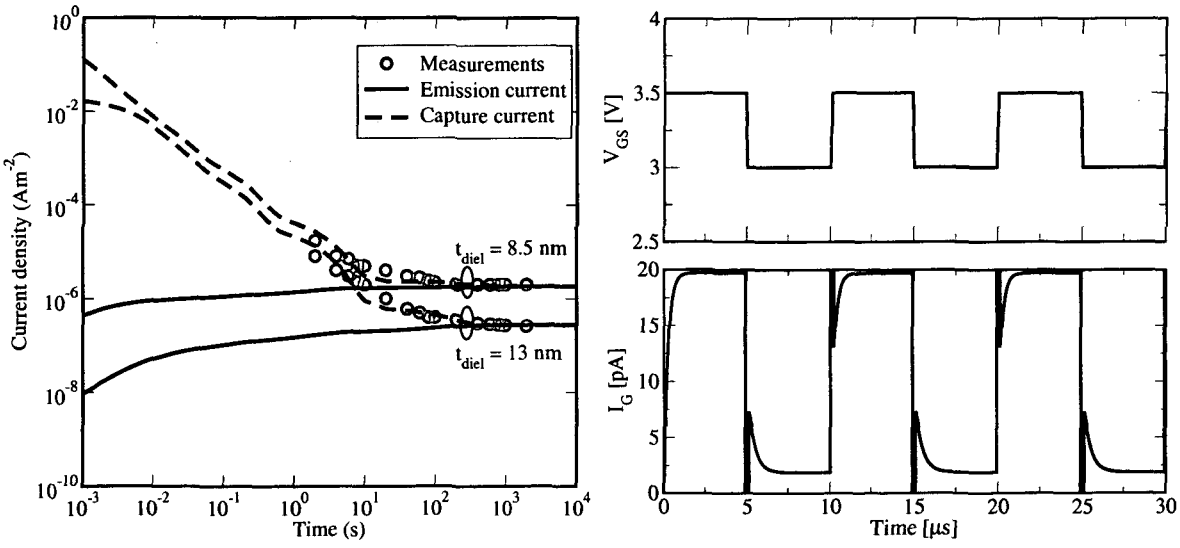


Figure 5.27: Transient capture and emission currents (left) of MOS capacitors at a gate bias of -5.8 V and -8.3 V. For the thinner dielectric, a trap energy of 2.5 eV and a trap concentration of 3×10^{18} cm^{-3} was used, while for the thicker dielectric, a trap concentration of 10^{18} cm^{-3} was found. The right figure shows transient simulation results of a MOS capacitor with a gate dielectric thickness of 3 nm and a trap energy level of 3 eV.

For EEPROM devices the charging and discharging characteristics are crucial: Programming and erasing should happen as fast as possible, therefore, high voltages are applied. The discharging current over time, on the other hand, determines the retention time and must be very low. To allow the simulation of these characteristics, the contact condition `Floating` was implemented in MINIMOS-NT. For floating contacts, the electrostatics in the device is acquired in the initial time step. Then, the current contact condition at the floating gate contact is set to zero (no out-flowing contact current) [226]. The tunneling current to or from the floating gate changes the charge and thus the voltage on the contact. For the simulation of the programming or erasing processes, first all voltages are set to zero (the voltage which is assumed to represent the empty state). Then, the charge at the floating gate is used as charge contact condition and the programming voltage is applied at the control gate. Fig. 5.28 shows the control gate voltage, floating gate voltage, floating gate charge, and tunneling current for the programming, storing, and erasing processes. It can be seen that the programming and erasing pulses must be carefully optimized to avoid over-erase, since the tunnel current density for positive and negative voltages on the floating gate is not equal. This is frequently addressed in the literature [278, 279].

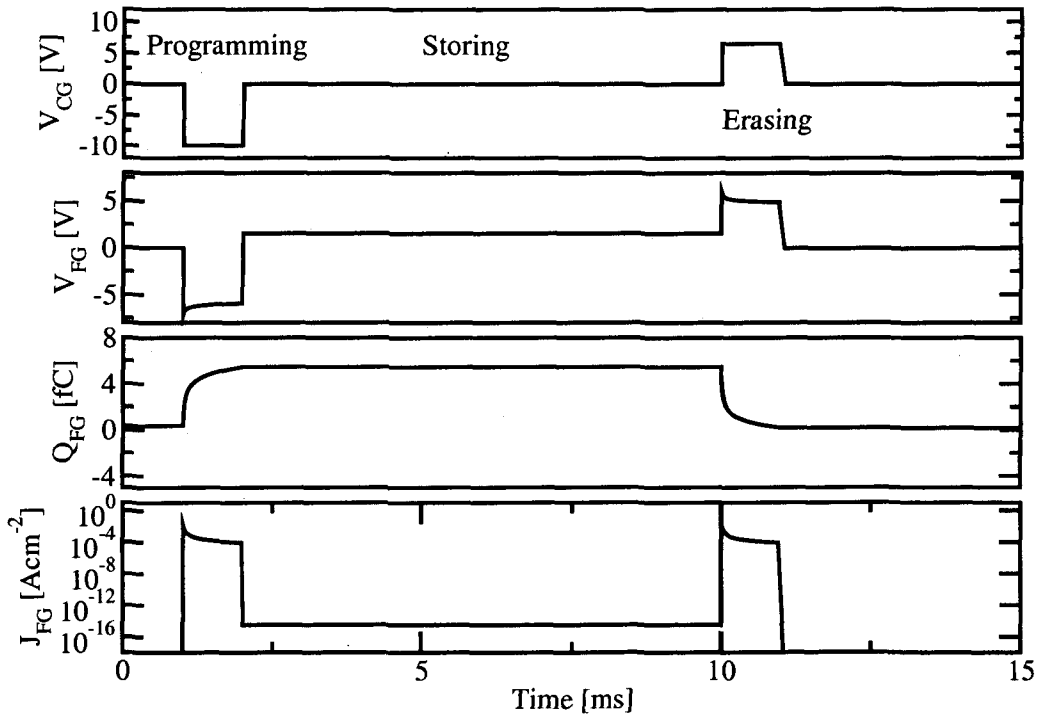


Figure 5.28: Discharging curve of an EEPROM. The floating gate is charged at a control gate voltage of -10 V and is then left floating at a control gate voltage of 0 V. Since the gate current density is not equal for positive and negative voltages, the program and erase pulses must be carefully chosen to avoid over-erase. Due to the low storing time in this example almost no charge is lost during the storing period.

5.2.2 Alternative Non-Volatile Memory Devices

Strong efforts are undertaken to improve the standard floating-gate EEPROM cell shown in Fig. 5.25 in terms of integration density, endurance, reliability, program time, erase time, and retention time. Some of these approaches are depicted schematically in Fig. 5.29 and Fig. 5.30. EEPROM devices with a tunnel window near the drain contact have been introduced to reduce the charge loss from the floating gate and thus reach higher retention time. However, due to the small area of the tunnel window high voltages have to be used at the drain contact which again reduces cell reliability.

Recently, CAYWOOD *et al.* proposed a device structure where non-selected cells are isolated from the drain and source contacts by two additional side gates [280]. In this device electrons tunnel from the inverted channel to the floating gate. The large area reduces programming and erasing time. Furthermore, the capacitive coupling between the control gate and the floating gate is higher than in the standard EEPROM cell which allows to use lower programming and erasing voltages. No drain-source bias is applied for charging, thus the power consumption is low and the injected electrons are less likely to cause degradation of the dielectric. The control gate functions as a select transistor which isolates unselected cells from the high voltages at the shared source and drain contacts during read and write access of neighboring cells.

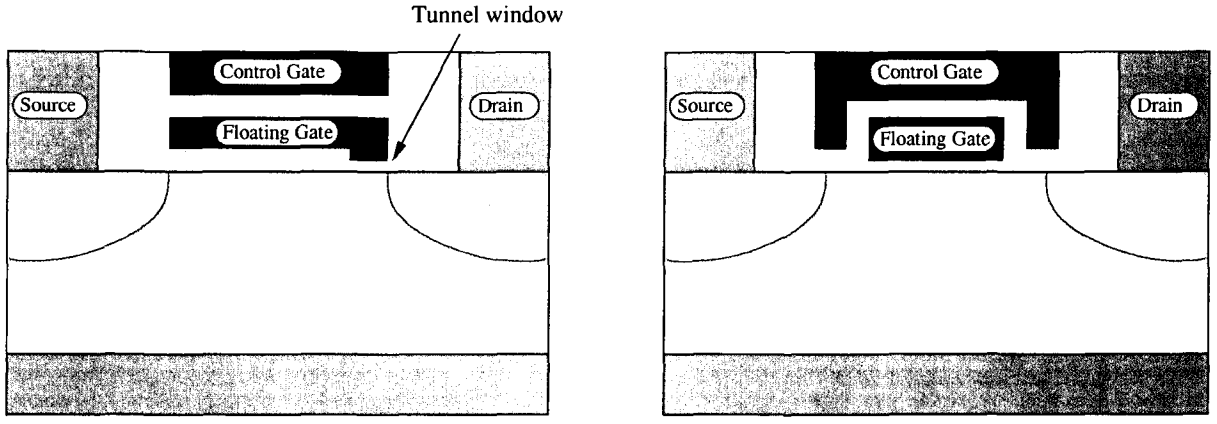


Figure 5.29: Alternative NVM structures: EEPROM with tunnel window (left), CAYWOOD memory device (right).

In contrast to the reduction of the cell footprint, integration density can also be increased by storing more than one bit on a standard EEPROM cell. This can be achieved by tailoring the programming and erasing pulses in such a way that the threshold voltage falls into one of 4, 8, or 16 voltage ranges. The different threshold voltages can be distinguished by the sensing circuits, resulting in two, three, or four bits which can be stored in the cell. However, charge loss must be extremely low over time and the threshold voltages have to be detected very precisely.

Single-poly devices as shown in the left part of Fig. 5.30 have been proposed to integrate NVM devices in standard CMOS logic processes, thus enabling an embedded memory. The control gate lies next to the floating gate and capacitive coupling is achieved by a layer of highly doped silicon. While such devices can readily be integrated into existing CMOS process flows, they come at the cost of a large footprint.

A different approach to store more than one bit in a single memory cell is to split the floating gate into two separate segments. If a non-uniform doping in the source and drain side of the channel is used, different amounts of charge can be stored in each floating gate. Such device structures are either achieved using separate metallic floating gates like the contacts FG1 and FG2 in the right part of Fig. 5.30 [182], or using a layer of trap-rich dielectric [281].

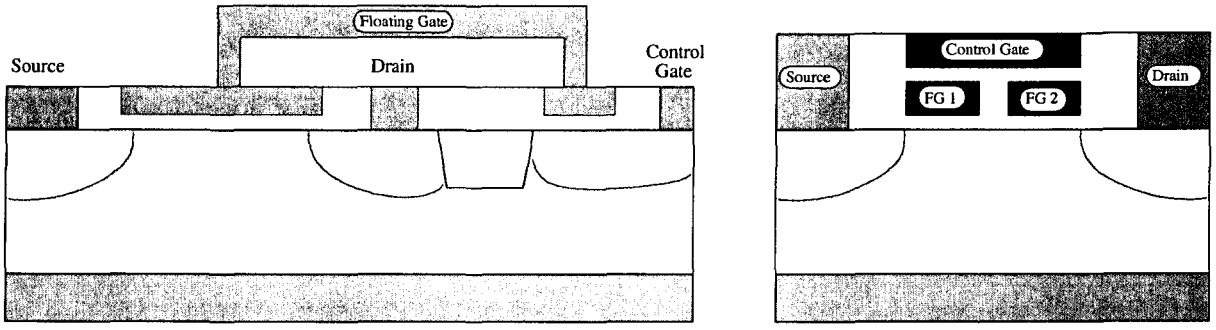


Figure 5.30: Alternative NVM structures: Single-poly EEPROM (left), split-gate EEPROM (right).

In the following sections three of the most promising alternative EEPROM devices will be studied in detail. These are

- **Quantum dot and trap-rich dielectric** based devices. In these devices, charging and discharging is achieved by tunneling of electrons to and from localized trapping centers in the dielectric.
- **Multi-barrier tunneling devices** consist of a floating gate — or memory node — which is separated from the control gate by several thin dielectric layers. By the use of a side gate, the tunneling current through these barriers can be controlled selectively. In contrast to EEPROMs the tunneling current flows from the floating gate to the control gate and not to the channel. Extremely high I_{on}/I_{off} ratios can be achieved because the tunneling current is controlled by a separate side gate contact.
- Devices where the tunnel dielectric consists of **stacked dielectrics** which are engineered in such a way that they block tunneling in the off-state, but allow strong tunneling in the on-state.

5.2.2.1 Non-Volatile Memory Devices Based on Trap-Rich Dielectrics

A SONOS (silicon-oxide-nitride-oxide-silicon) device is a non-volatile memory where the charge is stored in a layer of trap-rich dielectric material instead of a floating gate as in an EEPROM. Fig. 5.31 shows an example where a layer of Si_3N_4 is sandwiched between two layers of SiO_2. Electrons tunneling from the substrate are trapped and redistribute themselves in separate trapping centers. This has the advantage that the charge is stored independently in the traps. A leaky path in the tunnel dielectric cannot lead to full charge loss, as it is the case in conventional EEPROM devices. Therefore, reliability and retention time is increased [168, 209, 282–291].

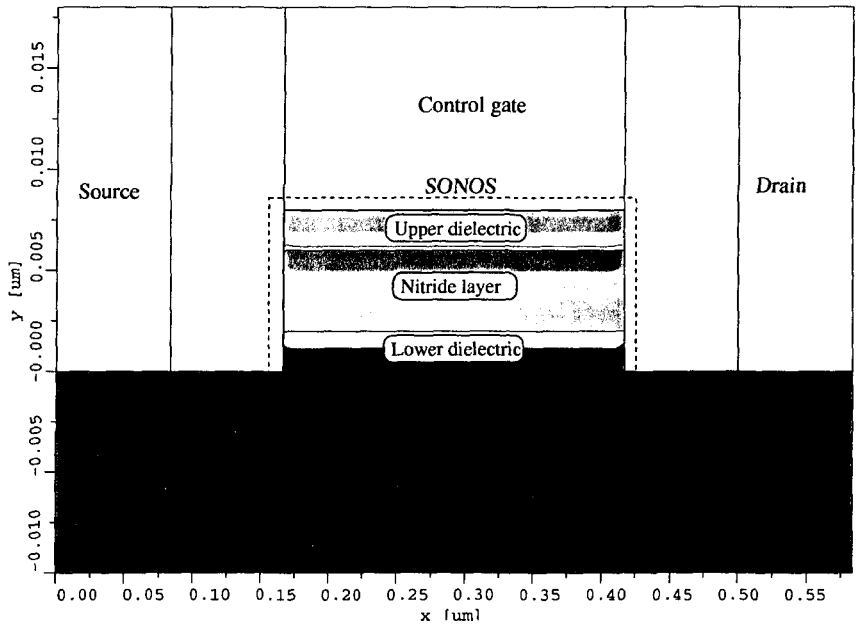


Figure 5.31: SONOS device structure. A layer of trap-rich dielectric, such as highly defective Si_3N_4, is sandwiched between two SiO_2 layers.

The band diagram along the dielectric of such a device is shown in Fig. 5.32 for the programming, storing, and erasing processes. By applying a positive voltage at the gate contact, electrons tunnel through the tunnel dielectric into the trap region. The traps are filled with electrons and become negatively charged. Because of the tunnel dielectric this charge is stored even if the bias is removed. To erase the memory cell, a negative voltage is applied on the gate contact, leading to a reduced potential barrier and a high tunneling current of electrons out of the traps. Important device parameters are the charging and discharging current through the dielectric, the drain current in the on- and off-state, and the retention time.

The trap-assisted tunneling model can be applied to simulate device characteristics of this device, where three layers of SiO_2 have been used and the trap concentration and trap energy level in the middle layer was chosen to resemble a layer of silicon nitride. The transient trap occupancy for a discharging process starting from an initial condition of 2 V at the gate contact is shown in Fig. 5.33. Initially the traps are filled. Over time, the electrons leak through the lower dielectric into the channel. After 10^9 s almost no more charge is stored in the trap-rich dielectric.

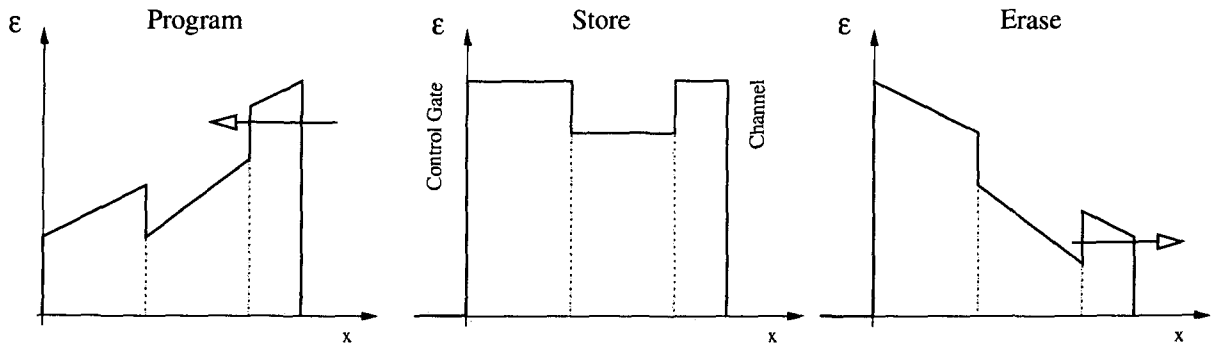


Figure 5.32: Conduction band edge in a SONOS device for the programming, storing, and erasing process.

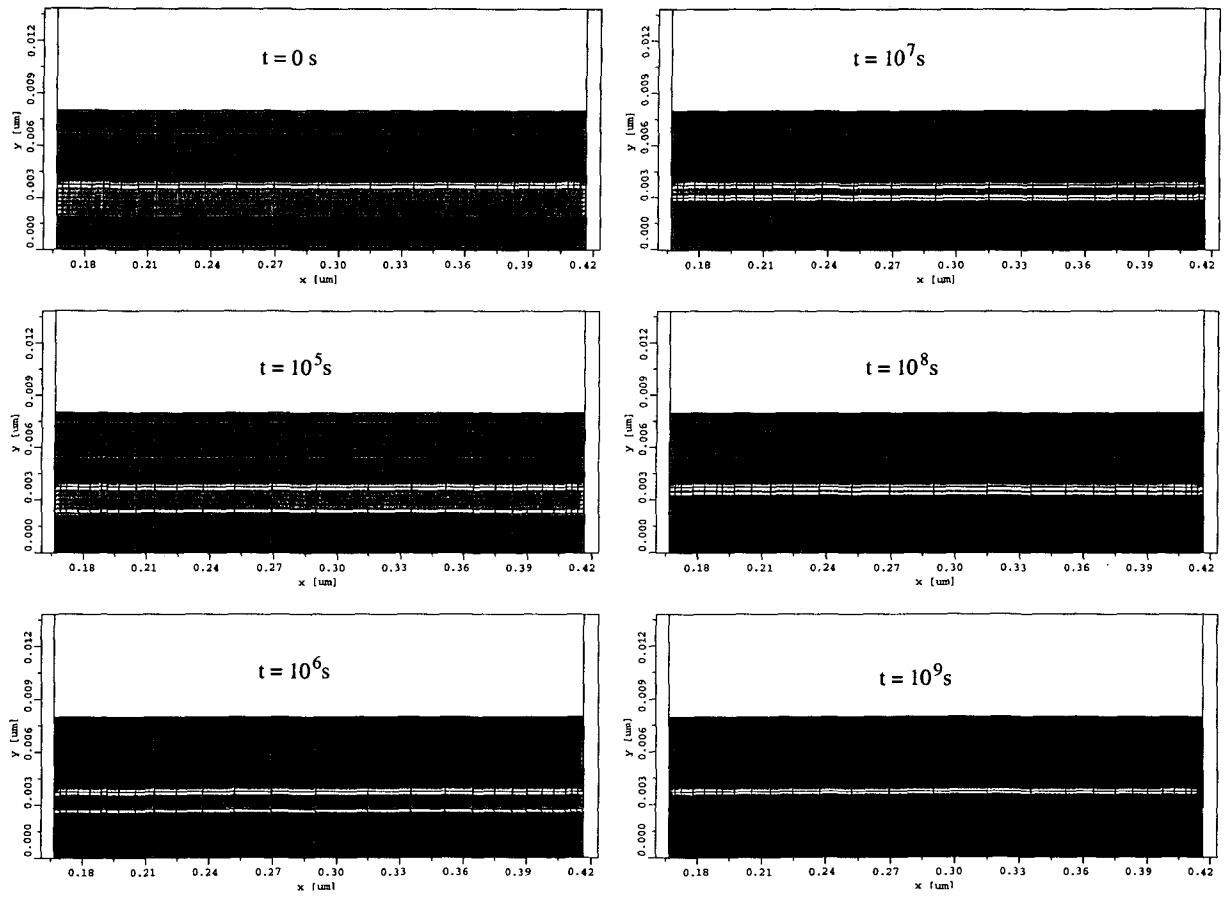


Figure 5.33: Transient trap occupancy in the trap-rich dielectric layer of a SONOS device which is discharged from t =0 s to $t = 10^9$ s.

5.2.2.2 Multi-Barrier Tunneling Devices

One of the main shortcomings of conventional EEPROM devices is that the current in the on-state and off-state — the programming and leakage currents — flow through the same tunnel dielectric and face the same energy barrier. They cannot be optimized independently: Increasing the thickness of the tunnel dielectric reduces the leakage, but also reduces the on-state current and thus increases the programming time. Multi-barrier tunneling devices offer a solution to this problem. Planar localized-electron device memory (PLEDM) cells have been presented by NAKAZATO *et al.* in [292], and promising results have been reported [293–296]. The principle of a PLEDM is to put a PLED transistor (PLEDTR) on top of the gate of a conventional MOSFET, as shown in Fig. 5.34. The charge on the memory node, which acts as a floating gate, is provided by tunneling of carriers through the PLED transistor which consists of a stack of Si_3N_4 barriers sandwiched between layers of intrinsic silicon. Upper and lower barriers prevent diffusion from the polysilicon contacts, while the middle barrier — the central shutter barrier (CSB) — blocks the tunneling current in the off-state. The PLED transistor has two side gates which are separated by a thin dielectric layer. In the on-state the energy barriers are heavily reduced by the voltage on the side gates, causing a strong tunneling current to flow at the interface to the side gate dielectric. In the off-state, however, the side gates are turned off and the energy barrier blocks the leakage current. As in a conventional EEPROM, the charge on the memory node is used to control the underlying MOS transistor. Only a small amount of charge has to be added to or removed from the memory node to change the state of the memory cell.

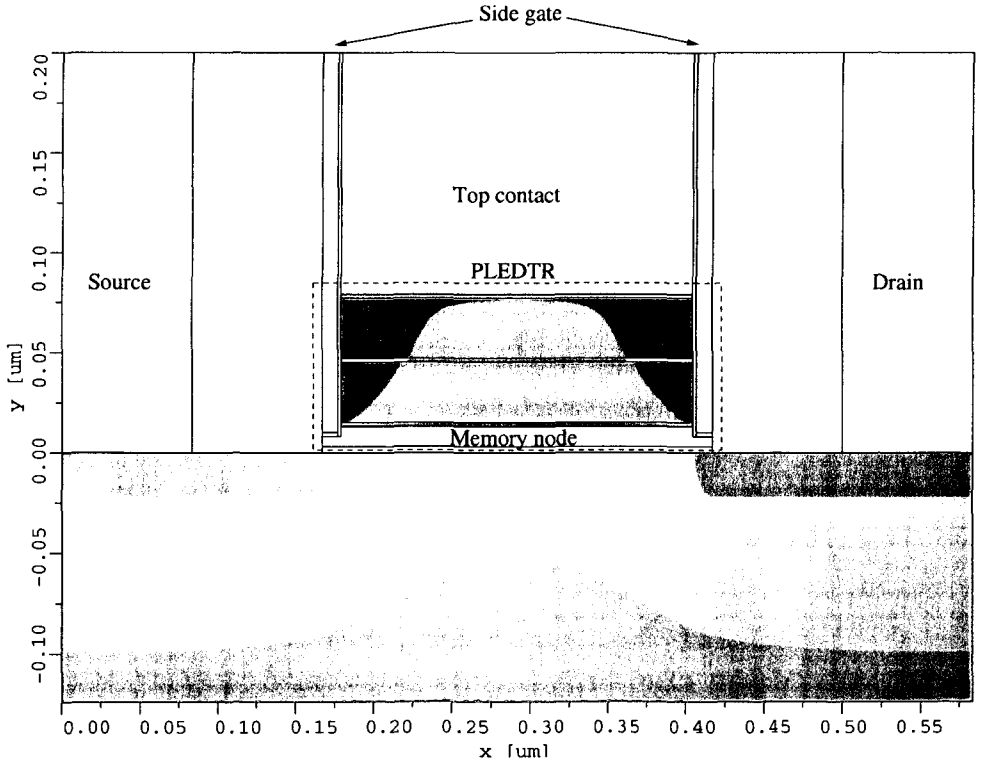


Figure 5.34: Conduction band edge energy in the PLEDM device.

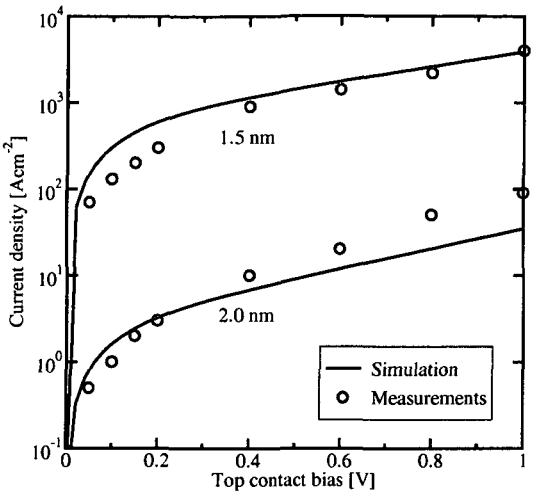


Figure 5.35: PLEDM calibration of the tunneling current density for a single Si_3N_4 layer with 1.5 nm and 2 nm thickness.

For the simulation of such devices measurement results for a single Si_3N_4 barrier diode [296] have been used to calibrate the model, as shown in Fig. 5.35. For calibration the carrier mass in the dielectric was used as a fit parameter. Electron and hole masses of $0.5\,m_0$ and $0.8\,m_0$ where found to reproduce the data. The Si_3N_4 barrier was modeled with a barrier height of 5 eV and a conduction band offset of 2 eV to the silicon conduction band edge with the relative dielectric permittivity being 7.5. Fig. 5.36 shows in the left part the conduction band edge along the PLEDTR and in the right part the electron wave function for the case of a top contact bias of 1 V and a side gate bias of 2 V. The wave function has been acquired using the QTB method described in Section 3.5.4.

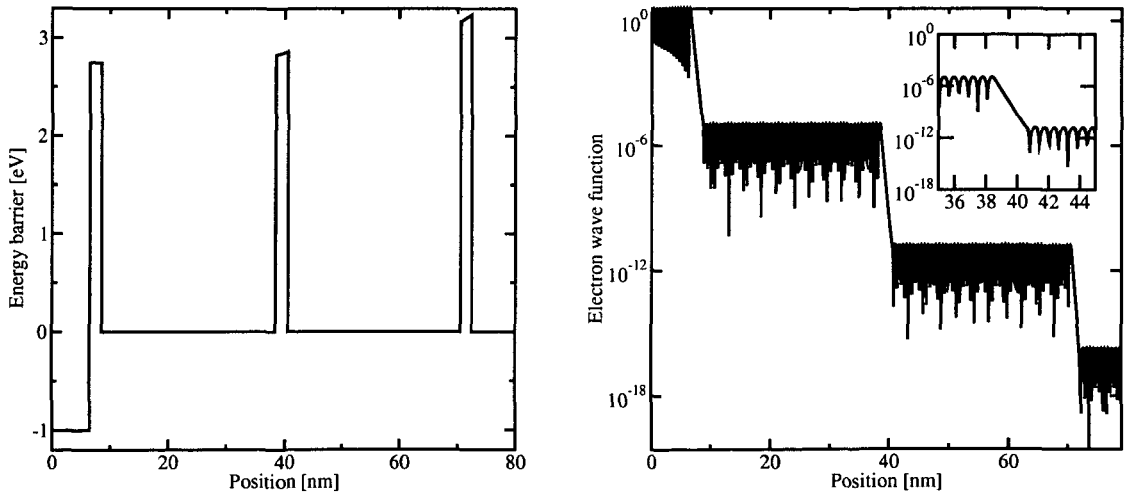


Figure 5.36: PLEDM conduction band edge (left) and electron wave function (right) for a top contact bias of 1 V and a side gate bias of 2 V.

The effect of the position and size of the central shutter barrier as well as the effect of shrinking the stack width have been investigated. Two cell states have been assumed: an on-state with 3 V applied on the top contact and the side gate, and an off-state with 0.8 V applied on the memory node and 0 V on the side gate. In both states the charging and discharging current was extracted. The PLEDTR had a stack width of 180 nm and a stack height of 100 nm. The thickness of the upper and lower barriers was set to 2 nm. The left part of Fig. 5.37 shows the effect of different CSB thicknesses on the on- and off-current of the device. While the on-current is hardly influenced by the different thicknesses, the off-current is very sensitive to it. Also, the position of the CSB is critical, because for a CSB located near the memory node, the energy barrier will be reduced in the off-state by the charge on the memory node. If, on the other hand, the CSB is placed near the top contact, the energy barrier is not suppressed in the off-state and the off-current is much lower. The on-current is also reduced by this effect, but the amount of reduction is much lower as compared to the off-current, due to the fact that the on-current mainly depends on the voltage of the side gate. Thus, the I_{on}/I_{off} ratio increases with the thickness of the central shutter barrier and is highest for a CSB located near the top contact. Such an asymmetry in the IV characteristics depending on the position of the central shutter barrier has already been observed experimentally [296].

In [297] the feasibility of very narrow silicon-insulator stacks is shown. This encourages the assumption that a reduction of the stack width is possible. Fig. 5.37 shows the on- and off-currents of the device with a CSB thickness of 10 nm for a stack width of 140 nm down to 20 nm. It can be seen that a reduction of the stack width leads to increasing on-currents and decreasing off-currents. The reason is that the current in the on-state, which mainly flows as a surface current near the side gate, is not reduced by the decreased width of the stack. It even increases for very low stack widths which may be due to the fact that the energy barriers at the side of the stack merge for very low stack widths. The off-current, on the other hand, is directly proportional to the stack area and can thus be directly downscaled by shrinking the stack width. For a stack width of 20 nm, I_{on}/I_{off} ratios of more than 10^{32} can be reached.

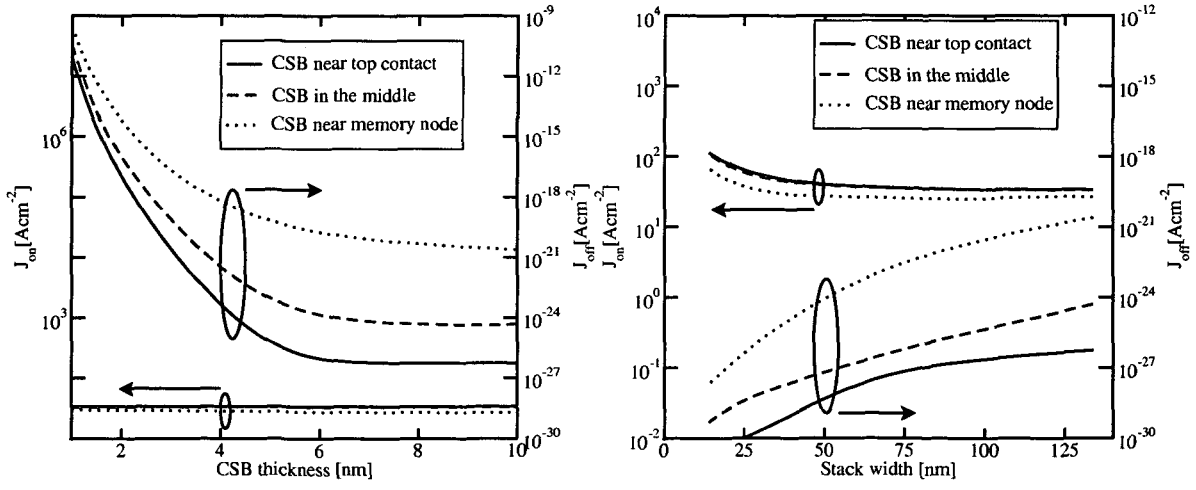


Figure 5.37: On-current density and off-current density as a function of the thickness of the central shutter barrier (left) and the stack width (right).

5.2.2.3 Non-Volatile Memory Devices Based on Crested Barriers

As shown in Section 5.2.2.2, one of the most important figures of merit of a non-volatile memory cell is its I_{on}/I_{off}-ratio: A high on-current leads to low programming and erasing times, and a low off-current increases the retention time of the device. This ratio can be increased if, for a given device, the tunneling current in the on-state (the charging/discharging current) is increased or, in the off-state (during the retention time), decreased. With a single-layer dielectric it is not possible to tune on- and off-current independently. However, if the tunnel dielectric is replaced by a dielectric stack of varying barrier height as shown in Fig. 5.38, it becomes possible. In this figure the device structure and the conduction band edge in the on- and off-state are shown. The device consists of a standard EEPROM structure, where the tunnel dielectric is composed of three layers. The middle layer has a higher energy barrier than the inner and outer layers. The flat-band case is indicated by the dotted lines.

In the on-state a high voltage is applied on the top contact. The middle energy barrier is strongly reduced and gives rise to a high tunneling current. If the dielectric would consist of a single layer, the peak of the energy barrier would not be reduced. Thus, the on-current is much higher for the layered dielectric. In the off-state a low negative voltage — due to charge stored on the memory node — is applied. The middle barrier is only slightly suppressed and blocks tunneling. The off-current is only slightly lower than for a single-layer dielectric. This behavior results in a high I_{on}/I_{off} ratio. A high suppression of the middle barrier in the on-state requires a low permittivity of the outer layers so that the potential drop in the outer layers is high [261]. This device design was first proposed by CAPASSO *et al.* in 1988 [298] based on AlGaAs-GaAs devices and later used by several authors [299, 300], where it became popular as *crested-barrier* memory or VARIOT (*varying oxide thickness* device).

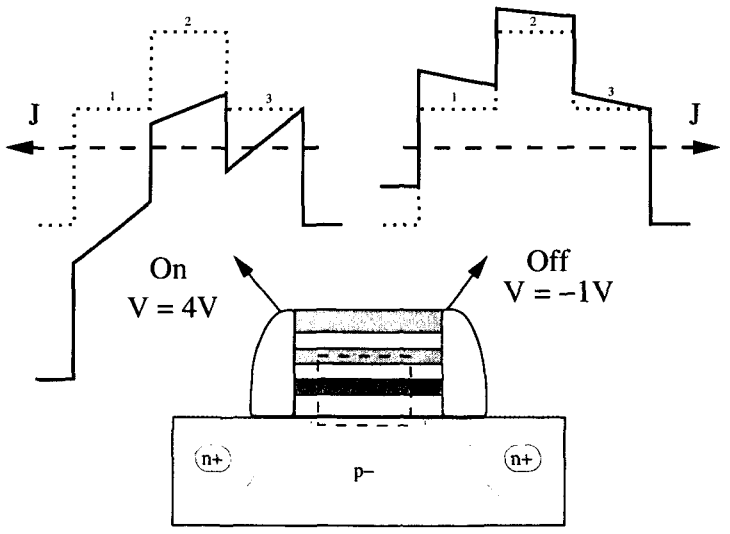


Figure 5.38: Device structure and operating principle of a non-volatile memory based on crested barriers.

The gate current density of the device depicted in Fig. 5.38 is shown as a function of the gate bias in the left part of Fig. 5.39. A stack thickness of 5 nm was chosen. Since the middle layers must have a high band gap, only few material combinations are possible. For the simulations middle layers of Al_2O_3 and SiO_2 have been chosen, with outer layers of Y_2O_3, Si_3N_4, and ZrO_2. For comparison full SiO_2 and Si_3N_4 stacks have also been simulated (the dotted and dash-dotted lines). While Y_2O_3 shows a very high off-current, stacks with outer layers of Si_3N_4 or ZrO_2 and Al_2O_3 as middle layer show good ratios between the on-state (positive gate bias) current density and the off-state (negative gate bias) current density.

The important figure of merit, however, is the I_{on}/I_{off}-ratio. In the right part of Fig. 5.39 the I_{on}/I_{off}-ratio is shown for Si_3N_4 and ZrO_2 stacks with SiO_2 and Al_2O_3 middle layers as a function of the thickness of the middle layer. Also shown is the ratio for a layer of SiO_2 and Si_3N_4 alone. It is obvious that the ratio strongly depends on the thickness of the middle layer, and both minima and maxima can be observed. Only outer layers of Si_3N_4 lead to a significantly increased performance as compared to full layers of SiO_2 or Si_3N_4. A middle layer thickness around 1–2 nm for the assumed 6 nm stack gives optimum performance.

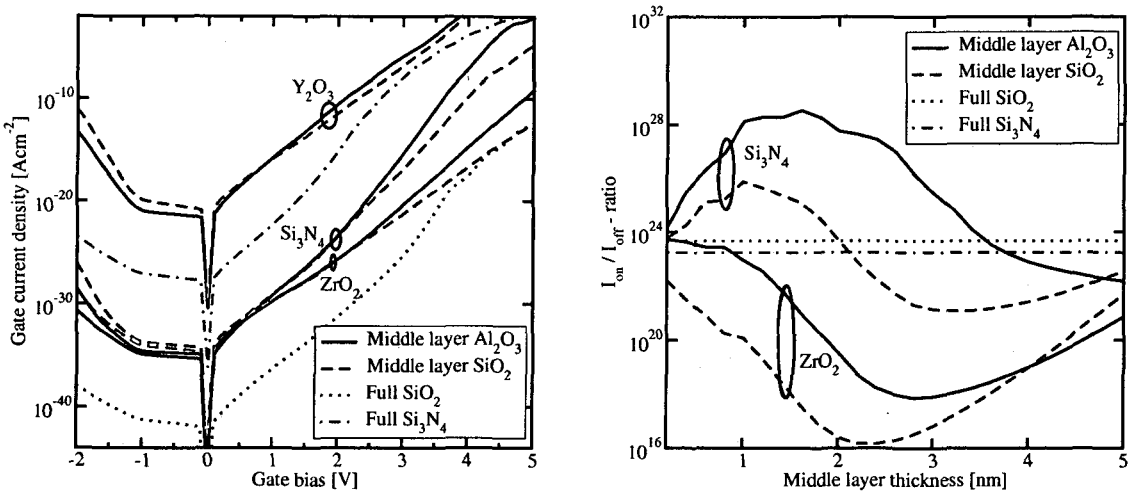


Figure 5.39: Gate current density as a function of the gate bias for different materials of the middle layer, compared to full SiO_2 and Si_3N_4 layers (left). Ratio between the on-current and the off-current as a function of the middle layer thickness for different materials of the outer layers (Si_3N_4 and ZrO_2) and middle layers (Al_2O_3 and SiO_2), compared to the resulting current density using full layers of SiO_2 and Si_3N_4 (right).

'At least some knowledge of the average pattern is the beginning of wisdom, and although we have not learnt as much as might be hoped, it is always worth remembering how little we knew when we were started.'

Jonathan Temple

Chapter 6

Summary and Conclusions

TUNNELING EFFECTS in semiconductor devices were investigated and simulated with MINIMOS-NT. Starting with an introduction to CMOS technology and semiconductor device simulation, a hierarchy of tunneling models was outlined. Three main properties were identified to influence the tunneling process: The carrier energy distribution function, the transmission coefficient, and the presence of traps in the dielectric layer.

The energetic distribution of carriers was investigated using different approximations, such as the frequently applied FERMI-DIRAC or MAXWELL-BOLTZMANN statistics. However, these approximations are only valid near equilibrium. Comparisons with the results from Monte Carlo simulations showed that in turned-on devices the distribution function strongly deviates from the ideal shape. Some non-MAXWELLian models were reviewed and it was found that a model which is based on the solution variables of a six-moments transport model accurately reproduces the Monte Carlo results.

The quantum-mechanical transmission coefficient can be computed from the solution of the stationary SCHRÖDINGER equation. Several approximations and analytical formulae were outlined. For a single-layer dielectric the analytical WKB approximation or GUNDLACH's formula can be used. For arbitrary-shaped energy barriers the numerical WKB, the transfer-matrix, or the quantum transmitting boundary method can be applied. It was found that the transfer-matrix method is prone to numerical problems due to the repeated matrix multiplications. The quantum transmitting boundary method turned out to be more robust.

Defects in the dielectric layer give rise to trap-assisted tunneling which leads to an additional tunneling current at low bias. After reviewing several models from the literature a recently presented inelastic trap-assisted tunneling model was adapted to avoid the numerical calculation of the overlap integral in the dielectric layer. This yielded a fully analytical model which was further developed to include transient trap charging and discharging effects.

All methods were implemented into the general-purpose device simulator MINIMOS-NT. The implementation was shortly described. Furthermore, a multi-dimensional SCHRÖDINGER solver was implemented to calculate the transmission coefficient and the energy eigenvalues of arbitrary energy barriers. This solver was designed in such a way that both open and closed boundary conditions can be applied on the same band diagram.

Several examples were studied where a general distinction between tunneling in MOS transistors, where it is a parasitic effect, and tunneling in non-volatile memory devices, where it is crucial for the device functionality, was made. Tunneling in MOS transistors was investigated, where special attention was paid on the investigation of the different tunneling paths from the gate to the channel and from the gate to the source and drain extension regions.

Furthermore, the importance of the carrier distribution functions for modeling of gate leakage in turned-on devices was shown. If a heated MAXWELLian approximation was used for the description of hot-carrier tunneling, the gate current density was heavily overestimated. This effect was found to be especially pronounced for devices with short gate lengths.

In future CMOS devices the use of alternative dielectric materials instead of SiO_2 will make the reduction of the effective oxide thickness possible. Several candidate materials were studied and it was found that they show a pronounced correlation between the barrier height and the permittivity. This makes optimization necessary to find the optimum layer composition. Furthermore, the investigation of a MOS capacitor with a ZrO_2 dielectric showed that the strong defect density makes the use of trap-assisted tunneling models a *sine qua non* for these materials.

In addition to MOS transistors non-volatile memory devices were studied. A general overview of non-volatile memory technology was followed by an investigation of three selected device structures: devices where the floating gate contact is replaced by a layer of trap-rich dielectric, multi-barrier tunneling devices, and devices which are based on crested barriers. Especially the multi-barrier tunneling devices allow an extremely high I_{on}/I_{off} ratio. The trap-rich dielectric devices, on the other hand, are easier to fabricate and have a smaller footprint. Devices which are based on crested barriers allow to tune the on- and off-current density independently. However, the I_{on}/I_{off}-ratio heavily depends on the thicknesses of the dielectric layers and simulation is necessary to find the optimum values. The investigated non-volatile memory applications are expected to show high performance, however, the bad quality of the interface between the dielectric layers may offset the advantage in the I_{on}/I_{off}-ratio.

The implementation of these direct and trap-assisted tunneling models allows the simulation and analysis of semiconductor devices where tunneling is either a parasitic effect or deliberately used as a part of the device functionality. Future work will concentrate on the coupling of the developed multi-dimensional SCHRÖDINGER solver to MINIMOS-NT to simulate quantization effects in MOSFET inversion layers and for the characterization of alternative dielectric materials. The numerical methods to calculate tunneling from quasi-bound states will be investigated in more detail. Finally, the developed tunneling models will be applied to the simulation of gate dielectric reliability issues.

Appendix A

The FOWLER-NORDHEIM Formula

The TSU-ESAKI expression (3.12) for the tunnel current density reads

$$J = \frac{4\pi \mathrm{q} m_{\mathrm{eff}}}{h^3} \int_{\mathcal{E}_{\min}}^{\mathcal{E}_{\max}} TC(\mathcal{E}_x)\,\mathrm{d}\mathcal{E}_x \int_0^{\infty} (f_1(\mathcal{E}) - f_2(\mathcal{E}))\,\mathrm{d}\mathcal{E}_\rho \;, \tag{A.1}$$

where the total energy is split into a longitudinal and a transversal energy

$$\mathcal{E} = \mathcal{E}_x + \mathcal{E}_\rho \;. \tag{A.2}$$

The goal is to find a simple approximation of (A.1) which avoids numerical integration. As a **first approximation**, $T \to 0$ is assumed [96]. This allows to replace the FERMI function $f(x)$ by the step function

$$\begin{aligned} f_1(\mathcal{E}) = f(\mathcal{E} - \mathcal{E}_{\mathrm{f},1}) &= \begin{cases} 1 & \text{for } \mathcal{E} \le \mathcal{E}_{\mathrm{f},1} \\ 0 & \text{for } \mathcal{E} > \mathcal{E}_{\mathrm{f},1} \end{cases} \\ f_2(\mathcal{E}) = f(\mathcal{E} - \mathcal{E}_{\mathrm{f},2}) &= \begin{cases} 1 & \text{for } \mathcal{E} \le \mathcal{E}_{\mathrm{f},2} \\ 0 & \text{for } \mathcal{E} > \mathcal{E}_{\mathrm{f},2} \end{cases} \end{aligned} \;. \tag{A.3}$$

Without loss of generality it can be assumed that $\mathcal{E}_{\mathrm{f},1} > \mathcal{E}_{\mathrm{f},2}$ (see Fig. A.1). The innermost integral can then be evaluated analytically for three distinct regions

$$\begin{aligned} \int_0^{\infty} (f(\mathcal{E} - \mathcal{E}_{\mathrm{f},1}) - f(\mathcal{E} - \mathcal{E}_{\mathrm{f},2}))\,\mathrm{d}\mathcal{E}_\rho &= \mathcal{E}_{\mathrm{f},1} - \mathcal{E}_{\mathrm{f},2} \quad \text{for} \quad \mathcal{E}_x \le \mathcal{E}_{\mathrm{f},2} \;, \\ &= \mathcal{E}_{\mathrm{f},1} - \mathcal{E}_x \quad \text{for} \quad \mathcal{E}_{\mathrm{f},2} \le \mathcal{E}_x \le \mathcal{E}_{\mathrm{f},1} \;, \\ &= 0 \quad \text{for} \quad \mathcal{E}_x > \mathcal{E}_{\mathrm{f},1} \;. \end{aligned} \tag{A.4}$$

This leads to the following expression for the current density:

$$J = \frac{4\pi \mathrm{q} m_{\mathrm{eff}}}{h^3} \left(\underbrace{\int_{-\infty}^{\mathcal{E}_{\mathrm{f},2}} TC(\mathcal{E}_x)(\mathcal{E}_{\mathrm{f},1} - \mathcal{E}_{\mathrm{f},2})\,\mathrm{d}\mathcal{E}_x}_{\approx 0} + \int_{\mathcal{E}_{\mathrm{f},2}}^{\mathcal{E}_{\mathrm{f},1}} TC(\mathcal{E}_x)(\mathcal{E}_{\mathrm{f},1} - \mathcal{E}_x)\,\mathrm{d}\mathcal{E}_x \right) \;. \tag{A.5}$$

The left integral represents tunneling current from electron states that are low in energy and face a high energy barrier. Hence, as a **second approximation**, the left integral is neglected. Still it is necessary to insert an expression for the transmission coefficient in the right integral. For a single-layer dielectric, two shapes are possible: triangular and trapezoidal. First, the formula will be derived assuming a triangular shape.

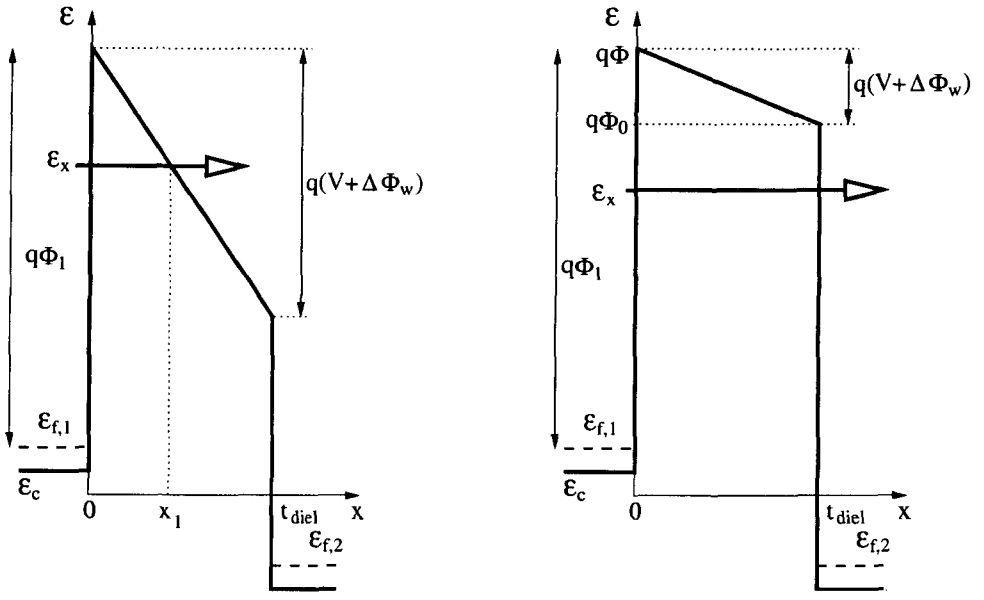


Figure A.1: Energy barrier in the FOWLER-NORDHEIM tunneling (left) and direct tunneling (right) regime.

A.1 Original FOWLER-NORDHEIM Formula

The original FOWLER-NORDHEIM formula assumes a triangular shape of the energy barrier. This is motivated by the fact that only tunneling at strong electric fields was studied. The WKB-approximation (3.57) for the transmission coefficient reads

$$TC(\mathcal{E}_x) = \exp\left(-\frac{2}{\hbar}\int_0^{x_1}\sqrt{2m_{\mathrm{diel}}(\mathcal{E}_\mathrm{c}-\mathcal{E}_x)}\,\mathrm{d}x\right) .$$

The classical turning point x_1 is (see the left part of Fig. A.1)

$$x_1 = \frac{\mathcal{E}_{\mathrm{f},1}+\mathrm{q}\Phi_1-\mathcal{E}_x}{\mathrm{q}E_{\mathrm{diel}}} ,$$

and the dielectric conduction band edge for a triangular barrier

$$\mathcal{E}_\mathrm{c}(x) = \mathcal{E}_{\mathrm{f},1}+\mathrm{q}\Phi_1-\mathrm{q}E_{\mathrm{diel}}x ,$$

where the electric field in the dielectric E_{diel} is caused by the different Fermi levels and the work function difference $\Delta\Phi_{\text{W}}$:

$$E_{\text{diel}} = \frac{\mathcal{E}_{\text{f},1} - \mathcal{E}_{\text{f},2} + \text{q}\Delta\Phi_{\text{W}}}{\text{q}t_{\text{diel}}} \ .$$

The **third approximation** is to assume equal materials for both electrodes, so that $\Delta\Phi_{\text{W}} = 0$. The WKB-based transmission coefficient can then be applied and yields

$$TC(\mathcal{E}_x) = \exp\left(-2\frac{\sqrt{2m_{\text{diel}}}}{\hbar}\int_0^{x_1}\sqrt{\mathcal{E}_{\text{f},1} + \text{q}\Phi_1 - \text{q}E_{\text{diel}}x - \mathcal{E}_x}\,\text{d}x\right) \tag{A.6}$$

$$= \exp\left(4\frac{\sqrt{2m_{\text{diel}}}}{3\hbar\text{q}E_{\text{diel}}}\left(\mathcal{E}_{\text{f},1} + \text{q}\Phi_1 - \text{q}E_{\text{diel}}x - \mathcal{E}_x\right)^{3/2}\Big|_0^{x_1}\right) \tag{A.7}$$

$$= \exp\left(4\frac{\sqrt{2m_{\text{diel}}}}{3\hbar\text{q}E_{\text{diel}}}\left(-\mathcal{E}_{\text{f},1} - \text{q}\Phi_1 + \mathcal{E}_x\right)^{3/2}\right) \tag{A.8}$$

$$= \exp\left(-4\frac{\sqrt{2m_{\text{diel}}}}{3\hbar\text{q}E_{\text{diel}}}\left(\text{q}\Phi_1 - (\mathcal{E}_x - \mathcal{E}_{\text{f},1})\right)^{3/2}\right) \ . \tag{A.9}$$

Using this expression in (A.5) the current density becomes

$$J = \frac{4\pi\text{q}m_{\text{eff}}}{h^3}\int_{\mathcal{E}_{\text{f},2}}^{\mathcal{E}_{\text{f},1}}\exp\left(-\frac{4\sqrt{2m_{\text{diel}}}}{3\hbar\text{q}E_{\text{diel}}}\left(\text{q}\Phi_1 - (\mathcal{E}_x - \mathcal{E}_{\text{f},1})\right)^{3/2}\right)(\mathcal{E}_{\text{f},1} - \mathcal{E}_x)\,\text{d}\mathcal{E}_x \ . \tag{A.10}$$

This integral cannot be solved analytically. Hence, the **fourth approximation** is to expand the square root into a first order TAYLOR[1] series around Φ_1:

$$\left(\text{q}\Phi_1 - (\mathcal{E}_x - \mathcal{E}_{\text{f},1})\right)^{3/2} \approx (\text{q}\Phi_1)^{3/2} + \frac{3}{2}(\mathcal{E}_x - \mathcal{E}_{\text{f},1})(\text{q}\Phi_1)^{1/2} \ . \tag{A.11}$$

Inserting this expression into (A.10) and setting $\epsilon = \mathcal{E}_x - \mathcal{E}_{\text{f},1}$ yields

$$J = \frac{4\pi\text{q}m_{\text{eff}}}{h^3}\exp\left(-\frac{4\sqrt{2m_{\text{diel}}}}{3\hbar\text{q}E_{\text{diel}}}(\text{q}\Phi_1)^{3/2}\right)\int_{\mathcal{E}_{\text{f},2}-\mathcal{E}_{\text{f},1}}^{0}\exp\left(\frac{2\sqrt{2m_{\text{diel}}}}{\hbar\text{q}E_{\text{diel}}}(\text{q}\Phi_1)^{1/2}\epsilon\right)\epsilon\,\text{d}\epsilon \ . \tag{A.12}$$

With

$$\int\epsilon\exp(\lambda\epsilon)\,\text{d}\epsilon = \frac{1}{\lambda^2}\exp(\lambda\epsilon)(\lambda\epsilon - 1) \tag{A.13}$$

and

$$a = -\frac{4\sqrt{2m_{\text{diel}}}}{3\hbar\text{q}E_{\text{diel}}}(\text{q}\Phi_1)^{3/2} \ , \qquad \lambda = \frac{2\sqrt{2m_{\text{diel}}}}{\hbar\text{q}E_{\text{diel}}}(\text{q}\Phi_1)^{1/2} \ , \tag{A.14}$$

[1] BROOK TAYLOR, British mathematician, 1685–1731.

the current density becomes

$$J = \frac{4\pi q m_{\text{eff}}}{h^3} \exp(a) \int_{\mathcal{E}_{\text{f},2}-\mathcal{E}_{\text{f},1}}^{0} \exp(\lambda\epsilon)\epsilon \, \mathrm{d}\epsilon \tag{A.15}$$

$$= \frac{4\pi q m_{\text{eff}}}{h^3} \exp(a) \frac{1}{\lambda^2} \exp\left(\lambda(\mathcal{E}_{\text{f},2} - \mathcal{E}_{\text{f},1})\right) \left(\lambda(\mathcal{E}_{\text{f},2} - \mathcal{E}_{\text{f},1}) - 1\right) \ . \tag{A.16}$$

The **fifth assumption** is now that $\mathcal{E}_{\text{f},1} \gg \mathcal{E}_{\text{f},2}$, leading to

$$J = \frac{4\pi q m_{\text{eff}}}{h^3} \exp(a) \frac{1}{\lambda^2} \ , \tag{A.17}$$

or

$$J = \frac{q^3 m_{\text{eff}}}{8\pi m_{\text{diel}} h q\Phi_1} E_{\text{diel}}^2 \exp\left(-\frac{4\sqrt{2m_{\text{diel}}(q\Phi_1)^3}}{3\hbar q E_{\text{diel}}}\right) \tag{A.18}$$

which is the equation commonly known as the FOWLER-NORDHEIM formula. Note that there is a difference between the effective electron mass in the electrode (m_{eff}) and the effective electron mass in the dielectric (m_{diel}).

A.2 Correction for Direct Tunneling

The equation derived above is only valid for triangular barriers, that is the case of high applied voltages. In [19] SCHUEGRAF proposed a correction to the FOWLER-NORDHEIM formula to account for tunneling in the direct tunneling regime. In this case the transmission coefficient is

$$TC(\mathcal{E}) = \exp\left(-\frac{2}{\hbar} \int_0^{t_{\text{diel}}} \sqrt{2m_{\text{diel}}(\mathcal{E}_{\text{c}} - \mathcal{E}_x)} \, \mathrm{d}x\right) \ ,$$

where t_{diel} is the dielectric thickness. The conduction band edge is again approximated by a linear shape

$$\mathcal{E}_{\text{c}}(x) = \mathcal{E}_{\text{f},1} + q\Phi_1 - qE_{\text{diel}}x \ .$$

The band edges $q\Phi$ and $q\Phi_0$ are given by (see the right part of Fig. A.1)

$$q\Phi = \mathcal{E}_{\text{f},1} + q\Phi_1 \ ,$$
$$q\Phi_0 = \mathcal{E}_{\text{f},1} + q\Phi_1 - qE_{\text{diel}}t_{\text{diel}} \ .$$

As for the triangular energy barrier, it is assumed that the electrodes have equal work functions: $\Delta\Phi_{\text{W}} = 0$. Using these expressions, the transmission coefficient becomes

$$TC(\mathcal{E}_x) = \exp\left(-2\frac{\sqrt{2m_{\text{diel}}}}{\hbar} \int_0^{t_{\text{diel}}} \sqrt{q\Phi - qE_{\text{diel}}x - \mathcal{E}_x} \, \mathrm{d}x\right) \tag{A.19}$$

$$= \exp\left(4\frac{\sqrt{2m_{\text{diel}}}}{3\hbar q E_{\text{diel}}} (q\Phi - qE_{\text{diel}}x - \mathcal{E}_x)^{3/2}\Big|_0^{t_{\text{diel}}}\right) \tag{A.20}$$

$$= \exp\left(-4\frac{\sqrt{2m_{\text{diel}}}}{3\hbar q E_{\text{diel}}} \left((q\Phi - \mathcal{E}_x)^{3/2} - (q\Phi_0 - \mathcal{E}_x)^{3/2}\right)\right) \ . \tag{A.21}$$

The exponent can be approximated using a first order TAYLOR series expansion around $q\Phi_1$ and $q\Phi_1 - qE_{diel}t_{diel}$, respectively:

$$(q\Phi - \mathcal{E}_x)^{3/2} = (\mathcal{E}_{f,1} + q\Phi_1 - \mathcal{E}_x)^{3/2} \tag{A.22}$$

$$= (q\Phi_1 - (\mathcal{E}_x - \mathcal{E}_{f,1}))^{3/2} \tag{A.23}$$

$$\approx (q\Phi_1)^{3/2} + \frac{3}{2}(\mathcal{E}_x - \mathcal{E}_{f,1})(q\Phi_1)^{1/2} , \tag{A.24}$$

$$(q\Phi_0 - \mathcal{E}_x)^{3/2} = (\mathcal{E}_{f,1} + q\Phi_1 - qE_{diel}t_{diel} - \mathcal{E}_x)^{3/2} \tag{A.25}$$

$$= ((q\Phi_1 - qE_{diel}t_{diel}) - (\mathcal{E}_x - \mathcal{E}_{f,1}))^{3/2} \tag{A.26}$$

$$\approx (q\Phi_1 - qE_{diel}t_{diel})^{3/2} + \frac{3}{2}(\mathcal{E}_x - \mathcal{E}_{f,1})(q\Phi_1 - qE_{diel}t_{diel})^{1/2} . \tag{A.27}$$

With the temporary variable η

$$\eta = (q\Phi - \mathcal{E}_x)^{3/2} - (q\Phi_0 - \mathcal{E}_x)^{3/2} \tag{A.28}$$

$$\approx -(q\Phi_1 - qE_{diel}t_{diel})^{3/2} + (q\Phi_1)^{3/2} - \frac{3}{2}(\mathcal{E}_x - \mathcal{E}_{f,1})\left((q\Phi_1)^{1/2} - (q\Phi_1 - qE_{diel}t_{diel})^{1/2}\right) ,$$

the tunnel current density becomes

$$J = \frac{4\pi q m_{eff}}{h^3} \int_{\mathcal{E}_{f,2}}^{\mathcal{E}_{f,1}} TC(\mathcal{E}_x)(\mathcal{E}_{f,1} - \mathcal{E}_x)\, d\mathcal{E}_x \tag{A.29}$$

$$\approx \frac{4\pi q m_{eff}}{h^3} \int_{\mathcal{E}_{f,2}}^{\mathcal{E}_{f,1}} \exp\left(-4\frac{\sqrt{2m_{diel}}}{3\hbar q E_{diel}}\eta\right)(\mathcal{E}_{f,1} - \mathcal{E}_x)\, d\mathcal{E}_x . \tag{A.30}$$

With the abbreviations

$$a = \frac{4\pi q m_{eff}}{h^3} , \tag{A.31}$$

$$b = -\frac{4\sqrt{2m_{diel}}}{3\hbar q E_{diel}}\left((q\Phi_1)^{3/2} - (q\Phi_1 - qE_{diel}t_{diel})^{3/2}\right) , \tag{A.32}$$

$$c = -\frac{2\sqrt{2m_{diel}}}{\hbar q E_{diel}}\left((q\Phi_1)^{1/2} - (q\Phi_1 - qE_{diel}t_{diel})^{1/2}\right) , \tag{A.33}$$

the tunnel current density can be written as

$$J = a\exp(b) \int_{\mathcal{E}_{f,1}}^{\mathcal{E}_{f,2}} \exp\left(c(\mathcal{E}_x - \mathcal{E}_{f,1})\right)(\mathcal{E}_{f,1} - \mathcal{E}_x)\, d\mathcal{E}_x . \tag{A.34}$$

With $\epsilon = \mathcal{E}_x - \mathcal{E}_{f,1}$ this yields

$$J = -a\exp(b) \int_{\mathcal{E}_{f,2}-\mathcal{E}_{f,1}}^{0} \exp(c\epsilon)\epsilon\, d\epsilon . \tag{A.35}$$

Using (A.13) this integral becomes

$$J = \frac{a\exp(b)}{c^2}\left(1-\exp\left(-c(\mathcal{E}_{\mathrm{f},1}-\mathcal{E}_{\mathrm{f},2})\right)\left(1+c(\mathcal{E}_{\mathrm{f},1}-\mathcal{E}_{\mathrm{f},2})\right)\right) \tag{A.36}$$

which, for $\mathcal{E}_{\mathrm{f},1} \gg \mathcal{E}_{\mathrm{f},2}$, simplifies to

$$J = \frac{a\exp(b)}{c^2}\ , \tag{A.37}$$

or, inserting the expressions for a, b, and c

$$J = \frac{\mathrm{q}^3 m_{\mathrm{eff}}}{8\pi h m_{\mathrm{diel}}\left((\mathrm{q}\Phi_1)^{1/2}-(\mathrm{q}\Phi_1-\mathrm{q}V_{\mathrm{diel}})^{1/2}\right)^2}E_{\mathrm{diel}}^2 \tag{A.38}$$

$$\exp\left(-\frac{4\sqrt{2m_{\mathrm{diel}}}}{3\hbar\mathrm{q}E_{\mathrm{diel}}}\left((\mathrm{q}\Phi_1)^{3/2}-(\mathrm{q}\Phi_1-\mathrm{q}V_{\mathrm{diel}})^{3/2}\right)\right) \tag{A.39}$$

which is the equation used in [19]. In some publications, the equation is rewritten to make it more similar to the FOWLER-NORDHEIM formula:

$$J = \frac{\mathrm{q}^3 m_{\mathrm{eff}}}{8\pi m_{\mathrm{diel}} h \mathrm{q}\Phi_1 B_1}E_{\mathrm{diel}}^2\exp\left(-\frac{4\sqrt{2m_{\mathrm{diel}}(\mathrm{q}\Phi_1)^3}B_2}{3\hbar\mathrm{q}E_{\mathrm{diel}}}\right)\ , \tag{A.40}$$

with the additional correction terms B_1, B_2 given as

$$\begin{aligned} B_1 &= \left(1-\left(1-\frac{\mathrm{q}E_{\mathrm{diel}}t_{\mathrm{diel}}}{\mathrm{q}\Phi_1}\right)^{1/2}\right)^2\ , \\ B_2 &= \left(1-\left(1-\frac{\mathrm{q}E_{\mathrm{diel}}t_{\mathrm{diel}}}{\mathrm{q}\Phi_1}\right)^{3/2}\right)\ . \end{aligned} \tag{A.41}$$

Appendix B

The WKB Approximation

The WENTZEL-KRAMERS-BRILLOUIN approximation is one of the most frequently applied approximations to solve SCHRÖDINGER's equation [127, 130, 131]. Starting from the time-independent SCHRÖDINGER equation (2.13), the one-dimensional case reads

$$\left(-\frac{\hbar^2}{2m}\frac{\mathrm{d}^2}{\mathrm{d}x^2} + W(x) - \mathcal{E}\right)\Psi(x) = 0\ . \tag{B.1}$$

If the following *Ansatz* is used for the wave function

$$\Psi(x) = R(x)\exp\left(\imath\frac{S(x)}{\hbar}\right)\ , \tag{B.2}$$

the equations

$$\frac{\mathrm{d}^2R}{\mathrm{d}x^2} - \frac{R}{\hbar^2}\left(\frac{\mathrm{d}S}{\mathrm{d}x}\right)^2 + \frac{2m\left(\mathcal{E} - W(x)\right)}{\hbar^2}R = 0 \tag{B.3}$$

and

$$R\frac{\mathrm{d}^2S}{\mathrm{d}x^2} + 2\frac{\mathrm{d}R}{\mathrm{d}x}\frac{\mathrm{d}S}{\mathrm{d}x} = 0 \tag{B.4}$$

for the real and imaginary part of (B.1) can be found. Equation (B.4) can be solved by

$$\frac{\mathrm{d}S}{\mathrm{d}x} = \frac{C}{R^2}\ , \tag{B.5}$$

where C is a constant. With (B.5) equation (B.3) becomes

$$\frac{1}{R}\frac{\mathrm{d}^2R}{\mathrm{d}x^2} - \frac{1}{\hbar^2}\left(\frac{\mathrm{d}S}{\mathrm{d}x}\right)^2 + \frac{2m\left(\mathcal{E} - W(x)\right)}{\hbar^2} = 0\ . \tag{B.6}$$

With the approximation

$$\frac{1}{R}\frac{\mathrm{d}^2R}{\mathrm{d}x^2} \ll \frac{1}{\hbar^2}\left(\frac{\mathrm{d}S}{\mathrm{d}x}\right)^2 \tag{B.7}$$

we can write

$$S(x) \approx \int \sqrt{2m(\mathcal{E} - W(x))}\,\mathrm{d}x\ , \tag{B.8}$$

and the wave function $\Psi(x)$ becomes

$$\Psi(x) = R(x)\exp\left(\frac{\imath}{\hbar}\int \sqrt{2m\,(\mathcal{E} - W(x))}\,\mathrm{d}x\right)\ . \tag{B.9}$$

Now we consider an energy barrier between the classical turning points x_1 and x_2 with an incoming wave Ψ_1 and a transmitted wave Ψ_2, and $x_2 > x_1$

$$\begin{aligned}\Psi_1(x \le x_1) &\sim \exp\left(\frac{\imath}{\hbar}\int_{-\infty}^{x_1} \sqrt{2m(\mathcal{E} - W(x'))}\,\mathrm{d}x'\right)\ ,\\ \Psi_2(x \ge x_2) &\sim \exp\left(\frac{\imath}{\hbar}\int_{-\infty}^{x_2} \sqrt{2m(\mathcal{E} - W(x'))}\,\mathrm{d}x'\right)\ .\end{aligned} \tag{B.10}$$

The transmission probability $TC(\mathcal{E})$ is proportional to $|\Psi_2(x_2)/\Psi_1(x_1)|^2$:

$$TC = \left|\frac{\exp\left(\frac{\imath}{\hbar}\int_{-\infty}^{x_2} \sqrt{2m(\mathcal{E} - W(x'))}\,\mathrm{d}x'\right)}{\exp\left(\frac{\imath}{\hbar}\int_{-\infty}^{x_1} \sqrt{2m(\mathcal{E} - W(x'))}\,\mathrm{d}x'\right)}\right|^2 = \left|\exp\left(\frac{\imath}{\hbar}\int_{x_1}^{x_2} \sqrt{2m(\mathcal{E} - W(x'))}\,\mathrm{d}x'\right)\right|^2 \tag{B.11}$$

$$= \exp\left(-\frac{2}{\hbar}\int_{x_1}^{x_2} \sqrt{2m(W(x') - \mathcal{E})}\,\mathrm{d}x'\right) \tag{B.12}$$

This expression can be evaluated for arbitrary barriers as shown in Section 3.5.1. In [130], however, it is shown that the WKB-approximation is only valid for

$$m\hbar\frac{\mathrm{d}W(x)}{\mathrm{d}x} \ll \sqrt{|2m(W(x) - \mathcal{E})|^3}\ . \tag{B.13}$$

This inequality is fulfilled for points where the variation of the energy barrier is small. The WKB approximation is therefore not valid in the close vicinity of the classical turning points.

Appendix C

Wave Function Normalization for a Triangular Potential

For the assumption of a triangular energy well, the wave function is approximately given as (see Section 3.6.1)

$$\Psi(x) = A\mathrm{Ai}(u(x)) \ , \tag{C.1}$$

with

$$u(x) = -\left(\frac{2m}{\hbar^2}\right)^{1/3}\left(\frac{x_1 - x_0}{W_1 - W_0}\right)^{2/3}(\mathcal{E} - W(x)) \ . \tag{C.2}$$

The square of the wave function is a probability, therefore the normalization can be written as [156]

$$\int_0^\infty |\Psi(u(x))|^2 \, \mathrm{d}x = 1 \ , \tag{C.3}$$

$$\int_0^\infty |A\mathrm{Ai}(u(x))|^2 \, \mathrm{d}x = 1 \ , \tag{C.4}$$

$$\int_0^\infty \mathrm{Ai}^2\left(-\left(\frac{2m}{\hbar^2}\right)^{1/3}\left(\frac{x_1 - x_0}{W_1 - W_0}\right)^{2/3}(\mathcal{E}_i - W(x))\right) \mathrm{d}x = \frac{1}{A^2} \ , \tag{C.5}$$

where an infinite barrier is assumed for $x < 0$. With $x_0 = 0$, $W_0 = 0$, and the electric field

$$E = \frac{W_1}{\mathrm{q}x_1} \ , \tag{C.6}$$

the integral becomes

$$\int_0^\infty \mathrm{Ai}^2\left(\left(\frac{2m\mathrm{q}E}{\hbar^2}\right)^{1/3}\left(x - \frac{\mathcal{E}_i}{\mathrm{q}E}\right)\right) \mathrm{d}x = \frac{1}{A^2} \ . \tag{C.7}$$

Substituting

$$\lambda(x) = \left(\frac{2m\mathrm{q}E}{\hbar^2}\right)^{1/3}\left(x - \frac{\mathcal{E}_i}{\mathrm{q}E}\right) \tag{C.8}$$

$$\mathrm{d}\lambda(x) = \left(\frac{2m\mathrm{q}E}{\hbar^2}\right)^{1/3}\mathrm{d}x \tag{C.9}$$

yields

$$\left(\frac{\hbar^2}{2m\mathrm{q}E}\right)^{1/3}\int_{\lambda(0)}^{\infty}\mathrm{Ai}^2(\lambda(x))\,\mathrm{d}\lambda(x) = \frac{1}{A^2}\,. \tag{C.10}$$

Using the expression [157]

$$\int_z^{\infty}\mathrm{Ai}^2(x)\,\mathrm{d}x = -z\mathrm{Ai}^2(z) + \mathrm{Ai}'^2(z) \tag{C.11}$$

and $\lambda(0) = \lambda_0$ the normalization constant becomes

$$A = \left(\frac{\left(\frac{2m\mathrm{q}E}{\hbar^2}\right)^{1/3}}{\mathrm{Ai}'^2(\lambda_0) - \lambda_0\mathrm{Ai}^2(\lambda_0)}\right)^{1/2}\,. \tag{C.12}$$

Appendix D

User Interface

This chapter describes the user interface for the tunneling models implemented in MINIMOS-NT. Several models can be chosen: The FOWLER-NORDHEIM model, the SCHUEGRAF model, the FRENKEL-POOLE model, and the TSU-ESAKI model with different methods to calculate the transmission coefficient. Additionally, a trap-assisted tunneling model accounting for inelastic tunneling of electrons via traps is available.

Tunneling is allowed for all dielectric – semiconductor or dielectric – ideal conductor interfaces. The keyword `tunnel` must be given in the `Phys` section to specify the dielectric segment where the tunneling model should be evaluated. Since the tunneling current is always evaluated between two boundaries, these boundaries have to be stated in the `tunnel` keyword. If the two boundaries considered for tunneling are the nearest non-touching two boundaries of the respective segment, the function `addNearestInterfaces()` can be used. This function returns the two boundaries of the segment that are nearest but do not touch each other.

D.1 Direct Tunneling

The segments and boundaries for which tunneling is calculated must be registered in the `tunnel` string located in the `Phys` section of the input deck. This can be performed in several ways. The first possibility is to state the segment and its boundaries manually like in the following example:

```
Phys
{  tunnel = "GateInsulator,Semiconductor_GateInsulator,GateInsulator_Gate";
   +GateInsulator
   {  Electron
      {
        tunnel = "FNPure";
      }
   }
}
```

Here, tunneling is turned on in the segment `GateInsulator` for the boundaries to the segments `Semiconductor` and `Gate`. Additionally, a tunneling model must be given for the respective segment. In the example above the model `FNPure` is used for electrons in the `GateInsulator` section, while hole tunneling is neglected.

In most cases tunneling will have to be evaluated for boundaries which are very close. If they are nearer than any other two non-adjacent boundaries of the considered segment, the function `addNearestInterfaces` can be used to find the respective tunneling boundaries:

```
Phys
{  tunnel = addNearestInterfaces("Device", "GateInsulator");
   +GateInsulator
   {  Electron
      {  tunnel = "FNPure";
      }
   }
}
```

Note that for all models, the electron and hole tunneling mechanisms can be stated separately. All tunneling models share a keyword `consistent` which can be used to turn self-consistent simulation on or off. If it is set to `no`, the additional electron and hole current is ignored in the continuity equation. This can be of use if only the order of magnitude of the tunneling current is of interest, since convergence is usually better when the continuity equation remains unchanged.

D.1.1 The Model FNPure

The tunnel current density is computed via expression (3.125) where A and B are fitting parameters. The model keywords are stated in Table D.1.

Symbol	Keyword	Type	Unit
A	`a`	`Quantity`	AV^{-2}
B	`b`	`Quantity`	mV^{-1}
	`consistent`	`Boolean`	

Table D.1: `FNPure` tunneling model keywords.

An example input deck is

```
Phys
{
   tunnel = addNearestInterfaces("Device", "GateInsulator");
   +GateInsulator
   {
      Electron
      {
         tunnel = "FNPure";
         Tunnel
         {
            FNPure
            {
               a = 9.946316e-7 "A/V^2";
               b = 2.635706e10 "V/m";
               consistent = no;
            }
         }
      }
      Hole
      {
         tunnel = "FNPure";
         Tunnel
         {
            FNPure
            {
               a = 4.013e-7   "A/V^2";
               b = 6.4216e12  "V/m";
            }
         }
      }
   }
}
```

where electron and hole tunneling is turned on. The values of `a` and `b` can be chosen to fit measurement results. The electron tunnel current is not entered into the continuity equation of the neighboring segments using the `consistent` keyword.

D.1.2 The Model FNLenzlingerSnow

This model is a generalization of the FOWLER-NORDHEIM model by giving the electron mass in the dielectric as a physically-based fitting parameter. The current density is calculated by expression (3.126). Since the electron mass in the dielectric m_{diel} is usually given in terms of the free electron mass m_0, the fitting parameters is now the ratio $m_{\text{diel}}/\text{m}_0$. Table D.2 shows the model keywords.

Symbol	Keyword	Type
$m_{\text{diel}}/\text{m}_0$	mOx	Real
	consistent	Boolean

Table D.2: FNLenzlingerSnow tunneling model keywords.

The electron or hole barrier height $\text{q}\Phi_\text{e}$ or $\text{q}\Phi_\text{h}$ is calculated from the band edge energies and cannot be given in the input deck.

D.1.3 The Model DTSchuegraf

While the FOWLER-NORDHEIM and the LENZLINGER-SNOW models are only valid in the case of a triangular barrier (high bias), the SCHUEGRAF model can be used for direct tunneling through a trapezoidal barriers (valid for low bias). It only differs from the LENZLINGER-SNOW model by two correction factors, see (3.127). For triangular barriers the FNLenzlingerSnow model is used per default, i.e. $B_1 = B_2 = 1$. The DTSchuegraf model has the same input deck parameters as the FNLenzlingerSnow model.

D.1.4 The Model FrenkelPoole

The FRENKEL-POOLE model can be used to describe trap-assisted tunneling for a highly defective dielectric. The tunneling current is given as a generalization of expression (3.137):

$$J = aE_{\text{diel}} \exp\left(\frac{b\sqrt{E_{\text{diel}}} - \mathcal{E}_\text{T}}{\text{k}_\text{B}T}\right) . \tag{D.1}$$

In this expression $\mathcal{E}_\text{T}$ is the trap energy level below the dielectric conduction band, and the values a and b can be used as fitting parameters. Table D.3 summarizes the model keywords.

Symbol	Keyword	Type	Unit
a	a	Real	
b	b	Real	
$\mathcal{E}_\text{T}$	trapNrg	Quantity	eV
	consistent	Boolean	

Table D.3: FrenkelPoole tunneling model keywords.

Note that the simple analytic models FNPure, FNLenzlingerSnow, DTSchuegraf, and FrenkelPoole should not be used for the case of a work function difference between the two materials

regarded for tunneling. In the case of a work function difference, the electrostatic field in the dielectric does not vanish for zero bias but only at the flat-band voltage. Hence, these models will show the minimum tunneling current if the flat band voltage is applied. The `TsuEsaki` model, however, takes the work function difference into account and should be used in that case.

D.1.5 The Model `TsuEsaki`

The tunneling current density is calculated by expression (3.13) which involves a numerical integration in the energy domain. The values $\mathcal{E}_{\min}$ and $\mathcal{E}_{\max}$ are found automatically for the ECB, HVB, and EVB processes. If the keyword `dfType` is set to `fermi`, the supply function is calculated using (3.14). Alternatively, the supply function can be calculated by numerical integration of the distribution function as described in Section 3.3 if the keyword `dfType` is set to `general`. With this model it is possible to simulate electron tunneling from the conduction band and hole tunneling from the valence band. The transmission coefficient can be calculated using numerical integration of the WKB expression (3.57) by setting `tcType` to `numericalWKB`, by the analytical expression for a linear energy barrier (see Section 3.5.1) by setting the keyword `tcType` to `analyticalWKB`, or using a SCHRÖDINGER solver based on the quantum transmitting-boundary method `qtbm`, see Section 3.5.4. If the keyword `imageForce` is set to `yes`, the energy barrier is corrected using the image force correction term described in Section 3.4.2. For the numerical integration, the step width can be given in the keyword `dNrg`. If the keywords `tat` and `direct` are set to `yes`, both direct and trap-assisted tunneling is calculated. The model keywords are summarized in Table D.4.

Symbol	Keyword	Type	Unit
	`direct`	`Boolean`	
	`tat`	`Boolean`	
	`consistent`	`Boolean`	
	`imageForce`	`Boolean`	
	`dNrg`	`Quantity`	eV
$h\omega$	`phononNrg`	`Quantity`	eV
S	`huangRhys`	`Real`	
$m_{\text{diel}}/\text{m}_0$	`mOx`	`Real`	

Table D.4: `TsuEsaki` tunneling model keywords.

The keywords which are related to trap-assisted tunneling are only relevant if `tat=yes` and are described in Section D.4. The possible values of the keywords `tcType` and `dfType` are given in Table D.5.

Keyword	Type	Description
`tcType`	`String`	`analyticalWKB`, `numericalWKB`, `qtbm`
`dfType`	`String`	`fermi`, `general`

Table D.5: `TsuEsaki` transmission coefficient and supply function keywords.

D.2 Stacked Segments

If the tunneling current through stacked segments is of interest, the explicit specification of all tunneling boundaries of the stack member segments may be quite cumbersome. Therefore, the input deck function `registerStack` was implemented which finds all boundaries for a number of given stack member segments. In the following example a stack is defined which consists of the two segments `SecondOxide` and `GateOxide`. The respective boundaries are found automatically by the `registerStack` function.

```
Phys
{  tunnel = registerStack("Device", "SecondOxide, GateOxide");
   +GateOxide { Electron { tunnel = "FNPure";}}
   +SecondOxide : GateOxide;
}
```

All stack member segments must share the same tunneling model. This can easily be done using the inheritance mechanism of the input deck: in the above input deck, the tunneling model in the `SecondOxide` segment is simply inherited from the `GateOxide` section. Note that it is also possible to evaluate tunneling in several independent stacks and segments simultaneously by concatenating the respective `tunnel` strings together:

```
  Phys
  {  tunnel = registerStack("Device", "LeftStackUpperOxide,
                                       LeftStackMiddleOxide,
                                       LeftStackLowerOxide") +
              registerStack("Device", "RightStackUpperOxide,
                                       RightStackMiddleOxide,
                                       RightStackLowerOxide") +
              "GateOxide,GateOxide_Semiconductor,GateOxide_FloatingGate";
     +GateOxide { Electron { tunnel = "TsuEsaki"; } }
     +LeftStackLowerOxide   : GateOxide;
     +LeftStackMiddleOxide  : GateOxide;
     +LeftStackUpperOxide   : GateOxide;
     +RightStackLowerOxide  : GateOxide;
     +RightStackMiddleOxide : GateOxide;
     +RightStackUpperOxide  : GateOxide;
  }}
Log
{  currentComponents = yes;
   tunnel            = yes;
}
```

In the `Log` section of the input deck the keywords `currentComponents` and `tunnel` can be set to print logging information to the standard output. If the keyword `currentComponents` in the `Log` section is set to `yes`, the electron, hole, and total tunneling currents are printed in the output as `IE`, `IH` and `It`.

If the `tunnel` keyword is set in the `Log` section of the input deck, some information about the chosen tunneling segments, boundaries, and the respective stacks is printed to `stdout` before the simulation is started:

```
                Tunneling Information for "Device"
-------------------------------------------------------------------------
Tunneling segment: LowerBarrier                          member of stack #0
   Using boundary: Gate_LowerBarrier
   Using boundary: LowerBarrier_LowerSemi
Tunneling segment: LowerSemi                             member of stack #0
   Using boundary: LowerBarrier_LowerSemi
   Using boundary: LowerSemi_MiddleBarrier
Tunneling segment: MiddleBarrier                         member of stack #0
   Using boundary: LowerSemi_MiddleBarrier
   Using boundary: MiddleBarrier_UpperSemi
Tunneling segment: UpperSemi                             member of stack #0
   Using boundary: MiddleBarrier_UpperSemi
   Using boundary: UpperSemi_UpperBarrier
Tunneling segment: UpperBarrier                          member of stack #0
   Using boundary: UpperSemi_UpperBarrier
   Using boundary: TopContact_UpperBarrier
-------------------------------------------------------------------------

                            Stack 0
-------------------------------------------------------------------------
 Member          Master          Reference Nbr    Opposite Nbr    Points
 LowerBarrier    UpperBarrier    Gate             LowerSemi       47 x 2
 LowerSemi       UpperBarrier    LowerBarrier     MiddleBarrier   47 x 8
 MiddleBarrier   UpperBarrier    LowerSemi        UpperSemi       47 x 2
 UpperSemi       UpperBarrier    MiddleBarrier    UpperBarrier    47 x 9
 UpperBarrier    UpperBarrier    UpperSemi        TopContact      47 x 2
 Inner stack reference segment : LowerBarrier
 Inner stack opposite  segment : UpperBarrier
 Outer stack reference segment : Gate
 Outer stack opposite  segment : TopContact
-------------------------------------------------------------------------
```

In the upper part of this logging information all tunneling segments with their tunneling boundaries are listed and it is stated, if they belong to a stack. In the lower part the stack members are listed for each stack. Each stack has a master segment (`UpperBarrier` in this case) and inner and outer reference and opposite segments which denote the direct neighbors of the stack. Also, each segment in a stack has a reference neighbor segment and an opposite neighbor segment. Furthermore, the number of grid points is given for each segment.

D.3 Oxide Traps

Trapped charge in insulator segments can be simulated using the model `oxideTrap` which must be specified in the `Phys` section of the input deck. In the following example the oxide trap model is evaluated in the segment `GateOxide` for a concentration of negative traps of $N_T = 10^{19}\,\mathrm{cm}^{-3}$ at an energy level of 2 eV below the conduction band edge in the dielectric and an occupancy of 0.1%. The model keywords are summarized in Table D.6.

```
Phys
{
   +GateOxide
   {
      oxideTrap = "Pure";
      OxideTrap
      {
         Pure
         {
            Nt        = 1e19 "cm^-3";  // trap concentration
            type      = "negative";    // charge state
            occupancy = 0.001;         // trap occupancy
            energy    = 2 "eV";        // trap energy level
         }
      }
   }
   +Gate
   {
      Contact { Ohmic { Ew = -0.5 eV; }}
   }
}
```

A charge state of -1 (`"negative"`), 0 (`"neutral"`), and $+1$ (`"positive"`) can be chosen. The trap charge is self-consistently considered in the POISSON equation. The possible keyword values are shown in Table D.7.

Symbol	Keyword	Type	Unit
N_T	`Nt`	`Quantity`	cm^{-3}
f_T	`occupancy`	`Real`	
$\mathcal{E}_T$	`energy`	`Quantity`	eV

Table D.6: `OxideTrap` model keywords.

Keyword	Type	Values
`type`	`String`	`"negative"`, `"neutral"`, `"positive"`

Table D.7: `OxideTrap` model trap charge state.

D.4 Trap-Assisted Tunneling

If the keyword `tat` is set in the `TsuEsaki` tunneling model, an additional trap-assisted tunneling current is calculated. The `oxideTrap` model must be used to specify the trap properties. Input deck parameters of this model are the electron mass in the dielectric, the emitted phonon energy $\hbar\omega$, and the Huang-Rhys factor S which can be used as a fitting parameter. The following code shows an example input deck. The model keywords are listed in Table D.4.

```
Phys
{
   tunnel = addNearestInterfaces("Device", "GateOxide");
   +GateOxide
   {
      oxideTrap = "Pure";
      OxideTrap
      {
         Pure
         {
            Nt        = 1e19 "cm^-3"; // trap concentration
            type      = "negative";   // charge state
            occupancy = 0.0;          // trap occupancy
            energy    = 3 "eV";       // trap energy level
         }
      }
      Electron
      {
         tunnel = "TsuEsaki";
         Tunnel
         {
            TsuEsaki
            {
               direct     = no;        // consider direct tunneling
               tat        = yes;       // consider trap-assisted tunneling
               mOx        = 0.5;       // electron mass in the dielectric
               consistent = yes;       // self-consistency
               tcType     = "qtbm";    // "analyticalWKB,qtbm"
               dfType     = "fermi";   // "general"
               dNrg       = 10 "meV";  // energy step for integration
               huangRhys  = 65;        // for trap-assisted tunneling
               phononNrg  = 0.03 "eV"; // for trap-assisted tunneling
               imageForce = no;        // image force correction
            }
         }
      }
   }
}
```

Bibliography

[1] J. E. Lilienfeld, "Method and Apparatus for Controlling Electric Currents." U. S. Patent #1.745.175, 1930.

[2] J. Bardeen and W. Brattain, "The Transistor, A Semiconductor Triode," *Physical Review*, vol. 74, no. 2, pp. 230–231, 1948.

[3] J. Bardeen and W. Brattain, "Physical Principles Involved in Transistor Action," *Physical Review*, vol. 75, no. 8, pp. 1208–1226, 1949.

[4] W. Shockley, M. Sparks, and G. K. Teal, "p-n Junction Transistors," *Physical Review*, vol. 83, no. 1, pp. 151–164, 1951.

[5] D. Kahng and M. Atalla, "Silicon-Silicondioxide Field Induced Surface Devices," in *Proc. IRE-AIEE Solid-State Device Res.Conf.*, 1960.

[6] G. E. Moore, "Cramming More Components onto Integrated Circuits," *Electronics*, vol. 38, no. 8, pp. 114–117, 1965.

[7] G. E. Moore, "Lithography and the Future of Moore's Law," in *Proc. Optical/Laser Microlithography VIII*, vol. 2440, pp. 2–17, SPIE, 1995.

[8] R. H. Dennard, F. H. Gaensslen, H.-N. Yu, V. L. Rideout, E. Bassous, and A. R. LeBlanc, "Design of Ion-Implanted MOSFETs with Very Small Physical Dimensions," *IEEE J.Solid-State Circuits*, vol. 9, no. 5, pp. 256–268, 1974.

[9] H.-S. P. Wong, D. J. Frank, P. M. Solomon, C. H. J. Wann, and J. J. Welser, "Nanoscale CMOS," *Proc.IEEE*, vol. 87, no. 4, pp. 537–570, 1999.

[10] G. Baccarani, M. Wordeman, and R. Dennard, "Generalized Scaling Theory and Its Application to a 1/4 Micrometer MOSFET Design," *IEEE Trans.Electron Devices*, vol. ED-31, no. 4, pp. 452–462, 1984.

[11] "International Technology Roadmap for Semiconductors - 2001 Edition," 2001. URL: `http://public.itrs.net`.

[12] S. Thompson, M. Alavi, M. Hussein, P. Jacob, C. Kenyon, P. Moon, M. Prince, S. Sivakumar, S. Tyagi, and M. Bohr, "130 nm Logic Technology Featuring 60 nm Transistors, Low-K Dielectrics, and Cu Interconnects," *Intel Technology Journal*, vol. 6, no. 2, pp. 5–13, 2002.

[13] B. Doyle, R. Arghavani, D. Barlage, S. Datta, M. Doczy, J. Kavalieros, A. Murthy, and R. Chau, "Transistor Elements for 30 nm Physical Gate Lengths and Beyond," *Intel Technology Journal*, vol. 6, no. 2, pp. 42–54, 2002.

[14] S. Thompson, P. Packan, and M. Bohr, "MOS Scaling: Transistor Challenges for the 21st Century," *Intel Technology Journal*, vol. 2, no. 3, pp. 1–19, 1998.

[15] D. J. Frank and Y. Taur, "Design Considerations for CMOS Near the Limits of Scaling," *Solid-State Electron.*, vol. 46, no. 3, pp. 315–320, 2002.

[16] D. Vasileska, I. Knezevic, R. Akis, S. Ahmed, and D. K. Ferry, "The Role of Quantization Effects on the Operation of 50 nm MOSFETs, 250 nm FIBMOS Devices and Narrow-Width SOI Device Structures," *Journal of Computational Electronics*, vol. 1, no. 4, pp. 453–465, 2002.

[17] B. Yu, C. H. J. Wann, E. D. Nowak, K. Noda, and C. Hu, "Short-Channel Effect Improved by Lateral Channel-Engineering in Deep-Submironmeter MOSFETs," *IEEE Trans.Electron Devices*, vol. 44, no. 4, pp. 627–634, 1997.

[18] F. Faggin and T. Klein, "Silicon Gate Technology," *Solid-State Electron.*, vol. 13, no. 8, pp. 1125–1144, 1970.

[19] K. F. Schuegraf and C. Hu, "Hole Injection SiO_2 Breakdown Model for Very Low Voltage Lifetime Extrapolation," *IEEE Trans.Electron Devices*, vol. 41, no. 5, pp. 761–767, 1994.

[20] P. Ranade, Y.-K. Choi, D. Ha, A. Agarwal, M. Ameen, and T.-J. King, "Tunable Work Function Molybdenum Gate Technology for FDSOI-CMOS," in *Proc. Intl. Electron Devices Meeting*, pp. 363–366, 2002.

[21] D. A. Buchanan, "Scaling the Gate Dielectric: Materials, Integration and Reliability," *IBM J.Res.Dev.*, vol. 43, no. 3, pp. 245–264, 1999.

[22] C. M. Osburn, I. Kim, S. K. Han, I. De, K. F. Yee, S. Gannavaram, S. J. Lee, C.-H. Lee, Z. J. Luo, W. Zhu, J. R. Hauser, D.-L. Kwong, G. Lucovsky, T. P. Ma, and M. C. Öztürk, "Vertically Scaled MOSFET Gate Stacks and Junctions: How Far are we Likely to Go?," *IBM J.Res.Dev.*, vol. 46, no. 2/3, pp. 299–315, 2002.

[23] K. A. Bowman, L. Wang, X. Tang, and J. D. Meindl, "A Circuit-Level Perspective of the Optimum Gate Oxide Thickness," *IEEE Trans.Electron Devices*, vol. 48, no. 8, pp. 1800–1810, 2001.

[24] J. H. Stathis, "Reliability Limits for the Gate Insulator in CMOS Technology," *IBM J.Res.Dev.*, vol. 46, no. 2/3, pp. 265–286, 2002.

[25] H.-S. P. Wong, "Beyond the Conventional Transistor," *IBM J.Res.Dev.*, vol. 46, no. 2/3, pp. 133–168, 2002.

[26] A. Burenkov and J. Lorenz, "On the Role of Corner Effects in FinFETs," in *Proc. 4th European Workshop on Ultimate Integration of Silicon*, pp. 31–34, 2003.

[27] A. Burenkov and J. Lorenz, "Corner Effect in Double and Triple Gate FinFETs," in *Proc. European Solid-State Device Research Conf.*, pp. 135–138, 2003.

[28] Y.-K. Choi, T.-J. King, and C. Hu, "Spacer FinFET: Nano-scale CMOS Technology for the Terabit Era," in *Proc. Intl. Semiconductor Device Research Symposium*, pp. 543–546, 2001.

[29] K. Kim, C.-G. Hwang, and J. G. Lee, "DRAM Technology Perspective for Gigabit Era," *IEEE Trans.Electron Devices*, vol. 45, no. 3, pp. 598–608, 1998.

[30] E. Bertagnolli, F. Hofmann, J. Willer, R. Maly, F. Lau, P. W. von Basse, M. Bollu, R. Thewes, U. Kollmer, U. Zimmermann, M. Hain, W. H. Krautschneider, A. Rusch, B. Hasler, A. Kohlhase, and H. Klose, "ROS: An Extremely High Density Mask ROM Technology Based on Vertical Transistor Cells," in *Proc. Intl. Electron Devices Meeting*, pp. 58–59, 1996.

[31] B. Goebel, E. Bertagnolli, and F. Koch, "Reliability of Vertical MOSFETs for Gigascale Memory Applications," in *Proc. Intl. Electron Devices Meeting*, pp. 939–942, 1998.

[32] T. Schulz, W. Rösner, L. Risch, A. Korbel, and U. Langmann, "Short-Channel Vertical Sidewall MOSFETs," *IEEE Trans.Electron Devices*, vol. 48, no. 8, pp. 1783–1788, 2001.

[33] C. K. Date and J. D. Plummer, "Increased Hot-Carrier Effects Using SiGe Layers in Vertical Surrounding-Gate MOSFETs," *IEEE Trans.Electron Devices*, vol. 48, no. 12, pp. 2690–2694, 2001.

[34] T. Schulz, W. Rösner, E. Landgraf, L. Risch, and U. Langmann, "Planar and Vertical Double Gate Concepts," *Solid-State Electron.*, vol. 46, no. 7, pp. 985–989, 2002.

[35] J. A. Mandelman, R. H. Dennard, G. B. Bronner, J. K. DeBrosse, R. D. Y. Li, and C. J. Radens, "Challenges and Future Directions for the Scaling of Dynamic Random-Access Memory (DRAM)," *IBM J.Res.Dev.*, vol. 46, no. 2/3, pp. 187–212, 2002.

[36] W. B. Choi, J. U. Chu, K. S. Jeong, E. Bae, and J.-W. Lee, "Ultrahigh-Density Nanotransistors by Using Selectively Grown Vertical Carbon Nanotubes," *Appl.Phys.Lett.*, vol. 79, no. 26, pp. 3696–3698, 2001.

[37] W. B. Choi, S. Chae, E. Bae, J.-W. Lee, B.-H. Cheong, J.-R. Kim, and J.-J. Kim, "Carbon-Nanotube-Based Nonvolatile Memory with Oxide-Nitride-Oxide Film and Nanoscale Channel," *Appl.Phys.Lett.*, vol. 82, no. 2, pp. 275–277, 2003.

[38] P. Avouris, J. Appenzeller, R. Martel, and S. J. Wind, "Carbon Nanotube Electronics," *Proc.IEEE*, vol. 91, no. 11, pp. 1772–1784, 2003.

[39] N. Sano, A. Hiroki, and K. Matsuzawa, "Device Modeling and Simulations Toward Sub-10 nm Semiconductor Devices," *IEEE Trans. Nanotechnology*, vol. 1, no. 1, pp. 63–71, 2002.

[40] A. Schenk, "Physical Modeling of Deep-Submicron Devices," in *Proc. European Solid-State Device Research Conf.*, pp. 9–16, 2001.

[41] Z. Yu, R. W. Dutton, and R. A. Kiehl, "Circuit/Device Modeling at the Quantum Level," *IEEE Trans.Electron Devices*, vol. 47, no. 10, pp. 1819–1825, 2000.

[42] Z. Yu, R. W. Dutton, and D. W. Yergeau, "Macroscopic Quantum Carrier Transport Modeling," in *Proc. Simulation of Semiconductor Processes and Devices*, pp. 1–9, 2001.

[43] R. Stratton, "Diffusion of Hot and Cold Electrons in Semiconductor Barriers," *Physical Review*, vol. 126, no. 6, pp. 2002–2014, 1962.

[44] K. Blotekjaer, "Transport Equations for Electrons in Two-Valley Semiconductors," *IEEE Trans.Electron Devices*, vol. 17, no. 1, pp. 38–47, 1970.

[45] C. Y. Chang and S. M. Sze, *ULSI Devices.* Wiley, 2000.

[46] S. Selberherr, *Analysis and Simulation of Semiconductor Devices.* Springer, 1984.

[47] T. Grasser, T.-W. Tang, H. Kosina, and S. Selberherr, "A Review of Hydrodynamic and Energy-Transport Models for Semiconductor Device Simulation," *Proc.IEEE*, vol. 91, no. 2, pp. 251–274, 2003.

[48] T. Grasser, H. Kosina, M. Gritsch, and S. Selberherr, "Using Six Moments of Boltzmann's Transport Equation for Device Simulation," *J.Appl.Phys.*, vol. 90, no. 5, pp. 2389–2396, 2001.

[49] W. Liang, N. Goldsman, I. Mayergoyz, and P. J. Oldiges, "2-D MOSFET Modeling Including Surface Effects and Impact Ionization by Self-Consistent Solution of the Boltzmann, Poisson, and Hole-Continuity Equations," *IEEE Trans.Electron Devices*, vol. 44, no. 2, pp. 257–267, 1997.

[50] N. Goldsman, C.-K. Lin, Z. Han, and C.-K. Huang, "Advances in the Spherical Harmonic-Boltzmann-Wigner Approach to Device Simulation," *Superlattices & Microstructures*, vol. 27, no. 2/3, pp. 159–175, 2000.

[51] Z. Han, C.-K. Lin, N. Goldsman, I. Mayergoyz, S. Yu, and M. Stettler, "Gate Leakage Current Simulation by Boltzmann Transport Equation and its Dependence on the Gate Oxide Thickness," in *Proc. Simulation of Semiconductor Processes and Devices*, pp. 247–250, 1999.

[52] C.-K. Huang and N. Goldsman, "2-D Self-Consistent Solution of Schrödinger Equation, Boltzmann Transport Equation, Poisson and Current-Continuity Equations for MOSFET," in *Proc. Simulation of Semiconductor Processes and Devices*, pp. 148–151, 2001.

[53] C.-K. Huang and N. Goldsman, "Non-Equilibrium Modeling of Tunneling Gate Currents in Nanoscale MOSFETs," *Solid-State Electron.*, vol. 47, no. 4, pp. 713–720, 2003.

[54] M. Gritsch, *Numerical Modeling of Silicon-on-Insulator MOSFETs.* Dissertation, Technische Universität Wien, 2002.
URL: `http://www.iue.tuwien.ac.at/phd/gritsch`.

[55] C. Jacoboni and L. Reggiani, "The Monte Carlo Method for the Solution of Charge Transport in Semiconductors with Applications to Covalent Materials," *Reviews of Modern Physics*, vol. 55, no. 3, pp. 645–705, 1983.

[56] M. V. Fischetti and S. E. Laux, "Monte Carlo Study of Electron Transport in Silicon Inversion Layers," *Physical Review B*, vol. 48, no. 4, pp. 2244–2274, 1993.

[57] A. Abramo, L. Baudry, R. Brunetti, R. Castagne, M. Charef, F. Dessenne, P. Dollfus, R. Dutton, W. L. Engl, R. Fauquembergue, C. Fiegna, M. V. Fischetti, S. Galdin, N. Goldsman, M. Hackel, C. Hamaguchi, K. Hess, K. Hennacy, P. Hesto, J. M. Higman, T. Iizuka, C. Jungemann, Y. Kamakura, H. Kosina, T. Kunikiyo, S. E. Laux, H. Lin, C. Maziar, H. Mizuno, H. J. Peifer, S. Ramaswamy, N. Sano, P. G. Scrobohaci, S. Selberherr, M. Takenaka, T.-W. Tang, K. Taniguchi, J. L. Thobel, R. Thoma, K. Tomizawa, M. Tomizawa, T. Vogelsang, S.-L. Wang, X. Wang, C.-S. Yao, P. D. Yoder, and A. Yoshii, "A Comparison of Numerical Solutions of the Boltzmann Transport Equation for High-Energy Electron Transport Silicon," *IEEE Trans.Electron Devices*, vol. 41, no. 9, pp. 1646–1654, 1994.

[58] S. E. Laux, M. V. Fischetti, and D. J. Frank, "Monte Carlo Analysis of Semiconductor Devices: The DAMOCLES Program," *IBM J.Res.Dev.*, vol. 34, no. 4, pp. 466–494, 1990.

[59] S. E. Laux and M. V. Fischetti, "Transport Models for Advanced Device Simulation – Truth or Consequences?," in *Proc. Bipolar/BiCMOS Circuits and Technology Meeting*, pp. 27–34, 1995.

[60] J. D. Bude and M. Mastrapasqua, "Impact Ionization and Distribution Functions in Sub-Micron nMOSFET Technologies," *IEEE Electron Device Lett.*, vol. 16, no. 10, pp. 439–441, 1995.

[61] A. Duncan, U. Ravaioli, and J. Jakumeit, "Full-Band Monte Carlo Investigation of Hot Carrier Trends in the Scaling of Metal-Oxide-Semiconductor Field-Effect Transistors," *IEEE Trans.Electron Devices*, vol. 45, no. 4, pp. 867–876, 1998.

[62] C. Jungemann and B. Meinerzhagen, *Hierarchical Device Simulation. The Monte Carlo Perspective.* Springer, 2003.

[63] E. Schrödinger, "Quantisierung als Eigenwertproblem," *Annalen der Physik*, vol. 79, pp. 361–376, 1926.

[64] R. Kassing, *Physikalische Grundlagen der elektronischen Halbleiterbauelemente.* AULA, 1997.

[65] A. Pacelli, "Self-Consistent Solution of the Schrödinger Equation in Semiconductor Devices by Implicit Iteration," *IEEE Trans.Electron Devices*, vol. 44, no. 7, pp. 1169–1171, 1997.

[66] A. Pacelli, *On the Modeling of Quantization and Hot Carrier Effects in Scaled MOSFETs and Other Modern Devices.* Dissertation, Polytechnic Institute of Milan, 1997.

[67] E. Wigner, "On the Quantum Correction for Thermodynamic Equilibrium," *Physical Review*, vol. 40, pp. 749–759, 1932.

[68] K. L. Jensen and A. K. Ganguly, "Simulation of Quantum Tunneling: Transmission Coefficient vs. Wigner Function Approaches," in *Proc. NASECODE VIII*, pp. 44–45, 1992.

[69] K. L. Jensen and A. K. Ganguly, "Numerical Simulation of Field Emission and Tunneling: A Comparison of the Wigner Function and Transmission Coefficient Approaches," *J.Appl.Phys.*, vol. 73, no. 9, pp. 4409–4427, 1993.

[70] M. G. Ancona and H. F. Tiersten, "Quantum Correction to the Equation of State of an Electron Gas in a Semiconductor," *Physical Review B*, vol. 39, no. 13, pp. 9536 – 9540, 1989.

[71] M. G. Ancona, "Macroscopic Description of Quantum-Mechanical Tunneling," *Physical Review B*, vol. 42, no. 2, pp. 1222 – 1233, 1990.

[72] M. G. Ancona, Z. Yu, R. W. Dutton, P. J. V. Voorde, M. Cao, and D. Vook, "Density-Gradient Analysis of Tunneling in MOS Structures with Ultra-Thin Oxides," in *Proc. Simulation of Semiconductor Processes and Devices*, pp. 235–238, 1999.

[73] M. G. Ancona, Z. Yu, R. W. Dutton, P. J. V. Voorde, M. Cao, and D. Vook, "Density-Gradient Analysis of MOS Tunneling," *IEEE Trans.Electron Devices*, vol. 47, no. 12, pp. 2310–2319, 2000.

[74] M. G. Ancona and B. A. Biegel, "Nonlinear Discretization Scheme for the Density-Gradient Equations," in *Proc. Simulation of Semiconductor Processes and Devices*, pp. 196–199, 2000.

[75] M. G. Ancona, "Equations of State for Silicon Inversion Layers," *IEEE Trans.Electron Devices*, vol. 47, no. 7, pp. 1449–1456, 2000.

[76] C. L. Gardner, "The Classical and Quantum Hydrodynamic Models," in *Proc. Intl. Workshop on Computational Electronics*, pp. 25–36, 1993.

[77] T. Hoehr, A. Schenk, A. Wettstein, and W. Fichtner, "On Density-Gradient Modeling of Tunneling Through Insulators," in *Proc. Simulation of Semiconductor Processes and Devices*, pp. 275–278, 2002.

[78] A. Asenov, G. Slavcheva, A. R. Brown, J. H. Davies, and S. Saini, "Increase in the Random Dopant Induced Threshold Fluctuations and Lowering in Sub-100 nm MOSFETs Due to Quantum Effects: A 3-D Density-Gradient Simulation Study," *IEEE Trans.Electron Devices*, vol. 48, no. 4, pp. 722–729, 2001.

[79] A. Asenov, A. R. Brown, and J. R. Watling, "Quantum Corrections in the Simulation of Decanano MOSFETs," *Solid-State Electron.*, vol. 47, no. 7, pp. 1141–1145, 2003.

[80] A. R. Brown, A. Asenov, and J. R. Watling, "Intrinsic Fluctuations in Sub 10 nm Double-Gate MOSFETs Introduced by Discreteness of Charge and Matter," *IEEE Trans. Nanotechnology*, vol. 1, no. 4, pp. 195–200, 2002.

[81] D. Connelly, Z. Yu, and D. Yergeau, "Macroscopic Simulation of Quantum Mechanical Effects in 2-D MOS Devices via the Density Gradient Method," *IEEE Trans.Electron Devices*, vol. 49, no. 2, pp. 619–626, 2002.

[82] K. Matsuzawa, S.-I. Takagi, M. Takayanagi, and H. Tanimoto, "Device Simulation of Surface Quantization Effect on MOSFETs with Simplified Density-Gradient Method," *Solid-State Electron.*, vol. 46, no. 5, pp. 747–751, 2002.

[83] A. Wettstein, *Quantum Effects in MOS Devices.* PhD thesis, ETH Zürich, 2000. URL: `http://www.iis.ee.ethz.ch/wettstae/papers/Dissertation`.

[84] A. Wettstein, A. Schenk, and W. Fichtner, "Quantum Device-Simulation with the Density-Gradient Model on Unstructured Grids," *IEEE Trans.Electron Devices*, vol. 48, no. 2, pp. 279–284, 2001.

[85] J.-R. Zhou and D. K. Ferry, "Simulation of Ultra-Small GaAs MESFET Using Quantum Moment Equations," *IEEE Trans.Electron Devices*, vol. 39, no. 3, pp. 473–478, 1992.

[86] H. Tsuchiya and T. Miyoshi, "Quantum Transport Modeling of Ultrasmall Semiconductor Devices," *IEICE Trans.Electron.*, vol. E82-C, no. 6, pp. 880–888, 1999.

[87] L. Shifren, C. Ringhofer, and D. K. Ferry, "A Wigner Function-Based Quantum Ensemble Monte Carlo Study of a Resonant Tunneling Diode," *IEEE Trans.Electron Devices*, vol. 50, no. 3, pp. 769–773, 2003.

[88] B. Winstead and U. Ravaioli, "A Quantum Correction Based on Schrödinger Equation Applied to Monte Carlo Device Simulation," *IEEE Trans.Electron Devices*, vol. 50, no. 2, pp. 440–446, 2003.

[89] H. Kosina, M. Nedjalkov, and S. Selberherr, "A Monte Carlo Method Seamlessly Linking Quantum and Classical Transport Calculations," in *Proc. Intl. Workshop on Computational Electronics*, 2003.

[90] Y. Li, T.-W. Tang, and X. Wang, "Modeling of Quantum Effects for Ultrathin Oxide MOS Structures with an Effective Potential," *IEEE Trans. Nanotechnology*, vol. 1, no. 4, pp. 238–242, 2002.

[91] S. S. Ahmed and D. Vasileska, "Threshold Voltage Shifts in Narrow-Width SOI Devices Due to Quantum Mechanical Size-Quantization Effects," in *Proc. Nanotech 2003 Vol. 2*, pp. 222–225, 2003.

[92] S. Datta, *Electronic Transport in Mesoscopic Systems.* Cambridge University Press, 1995.

[93] D. K. Ferry and S. M. Goodnick, *Transport in Nanostructures.* Cambridge University Press, 1997.

[94] S. Datta, "Nanoscale Device Modeling: the Green's Function Method," *Superlattices & Microstructures*, vol. 28, no. 4, pp. 253–278, 2000.

[95] S. Datta, "The Non-Equilibrium Green's Function (NEGF) Formalism: An Elementary Introduction," in *Proc. Intl. Electron Devices Meeting*, pp. 29.1.1–29.1.4, 2002.

[96] J. P. Shiely, *Simulation of Tunneling in MOS Devices.* Dissertation, Duke University, 1999.

[97] R. Clerc, *Etude des Effets Quantiques dans les Composants CMOS a Oxydes de Grille Ultra Minces — Modelisation et Caracterisation.* Dissertation, Institut National Polytechnique de Grenoble, 2001.

[98] C. B. Duke, *Tunneling in Solids.* Academic Press, 1969.

[99] R. Tsu and L. Esaki, "Tunneling in a Finite Superlattice," *Appl.Phys.Lett.*, vol. 22, no. 11, pp. 562–564, 1973.

[100] N. Ashcroft and N. Mermin, *Solid State Physics.* Harcourt College Publishers, 1976.

[101] D. Cassi and B. Riccó, "An Analytical Model of the Energy Distribution of Hot Electrons," *IEEE Trans.Electron Devices*, vol. 37, no. 6, pp. 1514–1521, 1990.

[102] A. Abramo and C. Fiegna, "Electron Energy Distributions in Silicon Structures at Low Applied Voltages and High Electric Fields," *J.Appl.Phys.*, vol. 80, no. 2, pp. 889–893, 1996.

[103] K.-I. Sonoda, M. Yamaji, K. Taniguchi, C. Hamaguchi, and S. T. Dunham, "Moment Expansion Approach to Calculate Impact Ionization Rate in Submicron Silicon Devices," *J.Appl.Phys.*, vol. 80, no. 9, pp. 5444–5448, 1996.

[104] C. Fiegna, F. Venturi, M. Melanotte, E. Sangiorgi, and B. Riccó, "Simple and Efficient Modeling of EPROM Writing," *IEEE Trans.Electron Devices*, vol. 38, no. 3, pp. 603–610, 1991.

[105] K. Hasnat, C.-F. Yeap, S. Jallepalli, S. A. Hareland, W.-K. Shih, V. M. Agostinelli, A. F. Tasch, and C. M. Maziar, "Thermionic Emission Model of Electron Gate Current in Submicron NMOSFETs," *IEEE Trans.Electron Devices*, vol. 44, no. 1, pp. 129–138, 1997.

[106] T. Grasser, H. Kosina, C. Heitzinger, and S. Selberherr, "Characterization of the Hot Electron Distribution Function Using Six Moments," *J.Appl.Phys.*, vol. 91, no. 6, pp. 3869–3879, 2002.

[107] T. Grasser, H. Kosina, and S. Selberherr, "Influence of the Distribution Function Shape and the Band Structure on Impact Ionization Modeling," *J.Appl.Phys.*, vol. 90, no. 12, pp. 6165–6171, 2001.

[108] E. Nicollian and J. Brews, *MOS (Metal Oxide Semiconductor) Physics and Technology.* Wiley, 1982.

[109] Y. Tsividis, *Operation and Modeling of the MOS Transistor.* McGraw-Hill, 1987.

[110] S. M. Sze, *Physics of Semiconductor Devices.* Wiley, second ed., 1981.

[111] M. Levinshtein, S. Rumyantsev, and M. Shur, *Handbook Series on Semiconductor Parameters*, vol. 1. World Scientific, 1996.

[112] Y.-C. Yeo, T.-J. King, and C. Hu, "Metal-Dielectric Band Alignment and its Implications for Metal Gate Complementary Metal-Oxide-Semiconductor Technology," *J.Appl.Phys.*, vol. 92, no. 12, pp. 7266–7271, 2002.

[113] W. Harrison, *Solid State Theory.* Dover, 1979.

[114] W. Harrison, *Electronic Structure and the Properties of Solids.* Dover Publications, 1989.

[115] E. H. Rhoderick and R. H. Williams, *Metal-Semiconductor Contacts.* Oxford Press, 1988.

[116] W. Franz, *Handbuch der Physik*, vol. XVII, p. 155. Springer, 1956.

[117] M. V. Fischetti, S. E. Laux, and E. Crabbé, "Understanding Hot-Electron Transport in Silicon Devices: Is There a Shortcut?," *J.Appl.Phys.*, vol. 78, no. 2, pp. 1058–1085, 1995.

[118] M. Kleefstra and G. C. Herman, "Influence of the Image Force on the Band Gap in Semiconductors and Insulators," *J.Appl.Phys.*, vol. 51, no. 9, pp. 4923–4926, 1980.

[119] F. Jiménez-Molinos, F. Gámiz, A. Palma, P. Cartujo, and J. A. Lopez-Villanueva, "Direct and Trap-Assisted Elastic Tunneling Through Ultrathin Gate Oxides," *J.Appl.Phys.*, vol. 91, no. 8, pp. 5116–5124, 2002.

[120] G. Yang, K. Chin, and R. Marcus, "Electron Field Emission Through a Very Thin Oxide Layer," *IEEE Trans.Electron Devices*, vol. 38, no. 10, pp. 2373–2376, 1991.

[121] A. Schenk and G. Heiser, "Modeling and Simulation of Tunneling through Ultra-Thin Gate Dielectrics," *J.Appl.Phys.*, vol. 81, no. 12, pp. 7900–7908, 1997.

[122] A. Schenk, *Advanced Physical Models for Silicon Device Simulation.* Springer, 1998.

[123] C. Fiegna, E. Sangiorgi, and L. Selmi, "Oxide-Field Dependence of Electron Injection from Silicon into Silicon Dioxide," *IEEE Trans.Electron Devices*, vol. 40, no. 11, pp. 2018–2022, 1993.

[124] Z. A. Weinberg, "On Tunneling in Metal-Oxide Silicon Structures," *J.Appl.Phys.*, vol. 53, no. 7, pp. 5052–5056, 1982.

[125] W.-Y. Quan, D. M. Kim, and M. K. Cho, "Unified Compact Theory of Tunneling Gate Current in Metal-Oxide-Semiconductor Structures: Quantum and Image Force Barrier Lowering," *J.Appl.Phys.*, vol. 92, no. 7, pp. 3724–3729, 2002.

[126] L. Larcher, A. Paccagnella, and G. Ghidini, "Gate Current in Ultrathin MOS Capacitors: A New Model of Tunnel Current," *IEEE Trans.Electron Devices*, vol. 48, no. 2, pp. 271–278, 2001.

[127] B. Majkusiak, "Gate Tunnel Current in an MOS Transistor," *IEEE Trans.Electron Devices*, vol. 37, no. 4, pp. 1087–1092, 1990.

[128] A. Hadjadj, G. Salace, and C. Petit, "Fowler-Nordheim Conduction in Polysilicon (n+)-Oxide-Silicon(p) Structures: Limit of the Classical Treatment in the Barrier Height Determination," *J.Appl.Phys.*, vol. 89, no. 12, pp. 7994–8001, 2001.

[129] S. Nagano, M. Tsukiji, E. Hasegawa, and A. Ishitani, "Mechanism of Leakage Current Through the Nanoscale SiO2 Layer," *J.Appl.Phys.*, vol. 75, no. 7, pp. 3530–3535, 1994.

[130] A. Messiah, *Quantenmechanik 1.* DeGruyter, 1991.

[131] S. Gasiorowicz, *Quantum Physics.* John Wiley & Sons, 1995.

[132] L. F. Register, E. Rosenbaum, and K. Yang, "Analytic Model for Direct Tunneling Current in Polycristalline Silicon-Gate Metal-Oxide-Semiconductor Devices," *Appl.Phys.Lett.*, vol. 74, no. 3, pp. 457–459, 1999.

[133] H. Y. Yang, H. Niimi, and G. Lucovsky, "Tunneling Currents Through Ultrathin Oxide/Nitride Dual Layer Gate Dielectrics for Advanced Microelectronic Devices," *J.Appl.Phys.*, vol. 83, no. 4, pp. 2327–2337, 1998.

[134] N. Yang, W. K. Henson, J. R. Hauser, and J. J. Wortman, "Modeling Study of Ultrathin Gate Oxides Using Direct Tunneling Current and Capacitance-Voltage Measurements in MOS Devices," *IEEE Trans.Electron Devices*, vol. 46, no. 7, pp. 1464–1471, 1999.

[135] M. I. Vexler, N. Asli, A. F. Shulekin, B. Meinerzhagen, and P. Seegebrecht, "Compact Quantum Model for a Silicon MOS Tunnel Diode," *Microelectronic Engineering*, vol. 59, no. 1–4, pp. 161–166, 2001.

[136] J. Zhang, J. S. Yuan, Y. Ma, and A. S. Oates, "Design Optimization of Stacked Layer Dielectrics for Minimum Gate Leakage Currents," *Solid-State Electron.*, vol. 44, no. 12, pp. 2165–2170, 2000.

[137] K. H. Gundlach, "Zur Berechnung des Tunnelstroms durch eine trapezförmige Potentialstufe," *Solid-State Electron.*, vol. 9, pp. 949–957, 1966.

[138] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions.* Dover, 1972.

[139] A. Shanware, J. P. Shiely, and H. Z. Massoud, "Extraction of the Gate Oxide Thickness of N- and P-Channel MOSFETs Below 20 Å from the Substrate Current Resulting from Valence-Band Electron Tunneling," in *Proc. Intl. Electron Devices Meeting*, pp. 815–818, 1999.

[140] M. O. Vassell, J. Lee, and H. F. Lockwood, "Multibarrier Tunneling in $Ga_{1-x}Al_xAs$/GaAs Heterostructures," *J.Appl.Phys.*, vol. 54, no. 9, pp. 5208–5213, 1983.

[141] Y. Ando and T. Itoh, "Calculation of Transmission Tunneling Current Across Arbitrary Potential Barriers," *J.Appl.Phys.*, vol. 61, no. 4, pp. 1497–1502, 1987.

[142] B. Zimmermann, E. Marclay, M. Ilegems, and P. Gueret, "Self-Consistent Calculations of Tunneling Currents in n+-GaAs/i-$Al_xGa_{1-x}As$/n+-GaAs Structures and Comparison with Measurements," *J.Appl.Phys.*, vol. 64, no. 7, pp. 3581–3588, 1988.

[143] G. Yong, "Quantum Magnetotransport of Electrons in Double-Barrier Resonant-Tunneling Structures," *Physical Review B*, vol. 50, no. 23, pp. 17249–17255, 1994.

[144] R. Clerc, A. Spinelli, G. Ghibaudo, and G. Pananakakis, "Theory of Direct Tunneling Current in Metal-Oxide-Semiconductor Structures," *J.Appl.Phys.*, vol. 91, no. 3, pp. 1400–1409, 2002.

[145] W. W. Lui and M. Fukuma, "Exact Solution of the Schrödinger Equation Across an Arbitrary One-Dimensional Piecewise-Linear Potential Barrier," *J.Appl.Phys.*, vol. 60, no. 5, pp. 1555–1559, 1986.

[146] K. F. Brennan, "Self-Consistent Analysis of Resonant Tunneling in a Two-Barrier-One-Well Microstructure," *J.Appl.Phys.*, vol. 62, no. 6, pp. 2392–2400, 1987.

[147] D. C. Hutchings, "Transfer Matrix Approach to the Analysis of an Arbitrary Quantum Well Structure in an Electric Field," *Appl.Phys.Lett.*, vol. 55, no. 11, pp. 1082–1084, 1989.

[148] J.-G. S. Demers and R. Maciejko, "Propagation Matrix Formalism and Efficient Linear Potential Solution to Schrödinger's Equation," *J.Appl.Phys.*, vol. 90, no. 12, pp. 6120–6129, 2001.

[149] B. A. Biegel, *Quantum Electronic Device Simulation.* Dissertation, Stanford University, 1997.

[150] J. N. Schulman and Y.-C. Chang, "Reduced Hamiltonian Method for Solving the Tight-Binding Model of Interfaces," *Physical Review B*, vol. 27, no. 4, pp. 2346–2354, 1983.

[151] D. Y. K. Ko and J. C. Inkson, "Matrix Method for Tunneling in Heterostructures: Resonant Tunneling in Multilayer Systems," *Physical Review B*, vol. 38, no. 14, pp. 9945–9951, 1988.

[152] T. Usuki, M. Saito, M. Takatsu, R. A. Kiehl, and N. Yokoyama, "Numerical Analysis of Ballistic-Electron Transport in Magnetic Fields by Using a Quantum Point Contact and a Quantum Wire," *Physical Review B*, vol. 52, no. 11, pp. 8244–8258, 1995.

[153] D. Z. Y. Ting, E. T. Yu, and T. C. McGill, "Multiband Treatment of Quantum Transport in Interband Tunnel Devices," *Physical Review B*, vol. 45, no. 7, pp. 3583–3592, 1992.

[154] W. R. Frensley and N. G. Einspruch, eds., *Heterostructures and Quantum Devices.* VLSI Electronics: Microstructure Science, Academic Press, 1994.

[155] C. S. Lent and D. J. Kirkner, "The Quantum Transmitting Boundary Method," *J.Appl.Phys.*, vol. 67, no. 10, pp. 6353–6359, 1990.

[156] A. P. Gnädinger and H. E. Talley, "Quantum Mechanical Calculation of the Carrier Distribution and the Thickness of the Inversion Layer of a MOS Field-Effect Transistor," *Solid-State Electron.*, vol. 13, no. 9, pp. 1301–1309, 1970.

[157] F. Stern, "Self-Consistent Results for n-Type Si Inversion Layers," *Physical Review B*, vol. 5, no. 12, pp. 4891–4899, 1972.

[158] W. Magnus and W. Schoenmaker, "On the Calculation of Gate Tunneling Currents in Ultra-Thin Metal-Insulator-Semiconductor Capacitors," *Microelectronics Reliability*, vol. 41, no. 1, pp. 31–35, 2001.

[159] E. Anemogiannis, E. N. Glytsis, and T. K. Gaylord, "Bound and Quasibound State Calculation for Biased/Unbiased Semiconductor Heterostructures," *IEEE J.Quantum Electronics*, vol. 29, no. 11, pp. 2731–2740, 1993.

[160] E. Cassan, "On the Reduction of Direct Tunneling Leakage through Ultrathin Gate Oxides by a One-Dimensional Schrödinger-Poisson Solver," *J.Appl.Phys.*, vol. 87, no. 11, pp. 7931–7939, 2000.

[161] A. T. M. Fairus and V. K. Arora, "Quantum Engineering of Nanoelectric Devices: the Role of Quantum Confinement on Mobility Degradation," *Microelectronics Journal*, vol. 32, no. 8, pp. 679–686, 2000.

[162] N. Matsuo, Y. Takami, and Y. Kitagawa, "Modeling of Direct Tunneling for Thin SiO_2 Film on n-Type Si (100) by WKB Method Considering the Quantum Effect in the Accumulation Layer," *Solid-State Electron.*, vol. 46, no. 4, pp. 577–579, 2002.

[163] S. Padmanabhan and A. Rothwarf, "Quantum Inversion Layer Mobility: Numerical Results," *IEEE Trans.Electron Devices*, vol. 36, no. 11, pp. 2557–2566, 1989.

[164] M. J. van Dort, P. H. Woerlee, and A. J. Walker, "A Simple Model for Quantisation Effects in Heavily-Doped Silicon MOSFETs at Inversion Conditions," *Solid-State Electron.*, vol. 37, no. 3, pp. 411–414, 1994.

[165] G. Gildenblatt, B. Gelmont, and S. Vatannia, "Resonant Behavior, Symmetry, and Singularity of the Transfer Matrix in Asymmetric Tunneling Structures," *J.Appl.Phys.*, vol. 77, no. 12, pp. 6327–6331, 1995.

[166] P. J. Price, "Resonant Tunneling via an Accumulation Layer," *Physical Review B*, vol. 45, no. 16, pp. 9042–9045, 1992.

[167] P. J. Price, "Electron Tunneling from Channel to Gate," *Appl.Phys.Lett.*, vol. 82, no. 13, pp. 2080–2081, 2003.

[168] A. Thean and J. P. Leburton, "3-D Computer Simulation of Single-Electron Charging in Silicon Nanocrystal Floating Gate Flash Memory Devices," *IEEE Electron Device Lett.*, vol. 22, no. 3, pp. 148–150, 2001.

[169] A. Ghetti, A. Hamad, P. J. Silverman, H. Vaidya, and N. Zhao, "Self-Consistent Simulation of Quantization Effects and Tunneling Current in Ultra-Thin Gate Oxide MOS Devices," in *Proc. Simulation of Semiconductor Processes and Devices*, pp. 239–242, 1999.

[170] S. H. Lo, D. A. Buchanan, Y. Taur, and W. Wang, "Quantum-Mechanical Modeling of Electron Tunneling Current from the Inversion Layer of Ultra-Thin-Oxide nMOSFETs," *IEEE Trans.Electron Devices*, vol. 18, no. 5, pp. 209–211, 1997.

[171] S. Mudanai, Y. Fan, Q. Ouyang, A. F. Tasch, and S. K. Banerjee, "Modeling of Direct Tunneling Current Through Gate Dielectric Stacks," *IEEE Trans.Electron Devices*, vol. 47, no. 10, pp. 1851–1857, 2000.

[172] S. Mudanai, L. F. Register, A. F. Tasch, and S. K. Banerjee, "Understanding the Effects of Wave Function Penetration on the Inversion Layer Capacitance of NMOSFETs," *IEEE Electron Device Lett.*, vol. 22, no. 3, pp. 145–147, 2001.

[173] F. Rana, S. Tiwari, and D. A. Buchanan, "Self-Consistent Modeling of Accumulation Layers and Tunneling Currents Through Very Thin Oxides.," *Appl.Phys.Lett.*, vol. 69, no. 8, pp. 1104–1106, 1996.

[174] W.-K. Shih, E. X. Wang, S. Jallepalli, F. Leon, C. M. Maziar, and A. F. Tasch, jr., "Modeling Gate Leakage Current in nMOS Structures due to Tunneling through an Ultra-Thin Oxide," *Solid-State Electron.*, vol. 42, no. 6, pp. 997–1006, 1998.

[175] E. Cassan, P. Dollfus, S. Galdin, and P. Hesto, "Semiclassical and Wave-Mechanical Modeling of Charge Control and Direct Tunneling Leakage in MOS and H-MOS Devices with Ultrathin Oxides," *IEEE Trans.Electron Devices*, vol. 48, no. 4, pp. 715–721, 2001.

[176] A. Dalla Serra, A. Abramo, P. Palestri, L. Selmi, and F. Widdershoven, "Closed- and Open-Boundary Models for Gate-Current Calculation in n-MOSFETs," *IEEE Trans.Electron Devices*, vol. 48, no. 8, pp. 1811–1815, 2001.

[177] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst, eds., *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide.* SIAM, Philadelphia, 2000.

[178] N. Arora, *MOSFET Models for VLSI Circuit Simulation.* Springer, 1993.

[179] C.-H. Choi, K.-H. Oh, J.-S. Goo, Z. Yu, and R. W. Dutton, "Direct Tunneling Current Model for Circuit Simulation," in *Proc. Intl. Electron Devices Meeting*, pp. 30.6.1–30.6.4, 1999.

[180] C.-H. Choi, K.-Y. Nam, Z. Yu, and R. W. Dutton, "Impact of Gate Direct Tunneling Current on Circuit Performance: A Simulation Study," *IEEE Trans.Electron Devices*, vol. 48, no. 12, pp. 2823–2829, 2001.

[181] Y.-S. Lin, H.-T. Huang, C.-C. Wu, Y.-K. Leung, H.-Y. Pan, T.-E. Chang, W.-M. Chen, J.-J. Liaw, and C. H. Diaz, "On the SiO_2-Based Gate-Dielectric Scaling Limit for Low-Standby Power Applications in the Context of a 0.13μm CMOS Logic Technology," *IEEE Trans.Electron Devices*, vol. 49, no. 3, pp. 442–448, 2002.

[182] H. Lin, J. T.-Y. Chen, and J.-H. Chang, "Investigation of Disturbance for the New Dual Floating Gate Multilevel Flash Cells," *Solid-State Electron.*, vol. 46, no. 8, pp. 1145–1150, 2002.

[183] S. Schwantes and W. Krautschneider, "Relevance of Gate Current for the Functionality of Deep Submicron CMOS Circuits," in *Proc. European Solid-State Device Research Conf.*, pp. 471–474, 2001.

[184] R. H. Fowler and L. Nordheim, "Electron Emission in Intense Electric Fields," *Proc.Roy.Soc.A*, vol. 119, pp. 173–181, 1928.

[185] M. Lenzlinger and E. H. Snow, "Fowler-Nordheim Tunneling into Thermally Grown SiO_2," *J.Appl.Phys.*, vol. 40, no. 1, pp. 278–283, 1969.

[186] K. F. Schuegraf, C. C. King, and C. Hu, "Ultra-Thin Silicon Dioxide Leakage Current and Scaling Limit," in *Proc. Symposium on VLSI Technology*, pp. 18–19, 1992.

[187] S. Aritome, R. Shirota, G. Hemink, T. Endoh, and F. Masuoka, "Reliability Issues of Flash Memory Cells," *Proc.IEEE*, vol. 81, no. 5, pp. 776–788, 1993.

[188] R. Moazzami and C. Hu, "Stress-Induced Current in Thin Silicon Dioxide Films," in *Proc. Intl. Electron Devices Meeting*, pp. 139–142, 1992.

[189] E. Rosenbaum and L. F. Register, "Mechanism of Stress-Induced Leakage Current in MOS Capacitors," *IEEE Trans.Electron Devices*, vol. 44, no. 2, pp. 317–323, 1997.

[190] S.-I. Takagi, N. Yasuda, and A. Toriumi, "A New I-V Model for Stress-Induced Leakage Current Including Inelastic Tunneling," *IEEE J.Solid-State Circuits*, vol. 46, no. 2, pp. 348–354, 1999.

[191] R. Rofan and C. Hu, "Stress-Induced Oxide Leakage," *IEEE Electron Device Lett.*, vol. 12, no. 11, pp. 632–634, 1991.

[192] J. Wu, L. F. Register, and E. Rosenbaum, "Trap-Assisted Tunneling Current Through Ultra-Thin Oxide," in *Proc. Intl. Reliability Physics Symposium*, pp. 389–395, 1999.

[193] B. Riccó, G. Gozzi, and M. Lanzoni, "Modeling and Simulation of Stress-Induced Leakage Current in Ultrathin SiO2 Films," *IEEE Trans.Electron Devices*, vol. 45, no. 7, pp. 1554–1560, 1998.

[194] K. Sakakibara, N. Ajika, K. Eikyu, K. Ishikawa, and H. Miyoshi, "A Quantitative Analysis of Time-Decay Reproducible Stress-Induced Leakage Current in SiO2 Films," *IEEE Trans.Electron Devices*, vol. 44, no. 6, pp. 1002–1008, 1997.

[195] A. Ghetti, E. Sangiorgi, J. Bude, T. W. Sorsch, and G. Weber, "Tunneling into Interface States as Reliability Monitor for Ultrathin Oxides," *IEEE Trans.Electron Devices*, vol. 47, no. 12, pp. 2358–2365, 2000.

[196] C.-M. Yih, Z.-H. Ho, M.-S. Liang, and S. S. Chung, "Characterization of Hot-Hole Injection Induced SILC and Related Disturbs in Flash Memories," *IEEE Trans.Electron Devices*, vol. 48, no. 2, pp. 300–306, 2001.

[197] A. I. Chou, K. Lai, K. Kumar, P. Chowdhury, and J. C. Lee, "Modeling of Stress-Induced Leakage Current in Ultrathin Oxides with the Trap-Assisted Tunneling Mechanism," *Appl.Phys.Lett.*, vol. 70, no. 27, pp. 3407–3409, 1997.

[198] K. Komiya and Y. Omura, "Spectroscopic Analysis of Stress-Induced Defects in Thin Silicon Oxide Films," *Microelectronic Engineering*, vol. 59, no. 1-4, pp. 61–65, 2001.

[199] T.-K. Kang, M.-J. Chen, C.-H. Liu, Y. J. Chang, and S.-K. Fan, "Numerical Confirmation of Inelastic Trap-Assisted Tunneling (ITAT) as SILC Mechanism," *IEEE Trans.Electron Devices*, vol. 48, no. 10, pp. 2317–2322, 2001.

[200] M. Lenski, T. Endoh, and F. Masuoka, "Analytical Modeling of Stress-Induced Leakage Currents in 5.1-9.6 nm-thick silicon-dioxide films Based on Two-Step Inelastic Trap-Assisted Tunneling," *J.Appl.Phys.*, vol. 88, no. 9, pp. 5238–5245, 2000.

[201] S.-I. Takagi, N. Yasuda, and A. Toriumi, "Experimental Evidence of Inelastic Tunneling in Stress-Induced Leakage Current," *IEEE Trans.Electron Devices*, vol. 46, no. 2, pp. 335–341, 1999.

[202] W. J. Chang, M. P. Houng, and Y. H. Wang, "Simulation of Stress-Induced Leakage Current in Silicon Dioxides: A Modified Trap-Assisted Tunneling Model considering Gaussian-Distributed Traps and Electron Energy Loss," *J.Appl.Phys.*, vol. 89, no. 11, pp. 6285–6293, 2001.

[203] W. J. Chang, M. P. Houng, and Y. H. Wang, "Electrical Properties and Modeling of Ultrathin Impurity-Doped Silicon Dioxides," *J.Appl.Phys.*, vol. 90, no. 10, pp. 5171–5179, 2001.

[204] L. Larcher, A. Paccagnella, and G. Ghidini, "A Model of the Stress Induced Leakage Current in Gate Oxides," *IEEE Trans.Electron Devices*, vol. 48, no. 2, pp. 285–288, 2001.

[205] D. Ielmini, A. S. Spinelli, M. A. Rigamonti, and A. L. Lacaita, "Modeling of SILC Based on Electron and Hole Tunneling - Part I: Transient Effects," *IEEE Trans.Electron Devices*, vol. 47, no. 6, pp. 1258–1265, 2000.

[206] D. Ielmini, A. S. Spinelli, M. A. Rigamonti, and A. L. Lacaita, "Modeling of SILC Based on Electron and Hole Tunneling - Part II: Steady-State," *IEEE Trans.Electron Devices*, vol. 47, no. 6, pp. 1266–1272, 2000.

[207] D. Ielmini, A. S. Spinelli, A. L. Lacaita, A. Martinelli, and G. Ghidini, "A Recombination- and Trap-Assisted Tunneling Model for Stress-Induced Leakage Current," *Solid-State Electron.*, vol. 45, no. 8, pp. 1361–1369, 2001.

[208] D. Ielmini, A. S. Spinelli, A. L. Lacaita, and G. Ghidini, "Modeling of Stress-Induced Leakage Current and Impact Ionization in MOS Devices," *Solid-State Electron.*, vol. 46, no. 3, pp. 417–422, 2002.

[209] D. Ielmini, A. S. Spinelli, A. L. Lacaita, and A. Modelli, "A New Two-Trap Tunneling Model for the Anomalous Stress-Induced Leakage Current (SILC) in Flash Memories," *Microelectronic Engineering*, vol. 59, no. 1-4, pp. 189–195, 2001.

[210] D. Ielmini, A. S. Spinelli, A. L. Lacaita, and A. Modelli, "Modeling of Anomalous SILC in Flash Memories Based on Tunneling at Multiple Defects," *Solid-State Electron.*, vol. 46, no. 11, pp. 1749–1756, 2002.

[211] A. Ghetti, "Characterization and Modeling of the Tunneling Current in Si-SiO_2 - Si Structures with Ultra-Thin Oxide Layer," *Microelectronic Engineering*, vol. 59, no. 1-4, pp. 127–136, 2001.

[212] C. Chaneliere, J. L. Autran, and R. A. B. Devine, "Conduction Mechanisms in Ta_2O_5/SiO_2 and Ta_2O_5/Si_3N_4 Stacked Structures on Si," *J.Appl.Phys.*, vol. 86, no. 1, pp. 480–486, 1999.

[213] M. Houssa, R. Degraeve, P. W. Mertens, M. M. Heyns, J. S. Leon, A. Halliyal, and B. Ogle, "Electrical Properties of Thin $SiON/Ta_2O_5$ Gate Dielectric Stacks," *J.Appl.Phys.*, vol. 86, no. 11, pp. 6462–6467, 1999.

[214] M. Houssa, M. Tuominen, M. Naili, V. Afanas'ev, A. Stesmans, S. Haukka, and M. M. Heyns, "Trap-Assisted Tunneling in High Permittivity Gate Dielectric Stacks," *J.Appl.Phys.*, vol. 87, no. 12, pp. 8615–8620, 2000.

[215] D. Caputo, F. Irrera, S. Salerno, S. Spiga, and M. Fanciulli, "Reliability of ZrO_2 Films Grown by Atomic Layer Deposition," in *Proc. 4th European Workshop on Ultimate Integration of Silicon*, pp. 89–92, 2003.

[216] B. DeSalvo, G. Ghibaudo, G. Pananakakis, B. Guillaumot, and G. Reimbold, "A General Bulk-Limited Transport Analysis of a 10 nm - thick Oxide Stress-Induced Leakage Current," *Solid-State Electron.*, vol. 44, no. 6, pp. 895–903, 2000.

[217] E. Kameda, T. Matsuda, Y. Emura, and T. Ohzone, "Fowler-Nordheim Tunneling in MOS Capacitors with Si-implanted SiO2," *Solid-State Electron.*, vol. 42, no. 11, pp. 2105–2111, 1998.

[218] J. A. López-Villanueva, J. A. Jiménez-Tejada, P. Cartujo, J. Bausells, and J. E. Carceller, "Analysis of the Effects of Constant-Current Fowler-Nordheim-Tunneling Injection with Charge Trapping Inside the Potential Barrier," *J.Appl.Phys.*, vol. 70, no. 7, pp. 3712–3720, 1991.

[219] F. Jiménez-Molinos, A. Palma, F. Gámiz, J. Banqueri, and J. A. Lopez-Villanueva, "Physical Model for Trap-Assisted Inelastic Tunneling in Metal-Oxide-Semiconductor Structures," *J.Appl.Phys.*, vol. 90, no. 7, pp. 3396–3404, 2001.

[220] A. Palma, A. Godoy, J. A. Jimenez-Tejada, J. E. Carceller, and J. A. Lopez-Villanueva, "Quantum Two-Dimensional Calculation of Time Constants of Random Telegraph Signals in Metal-Oxide-Semiconductor Structures," *Physical Review B*, vol. 56, no. 15, pp. 9565–9574, 1997.

[221] J. H. Zheng, H. S. Tan, and S. C. Ng, "Theory of Non-Radiative Capture of Carriers by Multiphonon Processes for Deep Centres in Semiconductors," *J.Phys.:Condensed Matter*, vol. 6, no. 9, pp. 1695–1706, 1994.

[222] W. B. Fowler, J. K. Rudra, M. E. Zvanut, and F. J. Feigl, "Hysteresis and Franck-Condon Relaxation in Insulator-Semiconductor Tunneling," *Physical Review B*, vol. 41, no. 12, pp. 8313–8317, 1990.

[223] M. Herrmann and A. Schenk, "Field and High-Temperature Dependence of the Long Term Charge Loss in Erasable Programmable Read Only Memories: Measurements and Modeling," *J.Appl.Phys.*, vol. 77, no. 9, pp. 4522–4540, 1995.

[224] Institut für Mikroelektronik, Technische Universität Wien, Austria, *MINIMOS-NT User's Guide*, 2002.

[225] Institut für Mikroelektronik, Technische Universität Wien, Austria, *MINIMOS 6 User's Guide*, 1994.

[226] C. Fischer, *Bauelementsimulation in einer computergestützten Entwurfsumgebung.* Dissertation, Technische Universität Wien, 1994.
URL: `http://www.iue.tuwien.ac.at/phd/fischer`.

[227] T. Simlinger, *Simulation von Heterostruktur-Feldeffekttransistoren.* Dissertation, Technische Universität Wien, 1996.
URL: `http://www.iue.tuwien.ac.at/phd/simlinger`.

[228] M. Knaipp, *Modellierung von Temperatureinflüssen in Halbleiterbauelementen.* Dissertation, Technische Universität Wien, 1998.
URL: `http://www.iue.tuwien.ac.at/phd/knaipp`.

[229] T. Grasser, *Mixed-Mode Device Simulation.* Dissertation, Technische Universität Wien, 1999.
URL: `http://www.iue.tuwien.ac.at/phd/grasser`.

[230] V. Palankovski, *Simulation of Heterojunction Bipolar Transistors.* Dissertation, Technische Universität Wien, 2000.
URL: `http://www.iue.tuwien.ac.at/phd/palankovski`.

[231] S. Wagner, *The Minimos-NT Linear Equation Solving Module.* Diplomarbeit, Technische Universität Wien, 2001.

[232] R. Klima, *Three-Dimensional Device Simulation with MINIMOS-NT.* Dissertation, Technische Universität Wien, 2002.
URL: `http://www.iue.tuwien.ac.at/phd/klima`.

[233] S. Duvall, "An Interchange Format for Process and Device Simulation," *IEEE Trans.Computer-Aided Design*, vol. 7, no. 7, pp. 741–754, 1988.

[234] T. Binder, *Rigorous Integration of Semiconductor Process and Device Simulators.* Dissertation, Technische Universität Wien, 2002.
URL: `http://www.iue.tuwien.ac.at/phd/binder`.

[235] R. Entner, *Three-Dimensional Device Simulation with MINIMOS-NT Using the Wafer-State-Server.* Diplomarbeit, Technische Universität Wien, 2003.

[236] IBM, *Open Visualization Data Explorer*, 2002.
URL: `http://www.research.ibm.com/dx`.

[237] M. Zohlhuber, *Visualisierung von Simulationsdaten.* Diplomarbeit, Technische Universität Wien, 2003.

[238] R. B. Lehoucq and J. A. Scott, "An Evaluation of Software for Computing Eigenvalues of Sparse Nonsymmetric Matrices," Technical Report MCS-P547-1195, Argonne National Laboratory, Argonne, IL, 1996.

[239] S. L. Moshier, "Cephes Mathematical Function Library," 1992.
URL: `http://www.netlib.org/cephes`.

[240] H. T. Lau, *A Numerical Library in C for Scientists and Engineers.* CRC Press, 1995.

[241] R. Lake, G. Klimeck, R. C. Bowen, and D. Jovanovic, "Single and Multiband Modeling of Quantum Electron Transport Through Layered Semiconductor Devices," *J.Appl.Phys.*, vol. 81, no. 12, pp. 7845–7869, 1997.

[242] R. C. Bowen, W. R. Frensley, G. Klimeck, and R. K. Lake, "Transmission Resonances and Zeros in Multiband Models," *Physical Review B*, vol. 52, no. 4, pp. 2754–2765, 1995.

[243] C. Bowen, C. L. Fernando, G. Klimeck, A. Chatterjee, D. Blanks, R. Lake, J. Hu, J. Davis, M. Kulkarni, S. Hattangady, and I.-C. Chen, "Physical Oxide-Thickness Extraction and Verification using Quantum Mechanical Simulation," in *Proc. Intl. Electron Devices Meeting*, pp. 35.1.1–35.1.4, 1997.

[244] G. Klimeck, R. Lake, C. Bowen, W. R. Frensley, and T. S. Moise, "Quantum Device Simulation with a Generalized Tunneling Formula," *Appl.Phys.Lett.*, vol. 67, no. 17, pp. 2539–2541, 1995.

[245] C. L. Fernando and W. R. Frensley, "An Efficient Method for the Numerical Evaluation of Resonant States," *J.Appl.Phys.*, vol. 76, no. 5, pp. 2881–2886, 1994.

[246] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C.* Cambridge University Press, 1997.

[247] W. R. Frensley, "Numerical Evaluation of Resonant States," *Superlattices & Microstructures*, vol. 11, no. 3, pp. 347–350, 1992.

[248] Raytheon TI Systems, *Nanotechnology Engineering Modeling Program (NEMO) Version 3.0*, 1996.

[249] J. Cai and C.-T. Sah, "Gate Tunneling Currents in Ultrathin Oxide Metal-Oxide-Silicon Transistors," *J.Appl.Phys.*, vol. 89, no. 4, pp. 2272–2285, 2001.

[250] H. Z. Massoud and J. P. Shiely, "The Role of Substrate Carrier Generation in Determining the Electric Field in the Oxide of MOS Capacitors Biased in the Fowler-Nordheim Tunneling Regime," *Microelectronic Engineering*, vol. 36, no. 1, pp. 263–266, 1997.

[251] Y. Shi, T. P. Ma, S. Prasad, and S. Dhanda, "Polarity Dependent Gate Tunneling Currents in Dual-Gate CMOSFETS," *IEEE Trans.Electron Devices*, vol. 45, no. 11, pp. 2355–2360, 1998.

[252] M. Städele, B. Tuttle, B. Fischer, and K. Hess, "Tunneling Through Thin Oxides - New Insights from Microscopic Calculations," *Journal of Computational Electronics*, vol. 1, pp. 153–159, 2002.

[253] M. Städele, F. Sacconi, A. D. Carlo, and P. Lugli, "Enhancement of the Effective Tunnel Mass in Ultrathin Silicon Dioxide Layers," *J.Appl.Phys.*, vol. 93, no. 5, pp. 2681–2690, 2003.

[254] F. Sacconi, M. Povolotskyi, A. D. Carlo, P. Lugli, and M. Städele, "Full-Band Approaches to the Electronic Properties of Nanometer-Scale MOS Structures," in *Proc. 4th European Workshop on Ultimate Integration of Silicon*, pp. 125–128, 2003.

[255] S. H. Lo, D. A. Buchanan, and Y. Taur, "Modeling and Characterization of Quantization, Polysilicon Depletion and Direct Tunneling Effects in MOSFETs with Ultrathin Oxides," *IBM J.Res.Dev.*, vol. 43, no. 3, pp. 327–337, 1999.

[256] S. Jallepalli, J. Bude, W.-K. Shih, M. R. Pinto, C. M. Maziar, and A. F. Tasch, jr., "Electron and Hole Quantization and Their Impact on Deep Submicron Silicon p- and n-MOSFET Characteristics," *IEEE Trans.Electron Devices*, vol. 44, no. 2, pp. 297–303, 1997.

[257] Synopsys, *MEDICI User's Manual*, 2003.

[258] L. Selmi, A. Ghetti, R. Bez, and E. Sangiorgi, "Trade-offs between Tunneling and Hot-Carrier Injection in Short Channel Floating Gate MOSFETs," *Microelectronic Engineering*, vol. 36, no. 1-4, pp. 293–296, 1997.

[259] E. M. Vogel, K. Z. Ahmed, B. Hornung, K. Henson, P. K. McLarty, G. Lucovsky, J. R. Hauser, and J. J. Wortman, "Modeled Tunnel Currents for High Dielectric Constant Dielectrics," *IEEE Trans.Electron Devices*, vol. 45, no. 6, pp. 1350–1355, 1998.

[260] M. LeRoy, E. Lheurette, O. Vanbesien, and D. Lippens, "Wave-Mechanical Calculations of Leakage Current Through Stacked Dielectrics for nanotransistor Metal-Oxide-Semiconductor Design," *J.Appl.Phys.*, vol. 93, no. 5, pp. 2966–2971, 2003.

[261] J. D. Casperson, L. D. Bell, and H. A. Atwater, "Materials Issues for Layered Tunnel Barrier Structures," *J.Appl.Phys.*, vol. 92, no. 1, pp. 261–267, 2002.

[262] G. D. Wilk, R. M. Wallace, and J. M. Anthony, "High-k Gate Dielectrics: Current Status and Materials Properties Considerations," *J.Appl.Phys.*, vol. 89, no. 10, pp. 5243–5275, 2001.

[263] J. Robertson, "Band Offsets of Wide-Bandgap Oxides and Implications for Future Electronic Devices," *J. Vac.Sci.Technol.*, vol. 18, no. 3, pp. 1785–1791, 2000.

[264] Y.-Y. Fan, R. E. Nieh, J. C. Lee, G. Lucovsky, G. A. Brown, L. F. Register, and S. K. Banerjee, "Voltage- and Temperature-Dependent Gate Capacitance and Current Model: Application to ZrO_2 n-Channel MOS Capacitor," *IEEE Trans.Electron Devices*, vol. 49, no. 11, pp. 1969–1978, 2002.

[265] S. Harasek, *Zirkoniumdioxiddünnfilme als hoch-ϵ Gateisolatoren für die Siliziumtechnologie.* Dissertation, Technische Universität Wien, 2003.

[266] P. Tanner, S. Dimitrijev, and H. B. Harrison, "Technique for Monitoring Slow Interface Trap Characteristics in MOS Capacitors," *Electron.Lett.*, vol. 31, no. 21, pp. 1880–1881, 1995.

[267] D. A. Antoniadis, I. J. Djomehri, K. M. Jackson, and S. Miller, ""Well-Tempered" Bulk-Si NMOSFET Device Home Page."
URL: `http://www-mtl.mit.edu/Well/`.

[268] W. D. Brown and J. Brewer, *Nonvolatile Semiconductor Memory Technology.* IEEE Press, 1998.

[269] A. Concannon, S. Keeney, A. Mathewson, R. Bez, and C. Lombardi, "Two-Dimensional Numerical Analysis of Floating-Gate EEPROM Devices," *IEEE Trans.Electron Devices*, vol. 40, no. 7, pp. 1258–1262, 1993.

[270] S. Keeney, R. Bez, D. Cantarelli, F. Piccinin, A. Mathewson, L. Ravazzi, and C. Lombardi, "Complete Transient Simulation of Flash EEPROM Devices," *IEEE Trans.Electron Devices*, vol. 39, no. 12, pp. 2750–2757, 1992.

[271] A. Kolodny, S. T. K. Nieh, B. Eitan, and J. Shappir, "Analysis and Modeling of Floating-Gate EEPROM Cells," *IEEE Trans.Electron Devices*, vol. 33, no. 6, pp. 835–844, 1986.

[272] K. T. San, C. Kaya, D. K. Y. Liu, T.-P. Ma, and P. Shah, "A New Technique for Determining the Capacitive Coupling Coefficients in Flash EPROM's," *IEEE Electron Device Lett.*, vol. 13, no. 6, pp. 328–331, 1992.

[273] R. Bouchakour, N. Harabech, P. Canet, P. Boivin, and J. M. Mirabel, "Modeling of a Floating-Gate EEPROM Cell Using a Charge Sheet Approach Including Variable Tunneling Capacitance Gate Depletion Effect," in *Proc. Intl. Symposium on Circuits & Systems*, pp. 822–825, 2001.

[274] R. Duane, A. Concannon, P. O'Sullivan, M. O'Shea, and A. Mathewson, "Extraction of Coupling Ratios for Fowler-Nordheim Programming Conditions," *Solid-State Electron.*, vol. 45, no. 2-3, pp. 235–242, 2001.

[275] D. Kahng and S. M. Sze, "A Floating Gate and Its Application to Memory Devices," *Bell Syst.Tech.J.*, vol. 46, no. 4, pp. 1288–1295, 1967.

[276] P. Pavan, R. Bez, P. Olivo, and E. Zanoni, "Flash Memory Cells - An Overview," *Proc.IEEE*, vol. 86, no. 8, pp. 1248–1271, 1997.

[277] P. Cappelletti, C. Golla, P. Olivo, and E. Zanoni, *Flash Memories.* Kluwer Academic Publishers, 2000.

[278] P. Canet, R. Bouchakour, N. Harabech, P. Boivin, J. M. Mirabel, and C. Plossu, "Study of Signal Programming to Improve EEPROM Cell Reliability," in *Proc. 43rd IEEE Midwest Symp. on Circuits and Systems*, pp. 1144–1147, 2000.

[279] M. K. Cho and D. M. Kim, "High Performance SONOS Memory Cells Free of Drain Turn-On and Over-Erase: Compatibility Issue with Current Flash Technology," *IEEE Electron Device Lett.*, vol. 21, no. 8, pp. 399–401, 2000.

[280] J. M. Caywood, C. J. Huang, and Y. J. Chang, "A Novel Nonvolatile Memory Cell Suitable for Both Flash and Byte-Writable Applications," *IEEE Trans.Electron Devices*, vol. 49, no. 5, pp. 802–807, 2002.

[281] B. Eitan, P. Pavan, I. Bloom, E. Aloni, A. Frommer, and D. Finzi, "NROM: A Novel Localized Trapping, 2-Bit Nonvolatile Memory Cell," *IEEE Electron Device Lett.*, vol. 21, no. 11, pp. 543–545, 2000.

[282] M. H. White, D. A. Adams, and J. Bu, "On the Go with SONOS," *IEEE Circuits & Devices*, no. 7, pp. 22–31, 2000.

[283] K.-T. Chang, W.-M. Chen, C. Swift, J. M. Higman, W. M. Paulson, and K.-M. Chang, "A New SONOS Memory Using Source-Side Injection for Programming," *IEEE Electron Device Lett.*, vol. 19, no. 7, pp. 253–255, 1998.

[284] G. Iannaccone and P. Coli, "Three-Dimensional Simulation of Nanocrystal Flash Memories," *Appl.Phys.Lett.*, vol. 78, no. 14, pp. 2046–2048, 2001.

[285] A. Thean and J. P. Leburton, "Three-Dimensional Self-Consistent Simulation of Silicon Quantum-Dot Floating-Gate Flash Memory Device," *IEEE Electron Device Lett.*, vol. 20, no. 6, pp. 286–288, 1999.

[286] J. J. Welser, S. Tiwari, S. Rishton, K. Y. Lee, and Y. Lee, "Room Temperature Operation of a Quantum-Dot Flash Memory," *IEEE Electron Device Lett.*, vol. 18, no. 6, pp. 278–280, 1997.

[287] B. DeSalvo, G. Ghibaudo, G. Pananakakis, P. Masson, T. Baron, N. Buffet, A. Fernandes, and B. Guillaumot, "Experimental and Theoretical Investigation of Nano-Crystal and Nitride-Trap Memory Devices," *IEEE Trans.Electron Devices*, vol. 48, no. 8, pp. 1789–1799, 2001.

[288] K. Han, I. Kim, and H. Shin, "Characteristics of P-Channel Si Nano-Crystal Memory," *IEEE Trans.Electron Devices*, vol. 48, no. 5, pp. 874–879, 2001.

[289] H. I. Hanafi, S. Tiwari, and I. Khan, "Fast and Long Retention-Time Nano-Crystal Memory," *IEEE Trans.Electron Devices*, vol. 43, no. 9, pp. 1553–1558, 1996.

[290] Y.-C. King, T.-J. King, and C. Hu, "A Long-Refresh Dynamic/Quasi-Nonvolatile Memory Device with 2 nm Tunneling Oxide," *IEEE Trans.Electron Devices*, vol. 20, no. 8, pp. 409–411, 1999.

[291] X. Tang, X. Baie, J.-P. Colinge, C. Gusting, and V. Bayot, "Two-Dimensional Self-Consistent Simulation of a Triangular P-Channel SOI Nano-Flash Memory Device," *IEEE Trans.Electron Devices*, vol. 49, no. 8, pp. 1420–1426, 2002.

[292] K. Nakazato, P. J. A. Piotrowicz, D. G. Hasko, H. Ahmed, and K. Itoh, "PLED - Planar Localised Electron Devices," in *Proc. Intl. Electron Devices Meeting*, pp. 179–182, 1997.

[293] H. Mizuta, K. Nakazato, P. J. A. Piotrowicz, K. Itoh, T. Teshima, K. Yamaguchi, and T. Shimada, "Normally-off PLED (Planar Localised Electron Device) for Non-Volatile Memory," in *Proc. Symposium on VLSI Technology*, pp. 128–129, 1998.

[294] N. Nakazato, K. Itoh, H. Mizuta, and H. Ahmed, "Silicon Stacked Tunnel Transistor for High-Speed and High-Density Random Access Memory Gain Cells," *Electronics Letters*, vol. 35, no. 10, pp. 848–850, 1999.

[295] K. Nakazato, K. Itoh, H. Ahmed, H. Mizuta, T. Kisu, M. Kato, and T. Sakata, "Phase-state Low Electron-number Drive Random Access Memory (PLEDM)," in *Proc. Intl. Solid-State Circuits Conf.*, p. TA 7.4, 2000.

[296] H. Mizuta, M. Wagner, and K. Nakazato, "The Role of Tunnel Barriers in Phase-State Low Electron-Number Drive Transistors (PLEDTR)," *IEEE Trans.Electron Devices*, vol. 48, no. 6, pp. 1103–1108, 2001.

[297] H. Fukuda, J. L. Hoyt, M. A. McCord, and R. F. W. Pease, "Fabrication of Silicon Nanopillars Containing Polycristalline Silicon/Insulator Multilayer Structures," *Appl.Phys.Lett.*, vol. 70, no. 3, pp. 333–335, 1997.

[298] F. Capasso, F. Beltram, R. J. Malik, and J. F. Walker, "New Floating-Gate AlGaAs/GaAs Memory Devices with Graded-Gap Electron Injector and Long Retention Times," *IEEE Electron Device Lett.*, vol. 9, no. 8, pp. 377–379, 1988.

[299] K. K. Likharev, "Layered Tunnel Barriers for Nonvolatile Memory Devices," *Appl.Phys.Lett.*, vol. 73, no. 15, pp. 2137–2139, 1998.

[300] B. Govoreanu, P. Blomme, M. Rosmeulen, J. V. Houdt, and K. D. Meyer, "VARIOT: A Novel Multilayer Tunnel Barrier Concept for Low-Voltage Nonvolatile Memory Devices," *IEEE Electron Device Lett.*, vol. 24, no. 2, pp. 99–101, 2003.

Own Publications

[1] F. Jiménez-Molinos, A. Palma, A. Gehring, F. Gámiz, H. Kosina, and S. Selberherr, "Static and Transient Simulation of Inelastic Trap-Assisted Tunneling," in *Proc. 14^{th} Workshop on Modeling and Simulation of Electron Devices*, (Barcelona, Spain), pp. 65–68, October 2003.

[2] A. Gehring, S. Harasek, E. Bertagnolli, and S. Selberherr, "Evaluation of ZrO_2 Gate Dielectrics for Advanced CMOS Devices," in *Proc. European Solid-State Device Research Conf.*, (Estoril, Portugal), pp. 473–476, September 2003.

[3] T. Ayalew, J.-M. Park, A. Gehring, T. Grasser, and S. Selberherr, "Silicon Carbide Accumulation-Model Laterally Diffused MOSFET," in *Proc. European Solid-State Device Research Conf.*, (Estoril, Portugal), pp. 581–584, September 2003.

[4] E. Ungersböck, A. Gehring, H. Kosina, S. Selberherr, B.-H. Cheong, and W. B. Choi, "Simulation of Carrier Transport in Carbon Nanotube Field Effect Transistors," in *Proc. European Solid-State Device Research Conf.*, (Estoril, Portugal), pp. 411–414, September 2003.

[5] T. Ayalew, A. Gehring, J.-M. Park, T. Grasser, and S. Selberherr, "Improving SiC Lateral DMOSFET Reliability under High Field Stress," *Microelectron.Reliab.*, vol. 43, no. 9-11, pp. 1889–1894, 2003.

[6] T. Ayalew, J.-M. Park, A. Gehring, T. Grasser, and S. Selberherr, "Modeling and Simulation of SiC MOSFETs," in *Proc. IASTED Intl. Conf. on Applied Simulation and Modeling*, (Marbella, Spain), pp. 552–556, September 2003.

[7] A. Gehring, F. Jiménez-Molinos, H. Kosina, A. Palma, F. Gámiz, and S. Selberherr, "Modeling of Retention Time Degradation Due to Inelastic Trap-Assisted Tunneling in EEPROM Devices," *Microelectron.Reliab.*, vol. 43, no. 9-11, pp. 1495–1500, 2003.

[8] A. Gehring, T. Grasser, H. Kosina, and S. Selberherr, "Energy Transport Gate Current Model Accounting for Non-Maxwellian Energy Distribution," *Electron.Lett.*, vol. 39, no. 8, pp. 691–692, 2003.

[9] A. Gehring, H. Kosina, and S. Selberherr, "Analysis of Gate Dielectric Stacks Using the Transmitting Boundary Method," in *Proc. Intl. Workshop on Computational Electronics*, (Rome, Italy), May 2003.

[10] A. Gehring, H. Kosina, T. Grasser, and S. Selberherr, "Consistent Comparison of Tunneling Models for Device Simulation," in *Proc. 4th European Workshop on Ultimate Integration of Silicon*, (Udine, Italy), pp. 131–134, March 2003.

[11] A. Gehring, T. Grasser, H. Kosina, and S. Selberherr, "An Energy Transport Gate Current Model Based on a Non-Maxwellian Energy Distribution," in *Proc. Nanotech 2003 Vol. 2*, (San Francisco, USA), pp. 48–51, February 2003.

[12] A. Gehring, H. Kosina, and S. Selberherr, "Transmission Coefficient Estimation for High-κ Gate Stack Evaluation," in *Proc. Advances in Simulation, Systems Theory and Systems Engineering*, (Skiathos, Greece), pp. 156–159, September 2002.

[13] A. Gehring, T. Grasser, H. Kosina, and S. Selberherr, "A New Gate Current Model Accounting for a Non-Maxwellian Electron Energy Distribution Function," in *Proc. Simulation of Semiconductor Processes and Devices*, (Kobe, Japan), pp. 235–238, September 2002.

[14] A. Gehring, T. Grasser, H. Kosina, and S. Selberherr, "Simulation of Hot-Electron Oxide Tunneling Current Based on a Non-Maxwellian Electron Energy Distribution Function," *J.Appl.Phys.*, vol. 92, no. 10, pp. 6019–6027, 2002.

[15] A. Gehring, T. Grasser, B.-H. Cheong, and S. Selberherr, "Design Optimization of Multi-Barrier Tunneling Devices Using the Transfer-Matrix Method," *Solid-State Electron.*, vol. 46, no. 10, pp. 1545–1551, 2002.

[16] T. Grasser, A. Gehring, and S. Selberherr, "Macroscopic Transport Models for Microelectronics Devices," in *Proc. Sixth World Multiconf. on Systemics, Cybernetics and Informatics*, (Orlando, Florida), pp. 1–8, July 2002.

[17] T. Grasser, A. Gehring, and S. Selberherr, "Recent Advances in Transport Modeling for Miniaturized CMOS Devices," in *Proc. Intl. Caracas Conf. on Devices, Circuits and Systems*, (Aruba, Dutch Caribbean), pp. 1–8, April 2002.

[18] A. Gehring, T. Grasser, and S. Selberherr, "Non-Parabolicity and Non-Maxwellian Effects on Gate Oxide Tunneling," in *Proc. Intl. Conf. on Modeling and Simulation of Microsystems*, (San Juan, Puerto Rico), pp. 560–563, April 2002.

[19] A. Gehring, F. Jiménez-Molinos, A. Palma, F. Gamiz, H. Kosina, and S. Selberherr, "Simulation of Non-Volatile Memory Cells by Accounting for Inelastic Trap-Assisted Tunneling Current," in *Proc. 3rd European Workshop on Ultimate Integration of Silicon*, (Munich, Germany), pp. 15–18, March 2002.

[20] A. Gehring, T. Grasser, and S. Selberherr, "Design Optimization of Multi-Barrier Tunneling Devices Using the Transfer-Matrix Method," in *Proc. Intl. Semiconductor Device Research Symposium*, (Washington D. C.), pp. 260–263, December 2001.

[21] A. Gehring, C. Heitzinger, T. Grasser, and S. Selberherr, "TCAD Analysis of Gain Cell Retention Time for SRAM Applications," in *Proc. Simulation of Semiconductor Processes and Devices*, (Athens, Greece), pp. 416–419, September 2001.

[22] A. Gehring, M. Steinbauer, I. Gaspard, and M. Grigat, "Empirical Channel Stationarity in Urban Environments," in *Proc. European Personal Mobile Communications Conf.*, (Vienna, Austria), February 2001.

[23] J. Bröker, A. Gehring, and T. Sauter, "Simulation und Analyse von Single-Clock CMOS Flip-Flops," in *Proc. Austrochip 2000*, (Graz, Austria), pp. 61–70, October 2000.

[24] A. Gehring and T. Neubauer, "Effect of Base Station Location on the Transmission Power Levels of an Indoor TDD System," in *COST 259 TD 00*, (Bergen, Norway), April 2000.

	Author	Co-Author	Total
Journals	4	1	5
Conferences	12	7	19
Total	16	8	24

Table 8: Publication Statistics.

Curriculum Vitae

February 5th, 1975
Born in Mistelbach, Austria.

June 1994
High school graduation (*Matura*) at the HTBLA Hollabrunn.

October 1994 – August 1995
Compulsory civil service.

October 1995
Enrolled in Electrical Engineering at the Vienna University of Technology, Austria.

March 2000
Received degree of *Diplom-Ingenieur* (M.Sc.) in Electrical Engineering from the Vienna University of Technology (with honors).

March 1998 and 2000
Received award *Leistungsstipendium* for extraordinarily fast and successful studies.

April 2000
Entered doctoral program at the Institute for Microelectronics, Vienna University of Technology, under the supervision of Prof. SIEGFRIED SELBERHERR.

April 2001
Entered the position of teaching assistant at the Institute for Microelectronics

July 2001 – August 2001
Held a position as a visiting researcher at the Samsung Advanced Institute of Technology (Seoul, South Korea).

July 2003 – August 2003
Held a position as a visiting researcher at Cypress Semiconductor (San Jose, USA).