QUANTUM MECHANICAL EFFECTS ON MOSFET SCALING LIMIT

A Dissertation
Presented to
The Academic Faculty

by

Lihui Wang

In Partial Fulfillment
of the Requirements for the Degree
DOCTOR OF PHILOSOPHY in the
School of ELECTRICAL AND COMPUTER ENGINEERING

Georgia Institute of Technology
AUGUST 2006

QUANTUM MECHANICAL EFFECTS ON MOSFET SCALING LIMIT

Approved by:

Dr. James D. Meindl, Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Ian F Akyildiz
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Russell Dupuis
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Phillip First
Physics Department
Georgia Institute of Technology

Dr. William R. Callen
School of Electrical and Computer Engineering
Georgia Institute of Technology

Date Approved: July 6, 2006

This Work Is Dedicated to My Parents for Having Faith in Me and Never Letting Me Give Up on Myself

ACKNOWLEDGEMENTS

This thesis is the result of six years of work whereby I have been accompanied and supported by many people. I am glad that I have now the opportunity to express my gratitude to all of them.

The first person I would like to thank is my advisor Dr. James D. Meindl. During these years, I have known Dr. Meindl as a knowledgeable and compassionate person. I owe him lots of gratitude for guiding me through the research, for his patience and for his understanding. Other than being an excellent advisor, Dr. Meindl is a good friend, giving me support throughout the course of this study.

I would like to thank Jennifer Root Tatham for keeping an eye on the progress of my work and motivating me to complete the work. Special thanks to Dr. Blanca Austin for being there when I needed her advises. I would also like to express my gratitude to committee members Dr. William Callen, Dr. Russell Dupuis, Dr. Ian Akyildiz and Dr. Phillip First for providing helpful suggestions for my thesis.

The work would not have been possible without the help from my fellow students and my friends. I thank them all for our many discussions and for having confidence in me.

Many thanks go to Semiconductor Reseach Corporation whose funding and assistance were vital for this research.

This has been a wonderful journey and I am grateful to all who had helped to make it that way.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

d	[cm] depletion depth
d_{cl}	[cm] depletion depth in classical model
d_{qm}	[cm] depletion depth in quantum model
g_m	transconductance
h	Planck constant, $6.62617*10^{-34}\ J{\cdot}s$
k	Boltzmann constant, $1.38066*10^{-23}\ J/K$
m	[g] mass
m_0	electron rest mass, $0.911*10^{-27}g$
m^*	[g] effective mass
n	free electron concentration
n_i	intrinsic electron density, $1.45*10^{10}\ cm^{-3}$ at 300 K
q	[C] electron charge
t_{ox}	[cm] gate oxide thickness
t_{Si}	[cm] silicon body thickness
v_{sat}	$[cm/s]$ saturation velocity
C_d	[F/cm^2] depletion layer capacitance
C_{inv}	[F/cm^2] inversion layer capacitance
C_g	[F/cm^2] gate capacitance
C_{ox}	[F/cm^2] oxide layer capacitance
D	tunneling probability
E_{Fm}	[eV] Fermi energy level of the gate material
E_{Fs}	[eV] Fermi energy level of substrate
E_g	[eV] silicon band gap, 1.12 eV at 300 K
E_b	[eV] potential barrier height
E_B	$[eV]$ average barrier height

I_{ds}	[A] drain current
J_{sm}	[A/cm^2] tunneling current density from channel to gate
J_{ms}	[A/cm^2] tunneling current density from gate to channel
J_T	[A/cm^2] net tunneling current density
L	[cm] channel length
L_G	[cm] gate length
N_A	[cm^{-3}] p-type doping concentration in substrate
N_D	[cm^{-3}] n-type doping concentration in substrate
$P_{dynamic}$	[$\mu W/gate$] dynamic power
P_{static}	[$\mu W/gate$] static power
P_{sub}	[$\mu W/gate$] subthreshold leakage power
P_{total}	[$\mu W/gate$] total power consumption
P_{tunnel}	[$\mu W/gate$] gate tunneling power
Q_{TH}	[cm^{-3}] sheet density of free carriers under threshold condition
S	[$mV/decade$] subthreshold swing
T	[K] absolute temperature
V_{ds}	[V] drain-to-source voltage
V_{dsat}	[V] saturation voltage
V_{FB}	[V] Flat-band voltage
V_{gs}	[V] gate-to-source voltage
V_{TH}	[V] threshold voltage
$V_{TH,CL}$	[V] threshold voltage in classical model
$V_{TH,QM}$	[V] threshold voltage in quantum-mechanical model
$V_{TH,long}$	[V] long-channel threshold voltage
ΔV_{TH}	[V] threshold voltage roll-off
$\Delta V_{TH,shift}$	[V] threshold voltage shift
ε_0	dielectric constant of vacuum, $8.854*10^{-12}\ F*m^{-1}$

ε_{Si}	[F/cm] dielectric constant of silicon,11.8ε_0
ε_{ox}	[F/cm] dielectric constant of silicon oxide, 3.9ε_0
ϕ_c	[V] channel potential
ϕ_s	[V] surface potential
ϕ_B	[V] difference between Fermi level and intrinsic level in silicon
μ_0	[$cm^2/V{\cdot}s$] low field mobility
μ_{eff}	[$cm^2/V{\cdot}s$] effective mobility
τ	[s] gate delay
χ	[V] modified electron affinity in silicon, 3.1 V
Ψ	electron wavefunction
ASIC	application-specific integrated circuit
BTBT	band-to-band tunneling
CMOS	complementary metal oxide semiconductor
DIBL	drain induced barrier lowering
FIBL	fringe induced barrier lowering
GIDL	gate-induced drain leakage
GSI	giga-scale integration
MOS	metal oxide semiconductor
MOSFET	metal-oxide-semiconductor field effect transistor
QME	quantum mechanical effect
SCE	short-channel effect
SDE	source and drain extension
SOI	silicon-on-insulator
UTB	ultra-thin-boday

SUMMARY

As CMOS technology continuous to be aggressively scaled, it approaches a point where classical physics is insufficient to explain the behavior of a MOSFET. At this classical physics limit, a quantum mechanical model becomes necessary to provide thorough assessment of the device performance and scaling. This thesis describes advanced modeling of nanoscale bulk MOSFETs incorporating critical quantum mechanical effects such as gate direct tunneling and energy quantization of carriers.

In the gate tunneling analysis, an explicit expression of gate direct tunneling for thin gate oxides has been developed by solving the Schrödinger equation analytically. In addition, the impact of different gate electrode as well as gate insulation materials on the gate direct tunneling is explored. This results in an analytical estimation of the potential solutions to excessive gate leakage current.

The energy quantization analysis involves the derivation of a quantum mechanical charge distribution model by solving the coupled Poisson and Schrödinger equations. Based on the newly developed charge distribution model, threshold voltage and subthreshold swing models are obtained. A transregional drain current model which takes into account the quantum mechanical correction on device parameters is derived. Results from this model show good agreement with numeric simulation results of both long-channel and short-channel MOSFETs, thus validating the analysis

The models derived here are used to project MOSFET scaling limits. These limits of bulk MOSFETs are predicted according to various criteria, including circuit power and delay, device leakage current and the system uniformity requirement. Tunneling and quantization effects cause large power dissipation, low drive current, and strong sensitivities to process variation, which greatly limit CMOS scaling. Developing new materials and structures is imminent to extend the scaling process.

CHAPTER 1

INTRODUCTION

1.1 Introduction and Background

For the last three decades, the semiconductor industry has strived to miniaturize the structure of the MOSFET, which is shown in Figure 1.1. Following Moore's Law [1], the size of the transistor is reduced by a factor of *0.7* each technology generation. According to the International Technological Roadmap for Semiconductors (ITRS) [2], the gate length (L_G) of the MOSFET transistor will shrink to *30 nm* in 2008, leading to *5* billion MOSFETs on one application-specific integrated circuit (ASIC) chip.

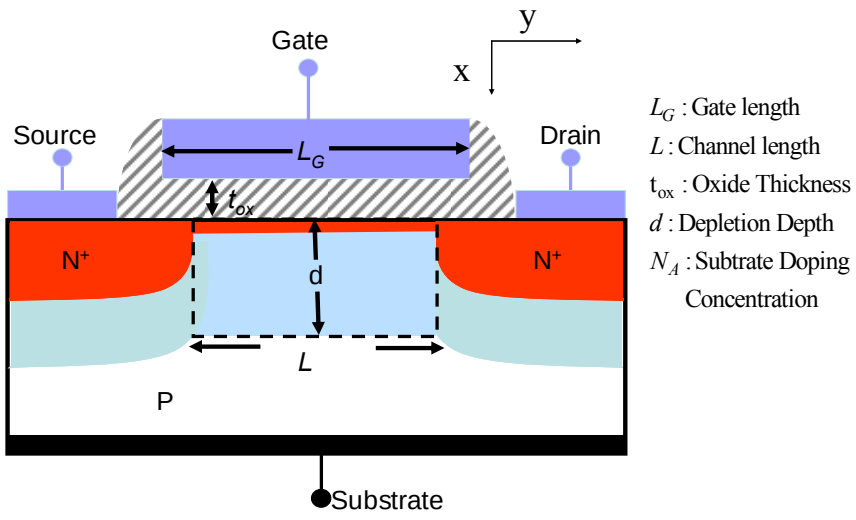

Figure 1.1
Structure of a bulk MOSFET transistor.

There are many reasons to pursue miniaturization including: (1) the cost per transistor is reduced as a MOSFET occupies less area; (2) more transistors can be

integrated on the chip, therefore, it can perform more complex functions; (3) capacitances are reduced, which in turn reduces the time and power required to switch a MOSFET. Despite its potential advantages, miniaturization presents a series of challenges to device design.

First, the power consumption of a system increases dramatically at high-frequency operation. Therefore, the heat generation by a high frequency silicon chip is tremendously high, and the cooling of the chip becomes difficult and, sometimes, practically impossible for some low-power applications. Supply voltage (V_{dd}) reduction is an effective method to reduce the power consumption per device. However, lowering power supply voltage reduces the operational speed of MOSFETs. Hence, controlling power consumption has been a primary concern in MOSFET scaling. It has been found that in order to maintain the switching speed of a MOSFET, its threshold voltage (V_{TH}) should be reduced at the same rate the supply voltage is reduced. However, low threshold voltage can lead to excessive subthreshold leakage current in a MOSFET. In addition, the threshold voltage is reduced by decreasing the oxide layer thickness (t_{ox}) in a bulk MOSFET, which causes leakage through the gate oxide. Thus, the relationship between power consumption and operation speed is critical in obtaining optimal scaled devices.

Second, the MOSFET characteristics degrade with the reduction in size. Two key characteristics are threshold voltage (V_{TH}) and subthreshold swing (S), known as short-channel effects (SCEs). Threshold voltage decreases and subthreshold swing increases because of two-dimensional (2-D) electrostatic charge sharing between the gate and the source-drain regions. Consequently, the on-to-off current ratio is reduced substantially, which results in a significant increase in standby power and compromised

overall performance. Additionally, SCEs exacerbate susceptibility to process variations. To scale down the channel length (L) without excessive SCEs, both the oxide thickness (t_{ox}) and the gate-controlled depletion depth (d) should be reduced. For the *90 nm* technology node, t_{ox} is *1.2 nm*, corresponding to six atomic layers of silicon oxide [2, 3]. Further reducing t_{ox} is increasingly more difficult and will cause severe gate leakage [4-6].

Finally, classical physics is insufficient to understand fully the behavior of MOSFETs at small dimensions. The channel length of modern MOSFETs is approaching the mean distance between carrier collisions, with the oxide layer thickness reaching the dimension of a few atomic layers [4-7]. In this situation, significant deviation from the classical calculation is observed in the behavior of MOSFETs, which must be explained by quantum theory.

1.2 Origin of the Problem

The fundamental distinction of quantum theory from classical physics is that infinitely small particles are treated as waves. Unlike the "solid billiards" with definite positions and velocities assumed in classical theory, particles in quantum theory are waves dispersed in space. This quantum mechanical description of a particle is represented by the wave function $\psi(\vec{r})$, such that the probability of finding the particle in the volume $d\vec{r}^3$ is equal to $|\psi(\vec{r})|^2 d\vec{r}^3$. The wavefunction of a carrier in a semiconductor satisfies the Schrödinger Equation [8, 9]:

$$-\frac{\hbar^2}{2m_0}\nabla^2\psi + \left[U_C(\vec{r}) + V_E(\vec{r})\right]\psi(\vec{r}) = E\psi(\vec{r}), \tag{1.1}$$

where $\hbar = h/\pi = 1.054\times10^{-34} J\bullet S$ is the reduced Planck constant, $m_0 = 0.911\times10^{-27}\ g$ is the electron rest mass, $U_C(\vec{r})$ is the periodic internal crystalline potential, $V_E(\vec{r})$ is the external potential resulting from an applied electric field, and E is the energy of the carrier. If the dimension of the crystal is large compared to the atomic dimension, the external potential can be considered a small perturbation on the crystalline potential. Considering the crystalline potential only results in

$$-\frac{\hbar^2}{2m_0}\nabla^2\psi_C + U_C(\vec{r})\psi_C(\vec{r}) = E\psi_C(\vec{r}). \tag{1.2}$$

From Bloch's Theorem [9, 10], since $U_C(\vec{r})$ is periodic with the periodicity of the lattice, $U_C(\vec{r}) = U_C(\vec{r}+\vec{R})$, there exists a wavevector $\vec{k}$ (in $[cm^{-1}]$) in the reciprocal lattice and a periodic function $\varphi_k(\vec{r})$ such that $\varphi_k(\vec{r}) = \varphi_k(\vec{r}+\vec{R})$ and ψ_C is of the form

$$\psi_C(\vec{r}) = \exp(i\vec{k}\cdot\vec{r})\varphi_k(\vec{k},\vec{r}). \tag{1.3}$$

The solution of equations (1.2) and (1.3) gives the relationship of energy versus the wavevector ($E-\vec{k}$ relationship) and, thereby, the band structure of semiconductors [9]. The movement of carriers follows the $E-\vec{k}$ relationship, so that they can be considered as classical particles with the effective mass m^* (in $[g]$) given by

$$m^* = \frac{1}{\hbar^2}\left(\frac{d^2E}{dk^2}\right)^{-1}. \tag{1.4}$$

With the effective mass, the motion of carriers under the applied field is handled by the classical method as

$$\vec{F} = -q\vec{E} = m^* \frac{d\vec{v}}{dt}, \tag{1.5}$$

where the force $\vec{F}$ (in $[N]$) is caused by the applied electric field, $q = 1.602 \times 10^{-19}\ C$ is the electron charge, $\vec{E}$ (in $[V/cm]$) is applied electric field, $\vec{v}$ (in $[cm/s]$) is the velocity of an electron, and t (in $[s]$) is the time. Therefore, in a relatively weak external field, carriers comply with the classical theory from a macroscopic view and the quantum nature of the carriers is concealed by the effective mass approximation [9]. In this way, equation (1.1) is solved by two steps: numerically solving equation (1.2) for the $E - \vec{k}$ relationship and the effective mass, and applying equation (1.5) for the motion of carriers in the external field.

However, such simplification does not apply when the dimension is extremely small and the external field is large. In this case, the external field cannot be considered as a small perturbation. Carriers exhibit their quantum mechanical properties in the external field and the wave-like behavior, such as tunneling through a potential barrier and energy quantization in a potential well, can be directly observed. Corrections should be introduced to account for these quantum-mechanical effects (QMEs) in the places where potential energy changes sharply. For that reason, it is necessary to deal with the external field quantum mechanically, which is given by [9] as

$$-\frac{\hbar^2}{2m^*}\nabla^2\psi_E + V_E(\vec{r})\psi_E(\vec{r}) = E\psi_E(\vec{r}). \tag{1.6}$$

Equation (1.1) is then simplified into two equations as

$$-\frac{\hbar^2}{2m_0}\nabla^2\psi_C + U_C(\vec{r})\psi_C(\vec{r}) = E\psi_C(\vec{r}), \tag{1.7}$$

and

$$-\frac{\hbar^2}{2m^*}\nabla^2\psi_E + V_E(\vec{r})\psi_E(\vec{r}) = E\psi_E(\vec{r}). \quad (1.8)$$

Therefore, the quantum effects induced by the external fields can be handled separately from crystalline potential, and the simplicity of effective mass approximation is preserved.

As MOSFETs are scaled down for Giga-Scale Integration (GSI) [11], quantum effects need to be considered in MOSFET design and modeling [2]. In today's CMOS technology, the gate oxide thickness of a MOSFET is less than *1.5 nm*, and the channel is doped as high as *1e18 cm*$^{-3}$ [3, 12]. For MOSFETs with heavily doped channels and ultra-thin oxide layers, the field in the oxide can reach a maximum of *5 MV/cm*, while the field in the silicon region routinely exceeds *1 MV/cm* [6]. The combination of the ultra-thin oxide layer and the heavily doped channel invalidates the accurate modeling of MOSFETs solely by classical physics, and the QMEs of the device must be taken into account. The ultra-thin oxide layer reduces the width of the energy barrier that separates the gate from the channel, thus making it easier for electrons to tunnel through the insulator layer [13-20]. Tunneling also occurs at the source/drain extension region overlapping with the gate, making a leakage in addition to the subthreshold current, as shown in Figure 1.2. This direct gate tunneling current could be the dominant source of device leakage, leading to faulty circuit operation and the increase in standby power in the MOSFET. Additionally, band-to-band tunneling (BTBT) is caused by electrons crossing a reverse biased p-n junction from the p-side valence band to the n-side conduction band [21, 22]. In a MOSFET it is manifested both as gate-induced drain leakage (GIDL) in the drain-gate overlap region and reverse-biased junction leakage in the halo-implanted region [23, 24]. As the channel length is scaled down to *10 nm* and

below, it is expected that the source-to-drain tunneling will dramatically affect performance. Theoretical studies and simulations show that the source-to-drain tunneling dominates off-current at $L < 10\ nm$ and sets an ultimate scaling limit [25, 26].

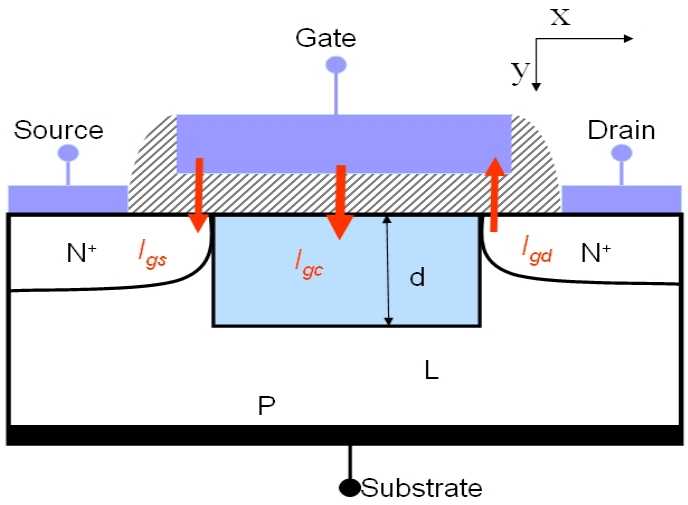

Figure 1.2
Tunneling current in a MOSFET. I_{gs}: tunneling between gate and source; I_{gc}: tunneling between gate and channel; I_{gd}: tunneling between gate and drain.

In addition, energy quantization occurs in the channel near the interface of the oxide layer and the silicon channel because of the presence of the strong electric field [5, 6, 27, 28]. As illustrated by Figure 1.3, quantization leads to the splitting of the continuous energy band and the formation of subbands with a two-dimensional (2-D) density of states in each one [4, 6, 28-31]. Because of the smaller density of states in the 2-D system, the net sheet charge density of carriers in the channel is lower than that calculated from the classical (3-D) case, as shown in Figure 1.4 [32]. Thus, it requires a

larger gate voltage to generate the same charge sheet density in the 2-D inversion layer as that in the corresponding 3-D case. Consequently, the threshold voltage increases when energy quantization is considered [29, 33]. The laws of quantum mechanics force the carrier density to vanish at the silicon/silicon oxide interface, whereas the carrier density reaches its peak value at the interface according to classical theory. Therefore, the overall distribution of carriers is effectively displaced toward the substrate by the quantization effect. This displacement results in a capacitance in series with the oxide layer capacitance. In sub-*90 nm* technologies, where the finite inversion layer thickness can be a significant fraction of the physical oxide thickness, the inversion layer capacitance causes a considerable discrepancy between the oxide capacitance and the measured gate capacitance [2]. For a MOSFET in the superthreshold region, the reduced gate capacitance [34] lowers the transconductance and the drive currents [35]. For MOSFETs in the subthreshold region, the reduced gate capacitance resulting from QME increases the short-channel effects. Therefore, since device behavior is noticeably affected, it is important to account for the quantum-mechanical effects in the design of sub-*90 nm* devices [36-39]. In this operational regime, classical models are inadequate and will lead to erroneous and misleading predictions of critical device structure and electrical behavior parameters, such as the physical oxide thickness, threshold voltage, drive current, gate capacitance, and subthreshold swing.

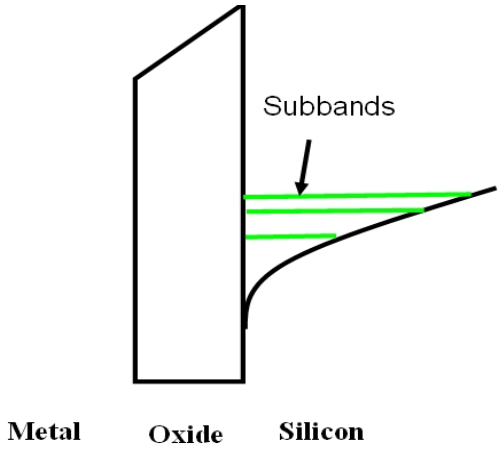

Figure 1.3
Discrete energy levels due to quantization.

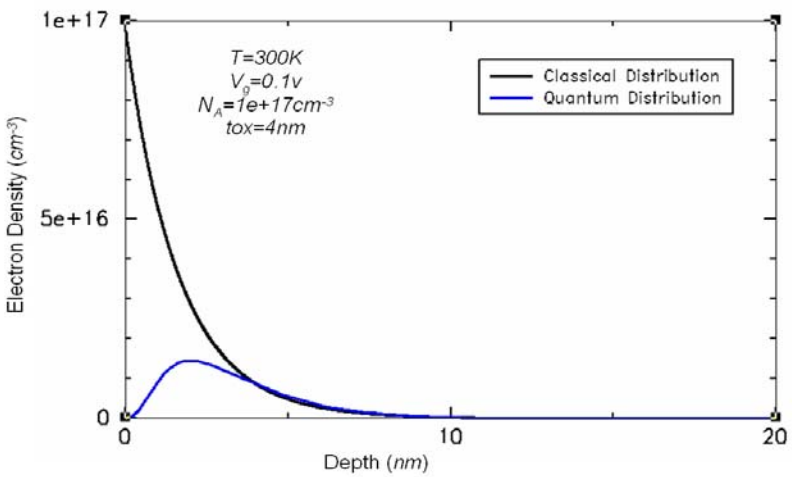

Figure 1.4
Electron density profile calculated classically and quantum mechanically by SCHRED [32].

1.3 Historical Review of QME Modeling

Accurately modeling tunneling and quantization in MOSFETs requires the multidimensional solution of the Schrödinger and Poisson equations. A lot of work [4, 5, 13, 40-45] has been devoted to develop algorithms for accurate numeric solutions of these equations. However, from a circuit modeling point of view, a one-dimensional analytical solution is sufficient to account for the quantum-mechanical correction in the classical transport framework of drift-and-diffusion models [46, 47]. Furthermore, analytical solutions are preferable because of their simplicity in format and fast computational speed. With these analytical solutions, compact physical models can be obtained that estimate QMEs on the device comprehensively, making it easy to predict device scalability and circuit performance for future technology generations.

1.3.1 Gate Direct Tunneling

The gate direct tunneling problem in MOSFETs was first addressed by H. S. Momose et al. [48] in 1994 when a high-performance MOSFET was fabricated with a *1.5 nm* thin oxide layer. Although high transconductance is obtained by reducing the gate oxide thickness in this device, large gate direct tunneling current is observed. Therefore, it is important to investigate the reduction of gate oxide thickness below the tunneling limit in small gate length MOSFETs, as it pertains to better current drive and transconductance [48].

Various gate direct tunneling models were developed using numeric methods such as Bardeen's approach [18], the resonant transfer matrix method [4, 5], and transparency-based approximations [16, 49]. All results reveal the exponential dependence of the gate direct tunneling on gate oxide thickness and show how the

standby power consumption restricts the gate direct tunneling current, presenting a severe limitation on the gate oxide thickness [6]. Although these models qualitatively indicate that excessive leakage aroused by gate direct tunneling potentially degrades circuit performance, their time-consuming numeric computations makes them impractical for a circuit simulation. Choi [13] studied in detail the impact of the tunneling current in different CMOS circuits by applying the macro-circuit model, which relies on the extracted tunneling current data from the device simulation of a single MOSFET. The results show increased delay and power consumption in both static and dynamic logic CMOS circuits. They also show erroneous switching in dynamic logic circuits caused by gate tunneling current. However, this model is inadequate for providing further information on how to optimize circuit and device design because of the time-consuming device simulation. Instead of using data extraction from numeric simulations, the semi-empirical tunneling model given in [14, 15, 50] is formulated as an analytical expression of terminal voltages. Although this model provides excellent accuracy for a large variety of operating conditions and is therefore convenient for circuit simulation and design, the empirical parameters are not consistent among technology generations. The deficiency of the detailed physics of tunneling makes these models unsuitable for long-term projection of the impact of gate direct tunneling on device scaling [51].

1.3.2 Energy Quantization

The phenomenon of inversion charge quantization has been observed for decades, but its influence on device performance has been addressed only recently. Early research on quantization in the 1970s [52] focused on the computation of electron energy levels and the inversion charge distribution of the 2-D gas on subbands. Fang and Howard [53]

use the single-electron assumption that all electrons in the channel are deemed equal and the energy levels of the electron gas are the same as those of one electron by accounting for the applied field and electron-electron repulsion. This approach simplifies the calculation and the problem is reduced to calculating the single electron energy levels in a potential well. They also approximate the channel as a triangular potential well to simplify the calculation of the quantized levels, regardless of the fact that the triangular shape does not resemble the potential distribution in the channel in a MOSFET. Stern [52] uses a more accurate variational method to solve the coupled Schrödinger and Poisson's equations, giving the analytical expression of the quantized energy levels. Their results show the two distinctive differences of between the quantum-mechanical solution and the classical models: (1) the channel carriers are distributed among discrete energy levels with 2-D density of states instead of the single continuous energy band with 3-D density of states; (2) the peak of the space carrier concentration is located some distance away from the surface of the substrate, which leads to a finite thickness of inversion layer. Although these works elucidate the fundamental changes in the carrier distribution induced by the quantization effect, it is unclear how the performance of a MOSFET is affected. Further research is needed to extend this physical analysis into MOSFET voltage and current models for device and circuit design. Consequently, it is necessary to incorporate quantization analysis in the capacitance-voltage ($C-V$) and current-voltage ($I-V$) characteristics.

1.3.3 Gate Capacitance Degradation

Inversion charges are characterized by the inversion layer capacitance C_{inv} (in $[F/cm^2]$), which is defined as the variation of the inversion charge sheet density Q_{inv}

(in $[C/cm^2]$) with respect to the surface potential of the channel ϕ_s (in $[V]$), i.e. $C_{inv} = \frac{\partial Q_{inv}}{\partial \phi_s}$. Classical theory assumes that inversion charges are concentrated beneath the gate oxide forming a very thin layer, so that C_{inv} is much larger than the gate oxide capacitance C_{ox} (in $[C/cm^2]$). Consequently, the gate capacitance C_g (in $[C/cm^2]$), the series combination of C_{inv} and C_{ox}, is almost equal to C_{ox}. However, it has been observed that C_g could be notably smaller than C_{ox} in MOSFETs with the ultra-thin oxide layers [28, 31, 35, 54]. This discrepancy can be explained by the highly reduced C_{inv}, which is induced by the finite inversion layer thickness, observed on the quantum mechanical analysis. It is proven by Takagi's experiments [54] that in strong inversion, C_{inv} changes linearly with only $Q_{inv}^{1/3}$, $C_{inv} \propto Q_{inv}^{1/3}$, not following $C_{inv} \propto Q_{inv}$ as predicted by the classical model [10]. The simulation results by Hareland [27, 55] coincide with the experiments. The gate capacitance attenuation in the superthreshold region is often referred to as transconductance (g_m) degradation [6, 31, 54-56], since g_m is approximately $g_m = \frac{W}{L}\mu C_g$. Similarly, C_g reduction is observed in the subthreshold region [30, 57, 58]. As a result, subthreshold swing (S) becomes larger, indicating worse turn-off characteristics.

An "effective oxide thickness" is introduced to compensate for the absence of inversion layer capacitance in current circuit simulation tools [2, 55, 59]. However, this effective oxide thickness is obtained from the measurement of the manufactured MOSFET and therefore it is difficult to project the future scalability of the device. Furthermore, such simplification fails to take into account the fact that C_{inv} varies with

gate voltage. An accurate physical model of C_{inv}, including the quantization effect, yields valuable insights, allowing for projections of the optimal scaling of device parameters such as voltage and oxide layer.

1.3.4 Threshold Voltage Shift

The 2-D carrier distribution of the subbands and discrete energy levels lead to a reduced charge sheet density compared to the classical calculation. Therefore, extra band bending is required for an increase in channel carrier density. Van Dort [29] found that the calculated threshold voltage in highly doped MOSFETs deviates from experiments. He observed that this threshold voltage deviation results from the energy quantization in the highly doped channel. Specifically, he observed that the lowest quantized energy states for inversion charges rise up from the conduction band bottom, causing an effective "band-gap broadening" [29]. Although this model properly attributes the threshold voltage shift to the energy quantization, it still treats the inversion charges as a 3-D system. The decreased density of states for inversion charge, i.e., from the higher 3-D density to the lower 2-D density, has not been accounted for in his model.

1.3.5 Short-Channel Effects and Quantum Mechanical Effects

When the channel length of the MOSFET is reduced, depletion charges in the channel region are apt to be influenced by the drain as much as by the gate, as shown in Figure 1.5. The barrier preventing the carriers in the source from leaking into the channel is lowered by the drain voltage, which is usually referred to as drain-induced barrier lowering (DIBL) [60, 61]. Because of the undesirable coupling between the channel and the drain region, the threshold voltage is reduced, and the subthreshold swing is increased

for a transistor with shorter channel length. Such short-channel effects, which degrade the off-state performance of MOSFETs and increase the sensitivity to dimensional variations, are critical to modern device design [6, 62, 63].

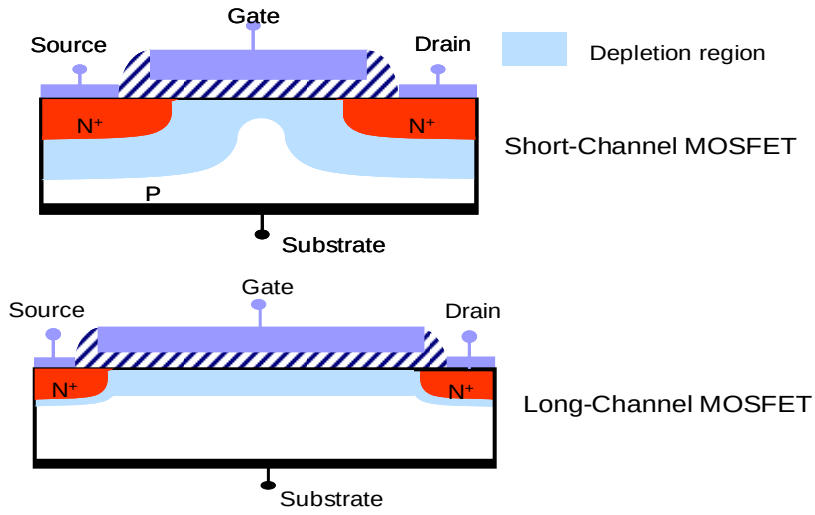

Figure 1.5
Depletion charge distribution influenced by the source/drain in the short-channel MOSFET.

The analysis of SCEs in miniaturized devices demands the 2-D electrostatic potential profile, which is obtained by solving the 2-D Poisson equation. Previous analytical solutions, without considering the QMEs [61, 63], indicate that the DIBL is highly dependent on the depletion depth and the gate oxide thickness. The energy quantization of carriers increases the depletion depth required at the threshold condition, giving rise to greater DIBL effects [64]. Although the 1-D solution of Poisson's equation and Schrödinger's equation leads to the predictive result that SCEs are influenced by the

presence of QMEs, quantitative analysis should deal with the 2-D Poisson equation while simultaneously considering the charge distribution quantum mechanically. The majority of the research on QMEs is based on the solution of the coupled 1-D Poisson equation and 1-D Schrödinger equation that does not handle the 2-D geometry of short-channel devices. Numeric solutions of the 2-D Poisson equation and Schrödinger equation [40, 44] provide results showing that QMEs aggravate SCEs quantitatively. However, there is still not enough insight into the interrelationship of QMEs and SCEs and their dependency on physical parameters.

1.3.6 I-V Characteristics

Because of prominent changes in device characteristics resulting from QMEs, it is useful to develop a compact MOSFET drain current ($I-V$) model for circuit simulations which accounts for the underlying QMEs. Recent numerical simulations [41, 44, 57, 65] as well as theoretical studies [56] clearly show that QMEs change the charge distribution and, thereby, play an important role in drain current for a state-of-the-art MOSFET. Unfortunately, these works fail to provide an explicit expression for the drain current, making them difficult to employ for the circuit simulation of future technology generations. These works, however, show no obvious alteration for electron transport resulting from QMEs. Thus, it is reasonable to develop a quantum-mechanical $I-V$ model by combining the quantum mechanical carrier distribution with the classical carrier transport based on the Boltzmann transport equation [46, 47, 66].

1.3.7 Scaling Limits Projection

As predicted by the ITRS roadmap [67], the MOSFET must be scaled down by various modes to meet the continuous demand for higher speed, integratability, and lower

power consumption, etc. The ITRS [67] specifies the different design targets as low power, high speed, low standby power for different applications of future systems. Different scaling methodologies have been developed for each criterion. The Minimum Power Methodology [68] has been proposed to minimize the power consumption under the constraint of circuit delay. Frank [62] found the multiple device scaling limits by minimizing SCEs with various constraints of low, medium, and high power consumption. Following these methodologies and criteria, the device parameters in future technology generations can be predicted according to the projected circuit/system performance.

The absence of QMEs in the previous scaling methodologies makes them inappropriate for the projection of sub-*90 nm* technologies, in which QMEs become dominant impeditive factors. Energy quantization reduces the control of SCEs, leading to degradation of the performance including worse system uniformity and severe leakage in the VLSI system [6]. The conventional scaling method, shrinking the insulator layer to suppress SCEs [6, 11, 62, 63], induces the prohibitively large gate tunneling current in MOSFETs. Therefore, the circuit/system functionality and performance in future generations must be re-examined, including the influence of QMEs. New technologies, such as high-permittivity (high-κ) dielectrics, proposed to cope with excessive tunneling current and to extend the scaling limit of MOSFETs, can also be evaluated through quantum-mechanical models.

1.4 Scope and Organization

With MOSFET scaling being challenged by quantum mechanical effects, nanoscale device modeling which incorporates them is critical to assist in device design as well as to understand the scaling limits. The compact physical model provides insight

into both the characteristics of modern semiconductor device and circuit performance under the influence of QMEs. Simulation based on the compact physical models reduces the cost of developing a novel technology and shortens the time-to-market. They may also be utilized to explore innovative device structures.

This thesis focuses on critical QMEs of bulk MOSFETs in the sub-*90 nm* regime. The main objective is to develop physics-based MOSFET device models including direct gate direct tunneling and energy quantization of the carriers. By applying the physical device models, we are able to investigate the impact of QMEs on device characteristics and circuit/system performance and therefore reveal the remaining potential of CMOS technologies under QMEs.

Chapter 2 is devoted to a thorough examination of gate tunneling in ultra-thin oxide MOSFETs. Gate tunneling equations are derived from the solution of the Schrödinger equation. This model considers both electron and hole tunneling in MOSFETs. In addition, the impact of the polysilicon gate depletion and high-κ gate insulators on gate tunneling is discussed.

Chapter 3 describes the energy quantization of the carriers in the channel. The quantization model is based on the analytical solution of coupled Schrödinger and Poisson equations by the variational method. Enforced by the quantiztion effect, carriers are distributed on the discrete subbands instead of the continuous energy band depicted by classical theory. This deviation from the classical theory critically affects the relationship between the density profile of inversion charges and the gate electrode voltage. A compact model of the *C-V* characteristic is presented by considering the energy quantization effect on carrier distribution.

In Chapter 4, compact models of various device parameters incorporated with the energy quantization effect are derived. Key device parameters, such as threshold voltage and subthreshold swing, are rederived based on the quantum-mechanical distribution of carriers. Their susceptibility to energy quantization effects is discussed in long and short channel MOSFETs. The influence of the energy quantization on SCEs is revealed by comparing the performance of short-channel and long-channel devices. These models are subsequently integrated into the comprehensive *I-V* model.

In Chapter 5, the different criteria according to various limiting factors in MOSFET scaling are exploited to predict the minimum size of MOSFETs. The hierarchical scaling limits at the device, circuit, and system levels are investigated using a quantum mechanical model. Predictions based on the classical and quantum models are compared to unveil the roles played by QMEs in future technological generations. The high-κ gate dielectric, which could potentially be used to reduce the tunneling current, is examined against silicon dioxide in MOSFET scaling. Extension of the scalability of bulk MOSFETs by adopting new materials and structures is also discussed.

Chapter 6 summarizes the findings of this research and suggests possible areas for further investigation.

CHAPTER 2

TUNNELING

2.1 Introduction and Background

MOSFETs are scaled down with the purpose of enhancing performance and accommodating more devices within the same solid-state real estate. This increases the device capacity per wafer, which cuts the manufacturing cost per transistor resulting in higher profits. For over thirty years, the feature size of a MOSFET has been reduced thanks to the progress in lithography at the rate of $0.7\times$ every three years [1, 39]. Such aggressive scaling of CMOS technology is becoming progressively more difficult because of undesirable physical effects in small devices.

The scaling of MOSFETs is performed in both the vertical and the lateral directions. The lateral shrinking is performed to obtain a shorter gate length and a higher packing density, while the vertical scaling is necessary to maintain the MOSFET's functionality in view of lateral scaling. When bulk CMOS technology evolves from one generation to the next, the channel doping concentration is increased and the gate oxide thickness is reduced to mitigate subthreshold leakage currents and manage SCEs [5, 69, 70]. This process of scaling CMOS has worked well for over the last couple of decades. For gate lengths below *90 nm*, the gate oxide thickness is estimated to be less than *2 nm* [2]. In a MOSFET with such ultra-thin oxides, the direct tunneling current is expected to contribute significantly to the leakage current [48].

In this chapter, an analytical model of gate direct tunneling is developed, for both electrons tunneling in the conduction band and the holes tunneling in the valence band. Section 2.2 introduces the basic theory of tunneling. In Section 2.3, the gate tunneling

model in an MOS structure for both n-MOSFETs and p-MOSFETs is derived. Section 2.4 presents the detail tunneling profile in a MOSFET, specifying the tunneling in the channel region and the source/drain region. In Section 2.5, the tunneling currents of metal gate and polycrystalline silicon gates are compared. Section 2.6 examines the tunneling current in a MOSFET with a high-permittivity gate dielectric expected to be used in future technology generations. Section 2.7 investigates junction tunneling such as the band-to-band tunneling and the gate-induced drain leakage in MOSFETs. The conclusion is given in Section 2.8.

2.2 Tunneling Theory

The quantum-mechanical concept states that all matter, including electrons, behaves like both particles and waves [8]. Particles can be described by a wavefunction $\psi(x,y,z)$, such that the probability of finding a particle in the volume $dxdydz$ is equal to $|\psi(x,y,z)|^2 dxdydz$. The electron wave function satisfies the Schrödinger equation,

$$-\frac{\hbar^2}{2m}\nabla^2\psi(x,y,z)+U(x,y,z)\psi(x,y,z)=E\psi(x,y,z), \qquad (2.1)$$

where $\hbar = h/2\pi$ is the reduced Planck constant, m is the electron mass, U is the potential energy, and E is the energy of the particle. The quantitative analysis of tunneling should be based on the solution of the Schrödinger equation. Figure 2.1 shows the simplest case where an electron of energy E is in the space with the rectangular potential barrier, where the barrier height is E_b (in [eV]) and the width is d (in [cm]). In the x direction, the Schrödinger equation can be written in different regions as follows:

$$d^2\psi/dx^2 + k_1{}^2\psi = 0 \qquad \text{for } x<0 \qquad (2.2)$$

$$d^2\psi / dx^2 + k_{II}{}^2\psi = 0 \qquad \text{for} \quad 0 \le x \le d \qquad (2.3)$$

$$d^2\psi / dx^2 + k_{III}{}^2\psi = 0 \qquad \text{for} \quad x \ge d \qquad (2.4)$$

where

$$k_I{}^2 = k_{III}{}^2 = \frac{2mE}{\hbar^2},$$

$$k_{II}{}^2 = n_b^2 k_I{}^2,$$

and

$$n_b^2 = \frac{(E - E_b)}{d}.$$

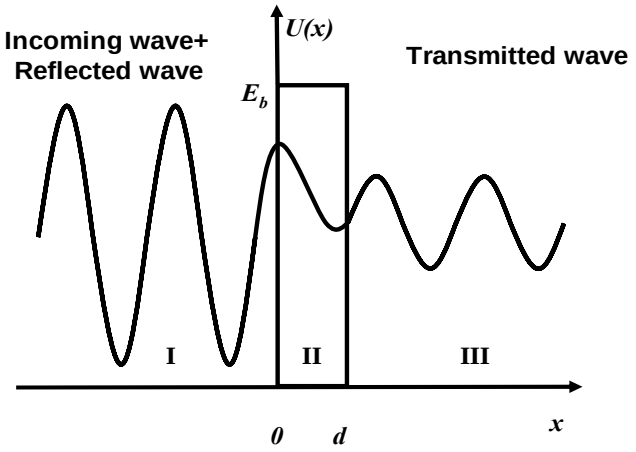

Figure 2.1
Particle tunneling through a rectangular potential barrier of height E_b and thickness d.

From the solution of the Schrödinger equation (see Appendix B), the tunneling probability is given by

$$D(E) = \exp\left\{\frac{-2\left[2m(E_b - E)\right]^{1/2} d}{\hbar}\right\}. \tag{2.5}$$

This formula shows that greater values of barrier height and width help to prevent tunneling, and particles with higher kinetic energy are more likely to tunnel through the barrier.

2.3 Tunneling in MOSFET

According to the tunneling theory, the width of the potential barrier is an important parameter determining the magnitude of the tunneling probability. In MOSFETs, the gate dielectric plays the role of a potential barrier separating the carriers in the channel from the gate. With a thick gate dielectric layer, the wavefunction of carriers cannot extend to the gate by penetrating the potential barrier. However, in MOSFET scaling the thickness of the gate dielectric must be decreased along with the channel length to enhance the gate control, avoiding short-channel effects and transconductance degradation. As devices continue to scale down, tunneling through the thin oxide has become a limiting factor. For the conventional bulk MOSFET using SiO_2 as the gate dielectric, with oxides thinner than *2 nm*, massive numbers of carriers can tunnel through and form a significant gate current. While the gate leakage current may be negligible compared to the drain current of a device, it will substantially increase the chip standby current [2, 6]. Moreover, gate tunneling occurs not only in the channel region, but also in the regions where gate overlaps with the source/drain regions. When the MOSFET is in the off-state, there is still considerable tunneling current along the leakage

path between the biased drain and gate electrodes. Such a leakage greatly degrades the performance of CMOS circuits [13, 15].

Figure 2.2 shows the three gate tunneling components as electron conduction band (ECB) tunneling, electron valence band (EVB) tunneling, and hole valence band (HVB) tunneling [14, 15]. ECB tunneling in the n-MOSFET and HVB tunneling in the p-MOSFET are the dominant tunneling sources, because electrons and holes are majority carriers in the inverted n-MOSFET and p-MOSFET channels, respectively [14]. EVB tunneling takes place only when the valence band transmitting the electrons overlaps with the conduction band receiving electrons, as indicated by Figure 2.2. The overlap requires a high gate voltage, exceeding the normal operating voltage of digital CMOS circuits [14, 15, 17], which makes EVB tunneling negligible for circuit simulation. Based on this reason, EVB tunneling is not included in the work presented in this thesis. For simplicity, ECB and HVB tunneling are simply referred to as electron and hole tunneling, respectively.

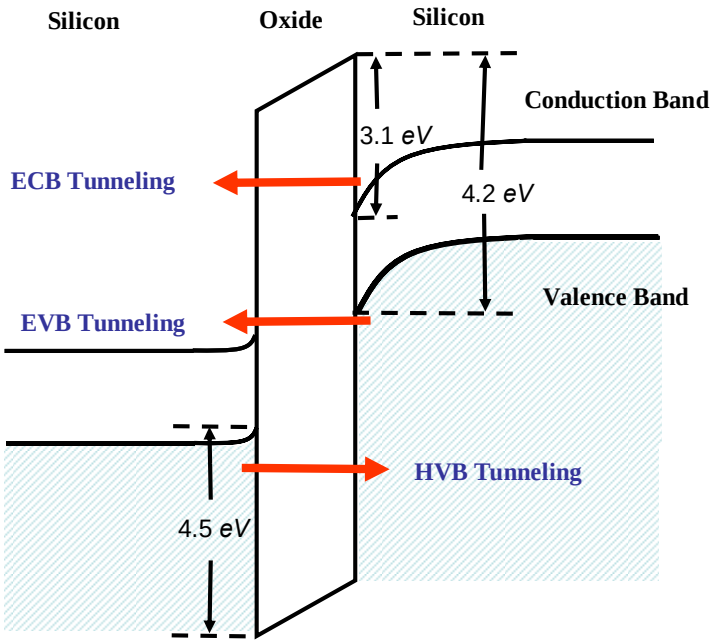

Figure 2.2
Energy band diagram for tunneling components in an MOS structure.

2.3.1 Electron Tunneling in MOS Structure

In this section, electron tunneling is analyzed from the channel to the gate in a metal-gate n-MOSFET. Tunneling in a polycrystalline silicon gate is discussed in section 2.5. The schematic potential diagram for electron tunneling is shown in Figure 2.3. The electron affinity, which is the energy difference between the conduction band edge (E_C) and the vacuum level, is smaller for the oxide layer than for silicon and metal [18]. Therefore, the oxide layer is a potential barrier for electrons in the gate and channel regions. The potential barrier height at the Si/SiO_2 interface is denoted by χ ($\chi = 3.1\ V$) [16, 17], so that $q\chi$ is the required energy to excite an electron from the silicon conduction band edge to the silicon oxide conduction band edge. Because of the

voltage drop across the gate oxide V_{ox} (in [V]), the potential barrier shape is trapezoidal. Therefore, as shown in Figure 2.4, V_{ox} is given by

$$V_{ox} = V_{gs} - V_{FB} - \phi_s , \tag{2.6}$$

where V_{gs} (in [V]) is the gate voltage referenced to the source, V_{FB} (in [V]) is the flat band voltage, and ϕ_S (in [V]) is the surface potential of the channel. In addition, the gate bias V_{gs} causes a difference in the Fermi energy of the silicon channel E_{Fs} (in [eV]) and the Fermi energy of the metal E_{Fm} (in [eV]), which is determined by

$$qV_{gs} = E_{Fs} - E_{Fm} . \tag{2.7}$$

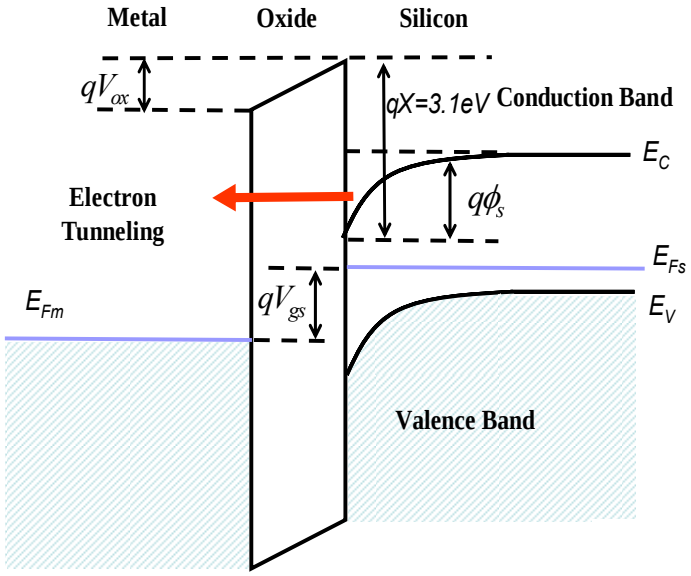

Figure 2.3
Energy band diagram for electron tunneling in an MOS structure.

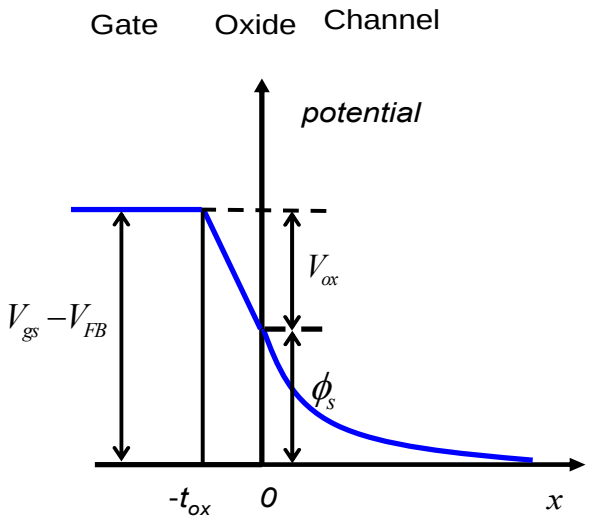

Figure 2.4
Potential distribution in the gate-to-channel direction for a metal-gate MOSFET.

Through the WKB method (see Appendix A), the tunneling probability associated with the trapezoidal potential barrier can be obtained from the integral

$$D(E_x) = \exp\left(-\frac{2}{\hbar}\int_{x_1}^{x_2} \sqrt{2m_{ox}\left(E_b(x) - E_x\right)}dx \right) \tag{2.8}$$

where $m_{ox} = 0.35m_0$ [71] is the effective electron mass in the oxide, and E_x is the electron kinetic energy in the x direction. As shown in Figure 2.3 and Figure 2.4, the turning points x_1 and x_2 are $x_1 = -t_{ox}$ and $x_2 = 0$, and the trapezoidal potential barrier is given by

$$E_b(x) = q\chi + qV_{ox}\cdot\frac{x}{t_{ox}} \tag{2.9}$$

The average barrier height E_B (in [eV]) over x can be expressed by

$$E_B = \frac{1}{t_{ox}} \int_{x_1}^{x_2} E_b(x)dx = q\chi - \frac{1}{2} qV_{ox} . \tag{2.10}$$

Thus, the integral in equation (2.8) can be approximated by

$$\int_{x_1}^{x_2} \sqrt{2m\left(E_b(x) - E_x\right)}dx \approx \int_{-t_{ox}}^{0} \sqrt{2m\left(E_B - E_x\right)}dx . \tag{2.11}$$

Substituting equation (2.11) into equation (2.8) results in

$$D(E_x) = \exp\left\{-\gamma (E_B - E_x)^{1/2}\right\}, \tag{2.12}$$

where

$$\gamma = \frac{4\pi t_{ox}\sqrt{2m_{ox}}}{h} . \tag{2.13}$$

Unlike the single-particle-tunneling case, there are a number of electrons distributed on both sides of the potential barrier in a MOSFET. In such a multiple-electron system, the complexity of the tunneling problem increases because electrons have various energies varying from 0 to ∞. The electrons' contribution to tunneling must be statistically accounted. This multiple-particle system is characterized by its degrees of freedom, which is the number of parameters completely describing the movement of each particle. These degrees of freedom for a system determine the density of available states for particles to fill in. From the basic law of statistical physics, the density of states is $\frac{m}{h}$ for every two degrees of freedom in the space characterized by the particle position (x, y, z) and velocity (v_x, v_y, v_z). For example, there are six degrees of freedom for a three-dimensional electron gas, namely, x, y, z, v_x, v_y, and v_z, implying $\frac{m^3}{h^3} dxdydzdv_x dv_y dv_z$ states in a volume $dxdydzdv_x dv_y dv_z$. For a two dimensional system

confined in the *x*-*y* plane, there are $\frac{m^2}{h^2} dxdydv_x dv_y$ states in a volume $dxdydv_x dv_y$. These states are available for particles to occupy; however, not every state is occupied by particles. For electrons, the states are filled as described by the Fermi-Dirac distribution,

$$f(E) = \frac{1}{1 + \exp(\frac{E - E_F}{kT})}, \tag{2.14}$$

where the distribution function $f(E)$ is the probability that an available state at the energy E will be occupied by an electron, k is the Boltzmann constant, T is the absolute temperature, and E_F is the Fermi energy of the electrons.

The tunneling current density J_{sm} (in [A/cm^2]) is formed by the tunneling electrons moving from the channel to the gate (in *x* direction). Similar to the flow of electrons in a metal and the flow of carriers in a semiconductor, the tunneling current density is written as

$$J_{sm} = q\overline{v}_x n, \tag{2.15}$$

where q is the electron charge, $\overline{v}_x$ (in [*cm/s*]) is the average electron tunneling velocity in *x* direction, and *n* is the tunneling electron density (number of electrons tunneling through the oxide per unit volume, measured in [cm^{-3}]). Noticing that the tunneling probability and the electron distribution in energy states vary with the electron's velocity (or kinetic energy) as in equations (2.14) and (2.19), the contribution of electrons at different velocities to the tunneling current must be determined statistically. In the velocity range from v_x to v_x+dv_x, the total number of electrons per unit volume in the channel can be written as

$$n(v_x) = \frac{2m_e^{*3}}{h^3} \iint f_s(E)\, dv_y dv_z \,, \tag{2.16}$$

where $n(v_x)$ (in $[cm^{-2}/s]$) is the volume of density of electrons associating with the v_x, $2m_e^{*3}/h^3$ is the state density, the effective electron mass $m_e^* = 0.19m_0$ is used to account for the silicon crystal structure's influence on electron state density [10], the factor *2* accounts for electron spin, and $f_s(E)$ is the Fermi-Dirac distribution function in the silicon channel given by

$$f_s(E) = \frac{1}{1+\exp(\frac{E-E_{Fs}}{kT})}, \tag{2.17}$$

in which E_{Fs} is the Fermi energy of the channel. Electrons in the channel can undergo a tunneling process with the prerequisite that there exists a vacancy state associated with exactly the same energy at the gate side. For the distribution function $f_m(E)$ in the metal gate,

$$f_m(E) = \frac{1}{1+\exp(\frac{E-E_{Fm}}{kT})}, \tag{2.18}$$

where E_{Fm} is the Fermi energy in the metal gate. The probability to find an empty energy state is $1-f_m(E)$. We can express the overall probability for an electron with energy E to tunnel from channel to gate as

$$P(E) = D(E_x)\left[1-f_m(E)\right], \tag{2.19}$$

where $D(E_x)$ is the single electron tunneling probability given in equation (2.12). Therefore, the volume density of tunneling electrons in the velocity range v_x to v_x+dv_x is given by $n(v_x)P(E)\, dv_x$. The total volume density of the tunneling electrons n is

$$n = \int_0^{\infty} n(v_x) P(E) dv_x \,, \tag{2.20}$$

and the average tunneling velocity is

$$\overline{v}_x = \frac{\int_0^{\infty} v_x n(v_x) P(E) dv_x}{\int_0^{\infty} n(v_x) P(E) dv_x} \,. \tag{2.21}$$

By substituting expressions of n and v_x into equation (2.15), J_{sm} is then given by

$$J_{sm} = \int_0^{\infty} q v_x n(v_x) P(E) dv_x \,, \tag{2.22}$$

which becomes,

$$J_{sm} = \int_0^{\infty} \frac{2qm_e^{*3}}{h^3} v_x D(E_x) \iint_{\infty} f_s(E)\left(1 - f_m\left(E\right)\right) dv_y dv_z dv_x \,, \tag{2.23}$$

which accounts for the tunneling current density from channel to gate. Similarly, the tunneling current density J_{ms} (in $[A/cm^2]$) from the gate to the channel is obtained as

$$J_{ms} = \int_0^{\infty} \frac{2qm_e^{*3}}{h^3} v_x D(E_x) \iint_{\infty} f_m(E)\left(1 - f_s\left(E\right)\right) dv_y dv_z dv_x \,. \tag{2.24}$$

The net tunneling current density J_T (in $[A/cm^2]$) includes both components as

$$J_T = J_{sm} - J_{ms} \,, \tag{2.25}$$

or

$$J_T = \frac{2qm_e^{*3}}{h^3} \int_0^{\infty} v_x D(E_x) \iint_{\infty} \left(f_s\left(E\right) - f_m\left(E\right)\right) dv_y dv_z dv_x \,. \tag{2.26}$$

Tunneling direction is determined by the larger one of the two components. The Fermi-Dirac distribution function in equation (2.14) can be approximated by the Maxwell-Boltzmann distribution function on the condition that $(E - E_F) >> kT$. Thus,

$$f(E) = \exp(-\frac{E - E_F}{kT}), \tag{2.27}$$

and the total tunneling current density is written as

$$J_T = \int_0^\infty \frac{2qm_e^{*3}}{h^3}\left[\exp(\frac{E_{Fs}}{kT}) - \exp(\frac{E_{Fm}}{kT})\right] v_x D(E_x) \iint_\infty \exp(-\frac{E}{kT}) dv_y dv_z dv_x . \tag{2.28}$$

Noticing that total kinetic energy has three components corresponding to the *x*, *y*, *z* directions,

$$E = E_x + E_y + E_z , \tag{2.29}$$

and $E_x = \frac{1}{2}mv_x^2$, $E_y = \frac{1}{2}mv_y^2$, and $E_z = \frac{1}{2}mv_z^2$, the double integral in equation (2.28) can be written as

$$\iint_\infty \exp(-\frac{E}{kT}) dv_y dv_z = \iint_\infty \exp(-\frac{\frac{1}{2}mv_x^2 + \frac{1}{2}mv_y^2 + \frac{1}{2}mv_z^2}{kT}) dv_y dv_z . \tag{2.30}$$

Equation (2.30) can be separated as two integrals with respect to v_x and v_y as

$$\iint_\infty \exp(-\frac{E}{kT}) dv_y dv_z = \exp\left(-\frac{E_x}{kT}\right) \int_0^\infty \exp\left(-\frac{1}{2}mv_y^2 / kT\right) dv_y \cdot \int_0^\infty \exp\left(-\frac{1}{2}mv_z^2 / kT\right) dv_z . \tag{2.31}$$

From $\int_0^\infty \exp\left(-x^2\right) dx = \sqrt{\pi}$, we have

$$\iint_\infty \exp(-\frac{E}{kT}) dv_y dv_z = \frac{2\pi kT}{m} \exp(-\frac{E_x}{kT}) . \tag{2.32}$$

Therefore, the total tunneling current density is given by

$$J_T = \frac{4\pi m_e^{*2} qkTv_x}{h^3}\left[\exp(\frac{E_{Fs}}{kT}) - \exp(\frac{E_{Fm}}{kT})\right] \int_0^\infty \exp(\frac{-E_x}{kT}) D\left(E_x\right) dv_x . \tag{2.33}$$

Since $E_x = \frac{1}{2}mv_x^2$ and $D(E_x) = \exp\left\{-\gamma(E_B - E_x)^{1/2}\right\}$, we have

$$J_T = \frac{4\pi m_e^* qkT}{h^3}\left[\exp(\frac{E_{Fs}}{kT}) - \exp(\frac{E_{Fm}}{kT})\right]\int_0^\infty \exp(\frac{-E_x}{kT})\exp\left\{-\gamma(E_B - E_x)^{1/2}\right\} dE_x \ . \quad (2.34)$$

The integral over E_x in equation (2.34) cannot be solved directly, however, it can be simplified. The term $\exp\left\{-\gamma(E_B - E_x)^{1/2}\right\}$ increases with E_x, implying that electrons with higher energy are more likely to tunnel. However, the density of electrons exponentially decreases with increasing of E_x, which is indicated by $\exp(\frac{-E_x}{kT})$. Typical magnitudes of these two terms are shown in Figure 2.5. We find that electrons with low E_x dominate electrons with high E_x in tunneling, because of the fast drop rate of $\exp(\frac{-E_x}{kT})\exp\left\{-\gamma(E_B - E_x)^{1/2}\right\}$. From a physical view, electron density decreases with E_x exponentially, while tunneling probability increases with E_x moderately. Thus, it is safe to consider the tunneling contribution of electrons concentrating near E_C only, where E_x is close to zero.

Therefore, the term $\exp\left\{-\gamma(E_B - E_x)^{1/2}\right\}$ can be expanded around $E_x = 0$ by the Taylor series as

$$\exp\left\{-\gamma(E_B - E_x)^{1/2}\right\} \approx \left(1 + \frac{1}{2}\frac{\gamma}{\sqrt{E_B}}E_x\right)\exp\left(-\gamma\sqrt{E_B}\right). \quad (2.35)$$

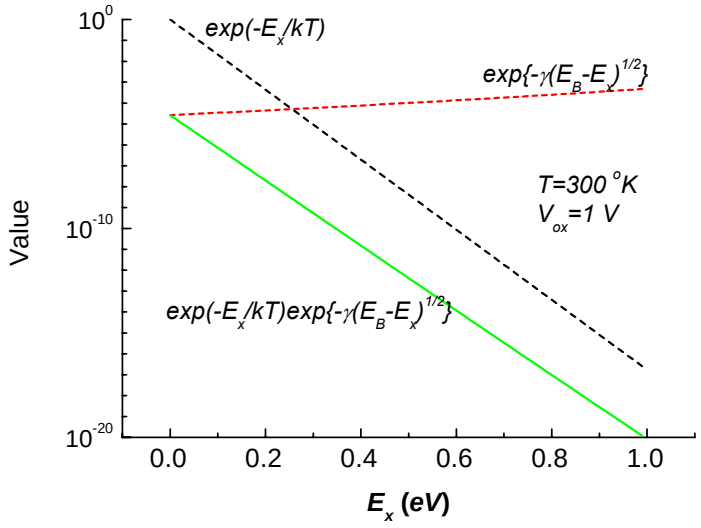

Figure 2.5
Comparison of two terms in the tunneling integral of equation (2.34).

Substituting the above equation in (2.34) gives

$$\int_0^{\infty} \exp(\frac{-E_x}{kT}) \exp\left\{-\gamma(E_B - E_x)^{1/2}\right\} dE_x \approx \exp\left(-\gamma\sqrt{E_B}\right) \int_0^{\infty} \exp(\frac{-E_x}{kT}) \left(1 + \frac{\gamma}{2\sqrt{E_B}} E_x\right) dE_x \quad (2.36)$$

The integral above breaks into the summation of two integrals as

$$\int_0^{\infty} \exp(\frac{-E_x}{kT}) \left(1 + \frac{\gamma}{2\sqrt{E_B}} E_x\right) dE_x = \int_0^{\infty} \exp(\frac{-E_x}{kT}) dE_x + \frac{\gamma}{2\sqrt{E_B}} \int_0^{\infty} \exp(\frac{-E_x}{kT}) E_x dE_x \,. \quad (2.37)$$

These integrals can be solved separately as

$$\int_0^{\infty} \exp(\frac{-E_x}{kT}) dE_x = kT \,, \quad (2.38),$$

and

$$\int_0^{\infty} \exp(\frac{-E_x}{kT}) E_x dE_x = -kT \exp\left(-\frac{E_x}{kT}\right) E_x \Big|_0^{\infty} - \int_0^{\infty} (-kT) \exp(\frac{-E_x}{kT}) dE_x = (kT)^2 \,. \quad (2.39)$$

Hence, equation (2.36) becomes

$$\int_0^{\infty} \exp(\frac{-E_x}{kT}) \exp\left\{-\gamma(E_B - E_x)^{1/2}\right\} dE_x \approx kT\left(1+\frac{\gamma kT}{2\sqrt{E_B}}\right)\exp\left(-\gamma\sqrt{E_B}\right). \qquad (2.40)$$

Therefore, the tunneling current density can be expressed as

$$J_T = \frac{4\pi m_e^* q}{h^3}\left(kT\right)^2\left[\exp(\frac{E_{Fs}}{kT}) - \exp(\frac{E_{Fm}}{kT})\right]\left(1+\frac{\gamma kT}{2\sqrt{E_B}}\right)\exp\left(-\gamma\sqrt{E_B}\right). \qquad (2.41)$$

To turn on an n-MOSFET, a positive voltage V_{gs} is applied on the gate, leading to shifts of the Fermi energy level in the gate E_{Fm} to values lower than the Fermi energy E_{Fs} of the channel by qV_{gs},

$$E_{Fs} - E_{Fm} = qV_{gs}. \qquad (2.42)$$

Noticing that E_{Fs} is referenced to the E_C at the surface silicon channel, its expression

$$E_{Fs} = q\phi_s - q\phi_B - E_g/2 \qquad (2.43)$$

can be derived from Figure 2.3, where $q\phi_B$ is the energy difference from Fermi level to the middle band gap at flat-band condition and $\phi_B = \frac{kT}{q}\ln\left(\frac{N_A}{n_i}\right)$ [10], E_g (in $[eV]$) is the energy band gap, N_A (in $[cm^{-3}]$) is the channel doping concentration, and n_i (in $[cm^{-3}]$) is the intrinsic electron density. Thus,

$$\exp(\frac{E_{Fs}}{kT}) - \exp(\frac{E_{Fm}}{kT}) = \left[1-\exp(-\frac{qV_{gs}}{kT})\right]\exp(\frac{q\phi_s - q\phi_B - E_g/2}{kT}). \qquad (2.44)$$

Then the electron tunneling density in the n-MOSFET can be expressed as an explicit function of the gate voltage as

$$J_T = \frac{4\pi m_e^* q}{h^3}(kT)^2(1+\frac{\gamma kT}{2\sqrt{E_B}})\left[1-\exp(-\frac{qV_{gs}}{kT})\right], \\ \cdot\exp(\frac{q\phi_s - q\phi_B - E_g/2}{kT})\exp(-\gamma\sqrt{E_B}) \tag{2.45}$$

with

$$\gamma = \frac{4\pi t_{ox}\sqrt{2m_{ox}}}{h}, \tag{2.46}$$

and

$$E_B = q\chi - \frac{1}{2}qV_{ox}. \tag{2.47}$$

Figure 2.6 shows that results from this model agree well with data from measurements and numeric simulations [4-6]. The gate tunneling current increases drastically when the thickness of gate oxide is reduced. As t_{ox} is reduced from *3.6 nm* to *1.5 nm*, the tunneling current density increases in magnitude by the order of 10^9. At $t_{ox} = 1.5\ nm$ and $V_{gs} = 1\ V$, the tunneling current density amounts to 1 A/cm^2, which is considered too much to handle in circuit design [4, 6]. Figure 2.6 also shows that tunneling current density is very sensitive to gate voltage. Gate voltage is given by $V_{gs} = V_{ox} + V_{FB} + \phi_s$ as indicated in Figure 2.4, so that the increase in gate voltage leads to larger V_{ox} and ϕ_s, consequently raising J_T as indicated by equations (2.45) to (2.47). From a physical perspective, larger V_{ox} induces a lower barrier and larger ϕ_s resulting in higher electron density in the channel. Both effects magnify electron tunneling.

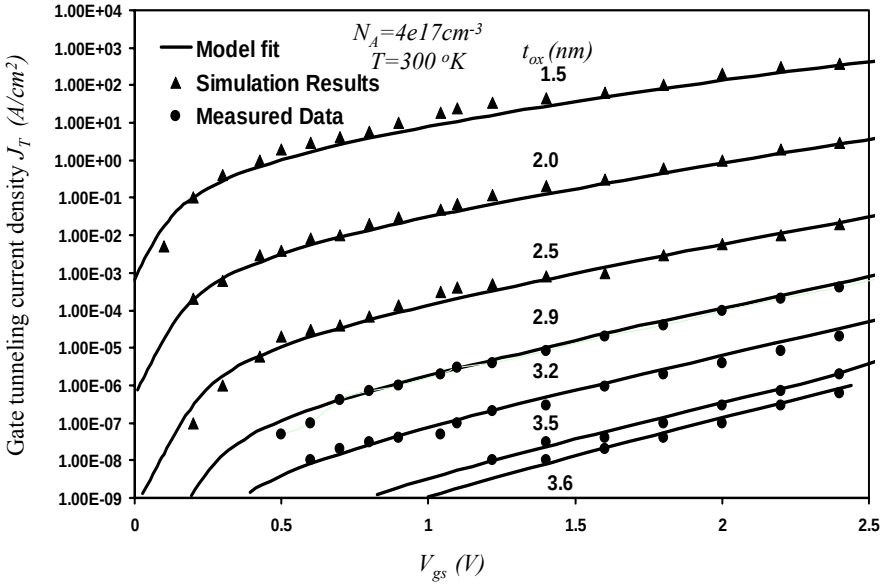

Figure 2.6
Validation of the compact tunneling current model against numeric simulation and measurement [4, 5].

Equation (2.41) can be applied to the electron tunneling in a p-MOSFET:

$$J_T = \frac{4\pi m_e^* q}{h^3}(kT)^2\left[\exp(\frac{E_{Fm}}{kT}) - \exp(\frac{E_{Fs}}{kT})\right]\left(1+\frac{\gamma kT}{2\sqrt{E_B}}\right)\exp\left(-\gamma\sqrt{E_B}\right), \qquad (2.48)$$

where E_{Fm} and E_{Fs} now represent Fermi energy levels in the gate side and the channel of the p-MOSFET, respectively. In p-MOSFETs, gate electrodes are biased to a lower voltage with respect to the source potential ($V_{gs} < 0$), so that the net flux for electron tunneling is from the gate side to the substrate side. When the channel is inverted, the Fermi energy level in the channel E_{Fs} is close to the valence band,

$$E_{Fs} \approx\sim -E_g . \qquad (2.49)$$

Therefore,

$$\exp\left(\frac{E_{Fm}}{kT}\right) - \exp\left(\frac{E_{Fs}}{kT}\right) \approx \left[1-\exp\left(\frac{qV_{gs}}{kT}\right)\right]\exp\left(-\frac{E_g}{kT}\right). \qquad (2.50)$$

Thus, the electron tunneling in the p-MOSFET can be written as

$$J_T = \frac{4\pi m_e^* q}{h^3}(kT)^2\left[1-\exp\left(\frac{qV_{gs}}{kT}\right)\right]\exp\left(-\frac{E_g}{kT}\right)\left(1+\frac{\gamma kT}{2\sqrt{E_B}}\right)\exp\left(-\gamma\sqrt{E_B}\right). \quad (2.51)$$

At room temperature, $300\ ^\circ K$, the band gap $E_g = 1.1\ eV$ which is almost $40kT$. With the additional term $\exp\left(-\frac{E_g}{kT}\right)$ in equation (2.51) as compared to equation (2.45), the magnitude of electron tunneling current density in a p-MOSFET is $10^8 \sim 10^{15}$ times smaller than it is in an n-MOSFET. Hence, in CMOS circuits consisting of equal numbers of n-MOSFETs and p-MOSFETs, the electron tunneling in the p-MOSFETs is virtually negligible.

2.3.2 Hole Tunneling in MOS Structure

The mechanism of the electron tunneling process in the conduction band is the same as that for holes in the valence band, as indicated by Figure 2.3. A most significant difference between electron tunneling and hole tunneling lies in the average barrier height. Holes face a barrier from the valence band edge of silicon to the valence band edge of gate oxide, which is *4.5 eV* high, as shown in Figure 2.2. In contrast, this value for electrons is only *3.1 eV*. Beside the barrier height, other parameters that differ are listed in Table 2.1. In spite of the different properties of the two kinds of carriers, the analysis for electron tunneling can be applied to hole tunneling. It leads to a similar equation for hole tunneling current as was derived for electron tunneling in equation (2.45). Thus, the hole tunneling current can be expressed as

$$J_T = \frac{4\pi m_h^* q}{h^3}(kT)^2(1+\frac{\gamma kT}{2\sqrt{E_B}})\left[1-\exp(\frac{V_{gs}}{kT})\right] \cdot \exp(\frac{q\phi_s - q\phi_B - E_g/2}{kT})\exp(-\gamma\sqrt{E_B}) \quad (2.52)$$

Table 2.1
Parameters for electron tunneling and hole tunneling.

	Effective mass in silicon	Effect mass in oxide	Barrier height
Electron tunneling	$m_e^* = 0.19m_0$	$m_I = 0.32m_0$	$E_B = 3.1eV - qV_{ox}$
Hole tunneling	$m_h^* = 0.45m_0$	$m_I = 0.30m_0$	$E_B = 4.5eV - qV_{ox}$

Results from the hole tunneling model shown in Figure 2.7 are validated by measurements [15] and an empirical model [14, 15, 17]. It can be shown that the current density of hole tunneling is typically smaller than that of the electron tunneling by an order of magnitude. The lower hole tunneling density is primarily due to the higher potential barrier for holes.

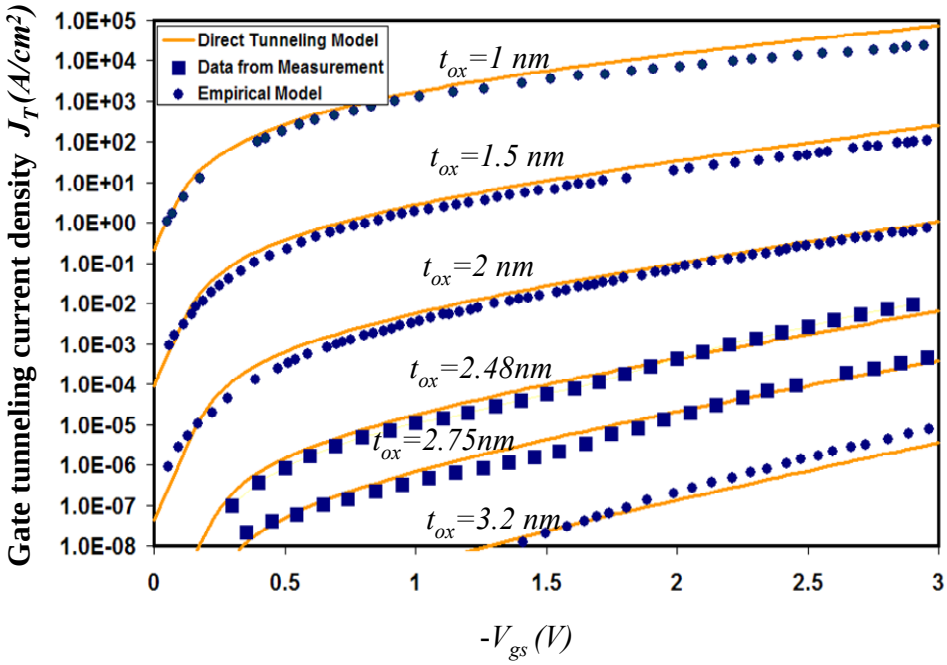

Figure 2.7
Tunneling in the p-MOSFET (substrate doping $N_D = 4e17\ cm^{-3}$, $T = 300\ ^{\circ}K$). The direct tunneling model is compared with an empirical model [15, 17] and measurement [15].

2.4 Tunneling in Different Regions

In a typical deep-submicron MOSFET, the heavily doped shallow drain extends underneath the gate. The overlap region between the gate and drain can be a path for tunneling current as shown in Figure 1.2. In short-channel devices, the length of the source and drain extension (SDE) area is comparable to the channel length [2], resulting in considerable leakage. Moreover, as will be shown, SDE tunneling causes leakage in the off-state of a MOSFET, under which condition the gate tunneling to the channel

region does not exist. Therefore, SDE tunneling is an important leakage source and must be considered in circuit design [13].

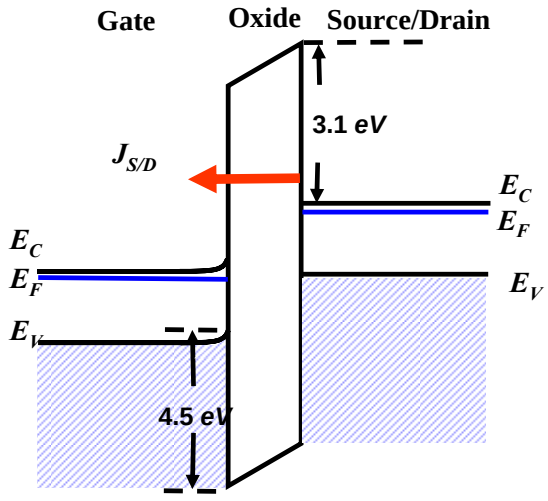

Figure 2.8
Energy band diagram from the gate to the source/drain of an n-MOSFET.

The band diagram for the gate-to-SDE tunneling is shown in Figure 2.8. Noticing that the gate and source/drain are both heavily doped by the same type of dopant, Fermi levels at both ends are approximately at the bottom of the conduction band in an n-MOSFET. Moreover, the voltage drop inside the source/drain is negligible because of heavy doping. As shown in Figure 2.9, voltage drop on the gate oxide is equal to the gate-to-source voltage,

$$V_{ox} = V_{gs} \tag{2.53}$$

The tunneling model for the MOS structure can be applied to SDE tunneling, by treating the source/drain as a highly doped channel. Denoting the Fermi energy levels in gate and

source/drain by E_{F1} and E_{F2} (both referenced to E_C at source/drain), equation (2.50) can be modified as

$$\exp(\frac{E_{F1}}{kT}) - \exp(\frac{E_{F2}}{kT}) \approx \left[1 - \exp(\frac{qV_{gs}}{kT})\right]. \tag{2.54}$$

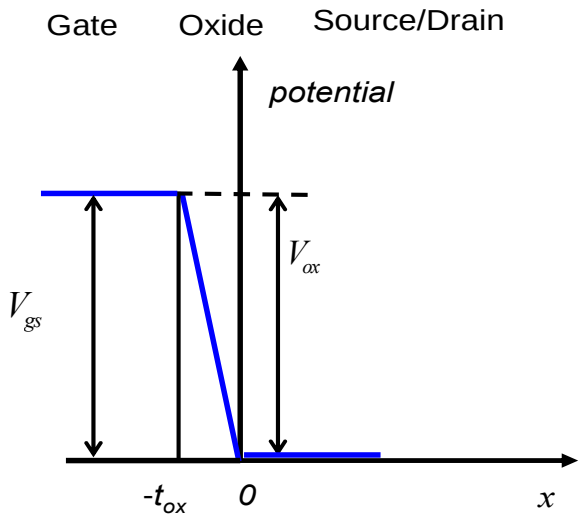

Figure 2.9
Potential distribution from gate to source/drain.

For a p-MOSFET, the source/drain is p-type doped and hole tunneling is dominant. SDE tunneling current density, denoted by $J_{S/D}$, can be modified from the tunneling model in equation (2.41) as

$$J_{S/D} = \frac{4\pi m_h^* q}{h^3}(kT)^2(1+\frac{\gamma kT}{2\sqrt{E_B}})\left[1-\exp(-\frac{q|V_{gs}|}{kT})\right]\exp(-\gamma\sqrt{E_B}). \tag{2.55}$$

For an n-MOSFET, electron tunneling is dominant, and SDE tunneling current density is given by

$$J_{S/D} = \frac{4\pi m_e^* q}{h^3}(kT)^2(1+\frac{\gamma kT}{2\sqrt{E_B}})\left[1-\exp(-\frac{q|V_{gs}|}{kT})\right]\exp(-\gamma\sqrt{E_B}). \qquad (2.56)$$

In equation (2.55) and (2.56), E_B and γ are defined by parameters for holes and electrons according to Table 2.1.

The magnitude of the tunneling current density in the channel region $J_{Channel}$, which is given by equations (2.45) and (2.52) for n-MOSFETs and p-MOSFETs respectively, are compared with the magnitudes of SDE tunneling in Figure 2.10. Because electrons tunnel through the gate oxide more easily than holes, in an n-MOSFET, the electron tunneling current density in SDE region is comparable to hole tunneling current density in the channel of a p-MOSFET.

The total tunneling current in the MOSFET can be expressed in terms of the tunneling currents of the channel region and the SDE regions. As a MOSFET is turned off, which refers to either the low voltage on the gate for an n-MOSFET or the high voltage on the gate of a p-MOSFET, the only tunneling path is between the drain and the gate. The off-state tunneling current $I_{tunnel,off}$ is given by

$$I_{tunnel,off} = J_{S/D} \cdot L_{overlap} \cdot W, \qquad (2.57)$$

where $L_{overlap}$ is the length of overlap region, and W is the width of the MOSFET.

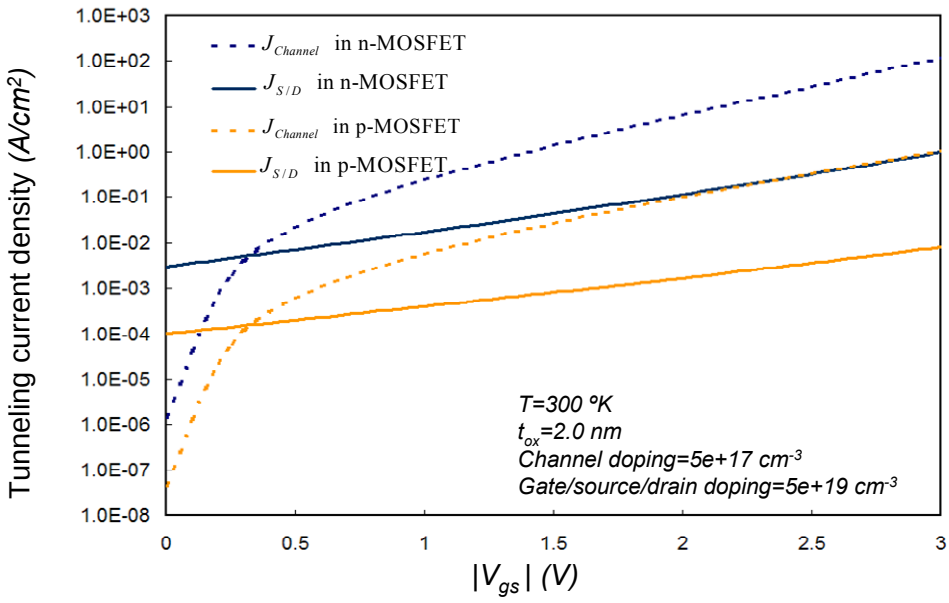

Figure 2.10
Comparison of tunneling density in channel and source/drain region.

If the device is turned on, the electrons can tunnel through the gate oxide layer to the channel as well as to the source and the drain. The overall on-state tunneling current $I_{tunnel,on}$ is given by accounting for both components as

$$I_{tunnel,on} = J_{Channel} \cdot L \cdot W + 2 \cdot J_{S/D} \cdot L_{overlap} \cdot W \,. \tag{2.58}$$

The different tunneling current components in a CMOS inverter with input "0" and "1" are shown in Figure 2.11 and Figure 2.12, respectively. The input "0" represents a low voltage in CMOS circuits, which turns the p-MOSFET on and the n-MOSFET off. Correspondingly, the input "1" represents a high voltage, which turns the p-MOSFET off and the n-MOSFET on. The tunneling current in CMOS circuits is obtained from equations (2.57) and (2.58), according to the on and off states of the devices.

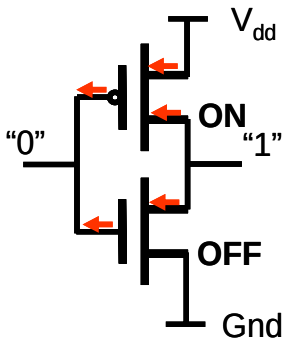

Figure 2.11
Tunneling current in a CMOS inverter with a "0" input.

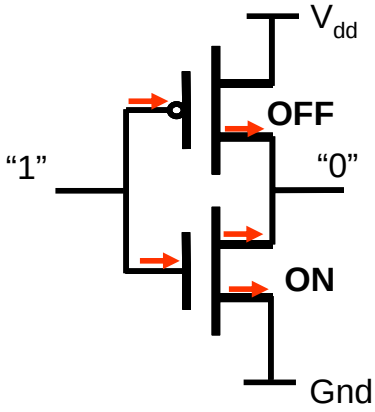

Figure 2.12
Tunneling current in a CMOS inverter with a "1" input.

2.5 Tunneling with Polysilicon Gate

For devices with a polysilicon gate, the voltage drop inside the gate must be considered. The electric field depletes the surface of electrons and forms a thin space-charge region in the polysilicon gate. Illustrated by Figure 2.13, voltage drops inside the gate and bands in the n^+-polysilicon gate bend upward near the interface of the polycrystalline silicon and the oxide. This effect leads to a lower voltage drop over the gate oxide that can be expressed as

$$V_{ox} = V_{gs} - V_{FB} - \phi_S - V_p , \tag{2.59}$$

where V_p is the voltage drop in the polysilicon. This polysilicon potential drop is given by

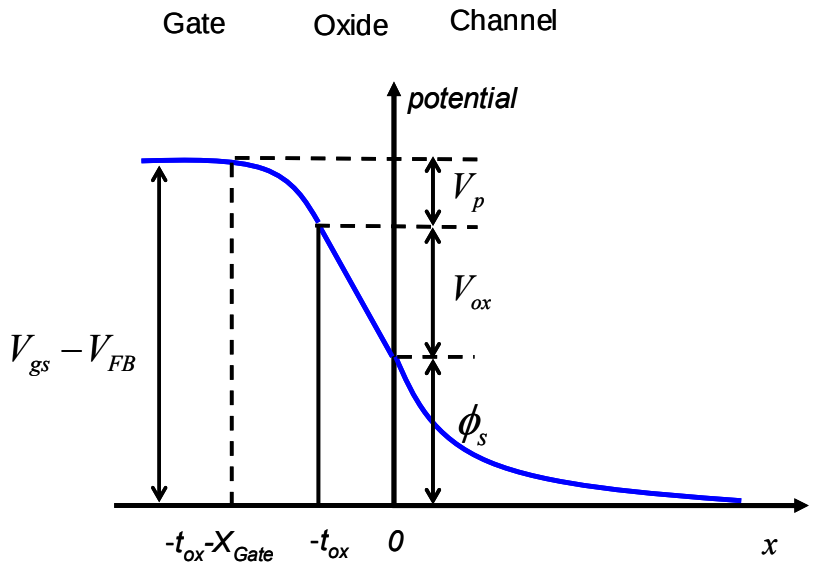

Figure 2.13
Potential diagram for the poly-silicon gate MOSFET.

$$V_p = \frac{1}{2}\frac{q}{\varepsilon_{Si}} N_{Gate} X_{Gate}^{\ 2}, \tag{2.60}$$

where ε_{Si} is the permittivity of silicon, N_{Gate} is the doping concentration of the polysilicon gate, and X_{Gate} is the gate depletion depth. When $V_{gs} > V_T$, X_{Gate} is derived as [71],

$$X_{Gate} = \left(\frac{\varepsilon_{Si}}{\varepsilon_{ox}}\right) t_{ox} \left\{ \sqrt{1 + \frac{2\varepsilon_{ox}^{\ 2}\left(V_{gs} - V_{FB} - \phi_s\right)}{qN_{Gate}\varepsilon_{Si}t_{ox}^{\ 2}}} - 1 \right\}, \tag{2.61}$$

where ε_{ox} is the permittivity of SiO_2. When the poly-depletion effect is accounted for, the potential drop on the gate oxide layer is reduced. For this reason, the tunneling current density in the polysilicon gate is less than it is in its metal counterpart at the same gate voltage level. This is shown in Figure 2.14.

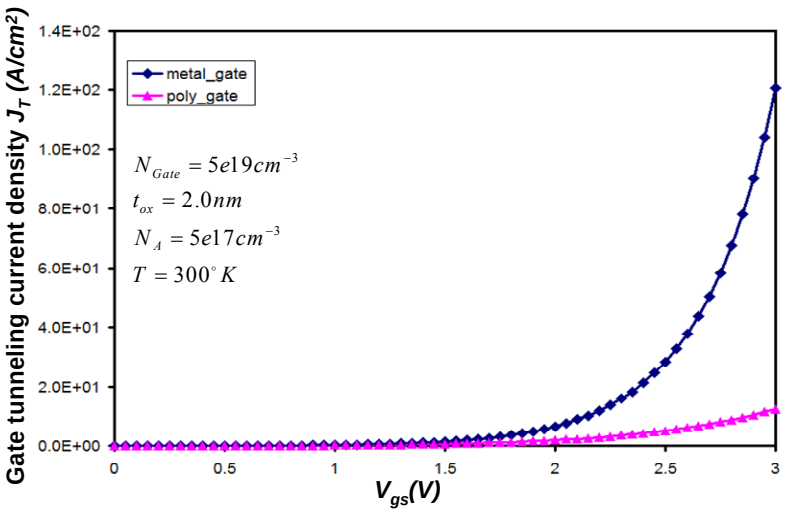

Figure 2.14
Comparison of tunneling current density for aluminum gate and polycrystalline silicon gate.

2.6 High-κ Gate Dielectrics

In order to limit the direct tunneling current and also to avoid reliability concerns [6, 48, 70], the physical thickness of the gate dielectric must be increased. However, the ultra-thin SiO_2 layer is utilized to generate a high electrical field in the vertical direction, which is critical to control SCEs and ensure sufficient drive current in bulk MOSFETs. Hence, the electrical thickness of the gate dielectric must be reduced while the physical thickness should be at least sustained. The only possible solution is replacing the silicon oxide with high dielectric constant (high-κ) materials for gate insulation.

Among various high-κ materials, HfO_2 and $HfSiO_4$ appear to be the most promising candidates for replacing silicon oxide [72, 73]. These high-κ dielectrics exhibit a trend of decreasing barrier height with increasing dielectric constant [72]. The dielectric constant and band gap offset referenced to silicon for high-κ dielectrics are shown in Figure 2.15 and Table 2.2. An equivalent oxide thickness (EOT) can be introduced for the high-κ dielectrics. It defines an equivalent thickness of silicon oxide needed to obtain the same gate capacitance as the one obtained by high-κ dielectrics,

$$EOT = \frac{\varepsilon_{ox}}{\kappa\varepsilon_0} \cdot t_I \tag{2.62}$$

where κ and t_I are the dielectric constant and physical thickness of high-κ dielectrics. Figure 2.16 gives the simulation results showing the reduction of tunneling current density by the $2 \sim 3$ orders of magnitude with the utilization high-κ gate dielectrics.

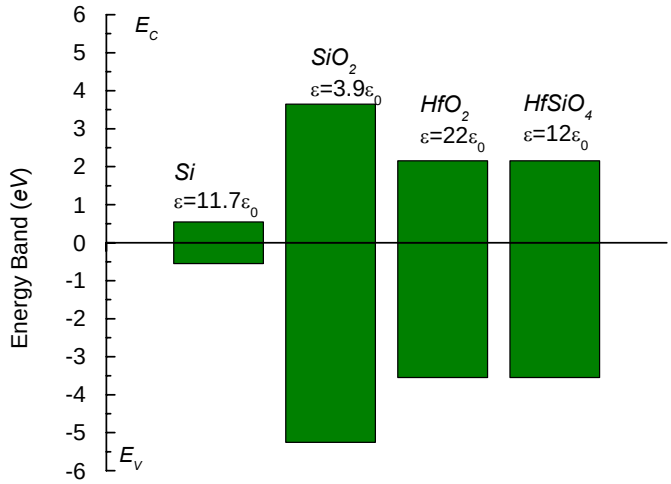

Figure 2.15
Band gaps of high-κ materials and silicon oxide [72].

Table 2.2
Summary of gate dielectric parameters [74].

Material	SiO_2	HfO_2	$HfSiO_4$
Permittivity	$3.9\,\varepsilon_0$	$22\,\varepsilon_0$	$12\,\varepsilon_0$
Conduction band offset (eV)	3.1	1.5	1.5

Although the gate tunneling current density is reduced by high-κ materials, new processing issues and device design concerns come into play. These include fringing-induced barrier lowering (FIBL) [70, 74] and interface defects [75, 76]. FIBL causes the off-state leakage current to increase and degrades the subthreshold

characteristics [77]. The interface defects reduce mobility in the channel and the drive current is decreased [76]. The trade-offs should be considered in device design [74, 78].

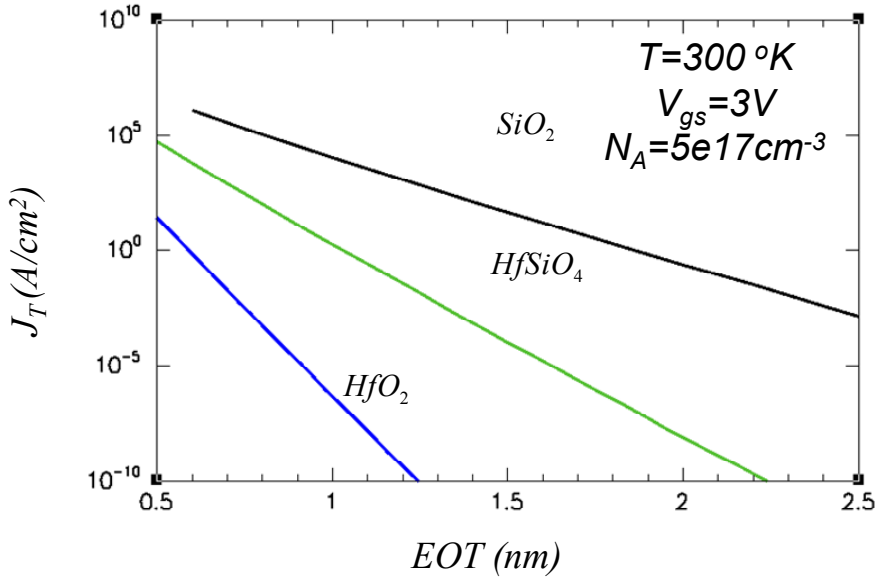

Figure 2.16
Comparison of tunneling currents in different gate insulation materials: SiO_2, $HfSiO_4$, and HfO_2.

2.7 Band-to-Band tunneling (BTBT) in MOSFET

Figure 2.17 shows electron tunneling from the valence band into the conduction band when a reverse bias is applied. Tunneling occurs when the voltage drop across the junction is sufficiently large that

$$V_{app} + \phi_b > \frac{E_g}{q} \tag{2.63}$$

where V_{app} (in $[V]$) is the applied reverse bias, ϕ_b (in $[V]$) is the build-in potential for the p-n junction and E_g is the bandgap. BTBT of an electron through a p-n junction is formally the same as that of a particle tunneling through a triangular barrier, as shown in Figure 2.18. The tunneling density is given by [79]

$$J_{BTBT} = \sqrt{\frac{2m^* q^3}{4\pi^3 \hbar^2}} \frac{E_{field} V_{app}}{E_g^{1/2}} \exp\left(-\frac{4\sqrt{2m^*}}{3q\hbar} \frac{E_g^{3/2}}{E_{field}} \right) \tag{2.64}$$

where E_{field} (in $[V/cm]$) is the electric field at the junction, and m^* is the effective mass of the electron. In a MOSFET, the BTBT usually occurs at the p-n junction formed by the drain and the substrate, inducing the substrate current. Furthermore, BTBT causes the leakage mechanism at the drain, which is referred to as gate induced drain leakage (GIDL).

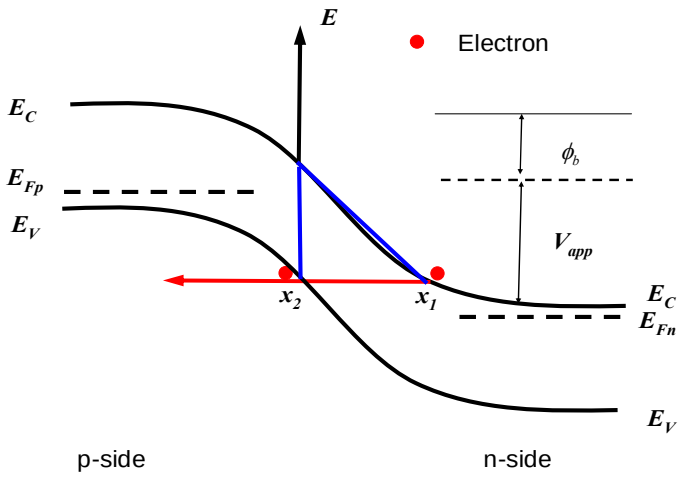

Figure 2.17
Electron tunneling through the p-n junction from conduction band to valence band. E_{Fn} and E_{Fp} are referred to as Fermi energy levels in n-side and p-side semiconductors, respectively.

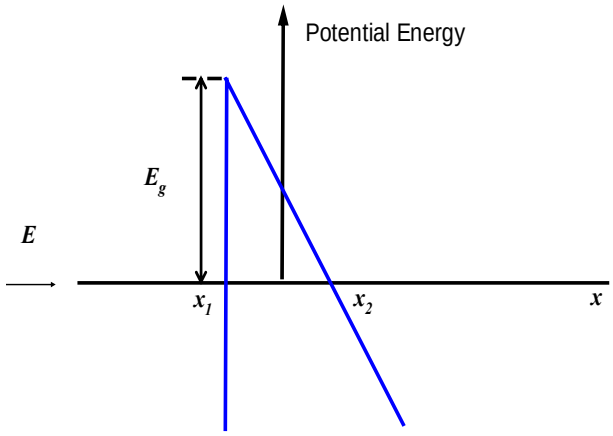

Figure 2.18
Potential energy diagram for BTBT tunneling.

GIDL occurs at the drain region of a MOSFET, where the gate overlaps the drain because of lateral diffusion. With the gate grounded and the drain biased in an n-MOSFET, the silicon surface is depleted, as shown in Figure 2.19 and Figure 2.20. Since the drain is heavily doped n^+, the depletion region is very small and the band bending is confined to a small spatial region. A very high field exists at the overlap region. An electron can tunnel from the valence band near the surface into the conduction band, leaving a hole behind, as shown in Figure 2.20. This is a carrier generation mechanism, with the holes swept into the bulk and the electrons into the drain, where they appear as a leakage current.

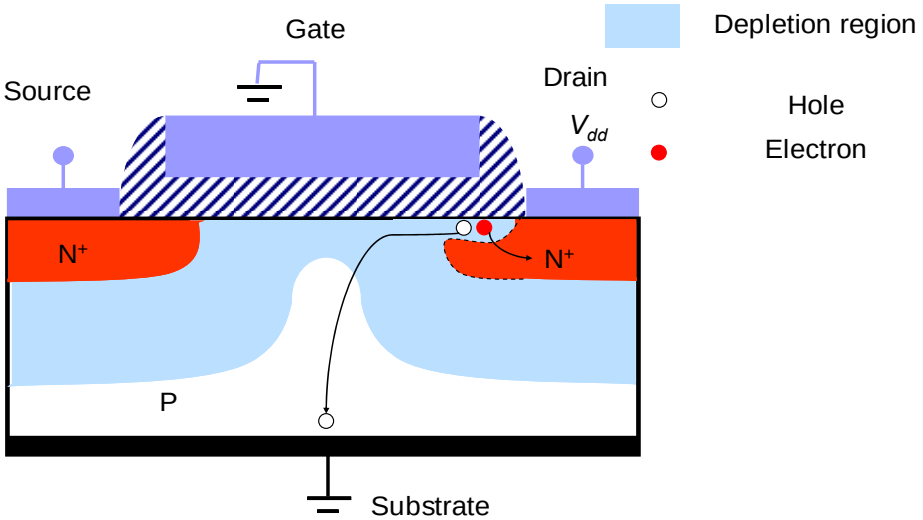

Figure 2.19
Carriers generated by BTBT forming GIDL.

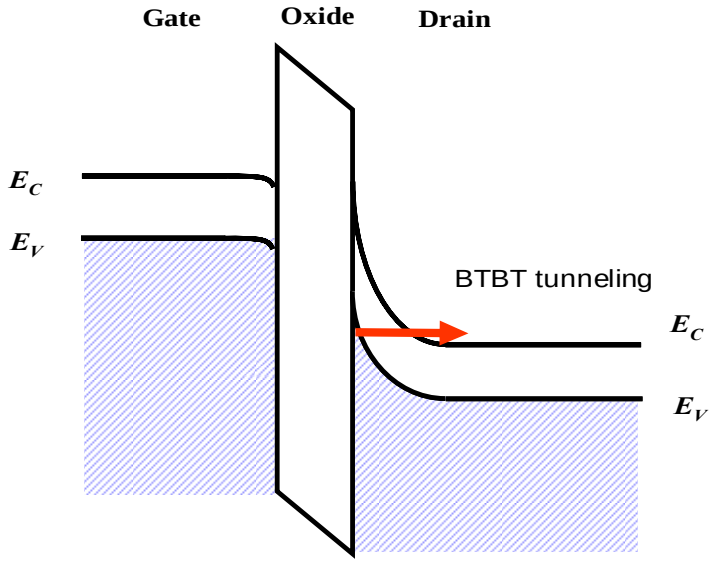

Figure 2.20
Origins of gate-induced drain leakage: BTBT at the gate overlap of the heavily doped drain.

At the normal operating mode of the MOSFET, BTBT and GIDL are not significant leakage components as gate direct tunneling and subthreshold leakage dominate [80, 81]. However, BTBT and GIDL can charge the substrate, inducing the body effect and coupling between isolated devices. BTBT is the major constraint for halo-doped MOSFETs [79, 81] and CMOS circuits employing reverse substrate bias [82].

2.8 Conclusion

In this chapter, an analytical model of gate direct tunneling is developed, for both electron tunneling in the conduction band and hole tunneling in the valence band. Electron and hole tunneling are dominant in n-MOSFETs and p-MOSFETs, respectively. In an n-MOSFET, the tunneling current density can be as high as *1.0 A/cm*, when the gate voltage is *1.0 V* and the oxide thickness is *1.5 nm.* The hole tunneling current density in a p-MOSFET is smaller than the electron tunneling current density in an n-MOSFET by a decade, because of the high barrier height and large effective mass for holes

The gate tunneling in a MOSFET is specified as the tunneling in the channel region and source/drain region in the bulk side. In CMOS circuits, the gate-to-channel and gate-to-source tunneling occur in the on-state and gate-to-drain tunneling usually occurs in the off-state of the devices.

The magnitude of tunneling current changes greatly with different materials. The use of a metal gate vs. a polycrystalline silicon gate makes a huge difference in gate tunneling because of the poly-depletion effect. High-permittivity dielectrics, which could potentially replace SiO_2 for gate insulation, can greatly reduce the tunneling current density at the same EOT as SiO_2. High-κ dielectrics provide a possible solution for the excessive tunneling current.

Besides gate tunneling, BTBT and GIDL cause considerable leakage in MOSFETs. They must be carefully considered in low-power designs.

CHAPTER 3

QUANTIZATION MODEL

3.1 Introduction and Background

To control SCEs, modern device technology uses the highly-doped channel and ultra-thin gate oxides in MOSFETs [36-39, 62]. In some cases, an even higher density of dopants is implanted in the channel near the source/drain regions, forming a halo structure [6]. All these methods used to control SCEs result in a high electric field in the direction vertical to the silicon/silicon oxide interface. Although the high electric field in the vertical direction can keep the charges in the channel under gate control against the influence of drain potential, it confines the movement of carriers in a narrow potential well. From quantum theory, the energy of the channel carriers can only take discrete values and not a continuous energy distribution as described by classical device physics [8]. Moreover, quantization also causes a redistribution of carrier density close to the Si/SiO_2 interface as compared to that of the classical prediction. Thus, it is critical to model accurately this quantization effect in the MOSFET and understand the relationship between the charge density and the gate bias.

The concept of quantization is explained in Section 3.2, and energy levels of the subbands are calculated in Section 3.3 by solving the Schrödinger and Poisson equations simultaneously. In Section 3.4, the relationship between charge density and gate bias is derived based on the solution obtained in Section 3.3. The gate capacitance models based on the classical theory and the quantum mechanics are compared in Section 3.4. Section 3.5 presents some concluding remarks.

3.2 Basic Concept of Quantization

When electrons are confined in a space comparable in size to the de Broglie wavelength, the quantum size effect becomes very relevant [8, 83]. In such confined space, quantum mechanics predicts that it is physically impossible to measure simultaneously the exact position and momentum of a particle. This principle of quantum mechanics was first introduced by physicist Werner Heisenberg and is known as the Heisenberg uncertainty principle [8]. Assuming that the uncertainties of measuring position and momentum are Δx and Δp, respectively, the uncertainty principle can be written as

$$\Delta p \cdot \Delta x \geq \frac{\hbar}{2}. \tag{3.1}$$

The above equation generalizes the uncertainty principle in quantum mechanics. If the uncertainty of measuring the position of a particle is very small, then the uncertainty for measuring its momentum must be large and vice-versa. This is because their product has a non-zero value. Furthermore, the uncertainties do not arise from random and systematic errors, but from the quantum structure of matter.

Consider an electron confined in a box as shown in Figure 3.1. Classically, the lowest allowed energy of the electron could be zero [9]. However, quantum mechanics doesn't allow the ground state to be zero [8]. If we assume that an electron has zero minimum energy and that we are able to locate its exact position in the box, then automatically we know its momentum is zero. That is a clear violation of the uncertainty principle.

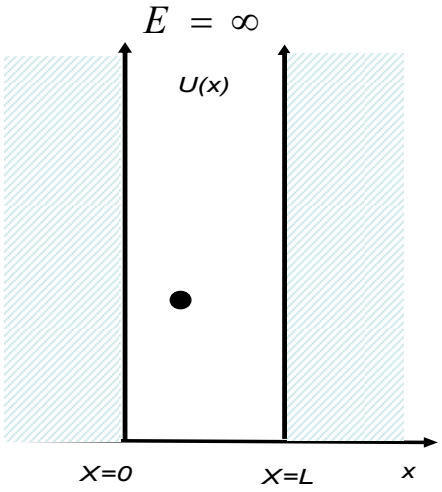

Figure 3.1
A confined particle in an infinite potential well.

The potential energy inside and outside the box is described by

$$U(x)=\begin{cases}\infty & x<0\\ 0 & 0<x<L\\ \infty & x>L\end{cases}. \tag{3.2}$$

The wavefunction of the particle $\psi(x)$ (in $[cm^{-1/2}]$) satisfies the time independent Schrödinger Equation

$$\frac{d^2\psi}{dx^2}+\frac{2mE}{\hbar^2}\psi=0, \tag{3.3}$$

and the boundary condition

$$\psi(0)=\psi(L)=0. \tag{3.4}$$

The solution for the equation (3.3) is

$$\psi(x)=\sqrt{\frac{2}{L}}\sin\left(\frac{x}{L}n\pi\right) \quad n=1,2,3,\ldots \quad . \tag{3.5}$$

The *n*th energy states E_n in the box is given by

$$E_n = \frac{\hbar^2 \pi^2 n^2}{2mL^2}. \tag{3.6}$$

The above equation shows the non-zero ground-state energy due to spatial confinement as

$$E_1 = \frac{\hbar^2 \pi^2}{2mL^2}. \tag{3.7}$$

Generally, if a particle is confined by a potential well of any shape, the particle can only take discrete energy levels and the ground energy level is not zero. This energy quantization is applicable to carriers in MOSFETs, when a strong confinement is induced by the electric field. In this situation, carriers have discrete energy levels, and the lowest allowable energy level is above the conduction band.

3.3 Quantization in MOSFET

The potential distribution $\phi(x,y)$ in the channel shown in Figure 3.2 can be described by the 2-D Poisson equation as

$$\frac{\partial^2 \phi(x,y)}{\partial x^2} + \frac{\partial^2 \phi(x,y)}{\partial y^2} = \frac{q}{\varepsilon_{Si}}\left(N_A(x) + n(x,y)\right), \tag{3.8}$$

where N_A (in [cm^{-3}]) is the depletion charge density, and $n(x,y)$ (in [cm^{-3}]) is the inversion charge density. In classical physics, the $n(x,y)$ is given by [10]

$$n(x,y) = \frac{n_i^2}{N_A}\exp\left(-q\phi / kT\right). \tag{3.9}$$

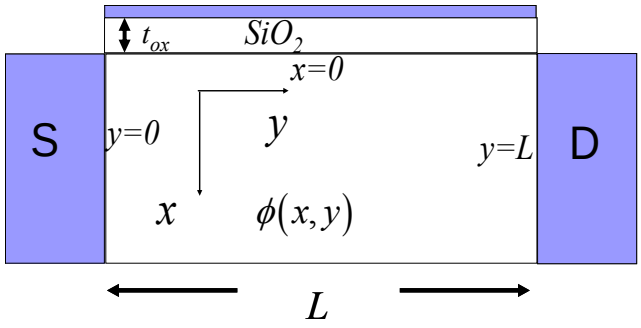

Figure 3.2
Schematic view of MOSFET channel region.

The boundary conditions of equation (3.8) are determined by the bias voltage on the four MOSFET electrodes, namely source, drain, gate, and substrate. Equation (3.9) is based on the assumption that particles are distributed on a continuous energy band, which is only valid for a slowly varying potential [9]. In comparison, the potential varies tremendously near the semiconductor-insulator interface in a MOSFET, where a potential well is formed by the interface barrier and the electrostatic potential in the semiconductor. In modern devices with highly doped channels and ultra-thin oxides, the potential well is as narrow as a few nanometers [27]. Consequently, carrier motion in the direction perpendicular to the interface (x direction) is confined, which results in energy quantization of carriers in the channel. The continuous conduction band is now split into subbands, according to the discrete energy values of the carriers. As a result, the volume density of inversion charges accounting for all the subbands can be expressed as

$$n(x,y) = -\sum_i \frac{Q_{inv,i}}{q} |\psi_i(x,y)|^2 , \tag{3.10}$$

where $Q_{inv,i}$ is the inversion charge sheet density (in [C/cm^2]) of the *i*th subband, and $\psi_i(x,y)$ is the carrier wavefunction on the *i*th subband. The wavefunction of each subband is given by the Schrödinger equation as

$$-\frac{\hbar^2}{2m}\nabla^2\psi_i(x,y)-q\phi(x,y)\psi_i(x,y)=E_i\psi_i(x,y), \tag{3.11}$$

and the eigenvalue E_i is the energy level associated with wavefunction ψ_i. Noticing that the confinement is most significant in the direction perpendicular to the gate, the one-dimensional Schrödinger equation is sufficient for the problem [27, 42]. In a systematic approach, an MOS structure with uniform potential distribution in the *y* direction is first assured. By removing variable *y* in equation (3.8) and (3.11), the problem is reduced to coupled one-dimensional Poisson and Schrödinger equations as

$$\frac{d^2\phi(x)}{dx^2}=\frac{qN_A(x)}{\varepsilon_{Si}}-\sum_i\frac{Q_{inv,i}}{q\varepsilon_{Si}}|\psi_i(x)|^2, \tag{3.12}$$

and

$$-\frac{\hbar^2}{2m}\frac{d^2\psi(x)}{dx^2}-q\phi(x)\psi(x)=E\psi(x), \tag{3.13}$$

where ϕ is related to gate voltage V_{gs} by

$$\phi(0)=V_{gs}-V_{FB}-V_{ox}. \tag{3.14}$$

Figure 3.3 compares the electron density distribution determined by both the classical and the quantum mechanical approaches.

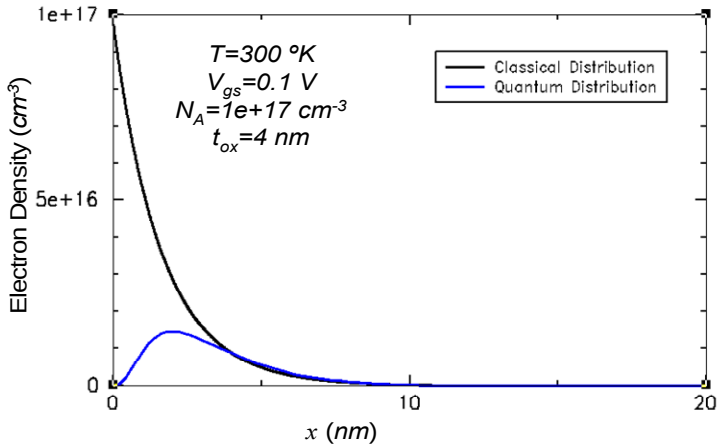

Figure 3.3
Illustration of quantum and classical electron distributions along the channel depth direction using numeric simulation from SCHRED [32].

In the Schrödinger equation (3.13), m is the effective mass of electrons, which is determined by the crystal orientation and band structure of the silicon. We assume that the *Si-SiO₂* interface is parallel to the [100] plane of the silicon, by which MOSFETs obtain best performance. In this orientation, the constant-energy surface of electrons forms six ellipsoids, as shown in Figure 3.4. Two of them are along the <100> axis with the longitudinal mass $m_l = 0.916m_0$ and the other four are transverse to the <100> axis with the transverse mass $m_t = 0.19m_0$. The subband energy varies according to the effective mass of the electrons.

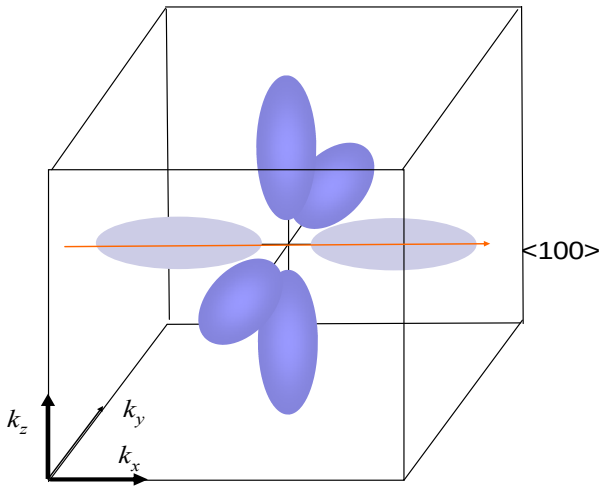

Figure 3.4
Constant-energy surface forming six ellipsoids in a cubic crystal cell.

As shown in Figure 3.5, subbands associated with $m_1 = 0.916m_0$ are grouped as valley one and a degeneracy of $g_1 = 2$ is used to count for the two ellipsoids on the <100> axis. In addition, subbands associated with $m_2 = 0.19m_0$ and a degeneracy $g_2 = 4$ are grouped as valley two, corresponding to the other four ellipsoids. The energy levels in the two valleys are denoted as $E_{1,1}$, $E_{2,1}$, ... and $E_{1,2}$, $E_{2,2}$, ... for valley one and valley two, respectively. The Schrödinger equations for valley one and valley two can be written as

$$-\frac{\hbar^2}{2m_1}\frac{d^2\psi_{i,1}(x)}{dx^2} - q\phi(x)\psi_{i,1}(x) = E_{i,1}\psi_{i,1}(x), \tag{3.15}$$

and

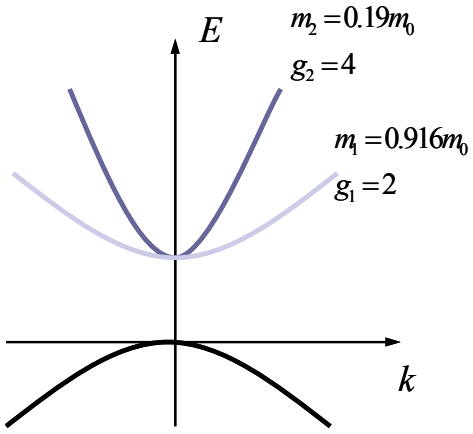

Figure 3.5
Illustration of different effective electron masses in two conduction band valleys.

$$-\frac{\hbar^2}{2m_2}\frac{d^2\psi_{i,2}(x)}{dx^2} - q\phi(x)\psi_{i,2}(x) = E_{i,2}\psi_{i,2}(x), \tag{3.16}$$

where $m_1 = 0.916m_0$ and $m_2 = 0.19m_0$ for valley one and two respectively. For the *i*th energy level, $E_{i,1} < E_{i,2}$ because of the larger effective mass of valley one as calculated from equation (3.6). Subbands in the conduction band are schematically shown in Figure 3.6.

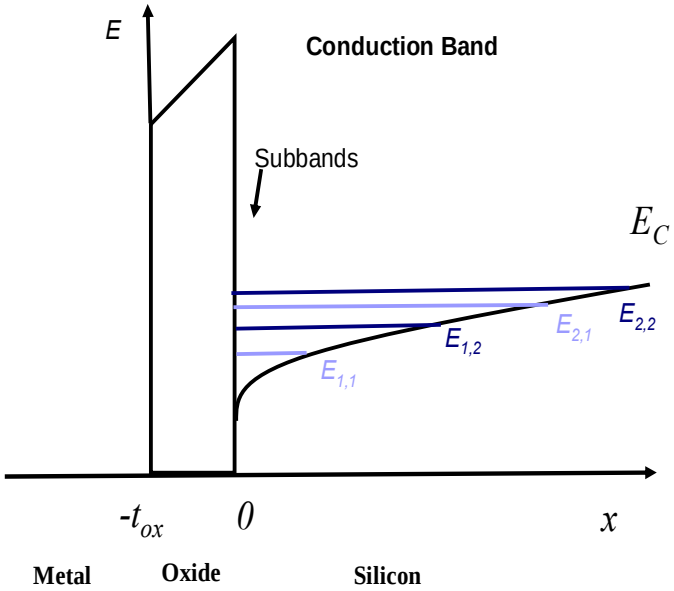

Figure 3.6
Schematic diagram of typical subbands formation for electrons with different effective masses.

Figure 3.7-Figure 3.10 show the numeric solution of equations (3.12), (3.15), and (3.16). Figure 3.7 shows the band bending, given by $-q\phi(x)$, forming a steep potential well near the channel surface. While Figure 3.8 shows the calculated subband energy levels in the conduction band, Figure 3.9 shows wavefunctions of first three subbands, where $\psi_{1,1}$ and $\psi_{2,1}$ are the wavefunctions of the first and second subbands in valley one and $\psi_{1,2}$ is the wavefunction of the lowest subband in valley two [32].

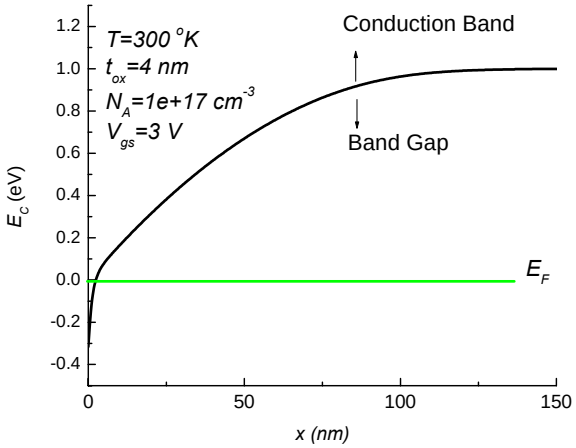

Figure 3.7
Numerical results for conduction band bending in the *x* direction from SCHRED [32].

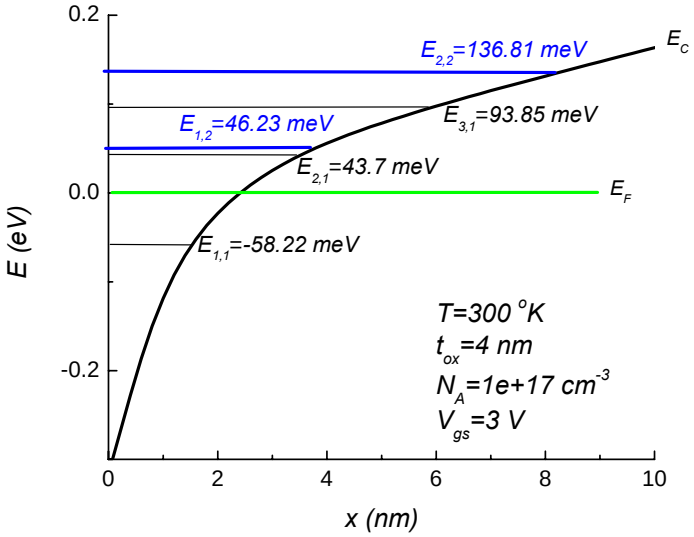

Figure 3.8
Numerical results for subband energy levels from SCHRED [32].

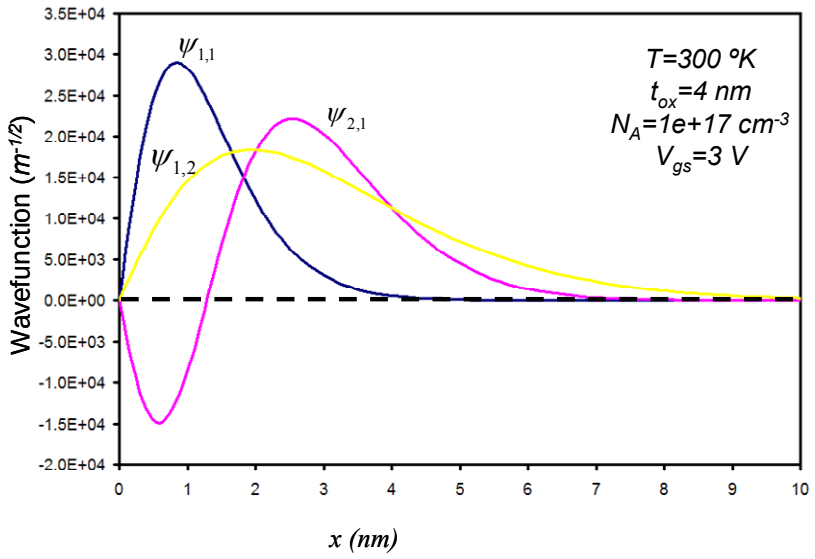

Figure 3.9
Wavefunctions of subbands from simulation results of SCHRED [32].

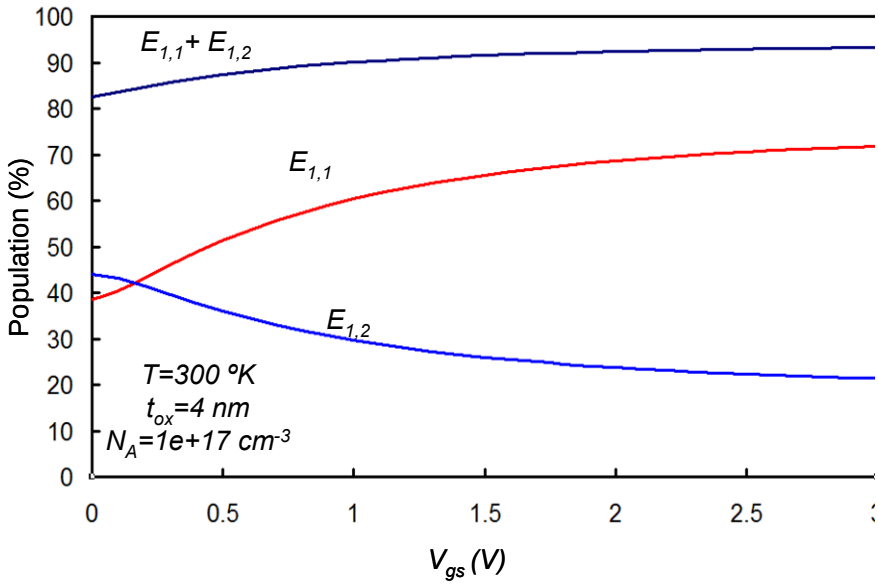

Figure 3.10
Numerical results for carrier population on lowest two subbands from SCHRED [32].

It is very difficult to give an analytical solution to the coupled Schrödinger and Poisson equations. However, numerical simulation [32] conveys several prospects. Most carriers reside on the two lowest subbands. For V_{gs} variations from *0-3 V*, the electron population on the lowest two subands accounts for *80-95%* of the total carrier population, as shown in Figure 3.10. Therefore, it is sufficient to consider only the lowest energy level in each valley in order to obtain an accurate analytical solution [27]. With this assumption, the Poisson and Schrödinger equations are reduced to

$$\frac{d^2\phi(x)}{dx^2} = \frac{qN_A(x)}{\varepsilon_{Si}} + \frac{Q_{inv,1}}{q\varepsilon_{Si}}\left|\psi_{1,1}(x)\right|^2 + \frac{Q_{inv,2}}{q\varepsilon_{Si}}\left|\psi_{1,2}(x)\right|^2, \tag{3.17}$$

$$-\frac{\hbar^2}{2m_1}\frac{d^2\psi_{1,1}(x)}{dx^2} - q\phi(x)\psi_{1,1}(x) = E_{1,1}\psi_{1,1}(x), \tag{3.18}$$

and

$$-\frac{\hbar^2}{2m_2}\frac{d^2\psi_{1,2}(x)}{dx^2} - q\phi(x)\psi_{1,2}(x) = E_{1,2}\psi_{1,2}(x) \tag{3.19}$$

where $Q_{inv,1}$ and $Q_{inv,2}$ (in [C/cm^2]) are the electron charge sheet densities for valley one and two, respectively.

3.3.1 Boundary Conditions

The depletion region charges are considered as uniformly distributed in a depletion charge layer of depth, d (in [cm]), as

$$\begin{aligned} N_A(x) &= N_A \qquad 0 \le x \le d \\ &= 0 \qquad x > d \end{aligned}. \tag{3.20}$$

The boundary condition for the Poisson equation (3.17) at the *Si-SiO₂* interface is that the normal component of displacement density vector is continuous across the *Si-SiO₂* interface [10] such that,

$$\varepsilon_{Si}\left(-\left.\frac{d\phi}{dx}\right|_{x=0}\right)=\varepsilon_{ox}E_{ox}\,, \tag{3.21}$$

where E_{ox} (in $[V/cm]$) is the vertical oxide dielectric field. For an applied gate voltage V_{gs}, E_{ox} is given by

$$E_{ox}=\frac{V_{ox}}{t_{ox}}\,. \tag{3.22}$$

According to equation (3.14) and Figure 2.4, E_{ox} can be written as

$$E_{ox}=\frac{V_{gs}-\phi(0)-V_{FB}}{t_{ox}}\,. \tag{3.23}$$

Substituting (3.23) into equation (3.21), we can write the boundary condition at the *Si-SiO₂* interface as

$$\varepsilon_{Si}\left(-\left.\frac{d\phi}{dx}\right|_{x=0}\right)=\frac{\varepsilon_{ox}\left(V_{gs}-\phi(0)-V_{FB}\right)}{t_{ox}}\,. \tag{3.24}$$

In the neutral substrate, the electrostatic potential stays constant. For convenience, it is chosen to be the reference potential. Thus, the potential boundary condition at infinity is given by

$$\phi(x=\infty)=0\,. \tag{3.25}$$

According to quantum mechanics, the wavefunctions for particles in the potential well are real [8]. Because the potential barrier at the oxide/silicon interface is relatively high (*3.1 eV* referred to the bottom of conduction band) for electrons in the potential well, it could be approximated as infinity [43, 55, 57]. This assumption simplifies the boundary condition at $x=0$ for the Schrödinger equation as

$$\psi(x=0)=0\,. \tag{3.26}$$

Moreover, inversion charges distribute near the surface of the channel, which indicates that both the wavefunction and its derivative should vanish in the infinite distance. This results in boundary conditions for the wavefunction at $x=\infty$ given by

$$\psi\left(x=\infty\right)=0\,, \tag{3.27}$$

and

$$\left.\frac{d\psi}{dx}\right|_{x=\infty}=0\,. \tag{3.28}$$

3.3.2 Solution by Variational Method

Using variational method (see Appendix C), electron wavefunctions for valley one and two are assumed to be $\psi_{1,1}\left(x\right)=\left(2\alpha_1^{\,3}\right)^{1/2}x\exp\left(-\alpha_1 x\right)$ and $\psi_{1,2}\left(x\right)=\left(2\alpha_2^{\,3}\right)^{1/2}x\exp\left(-\alpha_2 x\right)$ with the undetermined parameters being α_1 and α_2, respectively. These wavefunctions satisfy boundary conditions given by equations (3.26), (3.27) and (3.28). Moreover, these trial wavefunctions have similar forms with $\psi_{1,1}$ and $\psi_{1,2}$ as shown in Figure 3.9, ensuring good accuracy for the calculated energy levels. Noticing that

$$\frac{d\left|\psi_{1,1}\left(x\right)\right|^2}{dx}=\left(2\alpha_1^{\,3}\right)x\exp\left(-2\alpha_1 x\right)-\left(4\alpha_1^{\,4}\right)x^2\exp\left(-2\alpha_1 x\right) \tag{3.29}$$

leads to

$$\left.\frac{d\left|\psi_{1,1}\left(x\right)\right|^2}{dx}\right|_{x=\frac{1}{2\alpha_1}}=0\,, \tag{3.30}$$

it is found that these trial wavefunctions result in the electron density peak at $\frac{1}{2\alpha_1}$ and $\frac{1}{2\alpha_2}$ for valley one and valley two, respectively.

Given wavefunctions $\psi_{1,1}$ and $\psi_{1,2}$, integrating equation (3.17) from the bulk toward the surface leads to

$$\phi(x) = \iint_{\infty \to 0} \left(\frac{qN_A}{\varepsilon_{Si}} + \frac{Q_{inv,1}}{\varepsilon_{Si}} \left|\psi_{1,1}(x)\right|^2 + \frac{Q_{inv,2}}{\varepsilon_{Si}} \left|\psi_{1,2}(x)\right|^2 \right) dx^2 . \tag{3.31}$$

The terms of qN_A, $Q_{inv,1}$ and $Q_{inv,2}$ are the depletion charge density, and inversion charge densities from valley one and valley two, respectively. Specifying their contributions to the total potential as ϕ_{depl}, $\phi_{inv,1}$, and $\phi_{inv,2}$ leads to

$$\phi(x) = \phi_{depl} + \phi_{inv,1} + \phi_{inv,2} \tag{3.32}$$

In equation (3.32), ϕ_{depl} is the electrostatic potential generated by depletion charges

$$\phi_{depl} = \begin{cases} \frac{1}{2}\frac{qN_A}{\varepsilon_{Si}}(d-x)^2 & 0 \le x \le d \\ 0 & x > d \end{cases} . \tag{3.33}$$

ϕ_{inv1} is the potential induced by inversion charges in valley one, which is associated with wavefunction $\psi_{1,1}$ in Figure 3.9 and given by

$$\phi_{inv,1} = \frac{1}{2}\frac{Q_{inv,1}}{\alpha_1 \varepsilon_{Si}} \left(3 + 4\alpha_1 x + 2\alpha_1{}^2 x^2\right) \exp(-2\alpha_1 x), \tag{3.34}$$

and ϕ_{inv2} is the potential induced by inversion charges in valley two, which is associated with wavefunction $\psi_{1,2}$ in Figure 3.9, and given by

$$\phi_{inv,2} = \frac{1}{2}\frac{Q_{inv,2}}{\alpha_2 \varepsilon_{Si}} \left(3 + 4\alpha_2 x + 2\alpha_2{}^2 x^2\right) \exp(-2\alpha_2 x). \tag{3.35}$$

The lowest energy level is given by the expectation value of the Hamiltonian of the wavefunction (see Appendix C). For valley one,

$$E_{1,1} = \left\langle \psi_{1,1} \middle| \hat{H} \middle| \psi_{1,1} \right\rangle . \tag{3.36}$$

As shown in Appendix C, equation (3.36) can be replaced by the integral

$$E_{1,1} = \int_0^\infty \psi_{1,1}(x) \frac{\hbar^2}{2m_1} \frac{d}{dx^2} \psi_{1,1}(x)\, dx + \int_0^\infty q\phi(x) \psi_{1,1}(x)^2\, dx . \tag{3.37}$$

Substituting the expression $\psi_{1,1}(x) = \left(2\alpha_1^3\right)^{1/2} x \exp(-\alpha_1 x)$ into equation (3.37) leads to

$$\begin{aligned} E_{1,1} = {} & \frac{\hbar^2 \alpha_1^2}{2m_1} + \frac{3q^2 N_A}{2\varepsilon_{Si}\alpha_1}\left(d - \frac{1}{\alpha_1}\right) + \frac{3}{2}\frac{11 q Q_{inv,1}}{16\varepsilon_{Si}\alpha_1} \\ & + \frac{3}{2}\frac{qQ_{inv,2}}{\varepsilon_{Si}} \frac{\alpha_1^4 + 5\alpha_1^3\alpha_2 + 10\alpha_1^2\alpha_2^2 + 5\alpha_1\alpha_2^3 + \alpha_2^4}{(\alpha_1 + \alpha_2)^5} \end{aligned} . \tag{3.38}$$

As shown in Figure 3.7, the magnitude of depletion depth is in the order of hundreds of nanometers [10]. While Figure 3.3 and Figure 3.9 show that the peak of the electron density lies a few nanometers beneath the channel surface, namely $d >> \frac{1}{\alpha_1}$ and $\frac{1}{\alpha_2}$, we can approximate

$$d - \frac{1}{\alpha_1} \approx d . \tag{3.39}$$

Letting

$$\alpha_2 = \gamma \alpha_1 \tag{3.40}$$

and

$$f(\gamma) = \frac{16}{11} \frac{\gamma^4 + 5\gamma^3 + 10\gamma^2 + 5\gamma + 1}{(1+\gamma)^5}, \tag{3.41}$$

equation (3.38) can be rewritten as

$$E_{1,1} = \frac{\hbar^2 \alpha_1^{\,2}}{2m_1} + \frac{3q^2 N_A d}{2\varepsilon_{Si}\alpha_1} + \frac{3}{2}\frac{11qQ_{inv,1}}{16\varepsilon_{Si}\alpha_1} + \frac{3}{2}\frac{11qQ_{inv,2}}{16\varepsilon_{Si}\alpha_1} f(\gamma). \tag{3.42}$$

Likewise, the lowest energy in the valley two can be expressed as

$$E_{1,2} = \frac{\hbar^2 \alpha_2^{\,2}}{2m_2} + \frac{3q^2 N_A d}{2\varepsilon_{Si}\alpha_2} + \frac{3}{2}\frac{11qQ_{inv,2}}{16\varepsilon_{Si}\alpha_2} + \frac{3}{2}\frac{11qQ_{inv,1}}{16\varepsilon_{Si}\alpha_2} g(\gamma), \tag{3.43}$$

where

$$g(\gamma) = \frac{16}{11}\frac{\gamma\left(\gamma^4 + 5\gamma^3 + 10\gamma^2 + 5\gamma + 1\right)}{(1+\gamma)^5}. \tag{3.44}$$

The parameters α_1 and α_2 should minimize the energy level by the variational method, so that

$$\frac{dE_{1,1}}{d\alpha_1} = 0 \tag{3.45}$$

and

$$\frac{dE_{1,2}}{d\alpha_2} = 0. \tag{3.46}$$

Parameters α_1 and α_2 are obtained from equation (3.45) and (3.46) as

$$\alpha_1 = \left[\frac{3m_1 q}{2\varepsilon_{Si}\hbar^2}\left(Q_{depl} + \frac{11}{16}Q_{inv,1} + \frac{11}{16}f(\gamma)Q_{inv,2}\right)\right]^{1/3}, \tag{3.47}$$

and

$$\alpha_2 = \left[\frac{3m_2 q}{2\varepsilon_{Si}\hbar^2}\left(Q_{depl} + \frac{11}{16}Q_{inv,2} + \frac{11}{16}g(\gamma)Q_{inv,1}\right)\right]^{1/3}. \tag{3.48}$$

Therefore, γ is determined from equations (3.40), (3.47), and (3.48) as

$$\gamma^3 = \frac{m_2}{m_1}\frac{Q_{depl} + \frac{11}{16}Q_{inv,2} + \frac{11}{16}g(\gamma)Q_{inv,1}}{Q_{depl} + \frac{11}{16}Q_{inv,1} + \frac{11}{16}f(\gamma)Q_{inv,2}}. \tag{3.49}$$

The range of possible values of γ is given by considering two extreme situations. At low gate voltage, the conduction band bending in the channel is small, and depletion charges overwhelm inversion charges in equation (3.49). The equation is simplified as

$$\gamma^3 = \frac{m_2}{m_1}, \tag{3.50}$$

which gives $\gamma = 0.59$ by substituting $m_1 = 0.916m_0$ and $m_2 = 0.19m_0$. In the extreme situation that $Q_{inv,1}$ dominates Q_{depl} and $Q_{inv,2}$, the contribution of Q_{depl} and $Q_{inv,2}$ in equation (3.49) can be ignored. This extreme case corresponds to a MOSFET at high operating bias, in which inversion charges outnumber depletion charges. Meanwhile, Figure 3.10 shows that electrons in valley two become a smaller portion of the total inversion charges at high gate bias. By ignoring Q_{depl} and $Q_{inv,2}$, equation (3.49) becomes

$$\gamma^3 = \frac{m_2}{m_1}g(\gamma). \tag{3.51}$$

Equation (3.51) can be solved, resulting in $\gamma = 0.51$. Considering the two extreme cases $\gamma = 0.59$ and $\gamma = 0.51$, the value of γ is in the range 0.51~0.59. Evidently, γ does not vary much with $Q_{inv,1}$, Q_{depl} and $Q_{inv,2}$. Furthermore, Figure 3.11 shows $f(\gamma)$ and $g(\gamma)$ only change slightly in γ variance range. Approximating γ as the average of the two extreme cases results in

$$\gamma \approx 0.545, \tag{3.52}$$

and

$$\alpha_1 = \left[\frac{3m_1 q}{2\varepsilon_{Si}\hbar^2} \left(Q_{depl} + \frac{11}{16} Q_{inv} \right) \right]^{1/3} . \tag{3.53}$$

From equation (3.40), it is known that

$$\alpha_2 = 0.545\alpha_1 , \tag{3.54}$$

Therefore, $E_{1,1}$ is obtained as

$$E_{1,1} = 3\hbar^2 \alpha_1^{\,2} / \left(2m_1\right). \tag{3.55}$$

Energy level $E_{1,2}$ is similarly obtained, using equations (3.43) and (3.54), as

$$E_{1,2} = 3\hbar^2 \alpha_2^{\,2} / \left(2m_2\right) \approx 1.432 E_{1,1} \tag{3.56}$$

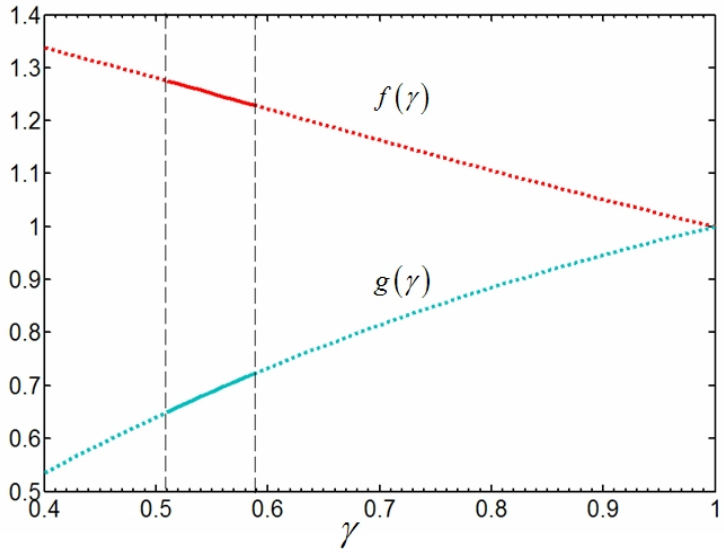

Figure 3.11
Range of values for $f(\gamma)$ and $g(\gamma)$.

Figure 3.12 shows good agreement between the derived model and the numerical Poisson-Schrödinger solver for MOSFETs operating in the range of $V_{gs} = 1 \sim 3V$. In equation (3.55) and (3.56), $E_{1,1}$ $E_{1,2}$ are given by the inversion charge and depletion charge density instead of gate voltage. For comparison purposes, Q_{inv}, Q_{depl} are extracted from the numerical simulation. The relationship between charge densities Q_{inv}, Q_{depl} and gate voltage is shown in the following section. From equation (3.56) and Figure 3.12, the separation between the energy levels of the two lowest subbands is enlarged at high bias, which explains the rapid increase of the relative occupation ratio ($Q_{inv,1} / Q_{inv}$) in the subband $E_{1,1}$ as shown in Figure 3.10.

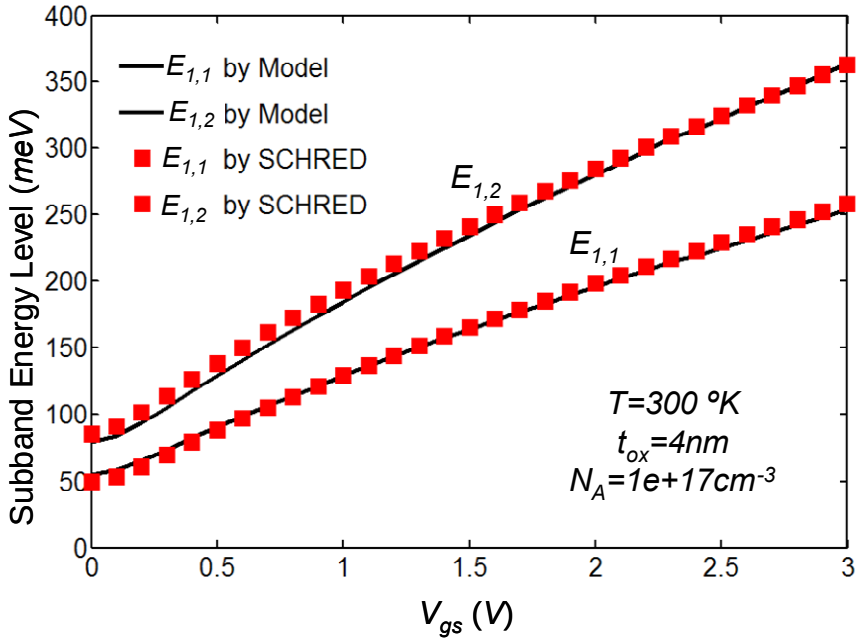

Figure 3.12
Comparison of the quantized energy levels given by the model and by numerical simulation from SCHRED [32].

3.4 Quantum Mechanical C-V Model

Although the charge profile in a MOSFET can be obtained from the subband energy levels, they do not explicitly describe how charge density responds with the applied gate voltage. The *C-V* characteristic [10] becomes a necessary tool for describing MOSFET behavior.

From the quantization effect analysis, it becomes clear that the distribution of inversion charge is quite different from that predicted by classical analysis. For example, in the classical model, inversion charge changes with potential exponentially, resulting in a maximum concentration at the channel surface [59]. In quantum mechanics, by considering the potential barrier of the gate oxide, the wavefunction diminishes at the interface, leading to zero charge density as shown in Figure 3.9. As a result, the density peak of inversion charges shifts away from the interface, which is shown in Figure 3.3. Meanwhile, inversion charges are distributed on the split subbands instead of having a continuous distribution, as shown in Figure 3.8. In the quantum analysis, inversion charge can be described as a two-dimensional gas at discrete energy levels [52, 84].

The appropriate charge profile must be considered for an accurate gate capacitance model. Through a gate capacitance model that incorporates the energy quantization effect, a better understanding of the behavior of MOSFETs in different operational regions can be achieved. Device parameters justified quantum mechanically can be further studied based on the gate capacitance model.

3.4.1 Gate Capacitance Components

Charges in the channel consist of depletion charges and inversion charges, namely

$$Q_T = Q_{depl} + Q_{inv}, \tag{3.57}$$

where Q_T (in [C/cm^2]) is the total charge sheet density in the channel. By assuming all voltages are referenced to the source of the MOSFET and the substrate is connected to the source, a MOSFET can be analyzed as an MOS capacitor with two electrodes: gate and grounded substrate. According to the variance of the charge density with respect to the channel surface potential ϕ_s (in [V]), two capacitances can be defined: inversion layer capacitance per unit area C_{inv} (in [F/cm^2]) and depletion layer capacitance per unit area C_d (in [F/cm^2]), which are given by

$$C_{inv} = \frac{\partial Q_{inv}}{\partial \phi_s}, \tag{3.58}$$

and

$$C_d = \frac{\partial Q_{depl}}{\partial \phi_s}. \tag{3.59}$$

The depletion layer capacitance be expressed as [9]

$$C_d = \frac{\varepsilon_{Si}}{d}, \tag{3.60}$$

where d [cm] is the depletion depth. Surface potential ϕ_s varies with gate voltage V_{gs} conforming with

$$Q_T = C_{ox}\left(V_{gs} - \phi_s - V_{FB}\right), \tag{3.61}$$

where C_{ox} (in [F/cm^2]) is the oxide layer capacitance per unit area given by [9]

$$C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}}. \tag{3.62}$$

The total gate capacitance per unit area C_g (in [F/cm^2]) is defined by

$$C_g = \frac{\partial Q_T}{\partial V_{gs}}. \tag{3.63}$$

Therefore, by combining equation (3.57) to (3.63),

$$C_g = \left(\frac{1}{C_{ox}} + \frac{1}{C_{inv} + C_d} \right)^{-1} \tag{3.64}$$

is obtained as illustrated in Figure 3.13.

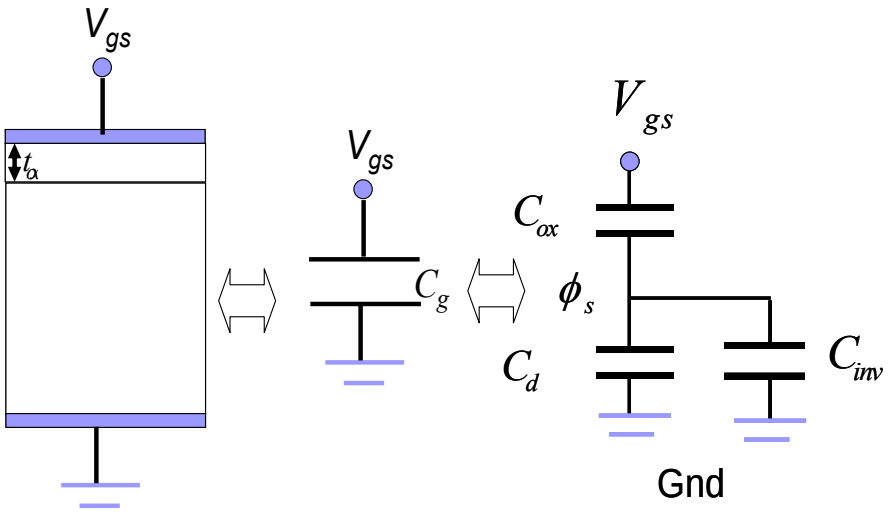

Figure 3.13
Schematic view of the generic gate capacitance model.

3.4.2 Classical Gate Capacitance Model

The inversion charge density described by classical physics is

$$n(x,y) = \frac{n_i^2}{N_A} \exp\left(-q\phi / kT\right). \tag{3.65}$$

The total electron sheet density Q_{inv} can be expressed as

$$Q_{inv} = \int_0^{\infty} \frac{n_i^2}{N_A} \exp\left(\frac{-q\phi(x)}{kT} \right) dx, \tag{3.66}$$

where $\phi(x)$ can be given solely by Poisson's equation (3.8), and C_{inv} can be obtained as

$$C_{inv} = \frac{\partial Q_{inv}}{\partial \phi_s}. \tag{3.67}$$

In the classical model (See Appendix D), the inversion layer capacitance is $\frac{qQ_{inv}}{kT}$ and $\frac{qQ_{inv}}{2kT}$ in the weak and strong inversion regions, respectively.

In the classical model, the inversion layer capacitance is intentionally ignored, because it assumes that inversion charges concentrate at the surface of the channel, which causes the inversion layer capacitance to be much larger than C_{ox}. However, in sub-*90 nm* devices, the inversion charge density peak is located inside the channel, and C_{inv} needs to be specified taking energy quantization into consideration. This results in a finite inversion layer capacitance. In cases when the oxide layer is thick and the inversion layer capacitance is much larger than the oxide layer capacitance, the error in total gate capacitance by ignoring the inversion layer capacitance is negligible [59]. However, oxide thickness is greatly reduced in sub-*90 nm* MOSFETs, where the oxide layer capacitance becomes comparable to the inversion layer capacitance. In this situation, the gate capacitance will be noticeably reduced by taking into account the quantum-induced inversion layer capacitance, which is demonstrated in Figure 3.14.

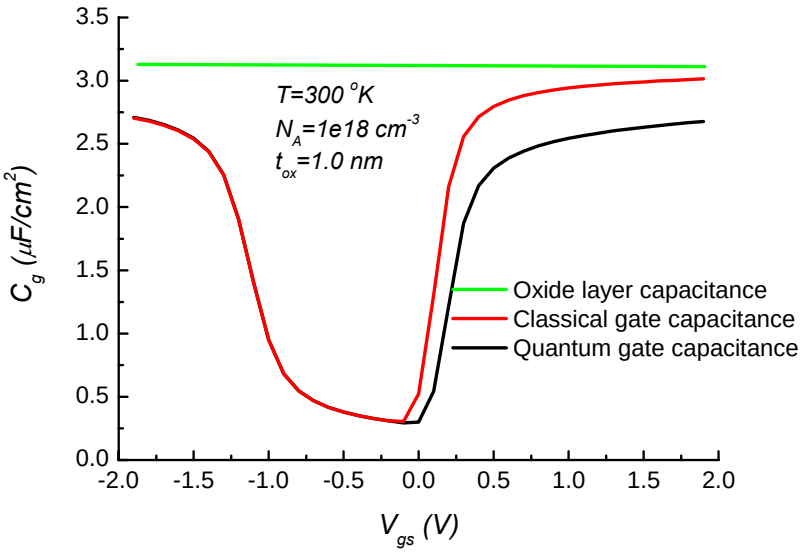

Figure 3.14
Comparison of gate capacitance as predicted by the classical and quantum simulations with SCHRED [32]. A metal gate n-MOSFET with Fermi energy -4.0 *eV* referenced to vacuum is used here.

3.4.3 Quantum Gate Capacitance Model

A quantum mechanical gate capacitance model accounts for the wave nature of carrier distribution. From the quantum mechanical carrier distribution on subbands, electron volume density in the inversion charge is described by

$$n(x) = -\sum_i \frac{Q_{inv,i}}{q} |\psi_i(x)|^2 \tag{3.68}$$

The distinction between the quantum mechanical and classical distributions has been clearly shown in Figure 3.3.

As previous sections state, most inversion charges stay in the lower subbands associated with effective masses of $m_1 = 0.916m_0$ and $m_2 = 0.19m_0$. The Q_{inv} is divided into two parts,

$$Q_{inv} = Q_{inv,1} + Q_{inv,2} \tag{3.69}$$

where $Q_{inv,1}$ and $Q_{inv,2}$ correspond to the inversion charge sheet density associated with valley one and valley two, respectively. By definition, inversion layer capacitance is the variance of Q_{inv} with respect to ϕ_s

$$C_{inv} = \frac{\partial Q_{inv}}{\partial \phi_s}. \tag{3.70}$$

Substituting equation (3.69) into the C_{inv} definition leads to

$$C_{inv} = \frac{\partial Q_{inv,1}}{\partial \phi_s} + \frac{\partial Q_{inv,2}}{\partial \phi_s}. \tag{3.71}$$

Inversion charges on subbands follow the two-dimensional distribution. As mentioned in Section 2.3.1, there are $\frac{m^2}{h^2} dydzdv_y dv_z$ states in a volume $dydzdv_y dv_z$. From the Boltzmann distribution,

$$f(E) = \exp(-\frac{E - E_F}{kT}), \tag{3.72}$$

inversion charges sheet density $Q_{inv,1}$ and $Q_{inv,2}$ can be given as

$$Q_{inv,1} = q\int_0^\infty \frac{g_1 m_{d1}^{*\,2}}{h^2} \exp\left(-\frac{E_1 - E_F}{kT}\right) dydzdv_y dv_z \tag{3.73}$$

and

$$Q_{inv,2} = q\int_0^\infty \frac{g_2 m_{d2}^{*\,2}}{h^2} \exp\left(-\frac{E_2 - E_F}{kT}\right) dydzdv_y dv_z, \tag{3.74}$$

where the degeneracy of states g_1 and g_2 are taken into account, and E_1 and E_2 are the energies of the inversion charges in valley one and valley two. Electron masses m_{d1}^* and m_{d2}^* are effective density-of-states masses in the *y*-*z* plane for valley one and valley two, respectively. From Figure 3.4, $m_{d1}^* = m_t = 0.19m_0$ and $m_{d2}^* = \sqrt{m_t m_l} = 0.42m_0$ [43].

The energies E_1 and E_2 in equations (3.73) and (3.74) can be written as the summation of their *x*, *y*, *z* components as

$$E_1 = E_{1,1} + E_y + E_z \tag{3.75}$$

and

$$E_2 = E_{1,2} + E_y + E_z \tag{3.76}$$

where $E_{1,1}$ and $E_{1,2}$ are given by equations (3.55) and (3.56) respectively, and E_y and E_z are kinetic energy components varying from zero to infinity. For inversion charges in valley one, $E_y = \frac{1}{2} m_{d1}^* v_y^{\,2}$ and $E_z = \frac{1}{2} m_{d1}^* v_z^{\,2}$. Thus, equation (3.73) can be written as

$$Q_{inv,1} = q \frac{g_1 m_{d1}^{*\,2}}{h^2} \exp\left(-\frac{E_{1,1} - E_F}{kT} \right) \int_0^\infty \exp\left(-\frac{1}{2kT} m_{d1}^* v_y^{\,2} \right) \exp\left(-\frac{1}{2kT} m_{d1}^* v_z^{\,2} \right) dv_y dv_z \,. \tag{3.77}$$

From $\int_0^\infty \exp\left(-x^2\right) dx = \sqrt{\pi}$, we have

$$Q_{inv,1} = q \frac{kT g_1 m_{d1}^*}{\pi \hbar^2} \exp\left(-\frac{E_{1,1} - E_F}{kT} \right). \tag{3.78}$$

The Fermi energy level E_F is referenced to the conduction band edge at the channel surface. According to equation (2.43) and Figure 2.3, it is given by

$$E_F = q\phi_s - q\phi_B - E_g / 2 \,. \tag{3.79}$$

Therefore, $Q_{inv,1}$ can be expressed as

$$Q_{inv,1} = q\frac{kTg_1 m_{d1}^{*}}{\pi\hbar^2}\exp\left(-\frac{q\phi_B + \frac{E_g}{2} - q\phi_s + E_{1,1}}{kT}\right). \tag{3.80}$$

In a manner identical to that used for $Q_{inv,1}$, $Q_{inv,2}$ is found as

$$Q_{inv,2} = q\frac{kTg_2 m_{d2}^{*}}{\pi\hbar^2}\exp\left(-\frac{q\phi_B + \frac{E_g}{2} - q\phi_s + E_{1,2}}{kT}\right). \tag{3.81}$$

By substituting equation (3.80) and (3.81) into (3.71), inversion layer capacitance is obtained as

$$C_{inv} = \frac{\partial}{\partial\phi_s}\left[\begin{array}{l}\frac{qkTg_1 m_{d1}^{*}}{\pi\hbar^2}\exp\left(-\frac{q\phi_B + \frac{E_g}{2} - q\phi_s + E_{1,1}}{kT}\right) \\ +\frac{qkTg_2 m_{d2}^{*}}{\pi\hbar^2}\exp\left(-\frac{q\phi_B + \frac{E_g}{2} - q\phi_s + E_{1,2}}{kT}\right)\end{array}\right], \tag{3.82}$$

which can be simplified as

$$C_{inv} = Q_{inv,1}\left(\frac{q}{kT} - \frac{1}{kT}\frac{\partial E_{1,1}}{\partial\phi_s}\right) + Q_{inv,2}\left(\frac{q}{kT} - \frac{1}{kT}\frac{\partial E_{1,2}}{\partial\phi_s}\right). \tag{3.83}$$

Using equation (3.53) and (3.55), $E_{1,1}$ can be written as

$$E_{1,1} = \frac{3\hbar^2}{2m_1}\left[\frac{3m_1 q}{2\varepsilon_{Si}\hbar^2}\left(Q_{depl} + \frac{11}{16}Q_{inv}\right)\right]^{2/3}. \tag{3.84}$$

Therefore,

$$\frac{\partial E_{1,1}}{\partial\phi_s} = \frac{3\hbar^2}{2m_1}\frac{2}{3}\left(\frac{3m_1 q}{2\varepsilon_{Si}\hbar^2}\right)^{2/3}\left(Q_{depl} + \frac{11}{16}Q_{inv}\right)^{-1/3}\left[\frac{\partial Q_{depl}}{\partial\phi_s} + \frac{11}{16}\frac{\partial Q_{inv}}{\partial\phi_s}\right]. \tag{3.85}$$

The above equation can be simplified by substituting $\alpha_1 = \left[\frac{3m_1 q}{2\varepsilon_{Si}\hbar^2}\left(Q_{depl} + \frac{11}{16}Q_{inv}\right)\right]^{1/3}$ as

$$\frac{\partial E_{1,1}}{\partial \phi_s} = \left(\frac{3q}{2\alpha_1 \varepsilon_{Si}}\right)\left[\frac{\partial Q_{depl}}{\partial \phi_s} + \frac{11}{16}\frac{\partial Q_{inv}}{\partial \phi_s}\right]. \tag{3.86}$$

From equation (3.56), $E_{1,2} = 1.432E_{1,1}$. Therefore, the above equation can be modified for $E_{1,2}$ as

$$\frac{\partial E_{1,2}}{\partial \phi_s} = 1.432\left(\frac{3q}{2\alpha_1 \varepsilon_{Si}}\right)\left[\frac{\partial Q_{depl}}{\partial \phi_s} + \frac{11}{16}\frac{\partial Q_{inv}}{\partial \phi_s}\right]. \tag{3.87}$$

By applying expressions (3.86) and (3.87) to equation (3.83), the inversion layer capacitance is obtained as

$$\begin{aligned} C_{inv} = {} & \frac{qQ_{inv}}{kT} - \frac{Q_{inv,1}}{kT}\left(\frac{3q}{2\alpha_1 \varepsilon_{Si}}\right)\left[\frac{\partial Q_{depl}}{\partial \phi_s} + \frac{11}{16}\frac{\partial Q_{inv}}{\partial \phi_s}\right] \\ & -1.432\frac{Q_{inv,2}}{kT}\left(\frac{3q}{2\alpha_1 \varepsilon_{Si}}\right)\left[\frac{\partial Q_{depl}}{\partial \phi_s} + \frac{11}{16}\frac{\partial Q_{inv}}{\partial \phi_s}\right] \end{aligned}. \tag{3.88}$$

Defining the effective inversion layer thickness X_{eff}

$$X_{eff} = \frac{33}{32}\frac{1}{\alpha_1}\frac{Q_{inv,1} + 1.432Q_{inv,2}}{Q_{inv}}, \tag{3.89},$$

equation (3.88) transforms to

$$C_{inv} = \frac{qQ_{inv}}{kT} - \frac{16}{11}\frac{qQ_{inv}}{kT}\frac{X_{eff}}{\varepsilon_{Si}}\frac{\partial Q_{depl}}{\partial \phi_s} - \frac{qQ_{inv}}{kT}\frac{X_{eff}}{\varepsilon_{Si}}\frac{\partial Q_{inv}}{\partial \phi_s}. \tag{3.90}$$

The terms $\frac{qQ_{inv}}{kT}$ and $\frac{\varepsilon_{Si}}{X_{eff}}$ can be considered as two equivalent capacitors denoted by C_{WI} and C_{SI},

$$C_{WI} = \frac{qQ_{inv}}{kT} \tag{3.91}$$

and

$$C_{SI} = \frac{\varepsilon_{Si}}{X_{eff}}. \tag{3.92}$$

Also noticing $C_d = \frac{\partial Q_{depl}}{\partial \phi_s}$ and $C_{inv} = \frac{\partial Q_{inv}}{\partial \phi_s}$, equation (3.90) becomes

$$C_{inv} = C_{WI} - \frac{16}{11}\frac{C_{WI}}{C_{SI}}C_d - \frac{C_{WI}}{C_{SI}}C_{inv}. \tag{3.93}$$

Solving equation (3.93) for C_{inv} results in

$$C_{inv} = \frac{1 - \frac{16}{11}\frac{C_d}{C_{SI}}}{\frac{1}{C_{WI}} + \frac{1}{C_{SI}}}. \tag{3.94}$$

From the definitions of C_d and C_{SI} in equation (3.60) and (3.92), it is known that

$$\frac{C_d}{C_{SI}} = \frac{\frac{\varepsilon_{Si}}{d}}{\frac{\varepsilon_{Si}}{X_{eff}}} = \frac{X_{eff}}{d}. \tag{3.95}$$

Referencing to equation (3.89), it can be easily found that X_{eff} and $\frac{1}{\alpha_1}$ are of the same order of magnitude. Comparing magnitudes of d and X_{eff}, as we have already done for d and $\frac{1}{\alpha_1}$ in Section 3.3.1, it is concluded that $\frac{C_d}{C_{SI}} = \frac{X_{eff}}{d} << 1$. For this reason, equation (3.94) is replaced with

$$C_{inv} = \left(C_{WI}^{-1} + C_{SI}^{-1}\right)^{-1}. \tag{3.96}$$

This expression is analogous to the case of two capacitors in series, as shown in Figure 3.15. This C_{inv} model is formulated in terms of charge densities: Q_{inv}, $Q_{inv,1}$, $Q_{inv,2}$, Q_{depl}, and the relationship between C_{inv} and gate voltage is not explicitly given. Derived from equation (3.58), (3.59) and (3.61), charge densities are related to the gate voltage as

$$V_{gs} = V_{FB} + \phi_s + \frac{Q_{depl} + Q_{inv}}{C_{ox}}, \tag{3.97}$$

$$Q_{inv} = \int C_{inv} d\phi_s , \tag{3.98}$$

and

$$Q_{depl} = \int C_d d\phi_s . \tag{3.99}$$

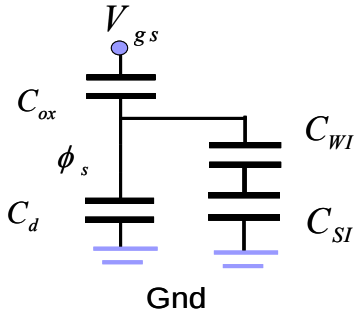

Figure 3.15
Inversion layer capacitance C_{inv} modeled as C_{WI} and C_{SI} in series.

Using equations (3.98)-(3.97) and the C_{inv} model in (3.96), the capacitance and charge density variations with gate voltage can be calculated. Figure 3.16 shows the inversion charge density calculation from the quantum mechanical C_{inv} point of view. Figure 3.17 illustrates the two components of C_{WI} and C_{SI} in the quantum mechanical

C_{inv} model. In Figure 3.17, C_{WI} is dominant in the weak inversion region, and C_{SI} is the major contributor to C_{inv} in strong inversion. At lower V_{gs}, the MOSFET is in weak inversion, and the inversion charge is vanishingly small, as shown in Figure 3.16. According to equation (3.91), C_{WI} is proportional to the inversion charge sheet density Q_{inv}, and has a small value in this situation. Since C_{WI} and C_{SI} are connected in series, the relatively small C_{WI} dominates C_{SI} in the total C_{inv}. At higher gate voltage, when the number of inversion charges increases and the channel surface is in strong inversion, C_{WI} becomes larger than C_{SI}, leading to $C_{inv} \approx C_{SI}$. By incorporating the C_{inv} described by equation (3.96) into C_g, the total gate capacitance is obtained as

$$C_g = \frac{1}{C_{ox}^{-1} + \left(C_d + C_{WI}C_{SI}/\left(C_{WI} + C_{SI}\right)\right)^{-1}}. \tag{3.100}$$

The analytical model of total gate capacitance is compared with numerical simulation [32] in Figure 3.18. A significant deviation of C_g from C_{ox} in the strong inversion region (i.e. at high V_{gs}) can be observed in Figure 3.18.

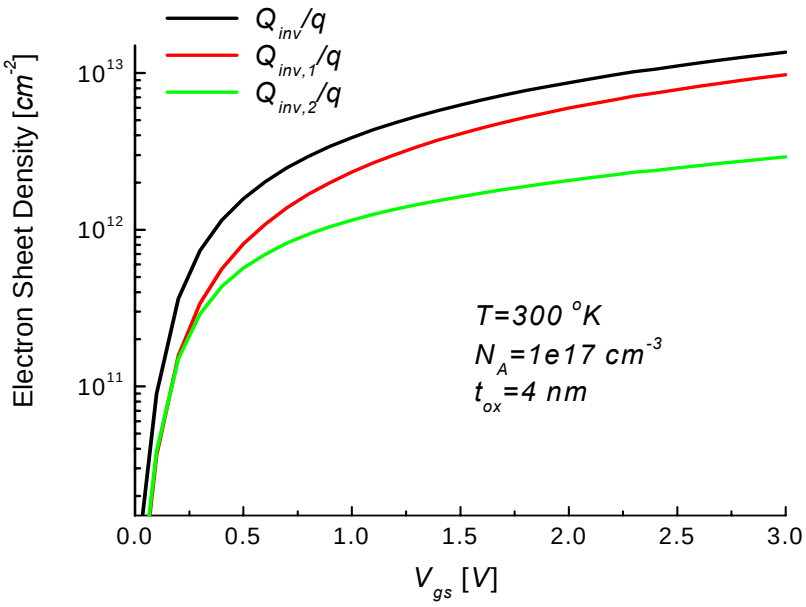

Figure 3.16
Electron sheet density in the channel vs. V_{gs} for n-MOSFET with metal gate workfunction *-4.0 eV*. Both substrate and drain are grounded.

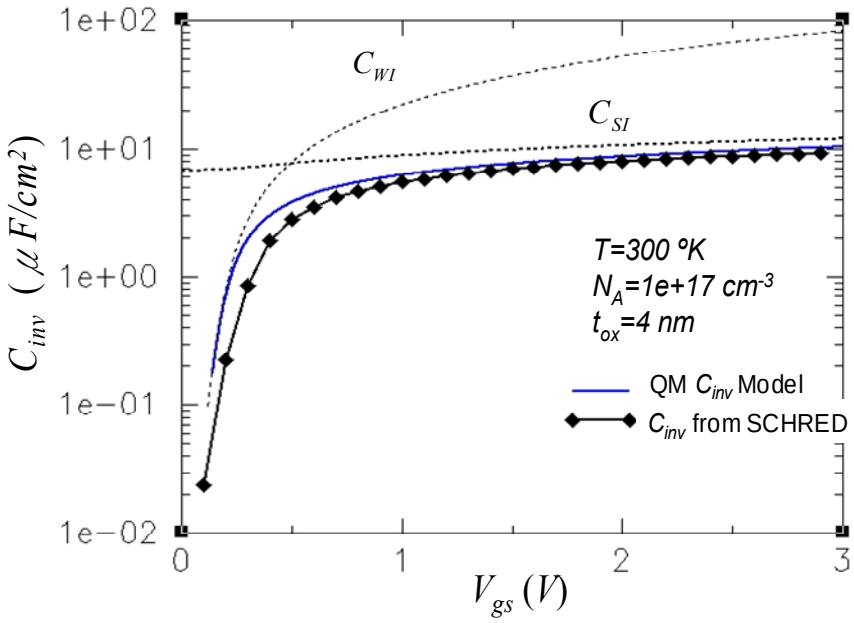

Figure 3.17
Inversion layer capacitance C_{inv} and its components, C_{WI} and C_{SI}. An n-MOSFET with metal gate of *-4.0 eV* workfunction is considered here. Both substrate and drain are grounded. The C_{inv} model is compared with simulation results from SCHRED [32].

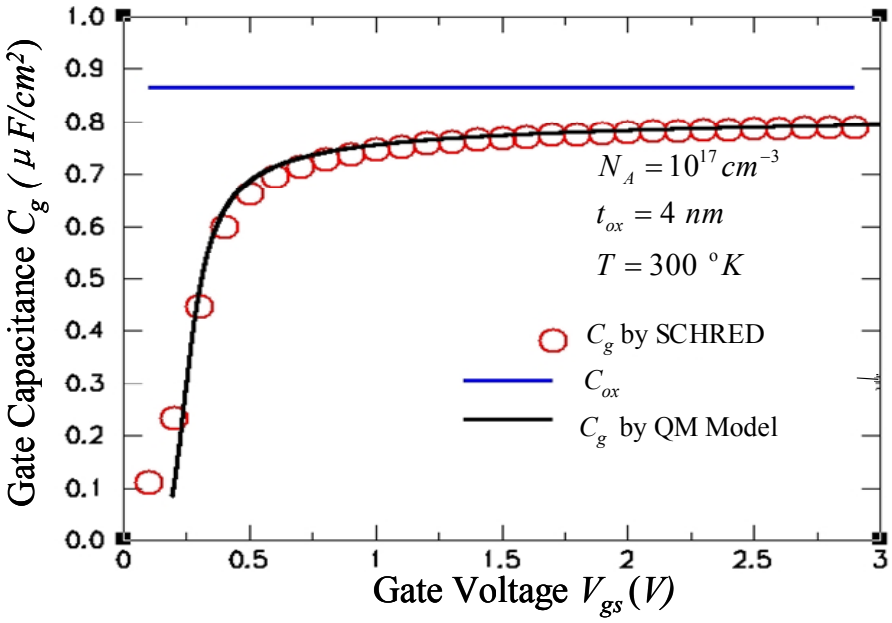

Figure 3.18
Total gate capacitance C_g dependency on gate voltage V_{gs}, validated by SCHRED [32].

3.5 Conclusion

In this chapter, the quantization effects on carrier distribution are studied. The carrier energy quantization leads to discrete subbands in the inversion layer of a MOSFET. Electrons are grouped into two conduction band valleys, owing to different effective masses. In each valley, the absolute majority of electrons is located on the lowest subband. By simultaneously solving the Schrödinger and the Poisson equations, the analytical expressions for energy levels of two lowest subbands are obtained. The relationship between charge density and gate bias in MOSFETs is then derived based on the subband energy levels. Results from the quantum mechanical model show that

classical theory greatly overestimates the value of inversion layer capacitance. This overestimation in turn results in the overestimation of gate capacitance in short-channel MOSFETs.

CHAPTER 4

QUANTUM MECHANICAL MOSFET MODEL

4.1 Introduction and Background

The presence of energy quantization significantly influences the manner in which electrons respond to the bias at electrodes. Levitated electron energy levels effectively reduce the inversion charge density, requiring extra gate voltage for the threshold condition. In addition, gate control of inversion charges is weakened, which is specified by the reduced inversion layer capacitance in the gate capacitance model considering quantization effect. Hence, device parameters, such as the threshold voltage and subthreshold swing, need to be adjusted by taking energy quantization into account. As channel lengths are reduced, devices become more susceptible to energy quantization. Suppression of SCEs is even more challenging, as gate control is weakened by both energy quantization and two-dimensional geometry. For this reason, QMEs must be considered in addition to the two-dimensional electrostatic potential distribution in the analysis of short-channel devices. By including key QMEs on device parameters, a comprehensive *I-V* characteristics model can be achieved that is appropriate for circuit simulation.

In the remainder of this chapter, threshold voltage and subthreshold swing models for long-channel MOSFETs are developed in Section 4.2 and 4.3, respectively. In Section 4.4, the quantization effect on SCEs is investigated. Section 4.5 develops an *I-V* characteristic model integrating key QMEs. An example of applying the developed model to study a CMOS circuit is given later, resulting in extra delays when QMEs are considered.

4.2 V_{TH} Model for Long-Channel MOSFET

The threshold voltage (V_{TH}) is the most important parameter in MOSFETs. It determines in which region of operation a MOSFET is functioning, namely the subthreshold (off-state) or the superthreshold (on-state) region. In the subthreshold region, there are only a few mobile charges in the channel, and fixed depletion charges are dominant. It can be assumed that the surface potential is determined solely by depletion charges. In the superthreshold region, there is a considerable amount of mobile charges due to surface inversion in the channel, and significant current can flow from the source to the drain. Inversion charges become dominant, while the number of depletion charges only varies slightly with gate voltage.

In classical models, the threshold voltage $V_{TH,CL}$ is defined as the gate voltage corresponding to the condition that the surface potential ϕ_s equals $2\phi_B$, where ϕ_B is the Fermi potential defined as

$$\phi_B = \frac{kT}{q}\ln\left(\frac{N_A}{n_i}\right). \tag{4.1}$$

As a consequence, the inversion charge density at the surface, (see equation (3.65)),

$$n(x=0) = \frac{n_i^2}{N_A}\exp\left(\frac{q\phi_s}{kT}\right) \tag{4.2}$$

is equal to the doping concentration at the threshold

$$n(x=0)\Big|_{V_{gs}=V_{TH,CL}} = N_A\,. \tag{4.3}$$

Thus, beyond this point, the channel surface is inverted [85].

The lowest subband energy level is well above the bottom of the conduction band when energy quantization is considered. Therefore, the density of inversion charges is

much lower than with the classical assumption, at the condition of surface potential $\phi_s = 2\phi_B$. In order to invert the channel, the band bending required is larger than $2\phi_B$. The extra gate voltage to generate the band bending can be considered as the threshold voltage shift. This becomes the quantum correction to the classical model for threshold voltage [45, 65].

The threshold condition of $\phi_s = 2\phi_B$ is not appropriate in the quantum model. Instead, the threshold condition is defined by counting inversion charges in the channel. For consistency with the classical model, the threshold voltage in the quantum mechanical model is defined as the gate voltage at which the inversion charge reaches the amount predicted by classical theory at $\phi_s = 2\phi_B$. This change can be expressed as

$$Q_{inv,QM}\left(V_{gs} = V_{TH,QM}\right) = Q_{inv,CL}\left(V_{gs} = V_{TH,CL}\right) \tag{4.4}$$

where $Q_{inv,QM}$ and $Q_{inv,CL}$ (both in [C/cm^2]) are the inversion charge number per unit area calculated by quantum and classical theory, respectively, and the quantum and classical threshold voltages are denoted as $V_{TH,QM}$ and $V_{TH,CL}$, respectively. The shift caused by the quantization effect is given by

$$\Delta V_{TH,shift} = V_{TH,QM} - V_{TH,CL}. \tag{4.5}$$

In the classical model, the inversion charge is given by

$$Q_{inv,CL} = \int_0^\infty \frac{qn_i^2}{N_A} \exp\left(-\frac{q\phi}{kT}\right) dx, \tag{4.6}$$

where the potential ϕ can be obtained from the one-dimensional Poisson equation as

$$\frac{d^2\phi(x)}{dx^2} = \frac{qN_A(x)}{\varepsilon_{Si}} + \frac{qn_i^2}{\varepsilon_{Si}N_A} \exp\left(-\frac{q\phi(x)}{kT}\right). \tag{4.7}$$

A rigorous derivation (See Appendix D) gives the inversion charge sheet density as,

$$Q_{inv,CL}(\phi_s) = \frac{\sqrt{2}kT\varepsilon_{Si}}{qL_D}\left\{\left[\beta\phi_s + \left(\frac{n_i}{N_A}\right)^2 \exp(\beta\phi_s)\right]^{1/2} - [\beta\phi_s]^{1/2}\right\}, \tag{4.8}$$

where $\beta = q/kT$ and L_D is the Debye Length (in [cm]) defined as

$$L_D = \sqrt{\frac{\varepsilon_{Si}}{qN_A\beta}}. \tag{4.9}$$

At the onset of threshold, the surface potential in the equation (4.8) can be replaced by $2\phi_B$. Since $\beta\phi_s >> \left(\frac{n_i}{N_A}\right)^2 \exp(\beta\phi_s)$ in the subthreshold region, applying the Taylor expansion $(1+x)^{1/2} \approx 1 + \frac{1}{2}x$, $x << 1$ to the above equation leads to

$$Q_{inv,CL}(\phi_s) = \frac{\sqrt{2}kT\varepsilon_{Si}}{qL_D}[\beta\phi_s]^{1/2}\left[\frac{1}{2}\left(\frac{n_i}{N_A}\right)^2 \frac{\exp(\beta\phi_s)}{\beta\phi_s}\right], \tag{4.10}$$

which can be simplified as

$$Q_{inv,CL}(\phi_s) = \frac{\sqrt{2}kT\varepsilon_{Si}}{2qL_DN_A}\left[\frac{n_i^2}{N_A}\frac{\exp(\beta\phi_s)}{[\beta\phi_s]^{1/2}}\right]. \tag{4.11}$$

With $n(x) = \frac{n_i^2}{N_A}\exp\left(\frac{q\phi(x)}{kT}\right)$, the inversion charge density becomes

$$Q_{inv,CL}(\phi_s) = \frac{q\sqrt{2}}{2}L_D\frac{n(0)}{[\beta\phi_s]^{1/2}}. \tag{4.12}$$

In the classical model, at the threshold, where the surface potential is $\phi_s = 2\phi_B$,

$$Q_{inv,CL}\Big|_{V_{gs}=V_{TH}} = \frac{qL_D^2}{d_{cl}}N_A \tag{4.13}$$

with

$$d_{cl} = \sqrt{\frac{2\varepsilon_{Si}\left(2\phi_B\right)}{qN_A}}\,. \tag{4.14}$$

In contrast, using the quantum-mechanical model results in the inversion charge sheet density given as

$$Q_{inv,1} = qN_{C2}\exp\left(\frac{E_F - E_{1,1}}{kT}\right), \tag{4.15}$$

where

$$N_{C2} = \frac{kTg_1 m_{d1}^*}{\pi\hbar^2} \tag{4.16}$$

is the 2-D state charge sheet density, and g_1 is the degeneracy of the energy subband. From the band diagram, $E_{1,1} - E_F$ is determined by the band-bending and the quantized energy level as

$$E_{1,1} - E_F = q\phi_B + \frac{E_g}{2} - q\phi_s + E_{1,1}\,, \tag{4.17}$$

where $E_{1,1}$ is energy of the lowest subband with respect to the conduction band minimum. Substituting equation (4.17) into equation (4.15) leads to

$$-kT\ln\left(\frac{Q_{inv,1}}{qN_{C2}}\right) = q\phi_B + \frac{E_g}{2} - q\phi_s + E_{1,1}\,, \tag{4.18}$$

where the surface potential is given by

$$\phi_s = \frac{qN_A d^2}{2\varepsilon_{Si}}\,, \tag{4.19}$$

the subband energy level is given by

$$E_{1,1} = \frac{9q^2 N_A d}{4\varepsilon_{Si}\alpha_1}\,, \tag{4.20}$$

and

$$\alpha_1 = \left[\frac{3m^* q^2}{2\varepsilon_{Si}\hbar^2} N_A d_{qm} \right]^{1/3} \tag{4.21}$$

as derived in Chapter 3. Inversion charges are neglected in equations (4.19) and (4.20) because at threshold, inversion charges are significantly lower than depletion charges and, consequently, subband energy levels and the surface potential are primarily set by depletion charges. This validates the simplification that is used in equations (4.19) and (4.20). Using the surface potential and the lowest energy level expressions, equation (4.18) changes to

$$\frac{q^2 N_A}{2\varepsilon_{Si}} \left(d_{qm}{}^2 - \frac{9}{2\alpha_1} \right) = q\phi_B + \frac{E_g}{2} + kT \ln\left(\frac{Q_{inv,1}}{qN_{C2}} \right), \tag{4.22}$$

where the d_{qm} (in [cm]) is the depletion depth in the quantum-mechanical model. Correspondingly, similar analysis with classical physics leads to [10]

$$\frac{q^2 N_A}{2\varepsilon_{Si}} d_{cl}{}^2 = q\phi_B + \frac{E_g}{2} + kT \ln\left(\frac{N_A}{N_C} \right), \tag{4.23}$$

where d_{cl} (in [cm]) is the depletion depth in the classical model, and N_C (in [cm^{-3}]) is the density-of-states for the three-dimensional electron gas. Thus, the relationship between d_{qm} and d_{cl} is given by the expression

$$\frac{q^2 N_A}{2\varepsilon_{Si}} \left(d_{qm}{}^2 - \frac{9 d_{qm}}{2\alpha_1} \right) = \frac{q^2 N_A}{2\varepsilon_{Si}} d_{cl}{}^2 + kT \ln\left(\frac{Q_{inv,1}}{qN_A} \frac{N_C}{N_{C2}} \right). \tag{4.24}$$

Noticing that $L_D = \sqrt{\frac{\varepsilon_{Si}}{qN_A\beta}}$ at threshold, equation (4.24) can be written as

$$\left(d_{qm}{}^2 - \frac{9 d_{qm}}{2\alpha_1} \right) = d_{cl}{}^2 + 2L_D{}^2 \ln\left(\frac{Q_{inv,1}}{qN_A} \frac{N_C}{N_{C2}} \right). \tag{4.25}$$

Equation (4.25) indicates the different depletion depths predicted by classical and quantum mechanical models. Although an explicit expression cannot be derived directly from equation (4.25), it can be further simplified based on several observations on the relative magnitudes of d_{qm}, α_1^{-1}, and L_D.

First, from the previous discussion, the magnitude of α_1^{-1} is only a few nanometers, and d_{qm} is the depletion depth, $\frac{1}{\alpha_1} << d_{qm}$. We can approximate $\left(d_{qm}{}^2 - \frac{9 d_{qm}}{2\alpha_1} \right)$ by $\left(d_{qm} - \frac{9}{4\alpha_1} \right)^2$ using a Taylor series expansion.

Second, the difference between the classical depletion depth and quantum mechanical depletion depth is small compared with the depletion depth itself, $\frac{d_{qm} - d_{cl}}{d_{cl}} << 1$. Hence, equation (4.21) can be written as

$$\alpha_1 = \left[\frac{3 m^* q^2}{2 \varepsilon_{Si} \hbar^2} N_A d_{cl} \right]^{1/3}. \tag{4.26}$$

Third, the Debye length L_D is negligibly small compared with the depletion depth. In the term $2 L_D{}^2 \ln\left(\frac{Q_{inv,1}}{q N_A} \frac{N_C}{N_{C2}} \right)$ in equation (4.25), $Q_{inv,1} \sim \frac{q L_D{}^2}{d_{cl}} N_A$, N_C, and N_{C2} depend on temperature only; L_D and d_{cl} vary with the substrate doping. At room temperature and substrate doping concentration in the range $1 \times 10^{15} \sim 1 \times 10^{20}\ cm^{-3}$, $2 L_D{}^2 \ln\left(\frac{L_D{}^2}{d_{cl}} \frac{N_A}{N_{C2}} \right)$ is no more than 5% of $d_{cl}{}^2$, as shown in Figure 4.1. Obviously, the term $2 L_D{}^2 \ln\left(\frac{Q_{inv,1}}{q N_A} \frac{N_C}{N_{C2}} \right)$ can be safely ignored in equation (4.25).

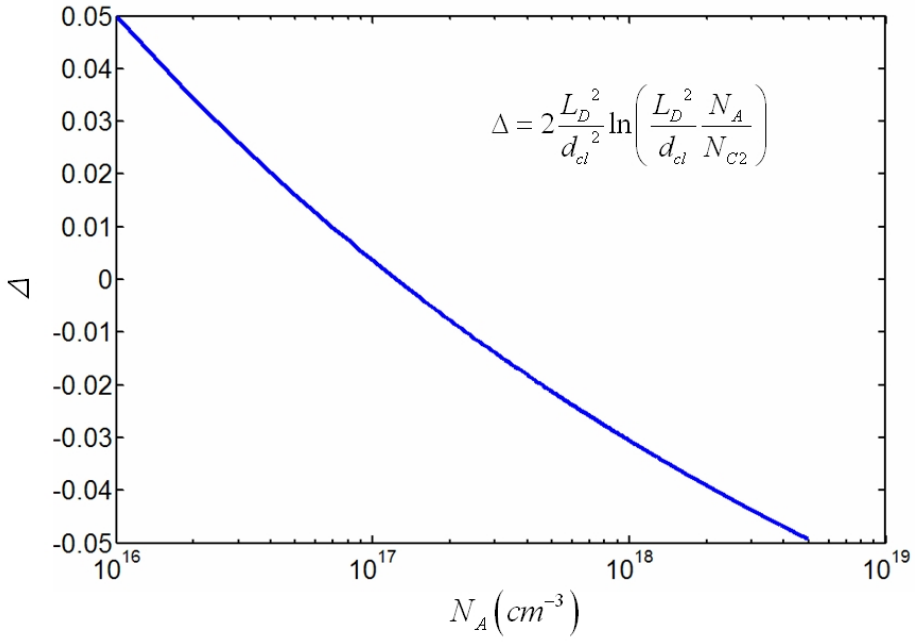

Figure 4.1

Ratio of $2L_D^{\ 2}\ln\left(\frac{L_D^{\ 2}}{d_{cl}}\frac{N_A}{N_{C2}}\right)$ and $d_{cl}^{\ 2}$ as a function of doping.

With these approximations, the expression for the quantum mechanical depth can be simplified as

$$d_{qm} \approx d_{cl} + \frac{9}{4\alpha_T} \tag{4.27}$$

where α_T is the value of α_1 at threshold

$$\alpha_T = \left[\frac{3m^* q^2}{2\varepsilon_{Si}\hbar^2} N_A d_{cl}\right]^{1/3} \tag{4.28}$$

Equation (4.27) implies that a greater depletion depth than what is predicted in classical analysis is required for the threshold condition. The extra depletion depth increases the

surface potential, allowing adequate inversion charges in the channel. The increase in the surface potential $\Delta\phi_s$ (in [V]) caused by the quantization effect is then given by

$$\Delta\phi_s = \frac{qN_A}{2\varepsilon_{Si}}\left(d_{cl} + \frac{9}{4\alpha_T}\right)^2 - 2\phi_B \,. \tag{4.29}$$

Using this, the gate voltage shift can be obtained from the surface potential difference as

$$\Delta V_{gs} = \left.\frac{dV_{gs}}{d\phi_s}\right|_{\phi_s=2\phi_B} \Delta\phi_s \,, \tag{4.30}$$

where

$$\left.\frac{dV_{gs}}{d\phi_s}\right|_{\phi_s=2\phi_B} = 1 + \frac{1}{2C_{ox}}\sqrt{\frac{\varepsilon_{Si}qN_A}{\phi_B}} \,. \tag{4.31}$$

From equations (4.29) and (4.30), the threshold voltage shift is obtained as

$$\Delta V_{TH,shift} = \left(1 + \frac{1}{2C_{ox}}\sqrt{\frac{\varepsilon_{Si}qN_A}{\phi_B}}\right)\left[\frac{qN_A}{2\varepsilon_{Si}}\left(d_{cl} + \frac{9}{4\alpha_T}\right) - 2\phi_B\right], \tag{4.32}$$

where

$$\alpha_T = \left[\frac{3m^* q^2}{2\varepsilon_{Si}\hbar^2} N_A d_{cl}\right]^{1/3} , \tag{4.33}$$

and

$$d_{cl} = \sqrt{\frac{2\varepsilon_{Si}(2\phi_B)}{qN_A}} \,. \tag{4.34}$$

This result (4.32) agrees well with measurement data [29], as shown in Figure 4.2.

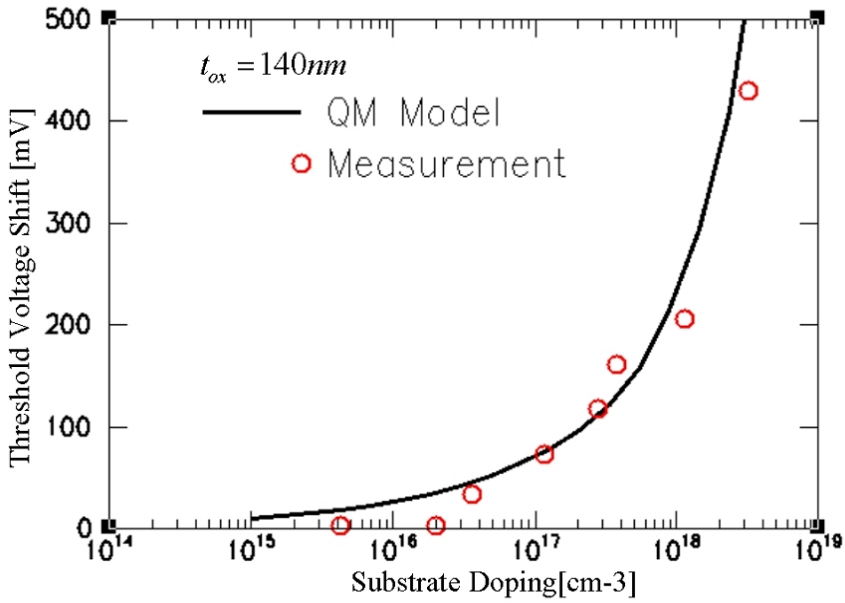

Figure 4.2
Comparison of quantum mechanical threshold voltage shift model with measurement data [29].

4.3 *S* Model for Long-Channel MOSFET

At $V_{gs} < V_{TH}$, the MOSFET is in its off-state and only conducts a leakage current from the source to the drain through the weakly inverted channel surface. This is referred to as the subthreshold current. The subthreshold current is the key leakage source in MOSFETs and, hence, the reduction of subthreshold current is a major concern in device design [10, 69, 86]. It is known that subthreshold current increases exponentially with gate voltage [85]. To measure the changing rate of subthreshold current with respect to the gate voltage, the subthreshold swing (*S*) is defined as the gate voltage swing needed to change the subthreshold drain current by a decade. By this definition, S is given by

$$S = \left(\frac{\partial \ln I}{\partial V_{gs}} \right)^{-1} \ln(10). \tag{4.35}$$

Physically speaking, the magnitude of subthreshold swing reflects the gate control of inversion charges in the channel in subthreshold. The smaller value of S devices show better subthreshold behavior reflected as lower leakage current.

Since the subthreshold drain current is proportional to the total amount of mobile charges diffusing to the drain $I \propto Q_{inv}$, then

$$\frac{\partial \ln I}{\partial V_{gs}} = \frac{\partial \ln Q_{inv}}{\partial V_{gs}}. \tag{4.36}$$

Classical physics describes the inversion charge as

$$Q_{inv} = \int_0^{\infty} \frac{n_i^2}{N_A} \exp\left(-\frac{q\phi(x)}{kT} \right) dx. \tag{4.37}$$

Thus, the subthreshold swing can be expressed as (See Appendix D),

$$S = \frac{1}{Q_{inv}} \frac{\partial Q_{inv}}{\partial \phi_s} \frac{\partial \phi_s}{\partial V_{gs}} \ln(10) = \frac{kT}{q} \ln(10) \frac{\partial \phi_s}{\partial V_{gs}}. \tag{4.38}$$

For a long-channel MOSFET, $\frac{\partial \phi_s}{\partial V_{gs}} = \left(1 + \frac{C_d}{C_{ox}} \right)$, and the subthreshold swing is simplified as

$$S = \left(1 + \frac{C_d}{C_{ox}} \right) \frac{kT}{q} \ln(10). \tag{4.39}$$

This expression can be described by the charge sheet model in which inversion charges at the bulk surface are determined by the ratio of the gate oxide and depletion layer capacitances.

In a quantum-mechanical model, the peak of inversion charges is displaced from the bulk surface, which reduces gate control. The subthreshold swing can be calculated from the expressions of inversion charges in equation (3.80) and (3.81). The total inversion charge amount is obtained by accounting for electrons in the two valleys as

$$Q_{inv} = Q_{inv,1} + Q_{inv,2} . \tag{4.40}$$

By definition,

$$S = \ln(10)\frac{1}{Q_{inv}}\left(\frac{\partial\left(Q_{inv,1}+Q_{inv,2}\right)}{\partial V_{gs}}\right)^{-1} \tag{4.41}$$

which can be written as

$$S = \ln(10)\frac{1}{Q_{inv}}\frac{\partial V_{gs}}{\partial \phi_s}\left(\frac{\partial Q_{inv,1}}{\partial \phi_s}+\frac{\partial Q_{inv,2}}{\partial \phi_s}\right)^{-1} . \tag{4.42}$$

In the same manner as equation (3.88) is derived, equation (4.42) becomes

$$S = \ln(10)\frac{kT}{q}\frac{\partial V_g}{\partial \phi_s}\left(1-\frac{3}{2\varepsilon_{Si}}C_d\left(\frac{Q_{inv,1}}{Q_{inv}\alpha_1}+\frac{Q_{inv,2}}{Q_{inv}\alpha_2}\right)\right)^{-1} \tag{4.43}$$

In the subthreshold region, the variational parameters α_1 and α_2 can be simplified as

$$\alpha_1 = \left[\frac{3m_1 q}{2\varepsilon_{Si}\hbar^2}Q_{depl}\right]^{1/3} \tag{4.44}$$

and

$$\alpha_2 = \left[\frac{3m_2 q}{2\varepsilon_{Si}\hbar^2}Q_{depl}\right]^{1/3} . \tag{4.45}$$

Letting

$$X_{CS} = \frac{3}{2}\left(\frac{Q_{inv,1}}{Q_{inv}\alpha_1}+\frac{Q_{inv,2}}{Q_{inv}\alpha_2}\right), \tag{4.46}$$

the subthreshold swing is then given by

$$S = 2.3\frac{kT}{q}\left(1+\frac{C_d}{C_{ox}}\right)\left(1-\frac{X_{CS}C_d}{\varepsilon_{Si}}\right)^{-1}. \tag{4.47}$$

If the effective depletion layer and the effective oxide layer capacitances are defined as

$$C_{d,eff} = \frac{\varepsilon_{si}}{d_{qm} - X_{CS}} \tag{4.48}$$

and

$$C_{ox,eff} = \left(\frac{1}{C_{ox}} + \frac{X_{CS}}{\varepsilon_{si}}\right)^{-1}, \tag{4.49}$$

then, the subthreshold swing can be written as

$$S = \ln(10)\frac{kT}{q}\left(1+\frac{C_{d,eff}}{C_{ox,eff}}\right), \tag{4.50}$$

which is a similar form to that of equation (4.39). The only significant difference between the two models is that the inversion “charge sheet" is displaced from the surface into the substrate by the depth of X_{CS}, where X_{CS} can be considered as the average inversion charge depth. Conforming with the classical charge-sheet model, the EOT can be effectively visualized by an increase of $\frac{X_{CS}\varepsilon_{ox}}{\varepsilon_{Si}}$. As shown in Figure 4.3, the increased EOT varies from *0.4 nm~ 1.1 nm* in the doping concentration range from 10^{16} cm^{-3} to 5×10^{18} cm^{-3}. As for modern MOSFETs, the effectively increased EOT by energy quantization is considerable, compared with the EOT of the insulation layer, which is less than *1.5 nm* in the sub-*90 nm* technologies [10, 36-38, 67].

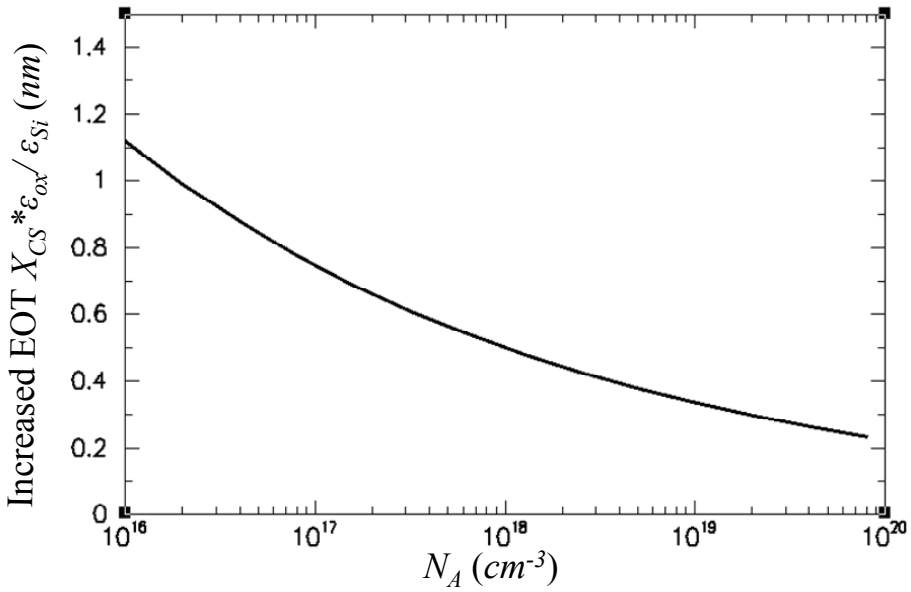

Figure 4.3
Increased EOT from energy quantization of inversion charges.

4.4 SCE Model for Short-Channel MOSFET

As the channel length is reduced, device threshold voltage decreases and the subthreshold swing increases because of drain induced barrier lowering (DIBL). To study the dependence of threshold voltage on channel length and other device parameters, it is essential to develop an expression for the two-dimensional channel potential distribution in the subthreshold region.

In the subthreshold region, the uniformly doped p-type silicon channel is virtually depleted of mobile carriers. The potential in the depleted region is determined by the 2-D Poisson equation,

$$\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} = \frac{qN_A}{\varepsilon_{Si}}. \tag{4.51}$$

Taking the channel intrinsic level under flat band condition as the reference potential, the boundary conditions are described below [61].

Left:

$$\begin{aligned}\phi(x,0) &= \phi_b \quad 0 \le x \le x_j \\ &= \frac{qN_A}{2\varepsilon_{Si}}\left(\alpha_s x - x_j\right)^2 + \phi_b - \left(\alpha_s x - x_j\right)\sqrt{\frac{2qN_A\phi_b}{\varepsilon_{Si}}}\end{aligned} \tag{4.52}$$

Right:

$$\begin{aligned}\phi(x,L) &= \phi_b + V_{ds} \quad 0 \le x \le x_j \\ &= \frac{qN_A}{2\varepsilon_{Si}}\left(\alpha_d x - x_j\right)^2 + \phi_b + V_{ds} - \left(\alpha_d x - x_j\right)\sqrt{\frac{2qN_A\phi_b}{\varepsilon_{Si}}}\end{aligned} \tag{4.53}$$

Top:

$$\left.\frac{\partial\phi}{\partial x}\right|_{(0,y)} - \frac{C_{ox}}{\varepsilon_{Si}}\phi(0,y) = -\frac{C_{ox}}{\varepsilon_{Si}}V_G' \tag{4.54}$$

$$V_G' = V_{gs} - V_{FB} \tag{4.55}$$

Bottom:

$$\left.\frac{\partial\phi}{\partial x}\right|_{(d,y)} = 0 \tag{4.56}$$

where x_j (in [cm]) is the junction depth, ϕ_b (in [V]) is the built-in potential of the one-side abrupt junction between substrate and source/drain, and V_{ds} (in [V]) is the applied drain voltage. The fitting parameters α_s and α_d are used to account for the higher electric field at the source/drain junction corners, which depend only on the drain voltage and can vary as much as *1.12-1.47* for the drain voltage variation of *0-1.5 V* [60, 61].

Equation (4.51) can be solved by superposition and separation of variables, resulting in [60, 61]

$$\phi(x,y) = U(x) + \varphi(x,y), \tag{4.57}$$

where

$$U(x) = \frac{qN_A}{2\varepsilon_{Si}}x^2 - \frac{qN_A d}{\varepsilon_{Si}}x + V_G' - \frac{qN_A d}{C_{ox}} \tag{4.58}$$

and

$$\begin{aligned}\varphi(x,y) = \sum_{n=1}^{\infty}&\left(\sin(\lambda_n x) + \frac{\varepsilon_{Si}\lambda_n}{C_{ox}}\cos(\lambda_n x)\right),\\ &\times\left\{D_n\exp(-\lambda_n y) + C_n\exp\left[-\lambda_n(y-L)\right]\right\}\end{aligned} \tag{4.59}$$

with

$$D_n \approx \frac{\frac{1}{\lambda_n}\left[\phi_b - V_G' + \frac{qN_A}{\varepsilon_{Si}\lambda_n^{\,2}} + B_n(x_j,\phi_b)\right]}{\frac{1}{2}\left[\frac{\varepsilon_{Si}}{C_{ox}} + d\left(1 + \frac{\varepsilon_{si}^{\,2}\lambda_n^{\,2}}{C_{ox}^{\,2}}\right)\right]}, \tag{4.60}$$

and

$$\begin{aligned}B_n(x_j,\phi_b) = 0 \quad & d \le x_j\\ = \cos(\lambda_n d)&\frac{qN_A\alpha_s}{C_{ox}}\left(1 + \frac{C_{ox}^{\,2}}{\varepsilon_{Si}^{\,2}\lambda_n^{\,2}}\right)(\alpha_s d - d_s - x_j)\\ &+\left\{\sin(\lambda_n x_j) - \frac{C_{ox}}{\varepsilon_{Si}\lambda_n}\cos(\lambda_n x_j)\right\}\cdot\\ &\frac{qN_A}{C_{ox}}\left(\frac{\alpha_s}{\lambda_n} - x_j\lambda_n(1-\alpha_s)\left\{\frac{(1-\alpha_s)}{2}x_j + d_s\right\}\right)\\ &+\left\{\frac{C_{ox}}{\varepsilon_{Si}\lambda_n}\sin(\lambda_n x_j) + \cos(\lambda_n x_j)\right\}\frac{qN_A\alpha}{C_{ox}}(d_s + x_j - \alpha_s x_j) \quad d > x_j\end{aligned} \tag{4.61}$$

Here, d_s and d_d are the depletion depth at the source and the drain, respectively, as described by

$$d_s = \sqrt{\frac{2\varepsilon_{Si}\phi_b}{qN_A}}, \tag{4.62}$$

and

$$d_d = \sqrt{\frac{2\varepsilon_{Si}\left(\phi_b + V_{ds}\right)}{qN_A}}. \tag{4.63}$$

The C_n is calculated by replacing D_n with C_n, $B_n\left(\phi_b\right)$ with $B_n\left(\phi_b + V_{ds}\right)$, d_s with d_d, ϕ_b with $\phi_b + V_{ds}$, and α_s with α_d in equation (4.60) and (4.61). The eigenvalues of λ_n are from the solutions of

$$\tan\left(\lambda_n d\right) = \frac{C_{ox}}{\varepsilon_{Si}\lambda_n}. \tag{4.64}$$

The most important terms in the φ series (4.59) are the lowest ones associated with λ_1, because the exponential terms decay too fast in the higher order series to have a significant effect on the channel potential. The term $L\lambda_1$ is a fundamental measurement of DIBL in a MOSFET. For $L\lambda_1 >> 1$, the MOSFET will behave ideally like a 1-D long channel device, but for small values of $L\lambda_1$ there will be strong SCEs. From the previous analysis of the long-channel devices, the quantization effect tends to increase the depletion depth at threshold as

$$d = d_{cl} + \frac{9}{4\alpha_T}. \tag{4.65}$$

The larger depletion depth in the quantum mechanical model leads to the decrease of λ_1, as shown in Figure 4.4, which means worse gate controllability when compared to the classical model.

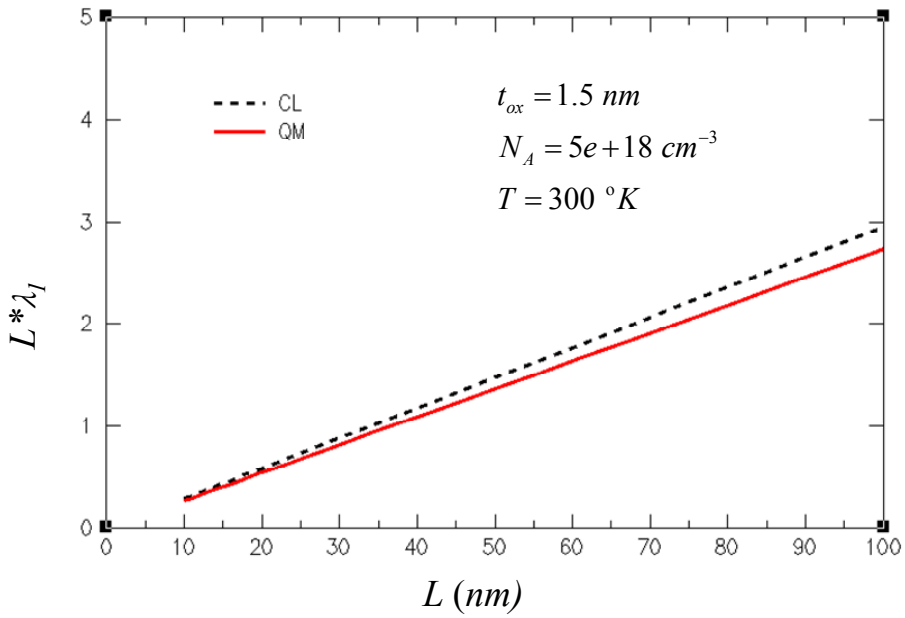

Figure 4.4
Comparison of $L\lambda_1$ magnitudes in the quantum mechanical model and the classical model as a function of L.

4.4.1 Short-Channel V_{TH} Model

In the classical model [60, 61], threshold voltage of a short-channel MOSFET is defined as the gate voltage at which the minimum surface potential in the channel is the same as channel potential at the threshold for a long-channel device, i.e. at threshold,

$$\phi(0, y_m) = 2\phi_B \tag{4.66}$$

where y_m is the minimum surface potential point in the channel. Using this definition, the short-channel threshold roll-off in the classical model [61] is given by

$$\Delta V_{TH} = \frac{4AB\exp\left(-\lambda_1 L/2\right)}{1+\frac{C_{ox}}{C_d}+\frac{C_d}{C_{ox}}\left(\lambda_1 d\right)^2 - 2\exp\left(\frac{-\lambda_1 L}{2}\right)\left[\frac{A}{B}+\frac{B}{A}\right]}, \tag{4.67}$$

where

$$A = \left[\phi_b + V_{ds} - 2\phi_B + 4\phi_B\left(\frac{1}{\left(\lambda_1 d\right)^2}-\frac{C_d}{C_{ox}}\right)+B_1\left(x_j, \phi_b + V_{ds}\right)\right]^{1/2}, \tag{4.68}$$

and

$$B = \left[\phi_b - 2\phi_B + 4\phi_B\left(\frac{1}{\left(\lambda_1 d\right)^2}-\frac{C_d}{C_{ox}}\right)+B_1\left(x_j, \phi_b\right)\right]^{1/2}. \tag{4.69}$$

From the previous discussion, in the subthreshold region, the depletion depth is increased by $9/4\alpha_T$ owing to QME. Moreover, the finite inversion layer capacitance induces the quantum-mechanical adjustment for the classical model as

$$\frac{\varepsilon_{si}}{d_{qm}} \to C_d \tag{4.70}$$

and

$$2\phi_B + \Delta\phi_s \to 2\phi_B . \tag{4.71}$$

Figure 4.5 shows the effective threshold voltage including the long-channel threshold voltage and the short-channel threshold voltage roll-off, which is validated against numerical simulator ISE TCAD [87].

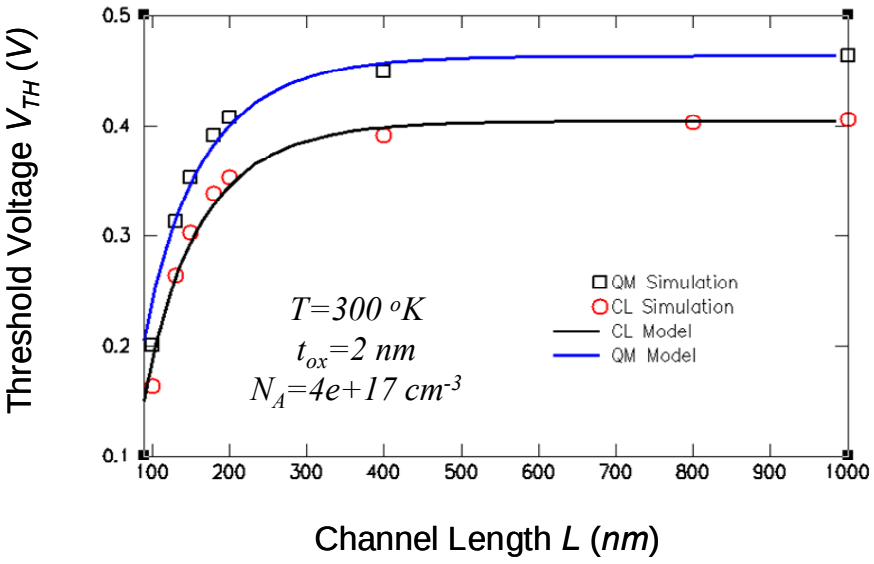

Figure 4.5
Comparison of quantum mechanical threshold voltage model with simulation data from ISE TCAD as a function of L [87].

4.4.2 Short-Channel *S* Model

In the classical model, charges in the channel are determined by the surface potential, and gate controllability is reflected by the surface potential change with change in gate voltage. Assuming that the short-channel MOSFET satisfies

$$V_{gs} - V_{TH} = \theta_S \left(\phi_s - 2\phi_B \right) \tag{4.72}$$

where the V_{TH} is the short channel threshold voltage and θ_S is given by

$$\frac{1}{\theta_S} = \left. \frac{\partial \phi_S}{\partial V_{gs}} \right|_{V_{gs}=V_{TH}}, \tag{4.73}$$

the subthreshold swing S is then given by

$$S = \frac{kT}{q}\theta_S \ln(10). \tag{4.74}$$

The surface potential for the short channel MOSFET is given as [61]:

$$\phi_s\big|_{y=\frac{L}{2}} = U(0) + \frac{\varepsilon_{Si}\lambda_1}{C_{ox}}\exp\left(-\frac{\lambda_1 L}{2}\right)(C_1 + D_1) \tag{4.75}$$

where

$$D_1 = \frac{\frac{1}{\lambda_1}\left[\phi_b - V_G' + \frac{qN_A}{\varepsilon_{Si}\lambda_1^2}\right]}{\frac{1}{2}\left[\frac{\varepsilon_{Si}}{C_{ox}} + d\left(1 + \frac{\varepsilon_{si}^2\lambda_1^2}{C_{ox}^2}\right)\right]} \tag{4.76}$$

and

$$C_1 = \frac{\frac{1}{\lambda_1}\left[\phi_b + V_{ds} - V_G' + \frac{qN_A}{\varepsilon_{Si}\lambda_1^2}\right]}{\frac{1}{2}\left[\frac{\varepsilon_{Si}}{C_{ox}} + d\left(1 + \frac{\varepsilon_{si}^2\lambda_1^2}{C_{ox}^2}\right)\right]}, \tag{4.77}$$

with λ_1 is the smallest λ from

$$\tan\lambda d = \frac{C_{ox}}{\varepsilon_{Si}\lambda}. \tag{4.78}$$

The surface potential can be simplified as

$$\phi_s\big|_{y=\frac{L}{2}} = U(0) + \exp\left(-\frac{\lambda_1 L}{2}\right)\left(\frac{2\left[\phi_b - V_G' + \frac{qN_A}{\varepsilon_{Si}\lambda_1^2}\right] + V_{ds}}{\frac{1}{2}\left[1 + \frac{C_{ox}}{\varepsilon_{Si}}d + \frac{\varepsilon_{Si}d\lambda_1^2}{C_{ox}}\right]}\right). \tag{4.79}$$

A simplified expression for the θ_S can be written as

$$\frac{1}{\theta_S} = \frac{C_{ox}}{C_{ox}+C_d} - \frac{\exp\left(-\lambda_1 L/2\right)}{1+\frac{C_{ox}}{C_d}+\frac{C_d}{C_{ox}}}$$
$$\left\{ \begin{array}{l} \dfrac{2\phi_b + V_{ds} - 2V_{TS} + \dfrac{qN_A}{\varepsilon_{Si}\lambda_1^{\,2}}}{1+\dfrac{C_{ox}}{C_d}+\dfrac{C_d}{C_{ox}}} \bullet \\ \left[\left(\dfrac{C_{ox}}{C_d}+\dfrac{C_d}{C_{ox}}\right)\left(\lambda_1' L + \dfrac{2d'}{d}\right) + \lambda_1' L + \dfrac{4C_d}{C_{ox}}\lambda_1 d^2 \lambda_1'\right] - 2 - 4\dfrac{qN_A}{\varepsilon_{Si}\lambda_1^{\,3}}\lambda_1' \end{array} \right\}, \qquad (4.80)$$

where

$$d' = \frac{1}{\frac{qN_A}{\varepsilon_{Si}}\left(d+\frac{\varepsilon_{Si}}{C_{ox}}\right)}, \qquad (4.81)$$

and

$$\lambda' = -\frac{\lambda_1 d'}{d+\frac{C_{ox}}{\varepsilon_{Si}\lambda}\cos^2\left(\lambda d\right)}. \qquad (4.82)$$

From the previous discussion, the inversion charge density can be calculated as a charge sheet at the average inversion depth X_{CS} by the quantum mechanical model. Therefore, $C_{ox,eff}$ and $C_{d,eff}$ must replace the oxide and depletion capacitances in the classical model, respectively. Moreover, the increased depletion depth at the threshold voltage increases the SCE. Taking the SCE into account, equation (4.72) is modified as

$$V_{gs} - V_{TH,QM} = \theta_S\left(\phi_s - 2\phi_B - \Delta\phi_s\right), \qquad (4.83)$$

in which $\Delta\phi_s$ is the surface potential difference at threshold for classical and quantum mechanical models defined by equation (4.29). In summary, the classical expression in equation (4.80) must be modified by replacing parameters $C_{ox,eff} \to C_{ox}$, $C_{d,eff} \to C_d$, $2\phi_B + \Delta\phi_s \to 2\phi_B$, $V_{TH,QM} \to V_{TH}$, as QMEs are taken into account.

Figure 4.6 and Figure 4.7 show that the new model shows close agreement with the numerical simulation results obtained from Dessis® of ISE TCAD [87] for short-channel MOSFETs with effective channel length down to *20 nm*. For the *20 nm* MOSFET with EOT=*0.5 nm*, QMEs cause *S* to increase from *74 mV/decade* to *105 mV/d*ecade, as shown in Figure 4.7. Meanwhile, with decreasing L, the value of subthreshold swing increases more rapidly when the QMEs are taken into account. Classical models underestimate MOSFETs' susceptibility to SCEs according to the effectively increased EOT from QMEs.

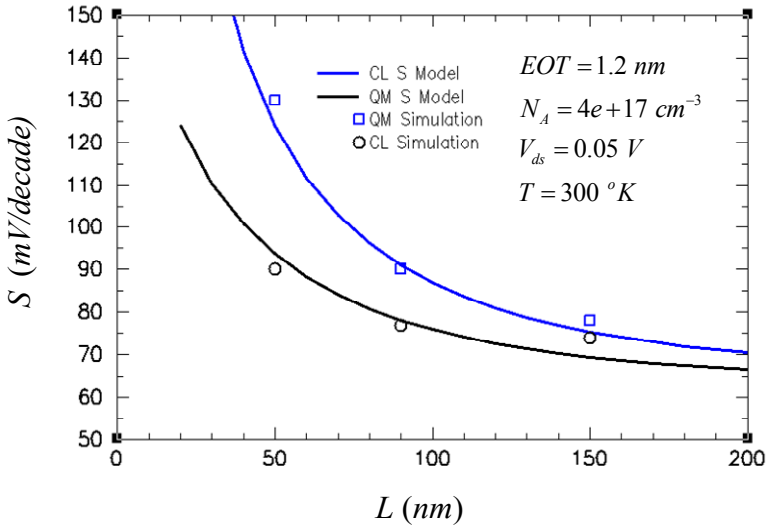

Figure 4.6
S model including QMEs for short-channel devices of EOT=*1.2 nm* as 2004 technology. Simulation data from ISE TCAD [87].

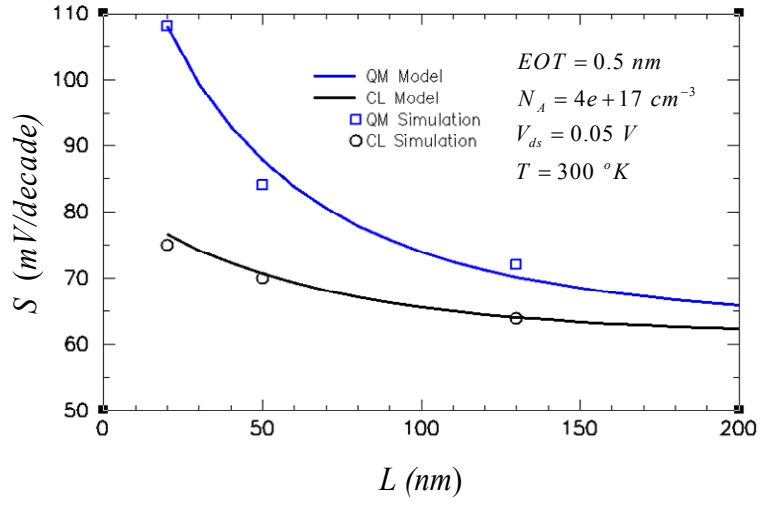

Figure 4.7
S model including QMEs for short-channel devices of EOT=*0.5 nm* as 2008 technology. Simulation data from ISE TCAD [87].

4.5 QME on *I-V* Characteristics

In order to evaluate the QMEs in circuit performance, an *I-V* characteristic model incorporating key QMEs is essential. The *I-V* model integrates two aspects of the charge density profile and carrier transport. Evidence [56, 88] shows that carriers following the quantum mechanical distribution conform with classical transport theory. In the following section, the charge density expression and device parameters with quantum corrections are substituted into the physical transregional model [89] to include QMEs in the MOSFETs' *I-V* characteristic.

4.5.1 Mobility Model

The transport behavior of the carriers in a MOSFET is affected by both the transverse and longitudinal electric fields. The transverse electric field in the vertical direction, which is caused by applied gate bias, has the effect of decreasing carrier mobility. The degraded mobility, denoted by the effective mobility μ_{eff} (in [$cm^2/V{\cdot}s$]), can be described by [89]

$$\begin{aligned} \mu_{eff} &= \mu_0 && V_{gs} < V_{TH}(\phi_c) \\ &= \frac{\mu_0}{1+\theta\left(V_{gs}-V_{TH}(\phi_c)\right)} && V_{gs} > V_{TH}(\phi_c) \end{aligned}, \tag{4.84}$$

where μ_0 (in [$cm^2/V{\cdot}s$]) is the low field mobility, V_{gs} is the voltage applied to gate, ϕ_c (in [V]) is the channel potential given by the potential difference between the electron quasi-Fermi level and bulk Fermi level, $V_{TH}(\phi_c)$ is the threshold voltage along the channel with different channel potential ϕ_c, and θ is the normal-field mobility degradation factor described as

$$\theta = \frac{\mu_0}{2t_{ox}v_{norm}}, \tag{4.85}$$

where v_{norm} is a constant determined as

$$v_{norm} = 2.2 \times 10^9 cm/s. \tag{4.86}$$

The longitudinal field induced by the drain potential causes the velocity saturation effect. As the drain voltage V_{ds} reaches the saturation voltage V_{dsat}, the carrier velocity departs from its linear relationship with the lateral field and becomes the constant saturation velocity v_{sat}. The carrier velocity dependence on the longitudinal field *E(y)* (in [*V/cm*]) is described by [89]

$$v(y) = \frac{\mu_{eff} E(y)}{1 + \frac{E(y)}{E_{Cr}}}, \tag{4.87}$$

where E_{Cr} is the critical field, as determined by

$$\begin{aligned} E_{Cr} &= 2E_{Cr0} && V_{ds} \le V_{dsat} \\ E_{Cr} &= E_{Cr0}\left[2 - \frac{(V_{ds} - V_{dsat})^2}{V_{ds}^2}\right] && V_{ds} > V_{dsat} \end{aligned}, \tag{4.88}$$

where

$$E_{Cr0} = \frac{v_{sat}}{\mu_{eff}}. \tag{4.89}$$

4.5.2 Quantum Mechanical Charge Model

As previously described, gate bias determines the region of operation of the MOS structure. The inversion charge density depends on which region the device is operating.

In the subthreshold, or off, region, the inversion charge density can be written as

$$Q_{inv}\left(V_{gs}\right)=Q_{inv}\big|_{V_{gs}=V_{TH}}\exp\left[\frac{S}{\ln 10}\frac{q}{kT}\left(V_{gs}-V_{TH}\right)\right], \tag{4.90}$$

as derived from equation (4.35). The threshold voltage is given by

$$V_{TH,QM}\left(\phi_c\right)=V_{FB}+\phi_{s,qm}+\phi_c+\frac{\sqrt{2q\varepsilon_{Si}N_A\left(\phi_c+\phi_{s,qm}\right)}}{C_{ox}} \tag{4.91}$$

where $\phi_{s,qm}$ is the surface potential at threshold given by the quantum model as

$$\phi_{s,qm}=2\phi_B+\Delta\phi_{s,qm}=\frac{qN_A}{2\varepsilon_{Si}}\left(d_{cl}+\frac{9}{4\alpha_T}\right)^2. \tag{4.92}$$

Denoting the inversion charge density at threshold by Q_{TH},

$$Q_{TH}=Q_{inv}\left(V_{gs}=V_{TH}\right), \tag{4.93}$$

the inversion charge density in the subthreshold region is expressed as

$$Q_{inv}\left(V_{gs}\right)=Q_{TH}\exp\left[\frac{S}{\ln 10}\frac{q}{kT}\left(V_{gs}-V_{TH,QM}\left(\phi_C\right)\right)\right]. \tag{4.94}$$

In the strong inversion region, the variance of the surface potential is small, and the inversion layer capacitance is nearly a constant. Therefore, from the first order approximation the inversion charge can be written as

$$Q_{inv}\left(\phi_s\right)=Q_{TH}+qC_{inv}\left(\phi_s-\phi_c-\phi_{s,qm}\right). \tag{4.95}$$

From the *C-V* relationship in the previous chapter, the inversion charge density is given as a function of applied gate voltage,

$$Q_{inv}\left(V_{gs}\right)=Q_{TH}+q\frac{\left(C_d+C_{inv}\right)C_{ox}}{C_d+C_{inv}+C_{ox}}\left[V_{gs}-V_{TH,QM}\left(\phi_c\right)\right]. \tag{4.96}$$

In the classical model, the inversion capacitance is much larger than the oxide capacitance, and, thus, ignored in the inversion charge expression. In the quantum model,

because of the energy level quantization, the inversion capacitance is limited by the finite inversion layer thickness. The inversion capacitance is given by

$$C_{inv} = \frac{\varepsilon_{Si}}{X_{eff}}. \tag{4.97}$$

By ignoring the small depletion capacitance and Q_{TH} in equation (4.96), the inversion charge density in the strong inversion region is given by

$$Q_{inv} = C_{inv} \frac{C_{ox}}{C_{inv} + C_{ox}} \left[V_{gs} - V_{TH,QM}(\phi_c) \right]. \tag{4.98}$$

In a summary, a simplified charge model incorporating the quantization effect is obtained as

$$Q_{inv} = Q_{TH} \exp\left\{ \frac{S}{\ln 10} \frac{q}{kT} \left[V_{gs} - V_{TH,QM}(\phi_c) \right] \right\} \tag{4.99}$$

in the weak inversion region and

$$Q_{inv} = q \frac{C_{inv} C_{ox}}{C_{inv} + C_{ox}} \left[V_{gs} - V_{TH,QM}(\phi_c) \right] \tag{4.100}$$

in the strong inversion region. The expressions above indicate that the inversion charge density keeps its exponential dependency on gate voltage at weak inversion and a linear dependency on the gate voltage at strong inversion.

4.5.3 Drain Current in the Triode Region

In the triode region, the channel is strongly inverted and the drain potential is less than the saturation voltage. The drain current can be written as

$$I_{ds,triode} = W Q_{inv}(y) v(y) \tag{4.101}$$

where W is the width of the device, $v(y)$ is the carrier velocity along the channel given by (4.87), and $Q_{inv}(y)$ is the charge in the inversion layer under strong inversion. The inversion layer charge is expressed in the charge sheet model as:

$$Q_{inv} = q\frac{C_{inv}C_{ox}}{C_{inv}+C_{ox}}\left[V_{gs} - V_{TH,QM}(\phi_c)\right]. \tag{4.102}$$

Substituting equation (4.87) into equation (4.101), the drain current can be written as

$$I_{ds,triode} = WQ_{inv}(y)\frac{\mu_{eff}E(y)}{1+\dfrac{E(y)}{E_{Cr}}}. \tag{4.103}$$

The lateral electric field is given by

$$E(y) = \frac{I_{ds,triode}}{WQ_{inv}(y)\mu_{eff} - \dfrac{I_{ds,triode}}{E_{Cr}}} = \frac{d\phi_c}{dy}, \tag{4.104}$$

and the drain voltage referenced to the source is V_{ds} (in [V]). The drain current can be obtained from the above equation as

$$I_{ds,triode} = \frac{1}{1+\dfrac{V_{ds}}{LE_{Cr}}}\frac{W}{L}\mu_{eff}\int_0^{V_{ds}} Q_{inv}(\phi_c)d\phi_c. \tag{4.105}$$

Substituting the inversion charge expression into equation (4.105), the drain current in the triode region can be written as

$$I_{ds,triode} = \frac{W\mu_0\left(1/C_{ox}+1/C_{inv}\right)^{-1}}{L\left(1+\theta\left[V_{gs}-V_{TH,QM}\left(V_{dd}/2\right)\right]\right)\left(1+\dfrac{V_{ds}}{LE_{cr}}\right)} \cdot\left(\left(V_{gs}-V_{TH,QM}\right)V_{ds} - \frac{V_{ds}^{\,2}}{2} + \frac{2}{3}\phi_{s,qm}\frac{Q_{BO,qm}}{C_{ox}}\left[\left(1+\frac{V_{ds}}{\phi_{s,qm}}\right)^{3/2} - \left(1+\frac{3}{2}\frac{V_{ds}}{\phi_{s,qm}}\right)\right]\right), \tag{4.106}$$

where $Q_{BO,qm}$ (in $[C/cm^2]$) is the depletion charge sheet density at the threshold condition

$$Q_{BO,qm} = N_A\left(d_{cl} + \frac{9}{4\alpha_T}\right). \tag{4.107}$$

4.5.4 Drain Current in the Saturation Region

The saturation voltage can be obtained by solving

$$\left.\frac{dI_{ds,triode}}{dV_{ds}}\right|_{V_{ds}=V_{dsat}} = 0, \tag{4.108}$$

which leads to the equation [89]

$$\begin{aligned}&\frac{1}{LE_{Cr}}\left(\left(V_{gs}-V_{TH,QM}\right)V_{dsat} - \frac{{V_{dsat}}^2}{2} + \frac{2}{3}\phi_{s,qm}\frac{Q_{BO,qm}}{C_{ox}}\left[\left(1+\frac{V_{dsat}}{\phi_{s,qm}}\right)^{3/2} - \left(1+\frac{3}{2}\frac{V_{dsat}}{\phi_{s,qm}}\right)\right]\right)\\ &= \left(1+\frac{V_{dsat}}{LE_{Cr}}\right)\left(\left(V_{gs}-V_{TH,QM}\right) - V_{dsat} + \frac{Q_{BO,qm}}{C_{ox}}\left[\left(1+\frac{V_{dsat}}{\phi_{s,qm}}\right)^{1/2} - 1\right]\right)\end{aligned}. \tag{4.109}$$

The saturation voltage is given as

$$V_{dsat} \approx LE_{Cr}\left\{\sqrt{1+\frac{2}{LE_{Cr}}\frac{\left(V_{gs}-V_{TH}\right)}{\left[1-\frac{1}{2\phi_{s,qm}}\frac{Q_{B0,qm}}{C_{ox}}\right]}} - 1\right\}. \tag{4.110}$$

For the saturation region $V_{ds} > V_{dsat}$, the current is determined by both drift and diffusion [89].

$$I_{ds,sat} = \frac{W\mu_{eff}\left(C_{ox}^{-1} + C_{inv}^{-1}\right)^{-1}}{L} \frac{1}{\left(1+\frac{V_{ds}}{LE_{Cr}}\right)} \cdot \left(\left(V_{gs} - V_{TH,QM}\right)V_{dsat} - \frac{V_{dsat}^{2}}{2} + \frac{2}{3}\phi_{s,qm}\frac{Q_{BO,qm}}{C_{ox}}\left[\left(1+\frac{V_{dsat}}{\phi_{s,qm}}\right)^{3/2} - \left(1+\frac{3}{2}\frac{V_{dsat}}{\phi_{s,qm}}\right)\right]\right). \quad (4.111)$$

4.5.5 Drain Current in the Subthreshold Region

In the subthreshold region, $V_{gs} < V_{TH,QM}$, the inversion charge density is given by

$$Q_{inv} = Q_{TH}\exp\left\{\frac{S}{\ln 10}\frac{q}{kT}\left[V_{gs} - V_{TH,QM}\left(\phi_c\right)\right]\right\}. \quad (4.112)$$

The subthreshold drain current can be derived in a similar manner as in [89]

$$I_{ds,off} = \frac{W}{L}\mu_0 C_{ox}\frac{\eta}{\beta^2}\left\{1-\exp\left[-\beta V_{ds}\right]\right\}\exp\left[\frac{\beta}{\eta}\left(V_{gs} - V_{TH,QM} - \frac{\eta}{\beta}\right)\right], \quad (4.113)$$

where the subthreshold slope factor η is defined as

$$\eta = \left(\frac{S}{\ln 10}\right)^{-1}. \quad (4.114)$$

By the quantum mechanical charge model, η is given by

$$\eta = 1 + \frac{C_{d,eff}}{C_{ox,eff}}. \quad (4.115)$$

The formulas of the quantum-mechanical *I-V* characteristic are listed in Figure 4.8. This model is consistent with the numerical simulation [87], as shown in Figure 4.9 and Figure 4.10, for long-channel devices as well as short-channel devices. Significant drain current reduction in the quantum mechanical model is observed, which results from the threshold voltage shift and gate capacitance degradation.

$V_{gs} < V_{\alpha}$:

$$I_{ds,off} \cong \frac{W}{L}\mu_0 C_{ox}\frac{\eta}{\beta^2}\left\{1-\exp\left[-\beta V_{ds}\right]\right\}\exp\left[\frac{\beta}{\eta}\left(V_{gs}-V_{TO,qm}-\frac{\eta}{\beta}\right)\right]$$

$V_{gs} > V_{\alpha}, V_{ds} < V_{dsat}$:

$$I_{ds,triode} = \frac{W\mu_0\left(1/C_{ox}+1/C_{SI}\right)^{-1}}{L\left(1+\theta\left[V_{gs}-V_{TH,QM}\left(V_{dd}/2\right)\right]\right)\left(1+V_{ds}/LE_{Cr}\right)}\left(\left(V_{gs}-V_{TH,QM}\right)V_{ds}-\frac{V_{ds}^{\ 2}}{2}+\frac{2}{3}\left(\phi_{s,qm}\right)\frac{Q_{BO,qm}}{C_{ox}}\left[\left(1+\frac{V_{ds}}{\phi_{s,qm}}\right)^{3/2}-\left(1+\frac{3}{2}\frac{V_{ds}}{\phi_{s,qm}}\right)\right]\right)$$

$V_{gs} > V_{\alpha}, V_{ds} > V_{dsat}$:

$$I_{ds,sat} = \frac{W\mu_{eff}\left(C_{ox}^{\ -1}+C_{SI}^{\ -1}\right)^{-1}}{L}\frac{1}{\left(1+V_{ds}/LE_{Cr}\right)}\left(\left(V_{gs}-V_{TH,QM}\right)V_{dsat}-\frac{V_{dsat}^{\ 2}}{2}+\frac{2}{3}\phi_{s,qm}\frac{Q_{BO,qm}}{C_{ox}}\left[\left(1+\frac{V_{dsat}}{\phi_{s,qm}}\right)^{3/2}-\left(1+\frac{3}{2}\frac{V_{dsat}}{\phi_{s,qm}}\right)\right]\right)$$

Transition:

$$V_{TL,qm}\left(\phi_C\right) = V_{FB}+\phi_C+\phi_{s,qm}-\frac{Q_{B,qm}\left(\phi_C\right)}{C_{ox}}; V_{TH,QM} = V_{TL,qm}+\Delta V_{TH,qm}; V_{TO,qm} = V_{TH,QM}\Big|_{\phi_C=0}$$

$$V_{\alpha} = V_{TH,qm}+\frac{2\eta_{qm}}{\beta\left(\sqrt{1+4\theta\eta_{qm}/\beta}-1\right)}-\frac{1}{\theta}; V_{dsat} \approx LE_{Cr}\left\{\sqrt{1+\frac{2}{LE_C}\frac{\left(V_{gs}-V_{TO,qm}\right)}{\eta_{qm}}}-1\right\}$$

Other parameters:

$$\phi_{s,qm} = \frac{qN_A}{2\varepsilon_{Si}}\left(d_{cl}+\frac{9}{4\alpha}\right)^2;\ \eta = 1+\frac{C_{d,eff}}{C_{ox,eff}};\ \theta = \frac{\mu_0}{2t_{ox}v_{norm}};\ \beta = \frac{q}{kT};\ C_d = \sqrt{\frac{q\varepsilon_{Si}N_A}{2\left(\phi_{s,qm}+\phi_c\right)}};\ C_{d0} = C_d\Big|_{\phi_C=0};$$

$$v_{norm} = 2.2\times10^9\, cm/s;\ E_{Cr} = v_{sat}/\mu_{eff};\ Q_{B,qm}\left(\phi_c\right) = \sqrt{2q\varepsilon_{Si}N_A\left(\phi_{s,qm}+\phi_c\right)};\ Q_{BO,qm} = Q_{B,qm}\Big|_{\phi_C=0}$$

Figure 4.8

Transregional current-voltage model [89] with quantum-mechanical modifications.

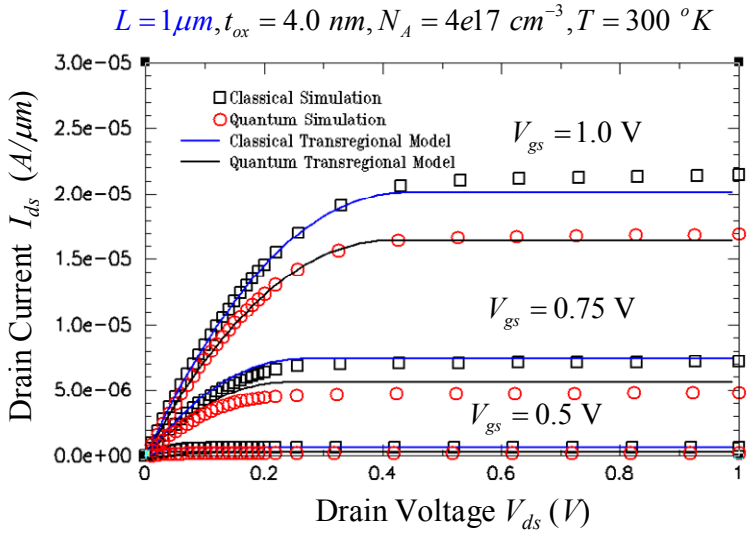

Figure 4.9
Quantum mechanical drain current model compared with simulation data from ISE TCAD [87].

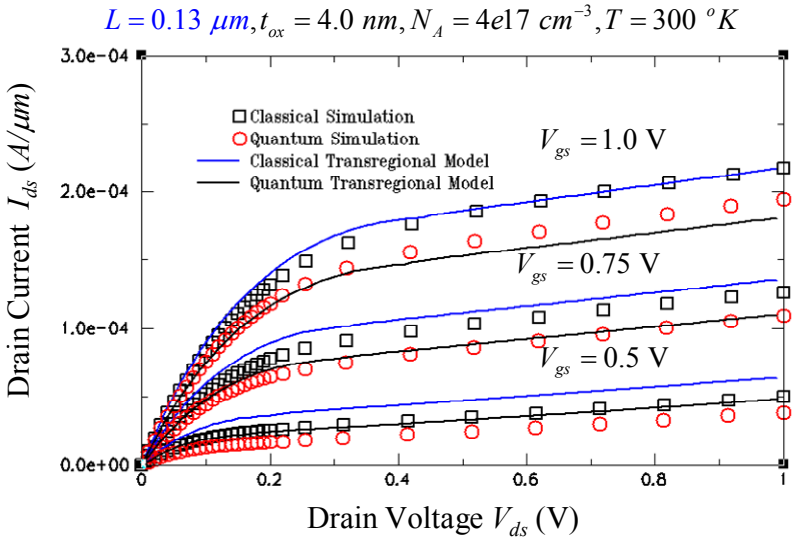

Figure 4.10
Quantum mechanical drain current model compared with simulation data from ISE TCAD [87].

4.5.6 Case Study

The drain current model is applied to estimate the performance degradation in a simple circuit from QMEs. As illustrated by Figure 4.11, an inverter with a load capacitance of 50 *fF* is analyzed for two technology generations with $t_{ox} = 4.0\ nm$ and $t_{ox} = 1.0\ nm$, respectively. In the two cases, the excess propagation delays caused by QMEs are about the same: *1.0 ps* and *1.1 ps*, as shown in Figure 4.12 and Figure 4.13 respectively. However, for $t_{ox} = 4.0\ nm$ technology, the extra delay time amounts to only a *19%* increase in the total delay while for $t_{ox} = 1.0\ nm$ it represents an increase of *115%*. Therefore, for integrated circuits in future technology generations, performance degradation resulting from QMEs must be considered in advance to avoid large deviation from design.

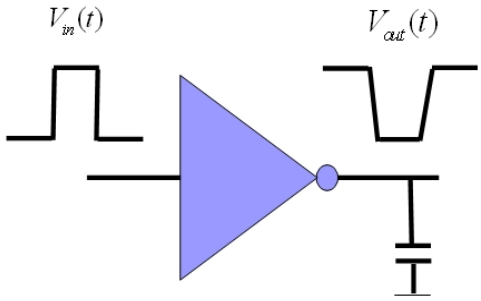

Figure 4.11

Study of the inverter delay.

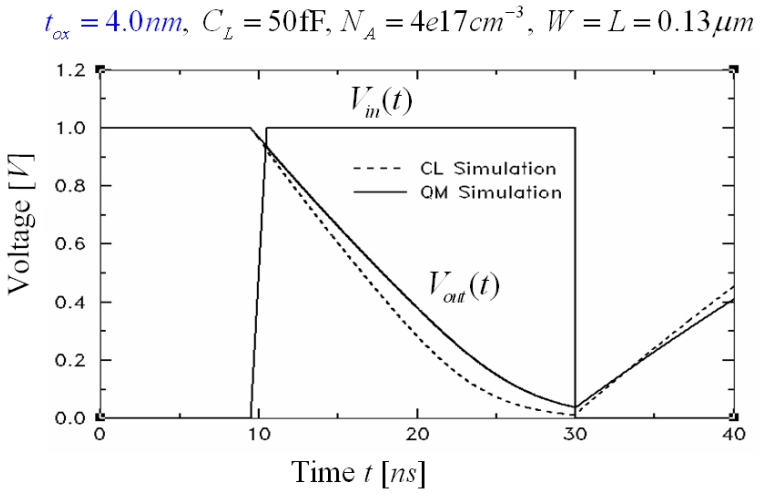

Figure 4.12
Input/output waveform of the inverter with t_{ox}=*4.0 nm* driving a fixed load capacitor. Propagation delay is measured as fall time from V_{dd} to $0.5V_{dd}$.

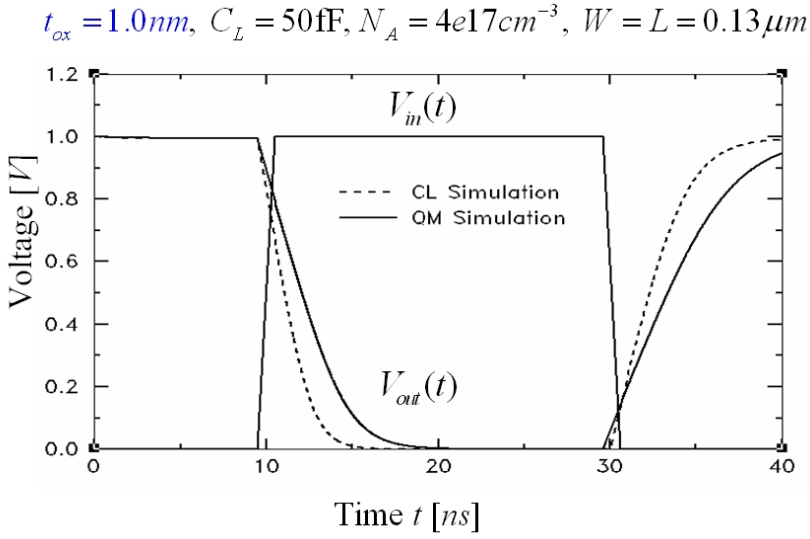

Figure 4.13
Input/output waveform of the inverter with t_{ox}=*1.0 nm* driving a fixed load capacitor. Propagation delay is measured as fall time from V_{dd} to $0.5V_{dd}$.

4.6 Conclusion

In this chapter, compact device models are developed to incorporate the energy quantization effect in device electrical parameters. Energy quantization causes the threshold voltage to shift to a greater value and an increase in effective oxide thickness, leading to a greater *S* value in long-channel MOSFETs. Short-channel V_{TH} and *S* models are derived including both two-dimensional electrostatic potential analysis and quantum correction to depletion depth, surface potential, etc. Increases in V_{TH} roll-off and *S* roll-up are observed by comparing the quantum mechanical model and the classical model, demonstrating that short-channel devices are more susceptible to the energy quantization. An *I-V* characteristic model integrating key QMEs is developed, showing great drive current loss due to the quantization effect. A case study using a CMOS inverter circuit shows extra delays as high as *115%* are obtained when QMEs are included in the *I-V* model for very thin oxide technology generation.

CHAPTER 5

MOSFET SCALING LIMIT

5.1 Introduction and Background

The semiconductor industry has been engaged in MOSFET scaling for over three decades. The minimum feature size, which is the smallest lateral geometric size of devices on an IC, has shrunk considerably over the past three decades. Consequently, the number of transistors on a chip increases over technology generations. Such trend is stated by Moore's law as the number of transistors per chip doubles every *18* months [1]. As a consequence, the channel length, the oxide thickness and the width of a MOSFET are simultaneously reduced [6].

When device dimensions are scaled down, not only the number of transistors per chip is increased, but also the performance of the devices is enhanced. The smaller area means that capacitance is also reduced, which enables the devices to be charged/discharged with less energy. This results in faster switching and reduced power dissipation [1].

However, the scaling process is challenged by several limiting factors [11, 90]. For MOSFET device design for digital applications, scaling challenges include controlling the leakage and SCEs, keeping or even increasing the drive current while reducing supply voltage, and maintaining the uniformity of device parameters within a chip and from chip to chip. These issues cannot be dealt without considering QMEs. Gate direct tunneling has become a significant source of leakage and greatly increased power dissipation. The carrier energy quantization effectively increases the oxide thickness, leading to less drive current and worse SCE control. Moreover, quantization effects

increase the threshold voltage dependency on channel length and substrate doping, which implies greater device sensitivities to process variations.

This chapter investigates the impact of QMEs on the MOSFET scaling, by various criteria, accommodating to the requirements on device, circuit, and system. Following this introduction, Section 5.2 briefly reviews the traditional scaling methodology and the criteria for estimating the scalability of bulk MOSFETs using QMEs. Section 5.3 discusses the device scaling limits from both SCE control and gate tunneling. Section 5.4 investigates the power and performance of CMOS circuits in future generations. Section 5.5 studies the QMEs on device parameter variation. Section 5.6 lists a few new materials and advanced structures of MOSFETs, which are proposed to extend the scaling process. Conclusion is given in Section 5.7.

5.2 Traditional Scaling Methods

MOSFETs are scaled down for size-reduction as well as performance improvement. The size-reduction is usually described by the scaling parameter (ξ). It is the pre-factor by which dimensions are reduced. For example, the channel length reduction from technology generation L to technology generation L' can be denoted as

$$L' = \xi L . \tag{5.1}$$

The performance improvement is roughly measured by the gate delay τ (in [s])

$$\tau = \frac{C_g V_{dd}}{I_{drive}} \tag{5.2}$$

where C_g is the gate capacitance, V_{dd} is the supply voltage, and I_{drive} is the drive current of the MOSFETs.

Historically, two scaling methods have been used: constant voltage scaling and constant field scaling [91]. A brief comparison between these two scaling methods is given in Table 5.1. For constant voltage scaling, the applied voltage maintains constant, while the dimensions of the MOSFET, including L, t_{ox}, and W are scaled. This method keeps C_g / W constant. The performance enhancement results primarily from the improvement of the drive current. Since the threshold voltage is scaled by the reduction of the oxide thickness, the resulting overdrive of $V_{dd} - V_{TH}$ in turn increases the drive current. However, since the supply voltage remains constant as the dimensions decrease, the electric fields increases with the size reduction. High fields and high currents tend to damage the gate oxide and lead to device deterioration. Thus, a main technology concern is to design reliable MOSFETs. Constant voltage scaling ended at $L_g = 0.5\ \mu m$ and t_{ox} near *10 nm* because of the problems described above [91].

For constant field scaling, the supply voltage is scaled along with the oxide such that the electric field in the oxide remains constant. The drive current per width remains constant, and the performance gain stems from the decreasing supply voltage. Gate delay decreases by *30%* per technology generation, nearly the same trend with constant voltage scaling [91]. However, this performance gain comes with a price: much higher off-state leakage due to the threshold voltage reduction [62, 90].

Table 5.1
Traditional scaling methods [91].

Parameter	Constant Voltage Scaling	Constant Field Scaling
Dimensions (*W, L*)	ξ	ξ
V_{dd}	1	ξ
Fields	ξ^{-1}	1
V_{TH}	1	ξ
I_{drive}/W	ξ^{-2}	1
C_g/W	1	1
τ	ξ^2	ξ
Power/circuit	ξ^{-1}	ξ^2

As MOSFET scaling continues, several factors weigh heavily against it. As more and more transistors are integrated onto a single chip and the operating frequency is increased, the overall power dissipation grows significantly Since there are so many transistors on a chip, the requirement on the power dissipation of a single MOSFET is stringent. A short-channel MOSFET tends to have large leakage current. Keeping the leakage/off-state power to a tolerable level is imperative in such devices in succeeding technology generations. In addition, highly scaled MOSFETs suffer from severe short-channel effects, namely, the threshold voltage roll-off and the subthreshold roll-up. Both of these effects result in a net increase in subthreshold leakage. Moreover, the severe short-channel effects cause the threshold voltage to be very sensitive to the

channel length variation. The non-uniformity of the devices on a wafer causes penalties in integration scale or results in requiring an increased supply voltage to ensure the delay time lies within the acceptable range.

Quantum mechanical effects must be addressed in MOSFET scaling issues associated with power dissipation, SCE control, and uniformity. Along with the required channel length reduction, oxide thickness needs to be shrunk aggressively in order to control SCE and threshold voltage scaling. However, this causes a sharp increase in tunneling current. The excessive tunneling current increases the stand-by power of MOSFETs, and even causes improper logic operations in circuits. Meanwhile, the effective oxide thickness becomes larger than the physical oxide thickness, because of energy quantization effect. The compact physical model developed in the previous chapter, which incorporates the quantum effect, is necessary to predict the scaling limit of the MOSFET.

Given the above facts, the prevailing scaling rules are not appropriate for scaling MOSFETs in and beyond the sub-*90 nm* regime [69, 92]. Instead, new scaling methods are needed to avoid excessive power dissipation and SCEs. In the past, different criteria have been proposed as benchmarks to project scaling limits [62]. In this work, these criteria are applied to investigate the role of QMEs in the MOSFET scaling problem, providing guidelines for the future technology developments.

5.3 Scaling Limit by Device Leakage

The scaling limit of bulk MOSFETs discussed in this section is based on two considerations: the SCE controllability and the tunneling current density constraint, which arises from the demand to control the subthreshold and tunneling leakage current.

The SCE control requires keeping S<100 *mV/decade*, which leads to the reduction of EOT along with the evolvement of technology towards shorter channel lengths [6, 62, 63]. The maximum allowed tunneling current density is specified by the ITRS as $J_{Tunnel} < 10 I_{sub} / L$ [2], where I_{sub} (in [*A/cm*]) is the subthreshold leakage current per width. Because it is desired to continue usage of SiO_2 as a gate oxide in order to delay a major technology transition to high-κ material, accurate EOT scaling projection incorporating QME is a must for exploring the ultimate scaling limit of SiO_2. As illustrated by Figure 5.1, for a *30 nm* MOSFET with the tolerable SCE as S<100 *mV/decade* and tunneling current density as $J_{Tunnel} < 10 I_{sub} / L$, the appropriate range for SiO_2 EOT is only *0.1 nm*. This value is less than the thickness of one atomic layer of SiO_2. This indicates that the fundamental "atom layer exhaustion limit" prevents further scale-down from using SiO_2 as the gate insulation in bulk MOSFETs. High-κ materials should be used to extend the scaling of the MOSFET beyond the *30 nm* technology generation, as shown in Figure 5.2. As the requirements of S<100 *mV/decade* and $J_{Tunnel} < 10 I_{sub} / L$ are applied for high-κ gate dielectric, the relaxed limitation from gate tunneling enables a larger suitable range of EOT for MOSFETs. Moreover, the insulation layer made by high-κ dielectrics has more atomic layer resources for scaling.

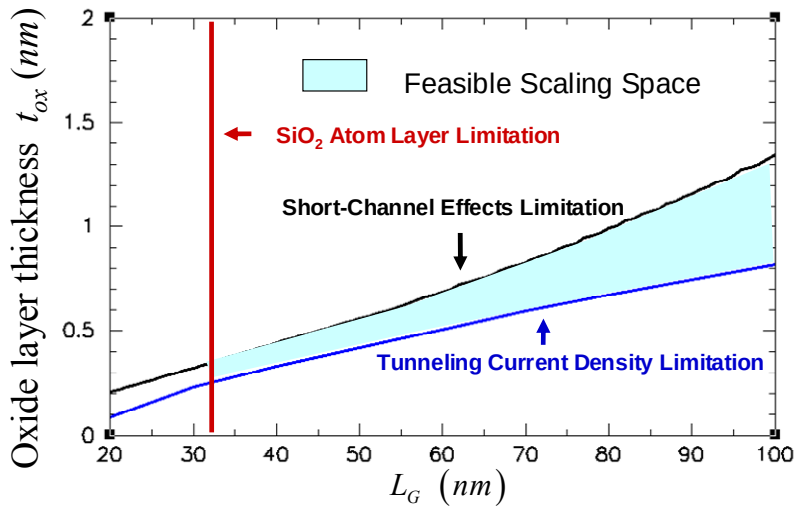

Figure 5.1
Design space for conventional MOSFETs in future technology generations.

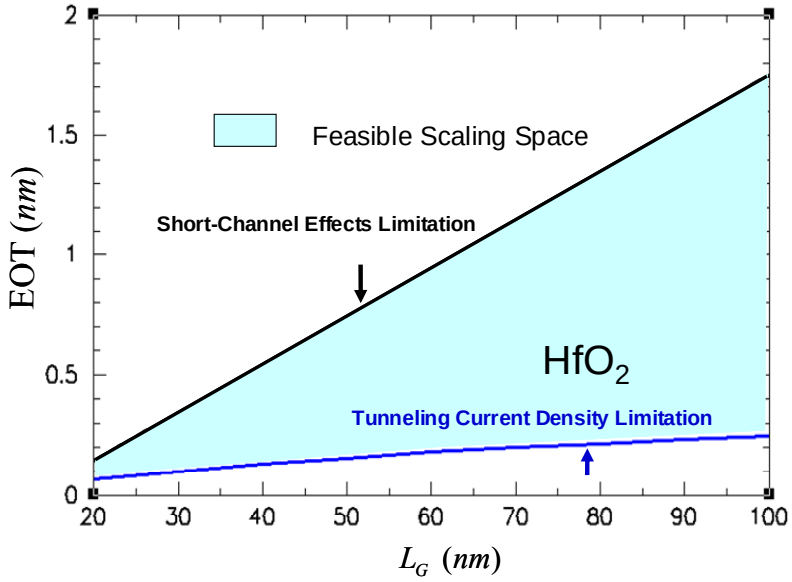

Figure 5.2
Design space for MOSFETs using high-κ gate dielectric in future technology generations.

5.4 Scaling Limit by Circuit Performance and Power Dissipation

The analysis contained in this section uses a total power minimization methodology [68] to determine optimal design parameters (V_{dd}, V_{TH}, t_{ox}, W) for specified technology generations. The methodology minimizes the power dissipation for a critical path consisting of seven stages of two-input static CMOS NAND gates with an average fan-out of three. The total power dissipation can be divided into static and dynamic components as

$$P_{total} = P_{dynamic} + P_{static} . \tag{5.3}$$

The dynamic power is the energy consumed by the charging and discharging capacitor when CMOS circuit switch between "0" and "1" and is described by

$$P_{dynamic} = \frac{1}{2} a C_L V_{dd}^{\ 2} f_{clk} , \tag{5.4}$$

where activity factor a is the average switching rate or the probability of a binary transition. This value is assumed to be *10%* for random logic networks. The load capacitance,

$$C_L = C_W + C_D + C_G \tag{5.5}$$

is the sum of the wiring, and device capacitance. The wiring capacitance C_W is based on an average interconnect length derived from a statistical distribution [12]. C_D is the drain capacitance and represents the device capacitance of an unloaded static logic gate. It is comprised of the n-MOSFET and p-MOSFET gate-drain overlap capacitance, the drain bottom junction capacitance, and the drain sidewall junction capacitance. C_G is the fan-out gate capacitance of the next logic stage, computed from the product of the

fan-out and the MOSFET gate capacitances of the n-MOSFET and p-MOSFET devices connected at the output.

The static component of power results from the MOSFET subthreshold leakage current and oxide tunneling current and is described as

$$P_{static} = P_{sub} + P_{tunnel}. \tag{5.6}$$

In more detail, the subthreshold leakage power and the tunneling power can be written as

$$P_{sub} = I_{sub} \cdot V_{dd} \tag{5.7}$$

and

$$P_{tunnel} = I_{tunnel} \cdot V_{dd}, \tag{5.8}$$

where I_{sub} is the subthreshold leakage current, and I_{tunnel} is the tunneling current.

As shown in Table 5.2, the technology evolvement including shorter channel length, higher transistor density and clock frequency is given by the ITRS [2]. The minimum power methodology [68] is employed to predict the power dissipation under transistor size and system performance constraints. Figure 5.3 shows the minimum total power consumed on the critical path that satisfies the requirement for circuit delay. From the results, it can be deduced that quantum effects induce significant increases in the projected total power (*39%~65%*), device aspect ratio (*21%~100%*), and necessary supply voltage (*36%~81%*) as shown in Figure 5.3, Figure 5.5, and Figure 5.6 respectively. In order to maintain the circuit performance, greater device area and supply voltage are necessary to compensate for the drive current loss and threshold voltage shift caused by quantum mechanical effects. Accordingly, larger gate area and higher supply voltage result in further power consumption. Figure 5.4 illustrates the scaling trend in oxide layer thickness as predicted by the quantum model and the classical model.

Decreasing the oxide thickness provides less circuit delay and subthreshold leakage. However, gate tunneling imposes a limit on the reduction of oxide thickness. The optimum oxide thickness is the result of the tradeoff between performance improvement and tunneling power increase. A higher optimum oxide thickness is projected by the quantum model than that predicted by the classical one (i.e., *1.5 nm* vs. *1.0 nm* in *65 nm node*). This implies the gate capacitance reduction due to the quantization effect. The performance degradation induced by quantization is another important QME on oxide scaling in addition to gate tunneling. It is worth mentioning that even if high-κ dielectrics are used to suppress the gate tunneling current, the diminishing performance enhancement with EOT reduction will remains a challenge for high performance and low power CMOS circuit design.

Table 5.2
CMOS scaling predicted by the ITRS [2].

Year	Technology node (*nm*)	Physical gate length (*nm*)	Transistor density (M/cm^2)	Local clock frequency (*MHz*)
2001	150	65	38.6	1684
2002	130	53	48.6	2317
2003	107	45	61.2	3088
2004	90	37	77.2	3990
2005	80	32	97.2	5173
2006	70	28	122.5	5631
2007	65	25	154.3	6739

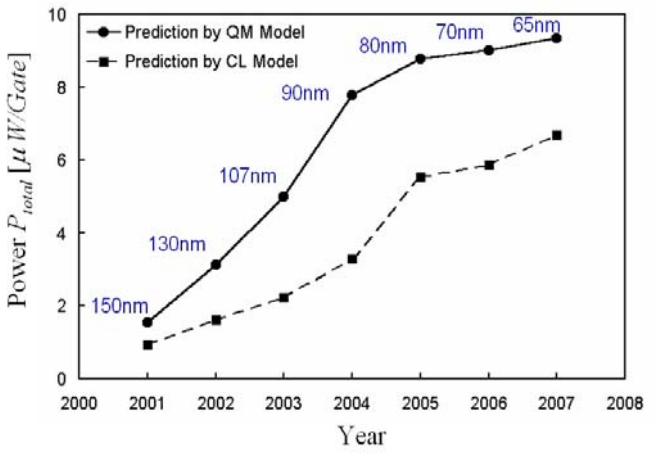

Figure 5.3
Minimum total power P_{total} ($\mu w/gate$) projected by the performance-constrained Minimum Power Methodology [68].

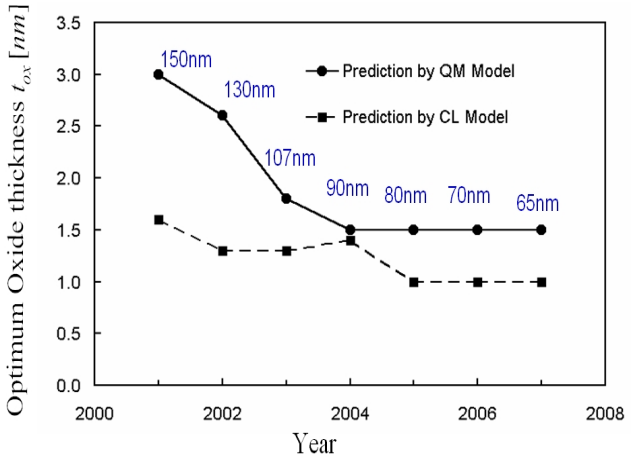

Figure 5.4
Optimum gate oxide thickness projected by the performance-constrained Minimum Power Methodology [68].

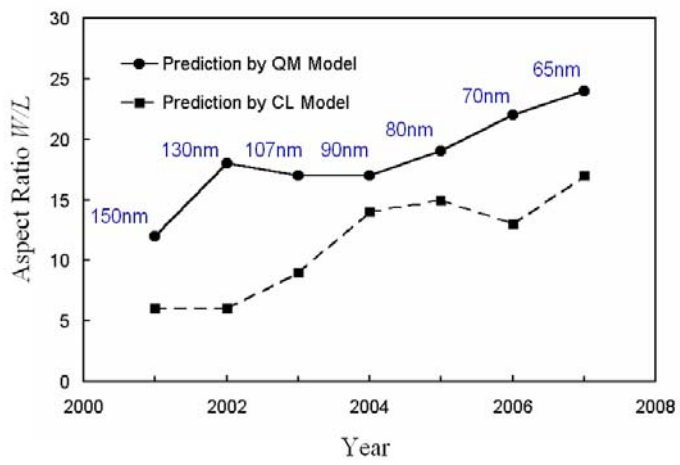

Figure 5.5
Optimum aspect ratio projected by the performance-constrained Minimum Power Methodology [68].

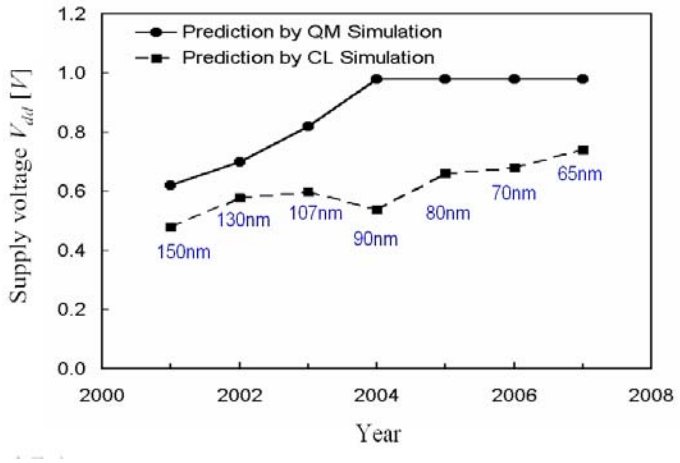

Figure 5.6
Optimum supply voltage projected by the performance-constrained Minimum Power Methodology [68].

For the 65 *nm* node and beyond, high-κ dielectrics are projected to replace the silicon oxide for the continuation of scaling. Scaling goals of increasing the transistor density and clock frequency are listed in Table 5.3. By applying the minimum power methodology [68], the minimum power dissipation to fulfill the performance requirement is specified in Figure 5.7 and Figure 5.8 corresponding to SiO_2 and HfO_2 gate dielectrics, respectively. With the SiO_2 gate dielectric, a significant 2.2× increase in power consumption from the *57 nm* node to the *40 nm* node is predicted, as shown in Figure 5.7. Gate tunneling makes a greater contribution to the power increase than other components. Tunneling power is only *9%* of the total power at the *57 nm* node and increases to *29%* at the *40 nm* node. Comparing Figure 5.7 with Figure 5.8, we find the power consumption can be greatly reduced, if the high-κ dielectric is used to suppress the gate tunneling. By employment of the high-κ dielectric, the total power can be cut by *40%* and *73%* at the *57 nm* and the *40 nm* nodes, respectively. As shown in Figure 5.9, gate tunneling strongly limits the reduction of the silicon oxide layer thickness, which is imperative to subthreshold leakage control. With high-κ dielectrics, the EOT can be further scaled down, seeking for better SCE and subthreshold control without the tunneling problem.

Table 5.3
CMOS scaling beyond 65 nm [67].

Year	Technology node (*nm*)	Physical gate length (*nm*)	Transistor density (M/cm^2)	Local clock frequency (*MHz*)
2008	57	23	194	9285
2009	50	20	245	10972
2010	45	18	309	12369
2011	40	16	389	15079

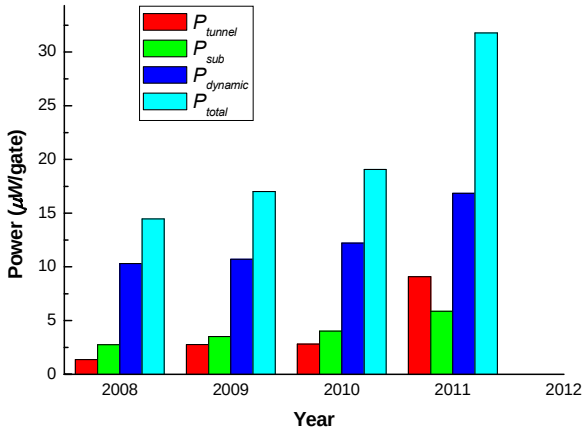

Figure 5.7
Power dissipation prediction for bulk MOSFETs with silicon oxide.

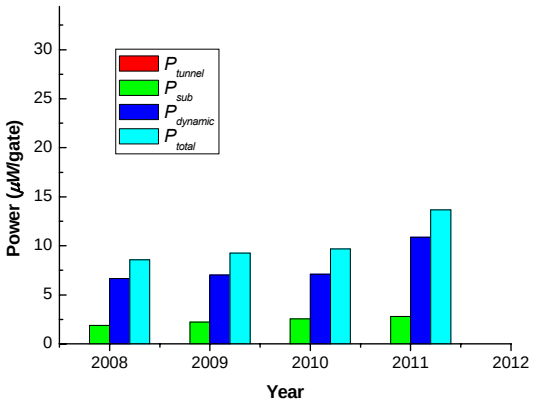

Figure 5.8
Power dissipation predictions for bulk MOSFETs with high-κ dielectrics.

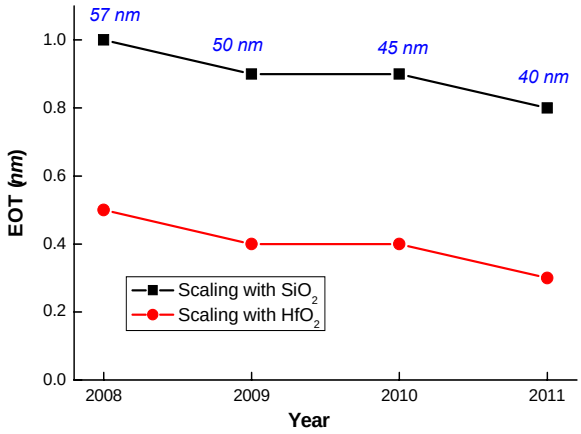

Figure 5.9
EOT scaling with SiO_2 and HfO_2 gate dielectrics.

5.5 Scaling Limits due to Parameter Variation

IC fabrication processes introduce variations in MOSFET dimensions and doping concentrations. As a result, the device electrical parameters, such as threshold voltage, subthreshold swing, subthreshold leakage current and superthreshold drain current, deviate from their nominal design values [67, 69, 93, 94]. With the continued aggressive scaling of MOSFET dimensions, statistical process variations have been the dominant factor in reducing the product yield and reliability. In device design, tolerance to process variations is an essential requirement particularly for gigascale integration [6, 62, 63].

Quantum mechanical V_{TH} models are hereby utilized to predict the limitations caused by parameter variations. Assuming *10%* variation in channel length and doping concentration, the variation in V_{TH} should not exceed *70 mV*. As shown in Figure 5.10 and Figure 5.11, greater threshold voltage variations are predicted by the quantum mechanical model (equations (4.32) and (4.67)) than by the classical model. At a channel length of *20 nm*, the *10%* change in channel length causes a variation in threshold voltage of *70 mV*, while for the *10%* change in doping concentration of $5e18\ cm^{-3}$, the threshold variation is *35 mV*. After a comprehensive examination of the fluctuations of L and N_A, the minimum channel length satisfying the constraint of system reliability is around *45 nm* with EOT=*1.5 nm.*

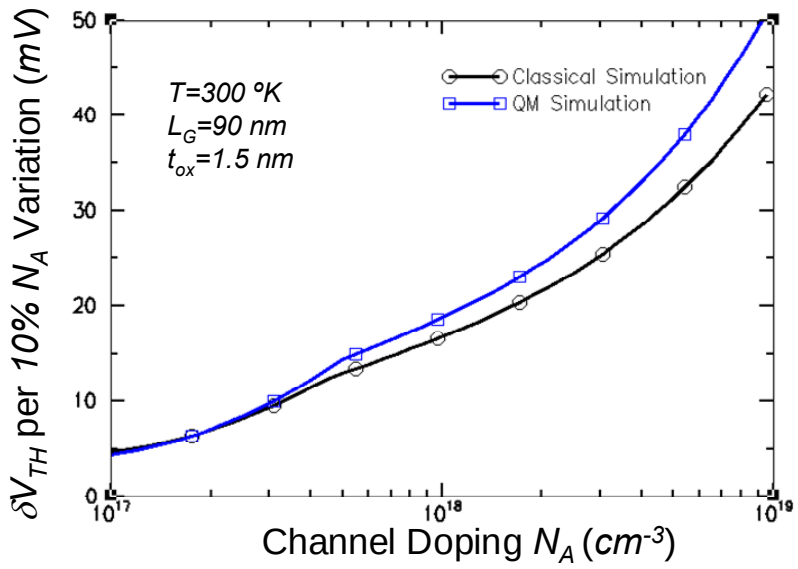

Figure 5.10
Threshold voltage changes with 10% channel doping variation.

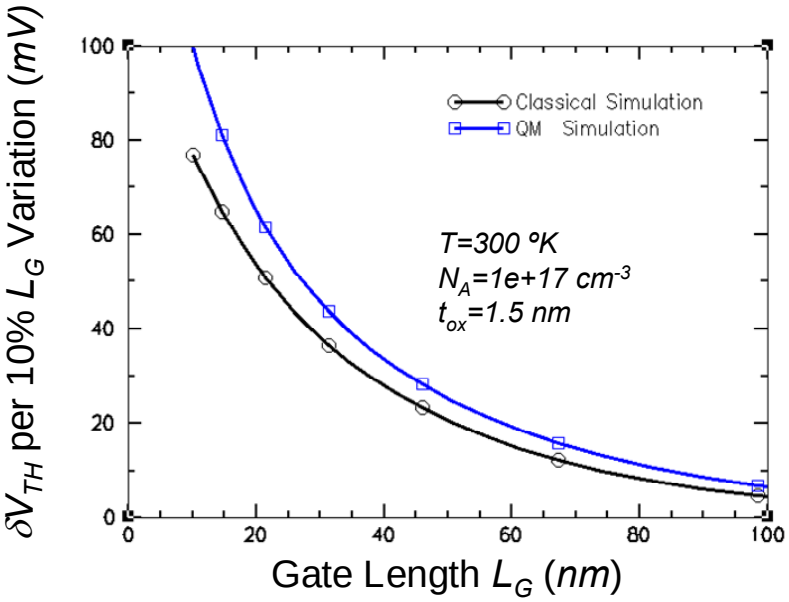

Figure 5.11
Threshold voltage changes with *10%* channel length variation.

5.6 CMOS Scaling with Advanced Materials and Structures

The rapid pace of MOSFET scaling makes it highly challenging in pursuing solutions to high drive current, low power consumption, tight control to SCE and device variations. Bulk MOSFET scaling entails an increase in vertical electrical field, leading to aggressively reduced gate oxide and increased channel doping. In current technology, as the gate insulation layer consists of only a few atomic layers of silicon oxide, gate direct tunneling and energy quantization pose stringent constraints on this approach. The ITRS has accelerated the introduction of new technologies including material enhancements and structural improvement to extend the scaling limit [67]. It is expected that numerous innovations on materials will be implemented in less than a decade [67, 78].

These new materials include those applied in the gate stack (high-κ dielectrics and metal electrodes), those used in the channel to boost carrier transport properties, and new materials used in the source/drain regions with reduced resistance and improved injection properties. High-κ gate dielectrics are utilized to prevent excessive gate tunneling and retain the high electric field in the vertical direction. Transport enhancement refers to the approaches to increase transistor drive current and improve circuit performance by enhancing the velocity and mobility of carriers in the channel. Alternative materials used to increase the mobility include strained-silicon [95-99], in which channel layers are mechanically strained, and high-mobility material such as silicon-germanium [100], germanium [101, 102] or *III-V* compound semiconductors [103]. Research on source/drain materials is proposed to address the issue that the source/drain resistance is an increasing fraction of the channel resistance as the channel length reduces. Metallic source/drain electrodes, which form low Schottky barrier heights in contact with silicon,

can be employed to reduce the parasitic series resistance and provide a sharp junction [104-108]. An overview for these materials is given in Table 5.4.

Table 5.4
New materials for MOSFET transistor.

	High-κ material	Transport enhanced material	Schottky Source/Drain
Concept	High dielectric constant material for gate insulation	Strained silicon, *SiGe*, *Ge*, or III-V semiconductor to improve carrier mobility	Metal source/drain electrodes forming a low Schottky barrier with silicon
Advantage	• Reducing gate leakage • No need for device structure modification	• High carrier mobility • No need for device architecture change	• Low source/drain resistance • No need for abrupt S/D doping • No need for ultra-shallow S/D
Weakness	• Incompatibility with silicon and poly silicon • Metal gate required • Interfacial defects causing mobility degradation	• Material defects • Process compatibility and thermal budget • Operation temperature	• Metal or silicide material not available for n-MOSFETs

In addition to the solution from new materials, new transistor structures, as alternatives to classical planar bulk MOSFETs, are investigated as possible approaches to successfully scale MOSFETs and meet device performance requirements. The new MOSFET structures offer better electrostatic properties, reduction of the gate control

dependency on the gate oxide layer, and improved drive current. Referred to as non-classical MOSFETs, these structural solutions include ultra-thin body (UTB) silicon-on-insulator (SOI), and various types of double-gate and multiple-gate MOSFET [67, 78].

The UTB SOI MOSFET [109-111], as shown in Figure 5.12, consists of a very thin (usually *<10 nm*) fully depleted (FD) transistor body to ensure good electrostatic control of the channel by the gate in the off-state. Typically, the ratio of the channel length to the channel thickness (t_{Si}) will be ≥ 3. It is shown that UTB SOI MOSFETs can be scaled down to *18 nm* gate length with an extremely thin (t_{Si}*<5 nm*) *Si* channel [112]. A lightly doped silicon body overcomes V_{TH} variations due to statistical dopant fluctuations and provides enhanced carrier mobility for higher drive current. This structure benefits from both a deep source/drain contact for low sheet resistance and the feature of improved electrostatics in SOI technologies [2].

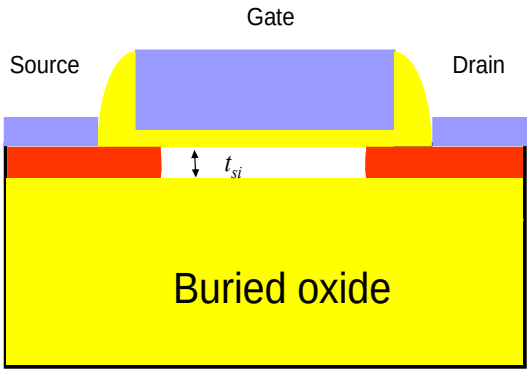

Figure 5.12
Schematic of UTB SOI MOSFET structure.

Several double-gate structures have been proposed to improve the electrostatics integrity and, in some cases, provide adjustable threshold voltage by isolated gates for low-power applications. The FinFet [113-116] is a tied double-gate, sidewall conduction structure, as shown in Figure 5.13. The width of the vertical silicon fin is smaller than the channel length to provide adequate control of short-channel effects. Implementation of a FinFET can take advantage of the bulk-like layout and process [114]. In fact, this structure can be realized on a bulk silicon substrate [115, 117]. However, the thin fins need to be a fraction (one third to one-half) of the gate length, thus requiring sub-lithographic techniques.

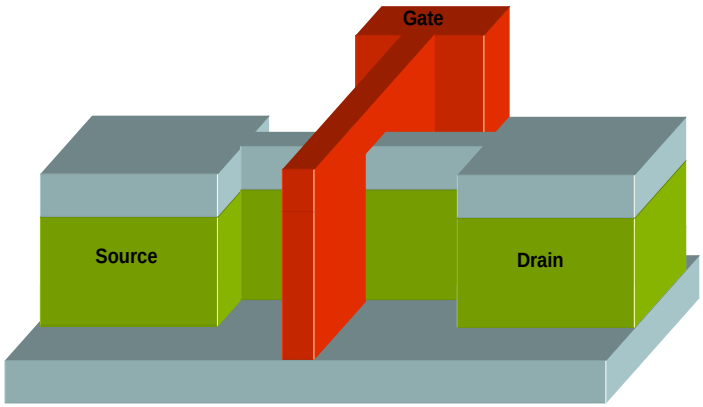

Figure 5.13
FinFET structure.

The second double-gate structure is the independently switched double gate (ground-plane) FET [118, 119], as shown in Figure 5.14. This structure is a planner FET, with the top and bottom gate electrodes electrically isolated to provide independent

biasing. The independent double gate MOSFET is attractive for dynamic threshold design, in which the top gate is typically used to switch the transistor on and off, and the bottom gate is used for threshold voltage adjustment [67, 120]. The presence of two gates helps to reduce SCE, which significantly reduces subthreshold leakage.

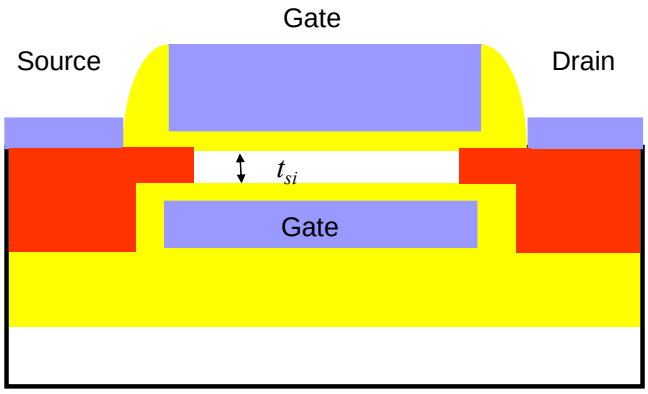

Figure 5.14
Planer double-gate structure.

Multiple-gate MOSFET structures [121-123], as shown in Figure 5.15, have been proposed and demonstrated to help manage electrostatic integrity in ultra-scaled CMOS. The large number (≥ 3) of gates provides for improved electrostatic control of the channel, so that the silicon body thickness and width can be larger for multiple-gate MOSFETs than that for the UTB SOI or double-gate FET structures. The principle advantage of the structure resides in the relaxation of the needs on the thickness of the silicon body or the vertical fin. The challenge is in slightly poorer electrostatic integrity than with double-gate structures, particularly in the corner regions of the channel [124].

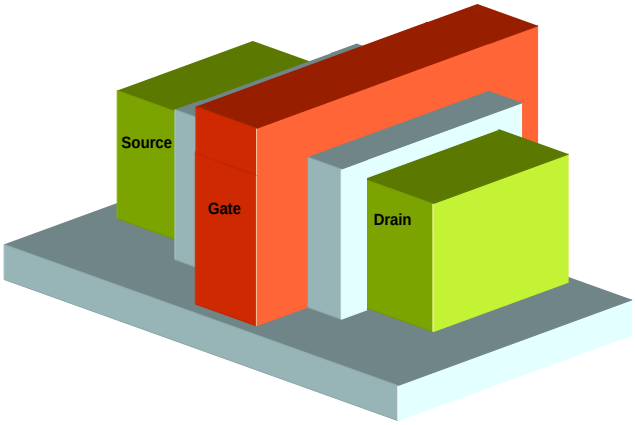

Figure 5.15
Tri-gate structure.

In spite of the performance enhancement by material and structure changes, QME is still a limiting factor in highly scaled MOSFETs. Although non-classical structures relax the stringent requirement for ultra thin gate oxide, gate tunneling is difficult to control without high-κ material in bulk MOSFETs as well as in non-classical MOSFETs [77]. Energy quantization, induced by the channel potential wells in bulk MOSFETs, can also be problematic in the non-classical structures. The ultra-thin channel layer results in spatial confinement to carriers [125, 126], similar to the case of a one-dimensional infinite potential well mentioned in Chapter 3. It has been demonstrated that threshold voltage of UTB SOI and double-gate MOSFETs is very sensitive to thickness variation of the silicon body. The statistical variation of the body thickness is projected to be a key issue for ultra-thin body MOSFET devices.

Successful realization of future ultra-scaled MOSFETs require advanced process technology to implement the material solution together with structure changes, i.e. the combination high-κ material and strained silicon with UTB SOI, the incorporation of

Schottky source/drain regions along with the FinFET to compensate the high series resistance in thinned source/drain [108]. The ITRS projects high-κ material and metal gate electrodes being required around year 2008 [67]. The UTB SOI MOSFETs could be introduced as early as year 2008 of the *57 nm* technology node and scaled down to the *25 nm* node in year 2015 [67]. Near mid-gap metal gate electrodes will be desirable to set the threshold voltage for UTB SOI. Following the UTB SOI, the double-gate MOSFET is predicted to be first manufactured in the *40 nm* technology node in 2011 [67]. Eventually, toward the end of the roadmap or beyond, high transport channel materials, such as germanium, III-V semiconductors, carbon nanotubes or nanowires, along with the non-classical structures will be utilized.

5.7 Conclusions

MOSFET scaling is challenged by power dissipation, short-channel effects, and parameter variations in a chip. QMEs on scaled MOSFETs are discussed, at the device, circuit, and system levels. It is observed that tunneling and quantization effects cause large power dissipation, low drive current and strong sensitivities to process variations, greatly limiting CMOS scaling. From the SCE controllability and tunneling power consumption requirement, the appropriate range of EOT for SiO_2 is only 0.1 *nm* and less than the thickness of one atomic layer of SiO_2 in a *30 nm* channel length MOSFET. In a circuit performance analysis, for 6*5 nm* technology, quantum effects cost *39%* increase in power dissipation and *41%* increase in device area compared with classical projections. Tunneling power reaches *29%* of the total power consumption at the *40 nm* node, which is higher than the subthreshold leakage power. Total power consumption can be greatly reduced by *73%* at the *40 nm* node, if the high-κ dielectric is used to suppress the gate

tunneling. Quantum models predict a greater variation in threshold voltage due to process variations than classical models do. The results suggest a minimum channel length of *45 nm* with EOT=*1.5 nm* by evaluating the threshold voltage instability induced by variations in both L and N_A. Developing new materials and structures offering better electrostatic properties, reducing the gate control dependency on the gate oxide layer, and improving drive current is imminent to extend MOSFET scaling.

CHAPTER 6

CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE WORK

6.1 Conclusions

6.1.1 Introduction

Traditionally, advances in CMOS technology have been obtained through continuous miniaturization of the devices. According to the scaling theory of MOSFETs, reduction of the channel length requires correspondent decrease in the gate dielectric thickness, in order to suppress SCEs and to achieve threshold voltage scaling. As MOSFETs are scaled down to the sub-*90 nm* range, the suitable equivalent oxide thickness is only about *1.2 nm*.For such thin oxides, gate leakage due to direct tunneling becomes unacceptably large. Gate tunneling gives rise to power consumption and results in less logic operating margins in CMOS circuits.

In addition, energy quantization leads to a significant increase in effective oxide thickness for an ultra-thin oxide layer. Consequently, the driving ability and electrostatic integrity of a MOSFET is degraded. Since the increased effective oxide thickness is non-scalable with the reduction of the physical oxide thickness, it is particularly important to consider the energy quantization effect in short-channel devices with ultra-thin gate dielectrics. For these reasons, quantum mechanical effects of gate tunneling and energy quantization have become critical issues in MOSFET scaling.

Modeling plays an important role in overcoming challenges brought by quantum mechanical effects. In this thesis, quantum mechanical effects on MOSFET scaling are

investigated by a modeling approach. Compact physical models are built from device physics and utilized for device design and technology projection.

6.1.2 Tunneling

As MOSFET gate oxide thicknesses rapidly approach their limit, accurate modeling of direct tunneling in ultra-thin oxides is essential. Both types of carriers, electrons and holes, can be involved in gate direct tunneling. The electron tunneling from the conduction band and hole tunneling from the valence band are the most significant tunneling components in an n-MOSFET and a p-MOSFET, respectively. A direct tunneling model for circuit simulation is developed from the solution of the Schrödinger equation. Simulated gate currents from this analytical model demonstrate good agreement with the results from a numerical solver and measured data for gate oxides with thicknesses ranging between *1.0-3.5 nm*. It was observed that the gate current density exceeds *1.0 A/cm*^{2} for t_{ox}*=1.5 nm* and V_{dd}*=1.0 V* in an n-MOSFET. The hole tunneling in p-MOSFETs is typically lower than electron tunneling in n-MOSFETs by an order of magnitude, which is due to the higher barrier and larger effective mass in hole tunneling. Besides gate-to-channel tunneling, the source/drain extension region provides the additional tunneling path in the source/drain-gate overlap. In short-channel devices, the length of the source and drain extension area can be comparable with the length of the channel itself, resulting in considerable leakage. Replacing silicon oxide with high dielectric constant material in the gate dielectric is the leading projected solution to reduce gate tunneling current to a tolerable level. By utilizing high-κ gate dielectrics, such as HfO_2 and $HfSiO_4$, the gate tunneling current density can be reduced by 2~3 orders of magnitude.

6.1.3 Energy Quantization

By energy quantization, carriers comply with a distribution on discrete subbands, instead of the continuous energy band assumed by the classical theory. The variational approach is applied to solve the coupled Schrödinger and Poisson equations for the quantized energy levels. Accordingly, carrier density on each subband can be computed from the subband energy levels and the two-dimensional state distribution function. From numerical simulation results, it is known that carriers on the lowest two subbands dominate the total electron population. Therefore, it is appropriate to obtain the total carrier density by considering the lowest two subbands only. A new MOS *C-V* model that accounts for the carrier energy quantization is developed based on the carrier distribution on subbands. From the quantum mechanical *C-V* model, it is observed that the inversion layer capacitance is much smaller than the value predicted in the classical model. As the inversion layer capacitance becomes comparable to the oxide layer capacitance, gate capacitance significantly deviates from the value predicted classically.

Compact models of device parameters incorporating the energy quantization effect are derived. Quantum mechanical models for the key parameters such as threshold voltage and subthreshold swing are also developed for both long-channel and short-channel MOSFET devices. In the long-channel device, the elevated energy level from the bottom of the conduction band requires a higher gate voltage to produce enough carrier density at threshold, exhibited as a larger value of V_{TH}, than the prediction from classical models. The subthreshold swing in a quantum mechanical model can be written in a similar form as that obtained from the classical model, by moving the position of the inversion charge sheet from the channel surface to its centroid below the surface. The

shift of the charge sheet position leads to the increase of the EOT, which in turn results in a corresponding increase in S. In short-channel devices, energy quantization weakens the gate control of the two-dimensional electrostatic potential distribution. Short-channel V_{TH} and S models are developed by incorporating the quantization effect into the electrostatic potential solution from the two-dimensional Poisson equation. It is observed that for the *20 nm* channel length MOSFET with an EOT of *0.5 nm*, QMEs lead to the increase of S from *74 mV/decade* to *105 mV/decade*. These models are subsequently incorporated into a transregional MOSFET *I-V* characteristic model, with the purpose of understanding the quantum mechanical impact on CMOS logic circuit performance. Results show that a significant drive current loss arises from the quantum mechanical effects. In a study of an inverter driving a load capacitance of 50 *fF* for two technology generations with $t_{ox} = 4.0\ nm$ and $t_{ox} = 1.0\ nm$, *19%* and *115%* differences in delay time are shown by simulation results from classical and quantum models. For integrated circuits with the ultra-thin oxide layer MOSFETs, the quantization effect must be included in device models to avoid large deviations from design.

6.1.4 MOSFET Scaling

MOSFETs are scaled down according to various requirements including power dissipation, SCE control and device uniformity. Complicated by quantum mechanical effects, these requirements have competing demands for device design. Traditional methods of constant voltage scaling and constant field scaling are unable to fulfill all the goals in MOSFET scaling. A more elaborate way is developing different criteria by constraints from device, circuit, and system levels, evaluating the MOSFET scaling limits

comprehensively. Quantum mechanical models are exploited to investigate the scaling limit, and assess the potential solutions to challenges in MOSFET scaling.

At the device level, constraints from SCE controllability and tunneling power consumption are investigated. Results show that for *30 nm* gate length MOSFETs with tolerable SCE and tunneling current density, the appropriate range of EOT for SiO_2 is only 0.1 *nm* and less than the thickness of one atomic layer of SiO_2. High-κ gate dielectrics must be used for scaling of bulk MOSFETs beyond the 30 *nm* technology generation.

In a circuit performance analysis, it is found that for the 65 *nm* technology, quantum effects induce a *39%* increase in power dissipation and a *41%* increase in device area when compared with the classical projection. It shows that the limits on the t_{ox} scaling from quantum mechanical effects become a critical constraint in high performance and low power CMOS circuit design. Beyond the 65 *nm* technology, scaling trends with SiO_2 and high-κ gate dielectrics are compared. Tunneling power reaches *29%* of the total power consumption at the *40 nm* node, which is higher than the subthreshold leakage power. Total power consumption can be greatly reduced, if the high-κ dielectric is used to suppress the gate tunneling. By the employment of high-κ dielectrics, the total power is cut by *73%* at the *40 nm* node, which is benefited from the EOT reduction without the tunneling power increase.

When considering the system variations, quantum models predict greater shifts in threshold voltage. Results suggest a minimum channel length of *45 nm* by evaluating the threshold voltage instability induced by variations in both L and N_A.

Various innovations on materials and MOSFET structures are projected to be implemented in less than a decade. These new materials include those applied in the gate stack, those used in the channel to boost carrier transport properties, and new materials used in the source/drain regions with reduced resistance and improved injection properties. Moreover, non-classical MOSFET structures, including UTB SOI, double-gate and multiple-gate MOSFETs, are proposed to offer better electrostatic properties, reduce the gate control dependency on the gate oxide layer, and improve drive current. Solutions from new materials and structures are mandatory for prolongation of MOSFET scaling.

6.2 Recommendations for Future Work

There are various modeling issues that await exploration in nanoscale MOSFETs, which include gate induce drain leakage (GIDL) and gate edge direct tunneling. Both of these affect off-state leakage current. Since GIDL is determined by both the vertical and lateral electric fields along the gate and the drain overlap region, the relationship between GIDL and direct tunneling current must be understood. In addition, along with the existing commercial device simulators that use band-to-band tunneling formulations, it is necessary to determine the appropriate model for trap-assisted tunneling and its dependence on processing, bias and ambient conditions.

The demand for new materials and technologies increases in the nanometer CMOS regime. Therefore, the understanding of mobility enhancement in germanium strained silicon, reliability and mobility degradation in high-κ gate dielectrics, and various device parameters for metal electrodes are required. Scaling CMOS toward the *25 nm* channel length generation requires innovative device structures to circumvent barriers due to the fundamental physics in conventional bulk MOSFETs. These may include UTB SOI, back-gate FETs, double gate FETs and FinFETs. Fundamental issues for these structures, such as the physics of the carrier transport in very thin silicon channels, must be further understood.

APPENDIX A

WKB METHOD

A.1 Hamilton-Jacobi Equation

Starting from the time-dependent Schrödinger equation [8] for a single particle in a potential well

$$i\hbar\frac{\partial}{\partial t}\psi\left(x,t\right)=\left[-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}+U\left(x\right)\right]\psi\left(x,t\right) \tag{A.1}$$

and using $\psi\left(x,t\right)=e^{iS(x,t)/\hbar}$:

$$-\frac{\partial S}{\partial t}\psi=\left[\frac{1}{2m}\left(\nabla S\right)^2-\frac{i\hbar}{2m}\left(\nabla^2 S\right)+U\right]\psi \tag{A.2}$$

is obtained. Assuming $\psi\neq 0$, this leads to an equation

$$-\frac{\partial S}{\partial t}=\frac{1}{2m}\left(\frac{\partial S}{\partial x}\right)^2-\frac{i\hbar}{2m}\left(\frac{\partial^2 S}{\partial x^2}\right)+U\,. \tag{A.3}$$

Taking the formal limit $\hbar\to 0$, equation (A.3) becomes the same as the classical Hamilton–Jacobi equation

$$-\frac{\partial S}{\partial t}=\frac{1}{2m}\left(\frac{\partial S}{\partial x}\right)^2+U \tag{A.4}$$

A.2 Classical Limit

To simplify the equation (A.4), we can use separation of variables as

$$S\left(x,t\right)=\tilde{S}\left(x,E\right)-Et \tag{A.5}$$

for the case where the Hamiltonian does not depend explicitly on time. The inverse Legendre transformation says

$$t = \frac{\partial \tilde{S}(x,E)}{\partial E}. \tag{A.6}$$

This condition can be viewed as an implicit equation to determine the position of the particle x as a function of time t. To form a wave packet, waves of slightly different energies need to be put together. It results in

$$\psi(x,t) = \int g(E) e^{i\left(\tilde{S}(x,E)-Et\right)/\hbar} dE. \tag{A.7}$$

The wave function is sizable only at special points where the phase factor is stationary with respect to E. Therefore, the position of the wave packet is determined by the stationary condition

$$\frac{\partial}{\partial E}\left(\tilde{S}(x,E) - Et\right) = \frac{\partial \tilde{S}(x,E)}{\partial E} - t = 0. \tag{A.8}$$

Thus, the wave packet follows the classical equation of motion.

A.3 $\hbar$ Expansion

Since $\hbar$ is small, it can be assumed that equation (A.4) is exact as long as $\psi \neq 0$. Using a Taylor series expansion yields,

$$S(x,t) = S_0 + \hbar S_1 + \hbar^2 S_2 + \ldots \quad . \tag{A.9}$$

This is an expansion in $\hbar$, and hence called $\hbar$-expansion or the semi-classical expansion. Expanding equation (A.4) results in

$$-\frac{\partial S_0}{\partial t} = \frac{1}{2m}\left(\frac{\partial S_0}{\partial x}\right)^2 + U, \tag{A.10}$$

$$-\frac{\partial S_1}{\partial t} = \frac{1}{2m}\left[-i\left(\frac{\partial^2 S_0}{\partial x^2}\right) + 2\frac{\partial S_0}{\partial x}\frac{\partial S_1}{\partial x}\right], \tag{A.11}$$

and similarly for higher terms in $\hbar$. The leading equation has only S_0, and it is the same as the Hamilton–Jacobi equation. Once these S_i equations are solved, the wave function is obtained as a systematic expansion in $\hbar$.

There is an important difference between the "classical limit" and the idea of expansion in $\hbar$. In the classical limit, S_0 is real. However, in the case of an expansion in $\hbar$, we do not know if S_0 has to be real. In fact, in many interesting cases, S_0 turns out to be complex.

A.4 WKB Approximation

The Wentzel, Kramers, and Brillouin (WKB) Approximation, keeps terms up to $O(\hbar)$ in the $\hbar$-expansion. It is used mostly for the time-independent case, for an eigenstate of energy E. In this case, the wavefunction has the ordinary time dependence $e^{-iEt/\hbar}$. This is also restricted to the one-dimensional problem. In terms of S, this corresponds to

$$S(x,t) = S(x) - Et. \tag{A.12}$$

Therefore, only S_0 has the time dependence $S_0(x,t) = S_0(x) - Et$, while higher order terms $S_i = S_i(x)$ for $i \neq 0$ do not depend on time.

The lowest order term S_0 satisfies the Hamilton–Jacobi equation,

$$E = \frac{1}{2m}(S_0')^2 + U(x). \tag{A.13}$$

The differential equation can be solved immediately as

$$S_0(x) = \pm\int^x \sqrt{2m(E - U(x'))}\,dx' = \int^x p(x')\,dx' \tag{A.14}$$

up to an integration constant, which can be determined only after imposing a boundary condition on the wavefunction. The notation $p(x) = \pm\sqrt{2m(E - U(x))}$ is used because it is the momentum of the particle in the classical sense. Once S_0 is known, S_1 can be solved for. Starting from equation (A.11), and using $\partial S_1 / \partial t = 0$, it is found that

$$2S_0'S_1' = iS_0'', \tag{A.15}$$

which has a solution given by

$$S_1(x) = i\int^x \frac{S_0''(x')}{2S_0'(x')}dx' = \frac{i}{2}\ln p(x) + \text{constant}. \tag{A.16}$$

Therefore, the general solution to the Schrödinger equation up to this order is

$$\begin{aligned}\psi(x,t) &= e^{i(S_0(x)+\hbar S_1(x))/\hbar}e^{-iEt/\hbar} \\ &= c\frac{1}{p(x)^{1/2}}\exp\left(\pm\frac{i}{\hbar}\int^x\sqrt{2m(E-U(x'))}dx'\right)e^{-iEt/\hbar},\end{aligned} \tag{A.17}$$

and the overall constant c is undetermined from this analysis. This solution makes it immediately clear that this approximation breaks down when $p(x)$ goes to zero. In other words, the approximation is not appropriate where the classical particle stops and turns because of the potential. Such points are called “classical turning points”.

A.5 Validity of the WKB Approximation

Using only the S_0 and S_1 terms, $\hbar$ expansion is valid only when S_1 is much smaller than S_0. In particular, it is required that

$$\left|(\nabla S)^2\right| \gg \hbar\left|\nabla^2 S\right|. \tag{A.18}$$

In the one-dimensional time-independent case discussed above, this becomes

$$p(x)^2 \gg \hbar \left| p'(x) \right| . \tag{A.19}$$

Using the definition of $p(x) = \pm\sqrt{2m\left(E - U(x)\right)}$, it is found that

$$\left| \frac{\hbar \dfrac{dU(x)}{dx}}{2\left(E - U(x)\right) p(x)} \right| \ll 1 . \tag{A.20}$$

Thus, the WKB approximation breaks down close to the classical turning point $U(x) = E$ (e.g., $p(x) = 0$). For example, in a harmonic oscillator $U(x) = \frac{1}{2} m\omega^2 x^2$. In this case, the validity condition equation (A.20) can be viewed as

$$8\left| E - \frac{1}{2} m\omega^2 x^2 \right|^3 \gg \left(\hbar\omega\right)^2 m\omega^2 x^2 \tag{A.21}$$

This inequality is always satisfied exactly at the origin $x = 0$, but once away from the origin, it is impossible to satisfy unless $E \gg \hbar\omega$. In this sense, we are indeed in the classical regime. However, even for a large $E \gg \hbar\omega$, the approximation is not valid close to the classical turning point of $E = \frac{1}{2} m\omega^2 x^2$. Conversely, the validity condition equation (A.20) may be satisfied even in the region where the particle cannot enter classically $E < U(x)$. For example, with the harmonic oscillator, the validity condition is always satisfied for large $x \gg \sqrt{2E/m\omega^2}$ for any value of E. In other words, the WKB approximation is good away from the classical turning points both where a classical particle exists and where a classical particle cannot exist. Thus, the WKB approximation is not really a classical limit as applies in the purely quantum mechanical

regime. In the classically forbidden region, the solution equation (A.17) needs to be modified to

$$\begin{aligned}\psi(x,t) &= e^{i(S_0(x)+\hbar S_1(x))/\hbar} e^{-iEt/\hbar} \\ &= c\frac{1}{\left(2m(U(x)-E)\right)^{1/4}} \exp\left(\pm\frac{1}{\hbar}\int^x \sqrt{2m(U(x')-E)}dx'\right) e^{-iEt/\hbar}\end{aligned} \tag{A.22}$$

by following the same steps as in the classically allowable region.

A.6 Matching

The WKB approximation can be good both in the region $E > U(x)$ and the region $E < U(x)$ but it is not good in regions close to the classical turning point $E = U(x_c)$. In order to utilize the WKB approximation, this limitation must be overcome. The standard method is to expand the Taylor series around x_c and solve for the wave function exactly. Then, the WKB solutions can be matched away from x_c to determine the entire wavefunction. The common method is to approximate the potential around the classical turning point x_c by a linear function as

$$U(x) = U(x_c) + U'(x_c)(x - x_c) + O(x - x_c)^2 \tag{A.23}$$

and ignore the second order term. By definition, $U(x_c) = E$. Therefore, the Schrödinger equation around this point is,

$$\left(-\frac{\hbar^2}{2m}\frac{d^2}{dx^2} + U(x) - E\right)\psi = \left(-\frac{\hbar^2}{2m}\frac{d^2}{dx^2} + U'(x_c)(x - x_c)\right)\psi = 0. \tag{A.24}$$

Using the new variable

$$u = \left(\frac{2m}{\hbar^2}\frac{dV}{dx}(x_c)\right)^{1/3}(x - x_c), \tag{A.25}$$

the differential equation simplifies to

$$\left(\frac{d^2}{du^2}-u\right)\psi=0\,. \tag{A.26}$$

The solution to this equation is known as the Airy function and it can be written as

$$Ai(u)=\frac{1}{\pi}\int_0^\infty \cos\left(\frac{1}{3}t^3+ut\right)dt\,. \tag{A.27}$$

This can be verified as follows. By substituting the differential operator in equation (A.26) on the definition of the Airy function, we find

$$\left(\frac{d^2}{du^2}-u\right)Ai(u)=-\frac{1}{\pi}\int_0^\infty dt\frac{d}{dt}\sin\left(\frac{1}{3}t^3+ut\right). \tag{A.28}$$

The boundary term at $t=0$ obviously vanishes. The behavior at $t=\infty$ is not so apparent. The argument of the sin grows as t^3 and oscillates more and more rapidly as $t\to\infty$. Therefore, for any infinitesimal interval of large t, the oscillation cancels the integrand except for a "left-over" that goes down as $\sim 1/t^2$. Therefore the boundary term for $t\to\infty$ can also be dropped. It can be shown that the asymptotic behavior of the Airy function smoothly matches to WKB solutions. The asymptotic behavior is given by

$$Ai(u)\sim\begin{cases}\frac{1}{2}\left(\frac{1}{\pi\sqrt{-u}}\right)^{1/2}\exp\left(-\frac{2}{3}u^{3/2}\right) & u\gg 0\\ \left(\frac{1}{\pi\sqrt{-u}}\right)^{1/2}\cos\left(\frac{2}{3}u\sqrt{-u}+\frac{\pi}{4}\right) & u\ll 0\end{cases}. \tag{A.29}$$

Note first that, for $u\ll 0$, the asymptotic behavior is

$$Ai(u)=\left(\frac{\hbar\left(2mU'/\hbar^2\right)^{1/3}}{\pi\sqrt{2m(E-U)}}\right)^{1/2}\cos\left(\frac{1}{\hbar}\int_{x_c}^{x}\sqrt{2m(E-U)}dx'+\frac{\pi}{4}\right). \tag{A.30}$$

Here, we used the linear expansion $U(x) = E + U'(x_c)(x - x_c)$ to relate powers of u to $\sqrt{2m(E-U)}$. This expression is consistent with the WKB solution for a particular choice of the overall constant. Similarly, for $u \gg 0$, the Airy expansion becomes

$$Ai(u) = \frac{1}{2}\left(\frac{\hbar\left(2mU'/\hbar^2\right)^{1/3}}{\pi\sqrt{2m(E-U)}}\right)^{1/2} \exp\left(-\frac{1}{\hbar}\int_{x_c}^{x}\sqrt{2m(E-U)}dx'\right). \qquad \text{(A.31)}$$

A.7 Tunneling

For the study of the tunneling process, there are three regions, a classically allowed region I, $x < a$, where the particle initially exists, a classically forbidden region II $a < x < b$, and a classically allowed region III, $x > b$ to where the particle tunnels. We follow the same matching procedure as in the bound state example. The boundary at $x = a$ is obtained as

$$\psi = \left(\frac{\hbar\left(2mU'/\hbar^2\right)^{1/3}}{\pi\sqrt{2m(E-U)}}\right)^{1/2} \cos\left(-\frac{1}{\hbar}\int_{x}^{a}\sqrt{2m(E-U)}dx' + \frac{\pi}{4}\right) \quad x < a, \qquad \text{(A.32)}$$

and

$$\psi \simeq \frac{1}{2}\left(\frac{\hbar\left(2mU'/\hbar^2\right)^{1/3}}{\pi\sqrt{2m(E-U)}}\right)^{1/2} \exp\left(-\frac{1}{\hbar}\int_{a}^{x}\sqrt{2m(U-E)}dx'\right) \qquad x > a. \qquad \text{(A.33)}$$

The matching at $x = b$, on the other hand, becomes

$$\psi \simeq C\frac{i}{2}\left(\frac{\hbar\left(2mU'/\hbar^2\right)^{1/3}}{\pi\sqrt{2m(E-U)}}\right)^{1/2} \exp\left(-\frac{1}{\hbar}\int_{x}^{b}\sqrt{2m(U-E)}dx'\right) \quad x < b, \qquad \text{(A.34)}$$

and

$$\psi \simeq C\frac{i}{2}\left(\frac{\hbar\left(2mU'/\hbar^2\right)^{1/3}}{\pi\sqrt{2m\left(E-U\right)}}\right)^{1/2}\sin\left(-\frac{1}{\hbar}\int_b^x\sqrt{2m\left(E-U\right)}dx'+\frac{\pi}{4}\right)\qquad x>b\,. \quad \text{(A.35)}$$

The overall normalization factor C is determined by the requirement that the behavior of the wave function is consistent for $a<x<b$ between two matching procedures. We therefore find

$$C=-i\left(\frac{U'(a)}{U'(b)}\right)^{1/6}\exp\left(-\frac{1}{\hbar}\int_a^b\sqrt{2m\left(U(x)-E\right)}dx\right)\quad . \qquad \text{(A.36)}$$

Comparing the two classically allowed regions, and taking advantage of a further normalization change, the matching reduces to

$$\psi \simeq \frac{1}{\left[2m\left(E-U\right)\right]^{1/4}}\cos\left(-\frac{1}{\hbar}\int_x^a\sqrt{2m\left(E-U\right)}dx'+\frac{\pi}{4}\right)\qquad x<a\,, \qquad \text{(A.37)}$$

and

$$\psi \simeq \frac{1}{2}\frac{1}{\left[2m\left(E-U\right)\right]^{1/4}}\exp\left(-\frac{1}{\hbar}\int_a^b\sqrt{2m\left(U(x)-E\right)}dx\right)\sin\left(-\frac{1}{\hbar}\int_b^x\sqrt{2m\left(E-U\right)}dx'+\frac{\pi}{4}\right)\qquad x>b \quad . \qquad \text{(A.38)}$$

In other words, the amplitude in the region $x>b$ due to tunneling from region $x>a$ is suppressed by

$$\frac{1}{2}\exp\left(-\frac{1}{\hbar}\int_a^b\sqrt{2m\left(U(x)-E\right)}dx\right), \qquad \text{(A.39)}$$

Therefore

$$\exp\left(-2\frac{1}{\hbar}\int_a^b\sqrt{2m\left(U(x)-E\right)}dx\right) \qquad \text{(A.40)}$$

Is a transmission coefficient normally referred to as a suppression factor for the tunneling rate (square of the amplitude).

APPENDIX B

TUNNELING THROUGH RECTANGULAR BARRIER

The electron wave function satisfies the Schrödinger equation, such that

$$-\frac{\hbar^2}{2m}\nabla^2\psi(x,y,z)+U(x,y,z)\psi(x,y,z)=E\psi(x,y,z). \tag{B.1}$$

Figure B.1 shows the simplest case where a particle of energy E in space with a rectangular potential barrier of height E_b and width d. In one dimension, the Schrödinger equation can be written in different regions as follows

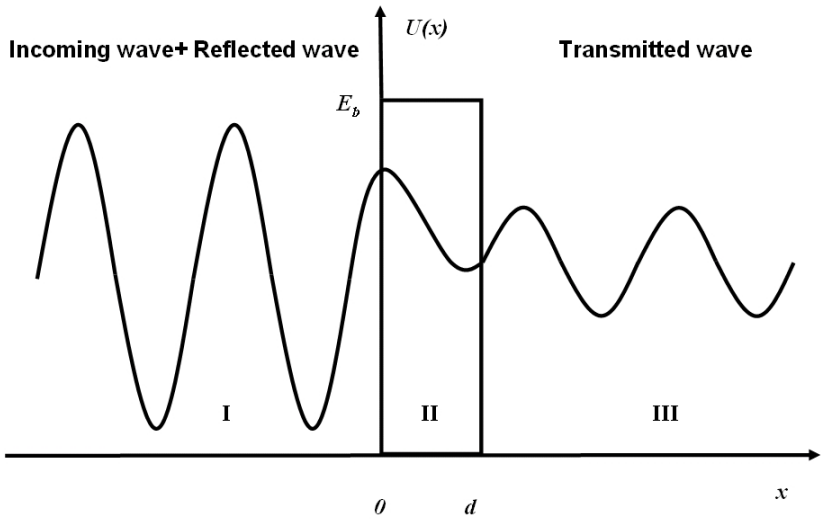

Figure B.1
Diagram of wavefunction of a particle with energy E tunneling through a rectangular potential barrier of height E_b and thickness d

$$d^2\psi / dx^2 + k_I{}^2\psi = 0 \qquad \text{for} \quad x < 0 \qquad \text{(B.2)}$$

$$d^2\psi / dx^2 + k_{II}{}^2\psi = 0 \qquad \text{for} \quad 0 \le x \le d \qquad \text{(B.3)}$$

$$d^2\psi / dx^2 + k_{III}{}^2\psi = 0 \qquad \text{for} \quad x \ge d\,, \qquad \text{(B.4)}$$

where

$$k_I{}^2 = k_{III}{}^2 = \frac{2mE}{\hbar^2},$$

$$k_{II}{}^2 = n_b^2 k_I{}^2,$$

and

$$n_b^2 = \frac{(E - E_b)}{E}.$$

In the region I, the solution of equation (B.2) can be written as

$$\psi_I = Ae^{ik_I x} + A'e^{-ik_I x}. \qquad \text{(B.5)}$$

Similarly, in $0 \le x \le d$,

$$\psi_{II} = Be^{ik_{II} x} + B'e^{-ik_{II} x} \qquad \text{(B.6)}$$

and in $x > d$

$$\psi_{III} = Ce^{ik_I x} + C'e^{-ik_I x}. \qquad \text{(B.7)}$$

In the region with $x > d$, there is no particle moving in the negative direction, thus

$$C' = 0. \qquad \text{(B.8)}$$

To satisfy the continuity condition at $x = 0$, it is required that

$$A + A' = B + B'. \qquad \text{(B.9)}$$

From $\left(\frac{d\psi_I}{dx}\right)_{x=0} = \left(\frac{d\psi_{II}}{dx}\right)_{x=0}$,

$$k_I A - k_I A' = k_{II} B - k_{II} B'. \qquad \text{(B.10)}$$

From $\psi_{II}\big|_{x=d} = \psi_{III}\big|_{x=d}$,

$$Be^{ik_{II}d} + B'e^{-ik_{II}d} = Ce^{ik_{I}d}, \tag{B.11}$$

and from $\left(\frac{d\psi_{II}}{dx}\right)_{x=d} = \left(\frac{d\psi_{III}}{dx}\right)_{x=d}$,

$$k_{II}Be^{ik_{II}d} - k_{II}B'e^{-ik_{II}d} = k_{I}Ce^{ik_{I}d}. \tag{B.12}$$

By solving the equations (B.9), (B.10), (B.11) and (B.12), it is obtained that

$$C = \frac{2k_{I}k_{II}e^{-ik_{I}d}}{\left(k_{I}^{\,2} - k_{II}^{\,2}\right)\sinh\left(-ik_{II}d\right) + 2k_{I}k_{II}\cosh\left(-ik_{II}d\right)}A. \tag{B.13}$$

For the case that $E < E_b$, k_{II} is a virtual number. The tunneling probability is

$$D = \frac{|C|^2}{|A|^2} = \frac{-4k_{I}^{\,2}k_{II}^{\,2}}{\left(k_{I}^{\,2} - k_{II}^{\,2}\right)^2 \sinh^2\left(-idk_{II}\right) - 4k_{I}^{\,2}k_{II}^{\,2}}. \tag{B.14}$$

If the energy of the particle is small and satisfies $dk_{II} >> 1$, then $e^{k_{II}d} >> e^{-k_{II}d}$ and $\sinh\left(-ik_{II}d\right)$ can be approximated by $\frac{1}{4}e^{-2ik_{II}d}$, or

$$\sinh^2\left(-ik_{II}d\right) = \left(\frac{e^{-ik_{II}d} - e^{ik_{II}d}}{2}\right) \approx \frac{1}{4}e^{-2ik_{II}d}. \tag{B.15}$$

Thus, equation (B.14) can be written as

$$D = \frac{4}{-\frac{1}{4}\frac{\left(k_{I}^{\,2} - k_{II}^{\,2}\right)^2}{k_{I}^{\,2}k_{II}^{\,2}}e^{-2idk_{II}} + 4}. \tag{B.16}$$

Since the magnitudes of k_I and k_{II} are of the same order, and $e^{-2ik_{II}d} >> 4$ for $dk_{II} >> 1$, the expression above can be written as

$$E = D_0 e^{-2ik_{II}d} = D_0 e^{-\frac{2}{\hbar}\sqrt{2m(E_b - E)}d}, \tag{B.17}$$

where D_0 is a constant whose order of magnitude is 1.

APPENDIX C

VARIATION METHOD

The variation method in quantum mechanics is used to obtain an approximate solution of the ground state for a given system described by Schrödinger equation

$$-\frac{\hbar^2}{2m}\frac{d^2\psi(x)}{dx^2}-U(x)\psi(x)=E\psi(x). \tag{C.1}$$

The solution is a series of eigenstates given by orthogonal wavefunctions ψ_1, $\psi_2\ldots$ with the eigen-energies E_1, E_2, … By definition, the eigen-energy state is given by the Hamiltonian of the wavefunction, which is obtained by applying the Hamiltonian operator on the given states. The Hamiltonian operator is denoted by H and the operation is denoted by $\langle\psi|H|\psi\rangle$. This is obtained from

$$\langle\psi|H|\psi\rangle=\int_{-\infty}^{+\infty}\psi^*(x)\left(-\frac{\hbar^2}{2m}\frac{d}{dx^2}+U(x)\right)\psi(x)\,dx, \tag{C.2}$$

and the energy states E_1, E_2, …are given by

$$E_i=\langle\psi_i|H|\psi_i\rangle. \tag{C.3}$$

Starting with a trial wavefunction ψ_T that satisfies all boundary conditions of (C.1). The potential can be denoted as a linear combination of different Hamiltonian eigenstates as

$$\psi_T(x)=\sum_i c_i\psi_i(x), \tag{C.4}$$

with normalized coefficients c_i satisfying

$$\sum_i|c_i|^2=1, \tag{C.5}$$

so that

$$\left|\psi_T\right|^2 = 1. \tag{C.6}$$

The expectation value of the energy with this trial wavefunction is

$$E = \left\langle \psi_T \middle| H \middle| \psi_T \right\rangle = \left|c_1\right|^2 E_1 + \sum_{i>1} \left|c_i\right|^2 E_i \geq \left|c_1\right|^2 E_1 + \sum_{i>1} \left|c_i\right|^2 E_1 = E_1 . \tag{C.7}$$

Thus, the expectation value must be always greater than or equal to the ground state energy. This provides an upper limit value for the ground state energy.

Given this observation, the ground state can be obtained by introducing parameters in the trial wavefunction, $\psi_T(\lambda)$. The expectation value of energy for $\psi_T(\lambda)$ is given by

$$E(\lambda) = \left\langle \psi_T(\lambda) \middle| H \middle| \psi_T(\lambda) \right\rangle . \tag{C.8}$$

Looking for the lowest possible expectation value,

$$\frac{dE(\lambda)}{d\lambda} = 0 \tag{C.9}$$

is solved for a particular value of λ. The success of this method depends greatly on having a good trial wavefunction and good parameter choices.

APPENDIX D

CLASSICAL CHARGE MODEL

The potential distribution $\phi(x)$ in the channel can be given by the 1-D Poisson equation:

$$\frac{d^2\phi(x)}{dx^2} = \frac{q}{\varepsilon_{Si}}\left(N_A + n\right) \tag{D.1}$$

where N_A is the volume density of depletion charges and $n(x)$ is the volume density inversion charges. In classical physics, the $n(x)$ is given by

$$n(x) = \frac{n_i^2}{N_A}\exp\left(-q\phi / kT\right). \tag{D.2}$$

The Poisson's equation changes to

$$\frac{d^2\phi(x)}{dx^2} = \frac{q}{\varepsilon_{Si}}\left(N_A + \frac{n_i^2}{N_A}\exp\left(\frac{q\phi}{kT}\right)\right). \tag{D.3}$$

Noting that

$$\frac{d^2\phi(x)}{dx^2} = \frac{d}{d\phi}\left(\frac{d\phi(x)}{dx}\right)\cdot\frac{d\phi}{dx}, \tag{D.4}$$

Equation (D.3) can be rewritten as

$$\frac{d\phi}{dx}d\left(\frac{d\phi(x)}{dx}\right) = \frac{2q}{\varepsilon_{Si}}\left(N_A + \frac{n_i^2}{N_A}\exp\left(\frac{q\phi}{kT}\right)\right)d\phi. \tag{D.5}$$

Integrating (D.5) from the bulk toward the surface leads to

$$\int_0^{\frac{d\phi}{dx}}\frac{d\phi}{dx}d\left(\frac{d\phi}{dx}\right) = \int_0^{\phi}\frac{2q}{\varepsilon_{Si}}\left(N_A + \frac{n_i^2}{N_A}\exp\left(\frac{q\phi}{kT}\right)\right)d\phi. \tag{D.6}$$

Given the relationship between the electric field and the potential ϕ, $E_{field} = -d\phi / dx$, the electric field can be obtained from equation (D.6) as

$$E_{Field} = \sqrt{\frac{2q}{\varepsilon_{Si}\beta}}\left[N_A\beta\phi + \frac{n_i^2}{N_A}\left(\exp(\beta\phi) - 1\right)\right]^{1/2}, \tag{D.7}$$

where $\beta = q / kT$. By Gauss' law, the total charge density is given by

$$Q_T = Q_{depl} + Q_{inv} = -\varepsilon_{Si}\, E_{Field}\big|_{x=0}. \tag{D.8}$$

With depletion charge density given by

$$Q_{depl} = \sqrt{2q\varepsilon_{Si}\phi_s}\,, \tag{D.9}$$

the inversion charge density can be obtained as

$$Q_{inv} = \sqrt{\frac{2q\varepsilon_{Si}N_A}{\beta}}\left\{\left[\beta\phi_s + \frac{n_i^2}{N_A^2}\left(\exp(\beta\phi_s) - 1\right)\right]^{1/2} - (\beta\phi_s)^{1/2}\right\}, \tag{D.10}$$

In the weak inversion region, $\phi_B < \phi_s < 2\phi_B$, $\beta\phi_s >> \left(\frac{n_i^2}{N_A^2}\right)\exp(\beta\phi_s)$, Q_{inv} is approximated by

$$Q_{inv} = \sqrt{\frac{q\varepsilon_{Si}N_A}{2\phi_s}}\,\frac{n_i^2}{N_A^2}\exp(\beta\phi_s), \tag{D.11}$$

and inversion layer capacitance is given by

$$C_{inv} = \frac{\partial Q_{inv}}{\partial\phi_s} \approx \beta Q_{inv}. \tag{D.12}$$

In the strong inversion region, $\phi_s > 2\phi_B$, $\left(\frac{n_i^2}{N_A^2}\right)\exp(\beta\phi_s) >> \beta\phi_s$, Q_{inv} is approximated by

$$Q_{inv} = \sqrt{\frac{q\varepsilon_{Si} n_i^2}{2\beta N_A} \exp(\beta\phi_s)}, \tag{D.13}$$

and inversion layer capacitance is given by

$$C_{inv} = \frac{\partial Q_{inv}}{\partial \phi_s} = \frac{1}{2}\beta Q_{inv}. \tag{D.14}$$

REFERENCES

[1] G. E. Moore, "Cramming more components onto integrated circuits", *Proceedings of the IEEE*, vol. 86, no. 1, pp. 82-85, 1998.

[2] "*International Technology Roadmap for Semiconductors (ITRS)*", SIA, 2003.

[3] S. E. Thompson, M. Armstrong, C. Auth, M. Alavi, M. Buehler, R. Chau, S. Cea, T. Ghani, G. Glass, T. Hoffman, C. H. Jan, C. Kenyon, J. Klaus, K. Kuhn, M. Zhiyong, B. Mcintyre, K. Mistry, A. Murthy, B. Obradovic, R. Nagisetty, N. Phi, S. Sivakumar, R. Shaheed, L. Shifren, B. Tufts, S. Tyagi, M. Bohr, and Y. El-Mansy, "A 90-nm logic technology featuring strained-silicon", *Electron Devices, IEEE Transactions on*, vol. 51, no. 11, pp. 1790-1797, 2004.

[4] 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 Journal of Research and Development*, vol. 43, no. 3, pp. 327-337, May1999.

[5] 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 nMOSFET's", *Electron Device Letters, IEEE*, vol. 18, no. 5, pp. 209-211, May1997.

[6] T. Yuan, D. A. Buchanan, C. Wei, D. J. Frank, K. E. Ismail, L. Shih-Hsien, G. A. Sai-Halasz, R. G. Viswanathan, H. J. C. Wann, S. J. Wind, and W. Hon-Sum, "CMOS scaling into the nanometer regime", *Proceedings of the IEEE*, vol. 85, no. 4, pp. 486-504, 1997.

[7] G. Timp, J. Bude, K. K. Bourdelle, J. Garno, A. Ghetti, H. Gossmann, M. Green, G. Forsyth, Y. Kim, R. Kleiman, F. Klemens, A. Kornblit, C. Lochstampfor, W. Mansfield, S. Moccio, T. Sorsch, D. M. Tennant, W. Timp, and R. Tung, "The ballistic nano-transistor", *IEDM Tech. Dig.* 1999, pp. 55-58.

[8] J.J.Sakurai, *Modern Quantum Mechanics,* 1995.

[9] Mark Lundstrom, *Fundamentals of Carrier Transport,* Second edition, Cambridge University Press, 2000.

[10] S.M.Sze, *Physics of Semiconductor Devices,* 2nd edition, New York: Wiley, 1981.

[11] J. D. Meindl, "Low power microelectronics: retrospect and prospect", *Proceedings of the IEEE*, vol. 83, no. 4, pp. 619-635, 1995.

[12] Z. Luo, A. Steegen, M. Eller, M. Mann, C. Baiocco, P. Nguyen, L. Kim, M. Hoinkis, V. Ku, V. Klee, F. Jamin, P. Wrschka, P. Shafer, W. Lin, S. Fang, A.

Ajmera, W. Tan, D. Park, R. Mo, J. Lian, D. Vietzke, C. Coppock, A. Vayshenker, T. Hook, V. Chan, K. Kim, A. Cowley, S. Kim, E. Kaltalioglu, B. Zhang, S. Marokkey, Y. Lin, K. Lee, H. Zhu, M. Weybright, R. Rengarajan, J. Ku, T. Schiml, J. Sudijono, I. Yang, and C. Wann, "High performance and low power transistors integrated in 65nm bulk CMOS technology", *IEDM Tech. Dig.* 2004, pp. 661-664.

[13] C. H. Choi, K. Y. Nam, Z. P. Yu, and R. W. Dutton, "Impact of gate direct tunneling current on circuit performance: A simulation study", *Electron Devices, IEEE Transactions on*, vol. 48, no. 12, pp. 2823-2829, Dec.2001.

[14] W. C. Lee, T. J. King, and C. M. Hu, "Evidence of hole direct tunneling through ultrathin gate oxide using P+ Poly-SiGe gate", *Electron Device Letters, IEEE*, vol. 20, no. 6, pp. 268-270, June1999.

[15] W. C. Lee and C. M. Hu, "Modeling CMOS tunneling currents through ultrathin gate oxide due to conduction- and valence-band electron and hole tunneling", *Electron Devices, IEEE Transactions on*, vol. 48, no. 7, pp. 1366-1373, July2001.

[16] 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", *Electron Devices, IEEE Transactions on*, vol. 46, no. 7, pp. 1464-1471, July1999.

[17] K. N. Yang, H. T. Huang, M. C. Chang, C. M. Chu, Y. S. Chen, M. J. Chen, Y. M. Lin, M. C. Yu, S. M. Jang, D. C. H. Yu, and M. S. Liang, "A physical model for hole direct tunneling current in P+ poly-gate PMOSFETs with ultrathin gate oxides", *Electron Devices, IEEE Transactions on*, vol. 47, no. 11, pp. 2161-2166, Nov.2000.

[18] R. Clerc, A. Spinelli, G. Ghibaudo, and G. Pananakakis, "Theory of direct tunneling current in metal-oxide-semiconductor structures", *Journal of Applied Physics*, vol. 91, no. 3, pp. 1400-1409, Feb.2002.

[19] J. Lee, G. Bosman, K. R. Green, and D. Ladwig, "Model and analysis of gate leakage current in ultrathin nitrided oxide MOSFETs", *Electron Devices, IEEE Transactions on*, vol. 49, no. 7, pp. 1232-1241, July2002.

[20] X. Y. Liu, J. F. Kang, and R. Q. Han, "Direct tunneling current model for MOS devices with ultra-thin gate oxide including quantization effect and polysilicon depletion effect", *Solid State Communications*, vol. 125, no. 3-4, pp. 219-223, Jan.2003.

[21] H. Ananthan, A. Bansal, and K. Roy, "Analysis of drain-to-body band-to-band tunneling in double gate MOSFET", *IEEE International SOI Conference* 2005, pp. 159-160.

[22] C. Chang-Hoon, Y. Shyh-Horng, G. Pollack, S. Ekbote, P. R. Chidambaram, S. Johnson, C. Machala, and R. W. Dutton, "Characterization of Zener-tunneling drain leakage current in high-dose halo implants", *SISPAD* 2003, pp. 133-136.

[23] L. Yo-Sheng, W. Chung-Cheng, C. Chih-Sheng, Y. Rong-Ping, C. Wei-Ming, L. Jhon-Jhy, and C. H. Diaz, "Leakage scaling in deep submicron CMOS for SoC", *Electron Devices, IEEE Transactions on*, vol. 49, no. 6, pp. 1034-1041, 2002.

[24] H. Nakajima, S. Yanagi, K. Komiya, and Y. Omura, "Off-leakage and drive current characteristics of sub-100-nm SOI MOSFETs and impact of quantum tunnel current", *Electron Devices, IEEE Transactions on*, vol. 49, no. 10, pp. 1775-1782, 2002.

[25] W. Jing and M. Lundstrom, "Does source-to-drain tunneling limit the ultimate scaling of MOSFETs?", *IEDM Dig.* 2002, pp. 707-710.

[26] M. Bescond, J. L. Autran, D. Munteanu, N. Cavassilas, and M. Lannoo, "Atomic-scale modeling of source-to-drain tunneling in ultimate Schottky barrier double-gate MOSFETs", *ESSDERC* 2003, pp. 395-398.

[27] S. A. Hareland, M. Manassian, W. K. Shih, S. Jallepalli, H. Wang, G. L. Chindalore, A. Tasch, and C. M. Maziar, "Computationally efficient models for quantization effects in MOS electron and hole accumulation layers", *Electron Devices, IEEE Transactions on*, vol. 45, no. 7, pp. 1487-1493, 1998.

[28] S. Takagi, M. Takayanagi, and A. Toriumi, "Impact of electron and hole inversion-layer capacitance on low voltage operation of scaled n- and p-MOSFET's", *Electron Devices, IEEE Transactions on*, vol. 47, no. 5, pp. 999-1005, 2000.

[29] M. J. van Dort, P. H. Woerlee, A. J. Walker, C. A. H. Juffermans, and H. Lifka, "Influence of high substrate doping levels on the threshold voltage and the mobility of deep-submicrometer MOSFETs", *Electron Devices, IEEE Transactions on*, vol. 39, no. 4, pp. 932-938, 1992.

[30] Q. Wuyun, D. M. Kim, and L. Hi-Deok, "Quantum C-V modeling in depletion and inversion: accurate extraction of electrical thickness of gate oxide in deep submicron MOSFETs", *Electron Devices, IEEE Transactions on*, vol. 49, no. 5, pp. 889-894, 2002.

[31] D. Vasileska, D. K. Schroder, and D. K. Ferry, "Scaled silicon MOSFETs: degradation of the total gate capacitance", *Electron Devices, IEEE Transactions on*, vol. 44, no. 4, pp. 584-587, 1997.

[32] "SCHRED 2.0", http://www.punch.purdue.edu: 5/2/04.

[33] L. Wang, Q. Chen, R. Murali, and J. D. Meindl, "Quantum mechanical effects on CMOS SOC performance", *SOC Conference Proc.* 2003, pp. 109-112.

[34] K. S. Krisch, J. D. Bude, and L. Manchanda, "Gate capacitance attenuation in MOS devices with thin gate dielectrics", *Electron Device Letters, IEEE*, vol. 17, no. 11, pp. 521-524, 1996.

[35] Y. Ohkura, "Quantum effects in Si n-MOS inversion layer at high substrate concentration", *Solid-State Electronics*, vol. 33, no. 12, pp. 1581-1585, Dec.1990.

[36] S. Yang, S. Ahmed, B. Arcot, R. Arghavani, P. Bai, S. Chambers, P. Charvat, R. Cotner, R. Gasser, T. Ghani, M. Hussein, C. Jan, C. Kardas, J. Maiz, P. McGregor, B. Mcintyre, P. Nguyen, P. Packan, I. Post, S. Sivakumar, J. Steigerwald, M. Taylor, B. Tufts, S. Tyagi, and M. Bohr, "A high performance 180 nm generation logic technology", *IEDM Tech. Dig.* 1998, pp. 197-200.

[37] P. Bai, C. Auth, S. Balakrishnan, M. Bost, R. Brain, V. Chikarmane, R. Heussner, M. Hussein, J. Hwang, D. Ingerly, R. James, J. Jeong, C. Kenyon, E. Lee, S. H. Lee, N. Lindert, M. Liu, Z. Ma, T. Marieb, A. Murthy, R. Nagisetty, S. Natarajan, J. Neirynck, A. Ott, C. Parker, J. Sebastian, R. Shaheed, S. Sivakumar, J. Steigerwald, S. Tyagi, C. Weber, B. Woolery, A. Yeoh, K. Zhang, and M. Bohr, "A 65nm logic technology featuring 35nm gate lengths, enhanced channel strain, 8 Cu interconnect layers, low-k ILD and 0.57 μm^2 SRAM cell", *IEDM Tech. Dig.* 2004, pp. 657-660.

[38] S. Tyagi, M. Alavi, R. Bigwood, T. Bramblett, J. Brandenburg, W. Chen, B. Crew, M. Hussein, P. Jacob, C. Kenyon, C. Lo, B. Mcintyre, Z. Ma, P. Moon, P. Nguyen, L. Rumaner, R. Schweinfurth, S. Sivakumar, M. Stettler, S. Thompson, B. Tufts, J. Xu, S. Yang, and M. Bohr, "A 130 nm generation logic technology featuring 70 nm transistors, dual Vt transistors and 6 layers of Cu interconnects", *IEDM Tech. Dig.* 2000, pp. 567-570.

[39] C. H. Jan, J. Bielefeld, M. Buehler, V. Chikamane, K. Fischer, T. Hepburn, A. Jain, J. Jeong, T. Kielty, S. Kook, T. Marieb, B. Miner, P. Nguyen, A. Schmitz, M. Nashner, T. Scherban, B. Schroeder, P. H. Wang, R. Wu, J. Xu, K. Zawadzki, S. Thompson, and M. Bohr, "90 nm generation, 300 mm wafer low k ILD/Cu interconnect technology", *IITC* 2003, pp. 15-17.

[40] C. S. Rafferty, B. Biegel, M. G. Ancona, J. Bude, and R. W. Dutton, "Multi-dimensional quantum effect simulation using a density-gradient model and script-level programming techniques", *Simulation of Semiconductor Processes and Devices 1998. SISPAD 98* Leuven, Belgium: Springer-Verlag/Wien, 1998, pp. 137-140.

[41] A. Abramo, A. Cardin, L. Selmi, and E. Sangiorgi, "Two-dimensional quantum mechanical simulation of charge distribution in silicon MOSFETs", *Electron Devices, IEEE Transactions on*, vol. 47, no. 10, pp. 1858-1863, 2000.

[42] S. Jallepalli, J. Bude, W. K. Shih, M. R. Pinto, C. M. Maziar, and A. F. Tasch, Jr.,

"Effects of quantization on the electrical characteristics of deep submicron p- and n-MOSFETs", *VLSI Technology, 1996. Digest of Technical Papers. 1996 Symposium on* 1996, pp. 138-139.

[43] T. Janik and B. Majkusiak, "Analysis of the MOS transistor based on the self-consistent solution to the Schrodinger and Poisson equations and on the local mobility model", *Electron Devices, IEEE Transactions on*, vol. 45, no. 6, pp. 1263-1271, 1998.

[44] A. Pirovano, A. L. Lacaita, and A. S. Spinelli, "Two-dimensional quantum effects in nanoscale MOSFETs", *Electron Devices, IEEE Transactions on*, vol. 49, no. 1, pp. 25-31, 2002.

[45] A. S. Spinelli, A. Benvenuti, and A. Pacelli, "Self-consistent 2-D model for quantum effects in n-MOS transistors", *Electron Devices, IEEE Transactions on*, vol. 45, no. 6, pp. 1342-1349, 1998.

[46] S. Takagi, A. Toriumi, M. Iwase, and H. Tango, "On the universality of inversion layer mobility in Si MOSFET's: Part I-effects of substrate impurity concentration", *Electron Devices, IEEE Transactions on*, vol. 41, no. 12, pp. 2357-2362, 1994.

[47] S. Takagi, A. Toriumi, M. Iwase, and H. Tango, "On the universality of inversion layer mobility in Si MOSFET's: Part II-effects of surface orientation", *Electron Devices, IEEE Transactions on*, vol. 41, no. 12, pp. 2363-2368, 1994.

[48] H. S. Momose, M. Ono, T. Yoshitomi, T. Ohguro, S. Nakamura, M. Saito, and H. Iwai, "Tunneling gate oxide approach to ultra-high current drive in small geometry MOSFETs", *International Electron Devices Meeting 1994. Technical Digest* 1994, pp. 593-596.

[49] N. Yang, W. K. Henson, and J. J. Wortman, "A comparative study of gate direct tunneling and drain leakage currents in N-MOSFET's with sub-2-nm gate oxides", *Electron Devices, IEEE Transactions on*, vol. 47, no. 8, pp. 1636-1644, Aug.2000.

[50] K. F. Schuegraf, C. C. King, and C. Hu, "Ultra-thin silicon dioxide leakage current and scaling limit", *VLSI Technology, 1992. Digest of Technical Papers. 1992 Symposium on* 1992, pp. 18-19.

[51] K.A.Bowman, L.Wang, T. Xinghai, and J.D.Meindl, "Oxide thickness scaling limit for optimum CMOS logic circuit performance", *ESSDERC 2000. Proceedings of the 30th European Solid-State Device Research Conference* 2000, pp. 300-303.

[52] F. Stern, "Self-consistent results for n-Type Si inversion layers", *Physics Review B*, vol. 5, no. 12, pp. 4891-4899, June1972.

[53] F.F.Fang and W.E.Howard, "Negative field-effect mobility on (100) Si surfaces", *Physics Review letters*, vol. 16, no. 18, p. 797, 1966.

[54] S. Takagi and A. Toriumi, "Quantitative understanding of inversion-layer capacitance in Si MOSFET's", *Electron Devices, IEEE Transactions on*, vol. 42, no. 12, pp. 2125-2130, 1995.

[55] S. A. Hareland, S. Krishnamurthy, S. Jallepalli, Y. Choh-Fei, K. Hasnat, A. F. Tasch, Jr., and C. M. Maziar, "A computationally efficient model for inversion layer quantization effects in deep submicron N-channel MOSFETs", *Electron Devices, IEEE Transactions on*, vol. 43, no. 1, pp. 90-96, 1996.

[56] B. K. Ip and J. R. Brews, "Quantum effects upon drain current in a biased MOSFET", *Electron Devices, IEEE Transactions on*, vol. 45, no. 10, pp. 2213-2221, 1998.

[57] C. Lallement, J. M. Sallese, M. Bucher, W. Grabinski, and P. C. Fazan, "Accounting for quantum effects and polysilicon depletion from weak to strong inversion in a charge-based design-oriented MOSFET model", *Electron Devices, IEEE Transactions on*, vol. 50, no. 2, pp. 406-417, 2003.

[58] A. Pacelli, A. S. Spinelli, and L. M. Perron, "Carrier quantization at flat bands in MOS devices", *Electron Devices, IEEE Transactions on*, vol. 46, no. 2, pp. 383-387, 1999.

[59] K. Yang, K. Ya-Chin, and H. Chenming, "Quantum effect in oxide thickness determination from capacitance measurement", *VLSI Technology, 1999. Digest of Technical Papers. 1999 Symposium on* 1999, pp. 77-78.

[60] B. Agrawal, V. K. De, J. M. Pimbley, and J. D. Meindl, "Short channel models and scaling limits of SOI and bulk MOSFETs", *Solid-State Circuits, IEEE Journal of*, vol. 29, no. 2, pp. 122-125, 1994.

[61] B. Agrawal, V. K. De, and J. D. Meindl, "Three-dimensional analytical subthreshold models for bulk MOSFETs", *Electron Devices, IEEE Transactions on*, vol. 42, no. 12, pp. 2170-2180, 1995.

[62] D. J. Frank, R. H. Dennard, E. Nowak, P. M. Solomon, Y. Taur, and P. W. Hon-Sum, "Device scaling limits of Si MOSFETs and their application dependencies", *Proceedings of the IEEE*, vol. 89, no. 3, pp. 259-288, 2001.

[63] D. J. Frank, Y. Taur, and H. S. P. Wong, "Generalized scale length for two-dimensional effects in MOSFETs", *Electron Device Letters, IEEE*, vol. 19, no. 10, pp. 385-387, 1998.

[64] Omura Y., Konishi H., and Sato S., "Quantum-Mechanical Suppression and Enhancement of Short-Channel Effects in Ultra-Thin SOI MOSFETs", PP ed 2006.

[65] F. Pregaldiny, C. Lallement, and D. Mathiot, "Quantum surface potential model suitable for advanced MOSFETs simulation", *IEEE International Conference on Simulation of Semiconductor Processes and Devices* 2003, pp. 227-230.

[66] M. V. Fischetti, S. E. Laux, and A. Kumar, "Simulation of quantum electronic transport in small devices: a master equation approach", *IEDM Tech. Dig.* 2003, p. 19.

[67] "*International Technology Roadmap for Semiconductors (ITRS)*", SIA, 2005.

[68] A. J. Bhavnagarwala, B. L. Austin, K. A. Bowman, and J. D. Meindl, "A minimum total power methodology for projecting limits on CMOS GSI", *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 8, no. 3, pp. 235-251, 2000.

[69] S. Borkar, "Design challenges of technology scaling", *Micro, IEEE*, vol. 19, no. 4, pp. 23-29, 1999.

[70] D. A. Buchanan, "Scaling the gate dielectric: materials, integration, and reliability", *Ibm Journal of Research and Development*, vol. 43, no. 3, pp. 245-264, May1999.

[71] K. A. Bowman, L. Wang, X. Tang, and J. D. Meindl, "A circuit-level perspective of the optimum gate oxide thickness", *Electron Devices, IEEE Transactions on*, vol. 48, no. 8, pp. 1800-1810, 2001.

[72] R. M. Wallace and G. D. Wilk, "High-κ Dielectric Materials for Microelectronics", *Critical Reviews in Solid State and Materials Sciences*, vol. 28, no. 4, p. 55, 2003.

[73] K. Yamamoto, W. Deweerd, M. Aoulaiche, M. Houssa, S. De Gendt, S. Horii, M. Asai, A. Sano, S. Hayashi, and M. Niwa, "Electrical and physical characterization of remote plasma oxidized HfO_2 gate dielectrics", *Electron Devices, IEEE Transactions on*, vol. 53, no. 5, pp. 1153-1160, 2006.

[74] P. M. Zeitzoff, "MOSFET scaling trends and challenges through the end of the roadmap", *Proc. IEEE Custom Integrated Circuits Conference* 2004, pp. 233-240.

[75] L. Kang, K. Onishi, Y. Jeon, H. L. Byoung, C. Kang, Q. Wen-Jie, R. Nieh, S. Gopalan, R. Choi, and J. C. Lee, "MOSFET devices with polysilicon on single-layer HfO_2 high-K dielectrics", *IEDM Tech. Dig.* 2000, pp. 35-38.

[76] K. Onishi, S. K. Chang, C. Rino, C. Hag-Ju, H. K. Young, S. Krishnan, M. S. Akbar, and J. C. Lee, "Performance of polysilicon gate HfO_2/MOSFETs on [100] and [111] silicon substrates", *Electron Device Letters, IEEE*, vol. 24, no. 4, pp. 254-256, 2003.

[77] C. Qiang, W. Lihui, and J. D. Meindl, "Impact of high-κ dielectrics on undoped double-gate MOSFET scaling", *IEEE International SOI Conference* 2002, pp. 115-116.

[78] T. Skotnicki, J. A. Hutchby, K. Tsu-Jae, H. S. P. Wong, and F. Boeuf, "The end of CMOS scaling: toward the introduction of new materials and structural changes to improve MOSFET performance", *Circuits and Devices Magazine, IEEE*, vol. 21, no. 1, pp. 16-26, 2005.

[79] Yuan, Taur and Ning, T. H., *Fundamentals of Modern VLSI Devices,* Cambridge University Press, New York, 1998.

[80] L. Yo-Sheng, W. Chung-Cheng, C. Chih-Sheng, Y. Rong-Ping, C. Wei-Ming, L. Jhon-Jhy, and C. H. Diaz, "Leakage scaling in deep submicron CMOS for SoC", *Electron Devices, IEEE Transactions on*, vol. 49, no. 6, pp. 1034-1041, 2002.

[81] A. Agarwal, S. Mukhopadhyay, C. H. Kim, A. Raychowdhury, and K. Roy, "Leakage power analysis and reduction: models, estimation and tools", *Computers and Digital Techniques, IEE Proceedings-*, vol. 152, no. 3, pp. 353-368, 2005.

[82] C. Ming-Jer, H. Huan-Tsung, H. Chin-Shan, and Y. Kuo-Nan, "Back-gate bias enhanced band-to-band tunneling leakage in scaled MOSFET's", *Electron Device Letters, IEEE*, vol. 19, no. 4, pp. 134-136, 1998.

[83] W. Jing, P. M. Solomon, and M. Lundstrom, "A general approach for the performance assessment of nanoscale silicon FETs", *Electron Devices, IEEE Transactions on*, vol. 51, no. 9, pp. 1366-1370, 2004.

[84] 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", *Electron Devices Meeting, 1997. Technical Digest., International* 1997, pp. 869-872.

[85] Robert F.Pierret, *Semiconductor Device Fundamentals,* Addison-Wesley Publishing Company, 1996.

[86] S. Narendra, V. De, S. Borkar, D. A. Antoniadis, and A. P. Chandrakasan, "Full-chip subthreshold leakage power prediction and reduction techniques for sub-0.18-μm CMOS", *Solid-State Circuits, IEEE Journal of*, vol. 39, no. 3, pp. 501-510, 2004.

[87] Integrated Systems Engineering Inc., "ISE TCAD Manuals", San Jose, CA 95113: 04.

[88] S. Krishnan and D. Vasileska, "A self-consistent quantum mechanical simulation of p-channel strained SiGe MOSFETs", *10th International Workshop on Computational Electronics* 2004, pp. 89-90.

[89] B. L. Austin, K. A. Bowman, T. Xinghai, and J. D. Meindl, "A low power transregional MOSFET model for complete power-delay analysis of CMOS gigascale integration (GSI)", *ASIC Conference 1998. Proceedings. Eleventh Annual IEEE International* 1998, pp. 125-129.

[90] T. Ghani, K. Mistry, P. Packan, S. Thompson, M. Stettler, S. Tyagi, and M. Bohr, "Scaling challenges and device design requirements for high performance sub-50 nm gate length planar CMOS transistors", *2000 Symposium on VLSI Technology. Digest of Technical Papers* 2000, pp. 174-175.

[91] D. L. Critchlow, "MOSFET Scaling-the Driver of VLSI Technology", *Proceedings of the IEEE*, vol. 87, no. 4, pp. 659-667, 1999.

[92] T. Karnik, S. Borkar, and V. De, "Sub-90 nm technologies-challenges and opportunities for CAD", *ICCAD* 2002, pp. 203-206.

[93] S. Haihua, F. Liu, A. Devgan, E. Acar, and S. Nassif, "Full chip leakage-estimation considering power supply and temperature variations", *ISLPED* 2003, pp. 78-83.

[94] A. Devgan and S. Nassif, "Power variability and its impact on design", *international Conference on VLSI Design* 2005, pp. 679-682.

[95] T. Mizuno, N. Sugiyama, T. Tezuka, T. Numata, and S. Takagi, "High performance CMOS operation of strained-SOI MOSFETs using thin film SiGe-on-insulator substrate", *Symposium on VLSI Technology Tech. Dig.* 2002, pp. 106-107.

[96] T. Tezuka, N. Sugiyama, T. Mizuno, and S. Takagi, "High-performance strained Si-on-insulator MOSFETs by novel fabrication processes utilizing Ge-condensation technique", *Symposium on VLSI Technology Tech. Dig.* 2002, pp. 96-97.

[97] X. Qi, G. Jung-Suk, J. Pan, Y. Bin, S. Ahmed, Z. John, and L. Ming-Ren, "Strained silicon NMOS with nickel-silicide metal gate", *Symposium on VLSI Technology Tech. Dig.* 2003, pp. 101-102.

[98] J. R. Hwang, J. H. Ho, S. M. Ting, T. P. Chen, Y. S. Hsieh, C. C. Huang, Y. Y. Chiang, H. K. Lee, L. Ariel, T. M. Shen, G. Braithwaite, M. Currie, N. Gerrish, R. Hammond, A. Lochtefeld, F. Singaporewala, M. Bulsara, Q. Xiang, M. R. Lin, W. T. Shiau, Y. T. Loh, J. K. Chen, S. C. Chien, and F. Wen, "Performance of 70 nm strained-silicon CMOS devices", *Symposium on VLSI Technology Tech. Dig.* 2003, pp. 103-104.

[99] S. E. Thompson, M. Armstrong, C. Auth, S. Cea, R. Chau, G. Glass, T. Hoffman, J. Klaus, M. Zhiyong, B. Mcintyre, A. Murthy, B. Obradovic, L. Shifren, S. Sivakumar, S. Tyagi, T. Ghani, K. Mistry, M. Bohr, and Y. El-Mansy, "A logic nanotechnology featuring strained-silicon", *Electron Device Letters, IEEE*, vol.

25, no. 4, pp. 191-193, 2004.

[100] K. Ikeda, Y. Yamashita, A. Endoh, T. Fukano, K. Hikosaka, and T. Mimura, "50-nm gate Schottky source/drain p-MOSFETs with a SiGe channel", *Electron Device Letters, IEEE*, vol. 23, no. 11, pp. 670-672, 2002.

[101] Chi On Chui, K. Hyoungsub, D. Chi, B. B. Triplett, P. C. McIntyre, and K. C. Saraswat, "A sub-400 ^{o}C germanium MOSFET technology with high-κ dielectric and metal gate", *IEDM Dig.* 2002, pp. 437-440.

[102] C. H. Huang, M. Y. Yang, C. Albert, W. J. Chen, C. X. Zhu, B. J. Cho, M. F. Li, and D. L. Kwong, "Very low defects and high performance Ge-on-insulator p-MOSFETs with Al_2O_3 gate dielectrics", *Symposium on VLSI Technology Tech. Dig.* 2003, pp. 119-120.

[103] P. D. Ye, G. D. Wilk, J. Kwo, B. Yang, H. J. L. Gossmann, M. Frei, S. N. G. Chu, J. P. Mannaerts, M. Sergent, M. Hong, K. K. Ng, and J. Bude, "GaAs MOSFET with oxide gate dielectric grown by atomic layer deposition", *Electron Device Letters, IEEE*, vol. 24, no. 4, pp. 209-211, 2003.

[104] J. Kedzierski, P. Xuan, E. H. Anderson, J. Bokor, K. Tsu-Jae, and H. Chenming, "Complementary silicide source/drain thin-body MOSFETs for the 20 nm gate length regime", *IEDM Tech. Dig.* 2000, pp. 57-60.

[105] R. Li, S. J. Lee, H. B. Yao, D. Z. Chi, M. B. Yu, and D. L. Kwong, "Pt-Germanide Schottky Source/Drain Germanium p-MOSFET with HfO_2 Gate Dielectric and TaN Gate Electrode", *Electron Device Letters, IEEE*, vol. 27, no. 6, pp. 476-478, 2006.

[106] D. Connelly, C. Faulkner, and D. E. Grupp, "Performance advantage of Schottky source/drain in ultrathin-body silicon-on-insulator and dual-gate CMOS", *Electron Devices, IEEE Transactions on*, vol. 50, no. 5, pp. 1340-1345, 2003.

[107] Z. Shiyang, H. Y. Yu, S. J. Whang, J. H. Chen, S. Chen, Z. Chunxiang, S. J. Lee, M. F. Li, D. S. H. Chan, W. J. Yoo, A. Du, C. H. Tung, J. Singh, A. Chin, and D. L. Kwong, "Schottky-barrier S/D MOSFETs with high-k gate dielectrics and metal-gate electrode", *Electron Device Letters, IEEE*, vol. 25, no. 5, pp. 268-270, 2004.

[108] T. Bing-Yue and L. Chia-Pin, "A novel 25-nm modified Schottky-barrier FinFET with high performance", *Electron Device Letters, IEEE*, vol. 25, no. 6, pp. 430-432, 2004.

[109] B. Doris, I. Meikei, T. Kanarsky, Z. Ying, R. A. Roy, O. Dokumaci, R. Zhibin, J. Fen-Fen, S. Leathen, W. Natzle, H. Hsiang-Jen, J. Mezzapelle, A. Mocuta, S. Womack, M. Gribelyuk, E. C. Jones, R. J. Miller, H. S. P. Wong, and W. Haensch, "Extreme scaling with ultra-thin Si channel MOSFETs", *IEDM Dig.*

2002, pp. 267-270.

[110] H. van Meer and K. De Meyer, "70 nm fully-depleted SOI CMOS using a new fabrication scheme: the spacer/replacer scheme", *Symposium on VLSI Technology Tech. Dig.* 2002, pp. 170-171.

[111] E. Suzuki, K. Ishii, S. Kanemaru, T. Maeda, T. Tsutsumi, T. Sekigawa, K. Nagai, and H. Hiroshima, "Highly suppressed short-channel effects in ultrathin SOI n-MOSFETs", *Electron Devices, IEEE Transactions on*, vol. 47, no. 2, pp. 354-359, 2000.

[112] C. Yang-Kyu, K. Asano, N. Lindert, V. Subramanian, K. Tsu-Jae, J. Bokor, and H. Chenming, "Ultrathin-body SOI MOSFET for deep-sub-tenth micron era", *Electron Device Letters, IEEE*, vol. 21, no. 5, pp. 254-255, 2000.

[113] C. Yang-Kyu, C. Leland, P. Ranade, L. Jeong-Soo, H. Daewon, S. Balasubramanian, A. Agarwal, M. Ameen, K. Tsu-Jae, and J. Bokor, "FinFET process refinements for improved mobility and gate work function engineering", *IEDM Tech. Dig.* 2002, pp. 259-262.

[114] Y. Bin, C. Leland, S. Ahmed, W. Haihong, S. Bell, Y. Chih-Yuh, C. Tabery, H. Chau, X. Qi, K. Tsu-Jae, J. Bokor, H. Chenming, L. Ming-Ren, and D. Kyser, "FinFET scaling to 10 nm gate length", *IEDM Tech. Dig.* 2002, pp. 251-254.

[115] T. Park, S. Choi, D. H. Lee, J. R. Yoo, B. C. Lee, J. Y. Kim, C. G. Lee, K. K. Chi, S. H. Hong, S. J. Hynn, Y. G. Shin, J. N. Han, I. S. Park, U. I. Chung, J. T. Moon, E. Yoon, and J. H. Lee, "Fabrication of body-tied FinFETs (Omega MOSFETs) using bulk Si wafers", *Symposium on VLSI Technology Tech. Dig.* 2003, pp. 135-136.

[116] G. Pei, J. Kedzierski, P. Oldiges, M. Ieong, and E. C. C. Kan, "FinFET design considerations based on 3-D simulation and analytical modeling", *Electron Devices, IEEE Transactions on*, vol. 49, no. 8, pp. 1411-1419, 2002.

[117] T. Ichimori and N. Hirashita, "Fully-depleted SOI CMOSFETs with the fully-silicided source/drain structure", *Electron Devices, IEEE Transactions on*, vol. 49, no. 12, pp. 2296-2300, 2002.

[118] K. W. Guarini, P. M. Solomon, Y. Zhang, K. K. Chan, E. C. Jones, G. M. Cohen, A. Krasnoperova, M. Ronay, O. Dokumaci, J. J. Bucchignano, C. Cabral, Jr., C. Lavoie, V. Ku, D. C. Boyd, K. S. Petrarca, I. V. Babich, J. Treichler, P. M. Kozlowski, J. S. Newbury, C. P. D'Emic, R. M. Sicina, and H. S. Wong, "Triple-self-aligned, planar double-gate MOSFETs: devices and circuits", *IEDM Tech. Dig.* 2001, p. 19.

[119] C. H. Lee, H. F. Luan, S. C. Song, S. J. Lee, B. Evans, and D. L. Kwong, "A manufacturable multiple gate oxynitride thickness technology for system on a chip", *IEDM Tech. Dig.* 1999, pp. 491-494.

[120] K. Roy, H. Mahmoodi, S. Mukhopadhyay, H. Ananthan, A. Bansal, and T. Cakici, "Double-gate SOI devices for low-power and high-performance applications", *Proc. VLSID* 2006, p. 8.

[121] Y. Fu-Liang, C. Hao-Yu, C. Fang-Cheng, H. Cheng-Chuan, C. Chang-Yun, C. Hsien-Kuang, L. Chi-Chuang, C. Chi-Chun, H. Huan-Tsung, C. Chih-Jian, T. Hun-Jan, Y. Yee-Chia, L. Mong-Song, and H. Chenming, "25 nm CMOS Omega FETs", *IEDM Tech. Dig.* 2002, pp. 255-258.

[122] B. Doyle, B. Boyanov, S. Datta, M. Doczy, S. Hareland, B. Jin, J. Kavalieros, T. Linton, R. Rios, and R. Chau, "Tri-Gate fully-depleted CMOS transistors: fabrication, design and layout", *IEEE International SOI Conference* 2003, pp. 133-134.

[123] B. S. Doyle, S. Datta, M. Doczy, S. Hareland, B. Jin, J. Kavalieros, T. Linton, A. Murthy, R. Rios, and R. Chau, "High performance fully-depleted tri-gate CMOS transistors", *Electron Device Letters, IEEE*, vol. 24, no. 4, pp. 263-265, 2003.

[124] M. Stadele, R. J. Luyken, M. Roosz, M. Specht, W. Rosner, L. Dreeskornfeld, J. Hartwich, F. Hofmann, J. Kretz, E. Landgraf, and L. Risch, "A comprehensive study of corner effects in tri-gate transistors", *ESSDERC* 2004, pp. 165-168.

[125] G. Lixin and J. G. Fossum, "A novel compact model of quantum effects in scaled SOI and double-gate MOSFETs", *IEEE International SOI Conference* 2000, pp. 114-115.

[126] C. Qiang, W. Lihui, and J. D. Meindl, "Quantum mechanical effects on double-gate MOSFET scaling", *IEEE International SOI Conference* 2003, pp. 183-184.