QUANTUM MODELING AND CHARACTERIZATION OF DEEP SUBMICRON MOSFETS

HOU YONG TIAN

NATIONAL UNIVERSITY OF SINGAPORE

2003

QUANTUM MODELING AND CHARACTERIZATION OF DEEP SUBMICRON MOSFETS

HOU YONG TIAN
(M. Sc., Peking University)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF ELECTRICAL AND COMPUTER

ENGINEERING

NATIONAL UNIVERSITY OF SINGAPORE

2003

ACKNOWLEDGEMENTS

Herewith I would like to express my sincere gratitude to my supervisor, Professor Li Ming-Fu, for his instruction, guidance and encouragement in both personal and academic matters. It is his firm theoretical background and expertise in semiconductors to assure the present project being conducted smoothly. Through my life, I will benefit from the experience and knowledge I gained in his group. I also deeply appreciate Professor Dim-Lee Kwong for his guidance, suggestions and valuable discussions throughout my research at NUS. I am further indebted to Dr. Jin Ying for his continuous help during all these years. Many thanks are also to other collaborators from CSM for their technical supports. The award of a research scholarship by the National University of Singapore is also gratefully acknowledged.

I wish to thank my fellow postgraduate students from SNDL, COE and CICFAR for their invaluable discussions and assistance. In particular, some topics in thesis were finished together with Mr. Yu Hong Yu and Mr. Tony Low, it is my pleasure to acknowledge their help and cooperation. Many thanks also to Dr. Jie Bin Bin, Dr. Guan Hao, Dr. Chen Gang and Ms. Jocelyn Teo for their kind help. I also wish to thank all staff members in SNDL, COE and CICFAR for their kind technical support and management throughout the project.

Finally, I would like to express my deeply gratitude to my family for their love, care, understanding, support and encouragement during all these years.

Table of Contents

Acknowledgement i

Table of Contents ii

Summary vii

List of Figures iv

List of Tables xvii

List of Abbreviations xviii

List of Symbols xx

Chapter 1. Introduction 1

1.1. Overview 1

1.2. Introduction to CMOS Transistor Scaling 3

1.3. Quantum Mechanical Effects in MOS Devices 6

1.3.1. Carrier Quantization in MOS devices 6

1.3.2. Capacitive Contribution due to Quantum Mechanical Effect 8

1.3.3. Threshold Voltage Shift due to Quantum Mechanical Effect 9

1.3.4. Models for Carrier Quantization in CMOS Devices 11

1.4. Direct Tunneling through Ultrathin Gate Dielectrics 15

1.4.1. Basics of Direct Tunneling 15

1.4.2. A Review of the Models for Tunneling Current 16

1.5. Alternative High Permittivity (High-K) Gate dielectrics 21

1.5.1. Scaling Limit of SiO_2 21

1.5.2. High-K Gate Dielectrics 22

1.6. Metal Gate Technology 28

1.6.1. Polysilicon Gate Depletion Effect 28

1.6.2. Metal Gate Technology 29

1.7. Novel Device Architectures on SOI Technology 31

1.8. Objective of this thesis 33

1.9. Major Achievements in this Thesis 36

Chapter 2. Hole Quantization in MOS devices 38

2.1. Introduction 38

2.2. Multi-band Effective Mass Approximation Model 40

2.3. A New Simple Model for Hole Quantization by Six-band Effective Mass Approximation 42

2.3.1. The Algorithm of the Model 42

2.3.2. Application to Electron Quantization 46

2.3.3. Application to Hole Quantization 48

2.4. Improved One-band Effective Mass Approximation 56

2.4.1. Empirical Effective Masses 56

2.4.2. Effective Field Triangular Well Approximation 58

2.4.3. Hole Quantization by Improved One-band Effective Mass Method 62

2.5. Conclusion 63

Chapter 3. Direct Tunneling Current through Ultra-thin Gate Oxides in CMOS Devices 64

3.1. Introduction 64

3.2 Conduction Mechanism in Dual Poly-Si Gate CMOS Transistors 66

3.2.1. Carrier Separation Measurement 66

3.2.2. Conduction Mechanism in n+ Poly-Si Gate NMOSFETs 68

3.2.3. Conduction Mechanism in p+ Poly-Si Gate PMOSFETs 70

3.2.4. Conduction in the Source/Drain Extension (SDE) Region 72

3.3. Physical Model for Tunneling Current 75

3.4. Experiments and C-V Characterization 78

3.5. Non-parabolic Effect in Hole Direct Tunneling Current 80

3.5.1. Dispersion Relationship in Oxide Energy Gap 81

3.5.2. Electron Tunneling in NMOSFETs 82

3.5.3. Simulation of Hole Tunneling Current Using Freeman-Dahlke Dispersion Form 86

3.6. Simulations of All Terminal Direct Tunneling Currents in CMOSFETs 92

3.6.1. Conduction Band Electron Tunneling Current 92

3.6.2. Valence Band Hole Tunneling Current 95

3.6.3. Valence Band Electron Tunneling Current 96

3.6.4. Tunneling in Source/Drain Extension Overlap Region 98

3.7. Conclusion 101

Chapter 4. Tunneling Currents and Scalability of High-K Gate Dielectrics in CMOS Technology 102

4.1. Introduction 102

4.2. Direct Tunneling through Si_3N_4 and Al_2O_3 Gate Dielectric Stacks 104

4.2.1. Tunneling Currents through Si_3N_4, Oxynitride Gate Stacks 104

4.2.2. Tunneling Current through Al_2O_3 Stacks 109

4.3. Direct Tunneling through HfO_2 and HfAlO Gate Stacks 111

4.4. Scalability of Gate Dielectrics in CMOS Technology 117

4.4.1. Scalability of Gate Dielectrics in High Performance Application 119

4.4.2. Scalability of Gate Dielectrics in Low Power Application 121

4.4.3. Interface Engineering on Gate Leakage of High-K Gate Stacks 123

4.5. Conclusion 125

Chapter 5. Metal Gate Engineering on Gate Leakage Characteristics of MOSFETs 126

5.1. Introduction 126

5.2. Tunneling Currents in Metal Gate CMOS Devices 128

5.3. Reduction of Gate Leakage by Metal Gate. 131

5.4. Metal Gate Work Function Engineering on Tunneling Characteristics of MOSFETs 134

5.4.1. Gate to Channel Tunneling 134

5.4.2. Gate to Source/Drain Extension (SDE) Tunneling 138

5.4.3. Advantage of Metal Double Gate MOSFETs on Leakage Current 141

5.5. Scalability of Metal Gate Advanced MOSFETs 144

5.6. Conclusion 145

Chapter 6: Conclusions and Recommendations 146

6.1. Conclusions 146

6.1.1. Hole Quantization in CMOS Devices 146

6.1.2. Direct Tunneling Currents through Ultra-thin Gate Dielectrics 147

6.2. Recommendations for Future Works 152

Reference 155

Appendix

Brief Descriptions of Simulation Programs 175

List of Publications 177

SUMMARY

The scope of this thesis emphasizes on studies of carrier quantization and direct tunneling through ultrathin gate dielectrics in deep submicron CMOS devices.

Quantum mechanical effects become increasingly important as CMOS device scales into deep submicron regime. For hole quantization, the traditional one-band effective mass approximation (EMA) is insufficient. In this thesis, we studied the hole quantization based on the six-band EMA to include the valence band mixing effect. The traditional one-band EMA is found to underestimate the subband density of states and resultantly overestimate the hole quantum mechanical effects. Based on the numerical results from six-band EMA, an improved one-band EMA was proposed. In conjunction with the introduction of an effective electric field, this simplified approach demonstrates its application to hole quantization with advantages of simplicity in formalism, efficiency in computation and accuracy in simulations.

In deep submicron CMOS devices, direct tunneling current is dramatically increased when gate dielectric thickness is scaled. In this thesis, direct tunneling is investigated both experimentally and theoretically. An efficient physical model for the direct tunneling current is demonstrated by the successful simulations of all terminal tunneling currents in CMOS transistors with ultrathin gate oxide. For hole tunneling current, instead of the traditional parabolic dispersion, a Freeman-Dahlke dispersion form is introduced, which takes the difference of conduction and valence band effective masses into account. Using this form, the agreement with the experimental data is significantly improved over a wide range of oxide thickness and gate voltage.

Alternative high dielectric constant (high-K) dielectrics have been explored because the scaling of SiO_2 thickness is approaching its physical limit. The modeling of tunneling current through high-K gate stack was conducted by using the physical model. The simulated gate tunneling currents in Si_3N_4, Al_2O_3 and HfO_2 gate stacks were in excellent agreements with experiments. The simulations were also used to analyze the scalability of these high-K dielectrics in future CMOS technology in term of gate leakage. It is found that a high-K material is urgently required in CMOS technology for low power application. Due to the low tunneling current, HfO_2 or HfAlO is demonstrated to be a viable dielectric replacing SiO_2 to the end of the roadmap. The simulations also show that the interfacial layer affects significantly the gate leakage of the high-K gate stacks. Guidelines for interface layer engineering were also provided.

To eliminate poly-Si gate depletion, metal gate has been suggested to replace the traditional poly-Si. A systematic study has been performed on metal gate MOSFETs to investigate the impact of metal gates on the tunneling leakage current. Metal gate has the advantage of an appreciable reduction of gate leakage over poly-Si, when at the same CET (capacitance equivalent oxide thickness at inversion). Moreover, in ultra-thin body silicon-on-insulator (SOI) structure, the use of mid-gap metal gate results in significant reduction of gate to source/drain extension tunneling, especially when high-K gate dielectric is used. As a result, ultrathin body SOI device with metal gate has much lower off-state leakage, indicating its superior capability in device scaling.

List of Figures

Fig. 1.1	Schematic illustration of (a) energy subbands and (b) carrier density distribution in the inversion layer of a nMOSFET.	7
Fig. 1.2	Equivalent circuit of the MOS capacitor at the inversion condition. C_{ox} is the oxide capacitance and C_{inv} the inversion layer capacitance.	8
Fig. 1.3	Illustrations of direct (left) and Fowler-Nordheim (F-N) (right) tunneling in a nMOS structure. V_{ox} is the oxide voltage drop and ΔE_C the conduction band offset of SiO_2/Si.	15
Fig. 1.4	Illustration of poly-Si gate depletion effect in nMOSFET. Cp, Cox and Cinv represent the capacitance from the poly depletion layer, gate oxide and substrate inversion layer, respectively.	28
Fig. 1.5	Cross-section schematic of a MOSFET fabricated on an SOI Wafer.	31
Fig. 2.1	The schematic of the multiple quantum wells with zigzag potential energy profile used in our model. $\Phi(z) = qF_s \cdot L$.	42
Fig. 2.2	The calculated (a) subband energies E_{ij} and (b) electron centroid z_{ij} in electron inversion layer of nMOSFET. The dotted curves are from the method of infinite triangular well. Results of the model using a periodic multiple quantum wells structure are show as solid lines. For subband index (ij), i = 1, 2 is the longitudinal and transverse valleys, respectively, while j represents the ladder number.	46
Fig. 2.3	The comparison between the calculated and experimental (a) capacitance of electron inversion layer C_{inv} versus surface charge density N_s and (b) threshold voltage shift ΔV_T as a function of channel doping concentration N_A. The experimental data are from [10] and [82], respectively.	48

Fig. 2.4 The in-plane dispersion of the first six subbands calculated by our model in hole inversion layer. The surface electric field is (a) F_s = 0.5 MV/cm and (b) F_s = 2 MV/cm. Both the dispersions along (100) and (110) in the k_z = 0 plane are shown. The dashed lines are dispersions of corresponding spin degenerate holes and the split is induced by external field. 49

Fig. 2.5 The variation of the heavy hole (hh), light hole (lh) and spin orbit split-off (so) hole components in the three lowest subbands (n=1, 2, 3) versus in-plane wave vector k. The surface electric field is (a) Fs = 0.5 MV/cm and (b) Fs = 2 MV/cm. P^n_j is the projection of the nth (n=1, 2, 3) subband wave function to the j (hh, lh or so) component defined by (2.7). 50

Fig. 2.6 The obtained density of states of the three lowest subbands in hole inversion layer. The relative energy ΔE is the subband energy referenced from the subband edge. The solid and dashed curves are for surface electric field F_s = 0.5 and 2 MV/cm respectively. The solid lines with open circles are the results from the traditional one-band effective mass approximation. 51

Fig. 2.7 The calculated (a) subband energies of the first 6 subbands, and (b) occupation factors of the three lowest subbands for hole inversion layer in pMOS device at various surface electric field. The substrate doping is 5×10^{17} cm^{-3}. The dashed curves are from the traditional one band effective mass approximation. The results of our six band model are shown as solid lines. Fermi energy is also added in (a) for reference (solid circles for our model, open circles for traditional one band effective mass approximation). 52

Fig. 2.8 The calculated surface potential (band bending) of pMOS structure at inversion. The substrate doping is 5×10^{17} cm^{-3}. The dashed curve is from the traditional one band effective mass approximation. The results of our six band model are shown as solid line. The solid line with open circles is that from classical calculation with Fermi-Dirac statistics. 52

Fig. 2.9 The threshold voltage shifts due to hole quantum effects, ΔV_T, at different channel doping concentrations N_D. Solid and lines are results from our model and the multi-band pseudopotential method [72] respectively, with oxide thickness 14 nm. 54

Fig. 2.10 The threshold voltage shifts due to hole quantum effects, ΔV_T, at different channel doping concentrations N_D. (a) and (b) are comparisons between our model (solid lines), the 55

traditional one-band EMA (dashed lines), and the experimental data [82](solid circles) for oxide thickness 23 and 15 nm, respectively.

Fig. 2.11 The comparison of calculated and experimental capacitance of hole inversion layer. The solid line is our model simulation while the solid circles are experimental data from [10]. 55

Fig. 2.12 The electric field F_s dependence of (a) empirical energy quantization effective mass m_z and (b) empirical DOS effective mass m_d , determined from the numerical results of six-band EMA calculations. 57

Fig. 2.13 Comparisons of (a) subband energies, (b) occupation factors and (c) surface potentials in the hole inversion layer calculated by the improved one-band effective mass approximation (solid lines) and the six-band effective mass theory (solid circles). The substrate doping concentration in the calculation is 5×10^{17} cm^{-3}. 58

Fig. 2.14 Weighting coefficient η as a function of surface electric field for n=1,2,3 hole subbands respectively for different doping concentration. 60

Fig. 2.15 Comparisons of the results from improved one-band triangular approximation using an effective field (η=0.77) and six-band self-consistent EMA. They are (a) subband energies, (b) occupation factor and (c) surface potential versus surface electric field, and (d) the hole carrier centroid at different surface potential V_s. 61

Fig. 2.16 Threshold voltage shifts ΔV_T at various substrate doping N_D . The numerical results by six-band effective mass model are shown as solid circles while solid line is those by improved one-band effective mass method using our new constant empirical effective masses. 62

Fig. 3.1 The cross-sectional schematic of the carrier separation measurement. 66

Fig. 3.2 Current-voltage (I-V) characteristics and band diagram of a n+ poly-Si gate nMOSFET at inversion. The transistor gate length and width are 20 and 0.5 μm, respectively, and the oxide thickness is ~ 2 nm. 68

Fig. 3.3 Current-Voltage (I-V) characteristics and band diagram of a n+ poly-Si gate nMOSFET at accumulation. The transistor gate length and width are 20 and 0.5 μm, respectively, and the oxide thickness is ~ 2 nm. 69

Fig. 3.4	Current-Voltage (I-V) characteristics and the band diagram of a p+ poly-Si gate pMOSFET at inversion. The transistor gate length and width are 20 and 0.5 um, respectively, and the oxide thickness is ~ 2 nm.	71
Fig. 3.5	Current-Voltage (I-V) characteristics and the band diagram of a p+ poly-Si gate pMOSFET at accumulation. The transistor gate length and width are 20 and 0.5 um, respectively, and the oxide thickness is ~ 2 nm.	72
Fig. 3.6	Current-Voltage (I-V) characteristics of n+ poly-Si gate nMOS short channel transistor ($10\times0.18\ \mu m^2$ and larger area ($10\times10\ um^2$) capacitor at accumulation and the band diagram at source/drain extension region overlapped with the gate. The oxide thickness is ~1.65 nm.	73
Fig. 3.7	Current-Voltage (I-V) characteristics of p+ poly-Si gate pMOS short channel transistor ($10\times0.18\ \mu m^2$ and larger area ($10\times10\ um^2$) capacitor at accumulation and the band diagram at source/drain extension region overlapped with the gate. The oxide thickness is ~1,65 nm.	74
Fig. 3.8	The measured capacitance-voltage(C-V) characteristics of the MOSFETs used in the simulations. The experimental data are shown as open circles and the solid lines are the fitting results using the QM-CV model of device group at UC Berkeley. The extracted oxide thickness is 1.85, 2.07, 2.44 and 2.74 nm, respectively.	79
Fig. 3.9	The electron direct tunneling currents in nMOSFETs. The open circles are the measurements. The solid and dashed lines are calculations by assuming the electron dispersion in SiO_2 energy gap to be Franz-type (m_{ox}=0.61 m_0) and parabolic (m_{ox}=0.50 m_0), respectively.	84
Fig. 3.10	The hole direct tunneling currents in pMOSFETs. The open circles are the measured data. The solid and dashed lines denote the calculated values by assuming the hole dispersion in SiO_2 band gap to be Franz-type (m_{ox}=0.55m_0) and parabolic (m_{ox}=0.40 m_0), respectively.	87
Fig. 3.11	The hole direct tunneling currents in pMOSFETs. The open circles are the measured data. The solid lines denote the calculated values by assuming a Freeman-Dahlke form dispersion in SiO_2 band gap with m_{cox}=0.50 m_0 and m_{vox}=0.80 m_0.	88
Fig. 3.12	The electron direct tunneling currents in nMOSFETs. The open circles are the measured values. The solid lines denote the calculated values by assuming a Freeman-Dahlke form	89

dispersion in SiO_2 band gap with m_{cox}=0.50 m_0 and m_{vox}=0.80 m_0.

Fig. 3.13 The calculated imaginary wave vector in the energy gap of SiO_2 by Freeman-Dahkle (solid line) and microscopic model [38] (open circles). The dashed lines are those determined from parabolic dispersion. 90

Fig. 3.14 The electron direct tunneling currents in nMOSFETs. The open circles are the measurements. The solid lines are the calculations by assuming the electron dispersion in SiO_2 band gap to be parabolic (m_{ox}=0.50 m_0). 93

Fig. 3.15 The electron direct tunneling currents in pMOSFETs. The open circles are the measurements. The lines are the calculations by assuming the electron dispersion in SiO_2 band gap to be parabolic (m_{ox}=0.50 m_0). 94

Fig. 3.16 The hole direct tunneling currents in pMOSFETs. The open circles are the measurements. The solid lines are the calculations by assuming the hole dispersion in SiO_2 band gap to be parabolic (m_{ox}=0.41 m_0). 95

Fig. 3.17 The hole direct tunneling currents in nMOSFETs. The open circles are the measurements. The solid lines are the calculations by assuming the hole dispersion in SiO_2 band gap to be parabolic (m_{ox}=0.41 m_0). The noise level at low voltage is due to the measurement limit of the analyzer. 96

Fig. 3.18 The direct tunneling currents from valence band electrons in nMOSFETs. The open circles are the measurements. The solid lines are the calculations by assuming the electron dispersion in SiO_2 band gap to be parabolic (m_{ox}=0.41 m_0). The left band diagram illustrates the possible tunneling from valence band electrons with energy above the surface valence band edge, which has not been included in the present calculations. 98

Fig. 3.19 The gate to source/drain extension (SDE) tunneling currents in CMOSFETs. The open circles are the measurements. The solid and dashed lines are the calculated tunneling currents in channel and SDE area, respectively. 98

Fig. 3.20 The simulations of all terminal direct tunneling currents in CMOSFETs. The open symbols are the measurements on $10\times0.18\ \mu m^2$ transistor by the carrier separation method. The solid lines are the calculations using the physical model. The oxide thickness is ~1.65 nm. 100

Fig. 4.1 Calculated (solid lines) (a) electron and (b) hole direct 106

tunneling currents through a Si_3N_4 film with EOT of 1.42 nm from inversion layer of n - and pMOSFETs. The measured data (open circles) are from [114].

Fig. 4.2 Calculated tunneling currents in CMOSFETs with oxynitride (SiON) versus the nitrogen composition. 107

Fig. 4.3 Simulated electron tunneling currents of nMOSFETs with Al_2O_3 gate dielectric. The experimental data are from [125, 126]. The tunneling effective mass is found to be 0.28 m_0 from overall fitting of all the data. The thickness values from best fitting to the measured data match well with those in [125, 126] from C-V method (in parenthesis). 109

Fig. 4.4 Simulated gate current of a n+ poly/HfO_2 /SiO_2/p-Si device. The measured data are from [141], the physical thickness are HfO_2(38Å)/IL(6Å) from HRTEM and interface layer (IL is likely SiO_2 from XPS [141]. The fitted tunneling mass m_{HfO} is 0.18 m_0. When the uncertainty of HRTEM is 1 Å, the resulted m_{HfO} error is ±0.02 m_0. The dashed lines are simulations with m_{HfO} = 0.20 and 0.16 m_0. 112

Fig. 4.5 Simulated tunneling current of n+ poly-Si/HfO_2 on NH_3 nitrided p-Si. The $V_G > 0$ data is from [137]. The interfacial layer (IL) is assumed as $(SiO_2)_{0.5}(Si_3N_4)_{0.5}$. Using m_{HfO}=0.18 m_0 and κ=22 for HfO_2, the effective IL physical thickness can be determined and their values are 6.5 Å and 9 Å, respectively. 114

Fig.4.6 Hoe tunnelling current simulation of p+ poly-Si gate pMOSFET with HfO_2 stack. An oxynitride interface layer is also concluded in the modeling. The experimental data is from [137]. The gate stack structure is same as the nMOSFET in Fig.4.5. 115

Fig.4.7 Calculated tunneling currents of $(HfO_2)_x(Al_2O_3)_{1-x}$ for various Hf compositions. 116

Fig.4.8 Simulated tunneling current of MOSFET versus EOT for various gate dielectrics. The substrate doping is 10^{18} cm^{-3}. Al_2O_3 mole fraction is 30% for HfAlO and Si_3N_4 mole fraction 40% for optimized SiON. For SiON, hole tunneling in pMOSFET, which determines the scalability, is shown. Electron tunneling in MOSFET is shown for other dielectrics. 118

Fig.4.9 The calculated gate leakage of high performance CMOS. The calculated high (low) gate leakage for each generation corresponds to the minimum (maximum) EOT proposed in ITRS 2001. For oxynitride(SiON), gate leakage is from hole tunnelling in pMOSFET and the Si_3N_4 mole fraction is 40%. 119

Fig.4.10 The calculated gate leakage for low standby power (LSTP) application. Here, an average value of the proposed maximum and minimum EOT from ITRS 2001 is used for each generation. Al_2O_3 mole fraction is 30% for HfAlO and Si_3N_4 mole fraction 40% for optimized SiON. 121

Fig.4.11 The calculated gate leakage for low standby power applications of HfAlO dielectric stacks with different interface layers (ILs). Physical 5 Å IL of SiO_2, optimized SiON, $HfSiO_4$ were presented. A minimum 3 Å SiO_2 and SiON ILs are also shown to demonstrate the limit of SiO_2-based dielectrics as an IL layer. 123

Fig.5.1 Tunneling currents in a TaN/SiO_2/p-Si capacitor. The experimental data is from [56]. $J_{VBH,S}$ and $J_{ME,G}$ represents the valence band hole tunneling from substrate and metal electron tunneling from the gate, respectively. 128

Fig.5.2 Tunneling currents in TaN gated MOSFETs with HfO_2 stacks. Solid lines are simulations assuming parabolic dispersion in HfO_2 with effective mass values listed in Table I. Using one fitting parameter of 8 Å oxide interfacial layer (IL), overall good agreements between simulations and experiments are obtained. Better agreement with experimental data at high V_G can be obtained by using Freeman-Dahlke dispersion for HfO_2 (dashed lines). 130

Fig.5.3 Gate leakage of metal and poly-Si gate nMOSFET with SiO_2 gate dielectric. The metal gate work function is assumed at Si conduction band edge. The EOT is the equivalent oxide thickness, and CET the capacitance equivalent thickness at inversion (V_G-V_T = 0.5V). 131

Fig.5.4 NMOS gate leakage in future generation CMOS for (a) high performance application using SiO_2 and (b) low stand-by power application using HfO_2/SiO_2 stack with SiO_2 interfacial layer of 3Å. In the calculations, the EOTs of gate dielectrics were selected to meet the required CET by ITRS 2001. 132

Fig.5.5 Band diagram schematics of tunneling in channel area of nMOSFET. It is similar for pMOSFET except the substrate Fermi energy. Labels: CBE: conduction band electron; VBE: valence band electron; VBH: valence band hole; ME: metal gate electron; G: gate and S: substrate electrode. 135

Fig.5.6 Tunneling currents of metal double gate (DG) nMOSFET with (a) SiO_2 and (b) HfO_2 stack as a function of V_G-V_{FB}. Solid lines are of metal work function Φ_B at Si conduction 135

band edge (E_C metal), while dashed lines are those with mid-gap metal. The band diagrams for various tunneling components are shown in Fig.5.5.

Fig.5.7 Gate current of nMOSFET at V_G-V_{FB}=-1V as a function of metal work function Φ_B. 136

Fig.5.8 Same as Fig.5.7 but for metal double gate (DG) pMOSFET. E_V metal means metal work function Φ_B is at Si valence band edge. 137

Fig.5.9 Schematics of band diagrams of gate to source-drain extension (SDE) tunneling (J_{SDE}) at accumulation bias. In SDE region, the metal electrons tunnel to n+ SDE in nMOSFET and valence band electrons tunnel from p+ SDE to metal gate in pMOSFET. 138

Fig.5.10 Gate to source-drain extension (SDE) tunneling in metal double gate (DG) MOSFETs with SiO_2 and HfO_2 stack. Solid lines are those with E_C metal (nMOS) and E_V metal (pMOS) gates, while dashed lines are those using mid-gap metal gates. 139

Fig.5.11 The effect of metal gate work function Φ_B on the gate to source-drain extension (SDE) tunneling J_{SDE} for various gate dielectrics 140

Fig.5.12 Comparison of metal double gate (DG) and bulk MOSFETs with oxynitride gate dielectrics at the same threshold voltage V_T (defined by inversion charge of 10^{11} cm^{-2}). In DG MOSFET, low body doping is assumed and V_T is adjusted by metal gate work function Φ_B while in bulk MOSFET, V_T is tuned by channel doping. 141

Fig.5.13 Circuits of an inverter (left) and a sample/hold with an nMOSFET switch (right). 142

Fig.5.14 Off-state leakage of metal gate nMOSFET contributed by gate to source-drain extension (SDE) tunneling estimated using 5 nm SDE dimension for (a) high performance application using SiO_2 and SiON, (b) low power application using HfO_2/SiO_2 stack. 144

List of Tables

Table 4.1 Material parameters used in tunneling simulations. $\Delta E_C/\Delta E_V$ (eV): conduction/valence band offsets with Si; m_e/m_h (m_0): electron/hole effective mass in dielectric, and K dielectric constant. 118

List of Abbreviations

hh	Heavy hole
lh	Light hole
so	Spin-orbit split-off hole
ALD	Atomic layer deposition
CBE	Conduction band electrons
CET	Capacitance equivalent thickness
CMOSFET	Complementary metal-oxide-semiconductor field-effect transistor
C-V	Capacitance-voltage
CVD	Chemical vapour deposition
DG	Double-gate
DOS	Density of states
DPN	Decoupled plasma nitridation
EMA	Effective mass approximation
EOT	Equivalent oxide thickness
FDSOI	Fully-depleted silicon on insulator
F-N	Fowler-Nordheim
HRTEM	High-resolution transmission electron microscopy
IL	Interfacial layer
ITRS	International technology roadmap for semiconductors
I-V	Current-voltage
JVD	Jet vapour deposition
LSTP	Low standby power
MOSFET	Metal-oxide-semiconductor field-effect transistor
PDA	Post deposition annealing

PDSOI	Partially-depleted silicon on insulator
PVD	Physical vapour deposition
QM	Quantum mechanical
SCE	Short channel effect
S/D	Source/drain
SDE	Source/drain extension
SOI	Silicon on insulator
TDDB	Time dependent dielectric breakdown
UTB	Ultra-thin body
VBE	Valence band electron
VBH	Valence band hole
WKB	Wentzel-Kramers-Brilliouin
XPS	X-ray photoelectron spectroscopy

List of Symbols

β	Empirical parameter of band-gap widening
ε_{Si}	Si permittivity
ε_{ox}	SiO_2 permittivity
ϕ	Potential in substrate of MOS device
ϕ_s	Surface potential
Φ_B	Metal work function
ΔE_C	Conduction band offset
ΔE_V	Valence band offset
κ	Particle imaginary wave number inside the barrier
μ	Mobility
τ	Life time of quasi-bound state
ν	Group velocity
γ_1, γ_2, γ_3	Luttinger parameters of valence band
ξ	Subband wave function
η	Weighting coefficient for effective electric field
C_{inv}	Inversion layer capacitance
C_{ox}	Gate oxide capacitance
C_P	Capacitance of depletion layer in poly-Si gate
$D_n(E)$	Density of states of the nth subband
E_C	Energy of conduction band edge
E_V	Energy of valence band edge
E_f	Fermi level energy
E_g	Energy gap of semiconductor
E	Energy

$\boldsymbol{F_s}$	Surface electric field
$\boldsymbol{I_{OFF}}$	Off-state leakage of a transistor
$\boldsymbol{J}$	Tunneling current density
$\boldsymbol{J_{SDE}}$	Tunneling current density between gate and SDE
$\boldsymbol{g}$	Band degenerate factor
$\boldsymbol{L}$	Gate channel length
$\boldsymbol{m_z}$	Effective mass perpendicular to surface
$\boldsymbol{m_d}$	Density of states effective mass
$\boldsymbol{n_i}$	Intrinsic carrier concentration
$\boldsymbol{N}$	Subband sheet charge density
$\boldsymbol{N_{depl}}$	Depletion sheet charge density
$\boldsymbol{N_{sub}}$	Substrate doping concentration
$\boldsymbol{N_A}$	Acceptor concentration
$\boldsymbol{N_D}$	Donor Concentration
$\boldsymbol{N_C}$	Effective density of states of Si conduction band
$\boldsymbol{N_V}$	Effective density of states of Si valence band
$\boldsymbol{P_j^n}$	Component of the *j*th bulk band in the *n*th subband wave function
$\boldsymbol{Q_{dep}}$	Depletion charge
$\boldsymbol{Q_s}$	Mobile charge
$\boldsymbol{Q_t}$	Total charge
OF	Carrier occupation factor of *n*th subband
Tox	Gate oxide thickness
$\boldsymbol{T_{WKB}}$	Transmission probability
$\boldsymbol{V}$	External potential
$\boldsymbol{T_R}$	Correction factor in modified WKB

V_G	Gate terminal voltage
V_T	Threshold voltage
V_D	Drain terminal voltage
V_{ox}	Oxide voltage drop
V_{DD}	Supply voltage
V_P	Voltage drop in poly-Si gate
V_{FB}	Flat-band voltage
W	Gate channel width
z_n	Classical turning point for the *n*th bound state

Chapter 1

Introduction

1.1. Overview

Since the invention of metal-oxide-semiconductor field-effect transistor (MOSFET) and its successful incorporation into integrated circuits, it has steadily emerged to be the main building blocks of today's electronics circuit. During the past 30 years, we have witnessed a tremendous progress in MOS device technology. One distinct characteristic in this evolution is the steady downscaling of the transistor geometry, particularly its channel length. This is because MOSFET scaling is able to yield a higher packing intensity and most importantly a faster circuit speed. The channel length in current manufacturing technology is now entering into the nanometer regime. Based on the most recent International Technology Roadmap for Semiconductors (ITRS2001) [1], MOSFETs with channel length down to ~ 10 nm are projected to enter production in 2016.

As the transistor feature size is scaled into nanometer scale, many physical phenomena, which are negligible in large-dimension MOSFET, are becoming more and more important [2-3]. For examples, the operation of MOSFET is now entering a regime in which quantum mechanical effects become noticeable and substantial tunneling current through the gate insulator takes place due to the aggressive scaling of the gate dielectric thickness. In addition, in order to maintain the rapid development for device performance improvement, the introduction of new materials and processing technologies is also needed [4-5].

In this chapter, we first discuss in Section 1.2 how the CMOS transistor is scaled and what challenges we will meet during the device scaling. Then, according to the contents of this thesis, quantum mechanical effects in CMOS devices and direct tunneling current through ultrathin gate dielectrics, as well as their impact on device performance, will be reviewed in Sections 1.3 and 1.4, respectively. In subsequent Sections 1.5-1.7, brief introductions will be given to the current research activities in several areas relevant to the topic of this thesis, including high permittivity (high-K) dielectric materials, metal gate technology and novel device architectures. After Section 1.8, a brief introduction to the objectives, a summary of the major achievements in this thesis will be presented in Section1.9.

1.2 Introduction to CMOS Transistor Scaling

The engine of MOS technology development is to improve the transistor drive current and maintain the off-state leakage current as low as possible. In gradual channel approximation, the drive current can be written as [2, 3]

$$I_D = \frac{W}{L}\mu C_{ox}\left(V_G - V_T - \frac{V_D}{2}\right)V_D \qquad (1.1)$$

where W is the channel width, V_G, V_D and V_T are the gate voltage, drain voltage and threshold voltage, respectively. It can be seen that high drive current can be obtained by reducing the transistor gate length L, increasing the gate capacitance C_{ox} or improving the channel carrier mobility μ.

The gate length scaling is the main stimulus to the development of the MOS technology. However, short-channel MOSFETs differ in many aspects from long-channel ones. One of the prominent features is the short-channel effects (SCEs) [2], manifested as V_T reduction when the channel length is reduced, or when the drain is highly biased. As a result, the subthreshold leakage current is dramatically increased. Therefore, the most difficult challenge in CMOS scaling is how to keep SCEs well controlled.

Gate Oxide thickness Scaling

The reduction of gate oxide thickness is efficient in enhancing the gate control over the channel, ensuring good short-channel behaviour. Oxide scaling also has an additional benefit of improving the driving current of MOSFETs. However, ultrathin

gate oxide is susceptible to direct tunneling, giving rise to high gate leakage current, which necessitates the efforts to explore an alternative dielectric material with higher permittivity than SiO_2 [4].

Well Engineering

In deep submicron MOSFETs, the well doping profiles in both vertical and lateral directions are engineered for the suppression of SCEs [5-6]. In vertical direction, super steep retrograde doping profile is used. The heavy doping beneath the channel surface allows for both the V_T adjustment and the well control of SCEs, while the doping is kept low at channel surface to avoid the degradation to channel carrier mobility. In lateral direction, a halo structure is created by implanting extra dopants into the local regions surrounding the edges of the source/drain extensions. These halo implants provide a reduction of the subthreshold leakage current. The highly non-uniform profile in the lateral direction sets up a higher effective doping concentration toward shorter devices, which counteracts short channel effects. However, the production of highly non-uniform doping profiles required in deep submicron devices presents new challenges for ion implantation technology.

Source/Drain Engineering

Shallow source/drain extension depth is another effective method to suppress the SCEs by reducing the amount of channel depletion charges controlled by the drain [5-6]. However, the increased series resistance limits the scaling of source/drain junction depth. As a viable solution, raised or elevated source/drain structure may be used in future MOSFETs.

Channel Engineering

MOSFET performance improvement can also be achieved by increasing the channel carrier mobility. MOSFETs with high mobility using strain Si, Ge or SiGe channels have been demonstrated [7]. In particular, strain-Si channel has already been implemented into the current leading edge manufacturing technology.

1.3 Quantum Mechanical Effects in MOS Devices

As discussed in Section 1.2, the scaling of transistor gate length is accompanied by the decrease of the oxide thickness and the increase of the substrate doping concentration. However, the supply voltage is less aggressively scaled than the gate oxide thickness. As a result, the operating electric field in silicon substrate becomes higher and higher. The electric field in Si substrate at operation of a MOSFET is as high as 2 MV/cm at the present 130 nm technology, and it is expected to be continually increased in the future according to the ITRS 2001 [1]. In the presence of such a large electric field, significant carrier quantization is observable in operating transistor. Another concern is the threshold characteristic. The substrate doping in deep submicron transistor has been increased in order to achieve proper threshold voltage as well as to suppress the short channel effects. For a doping concentration of 10^{18} cm^{-3} or above used in deep submicron CMOS devices, the electric field in the substrate exceeds 0.5 MV/cm at V_T and quantum mechanical effects cannot be ignored even at threshold region.

1.3.1 Carrier Quantization in MOS devices

The carrier quantization in MOS structures has been extensively studied since 1970's [8]. As a typical example, Fig.1.1 illustrates the carrier quantization phenomenon in nMOSFET at inversion. Due to the existence of an electric field perpendicular to the Si surface, the energy band is bent strongly near the semiconductor-oxide interface, a potential well is thus formed by the oxide barrier and the eletrostatic potential in the semiconductor, as shown in Fig.1.1(a). Instead of

three dimensional (3-D) continuous states in classical physics, from a quantum mechanical picture, the electrons are confined in this potential well and form discrete subbands. From Fig.1.1(a), the lowest subband energy is lifted from the bottom of the Si conduction band to a higher energy level. Another important feature of carrier quantization is the different density distribution from the classical one. Figure 1.1(b) shows schematically the electron density distributions in the substrate under quantum mechanical (2-D) and classical (3-D) schemes. From quantum mechanical point of view, the carrier density must be zero at the boundary and the peak carrier density is beneath the dielectric/substrate interface, while it peaks near the surface in the classical case. The same analysis is also applied to accumulation layer or hole quantization. Due to the carrier quantization, the lowest subband lies above the bottom of the bulk band by a finite energy and the density of state (DOS) is also lower than the classical one. This will lead to a decrease in gate capacitance and an increase in threshold voltage.

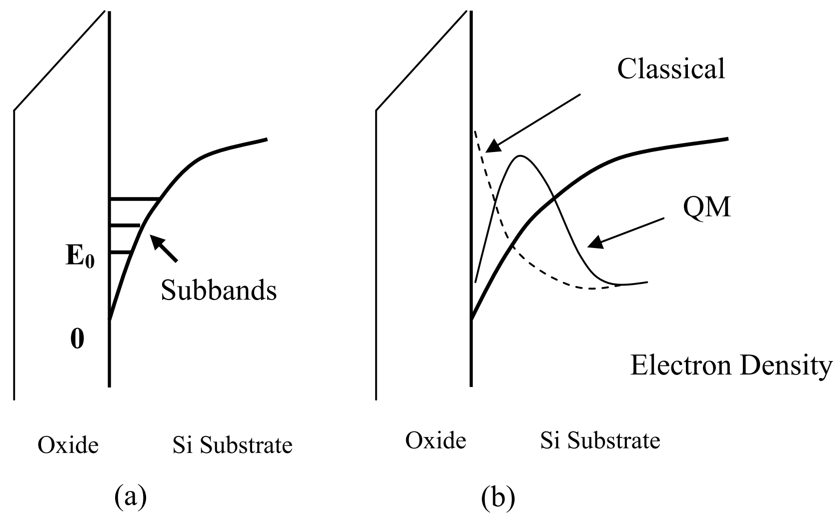


Fig.1.1: *Schematic illustration of (a) energy subbands and (b) carrier density distribution in the inversion layer of a nMOSFET.*

1.3.2. Capacitive Contribution due to Quantum Mechanical Effect

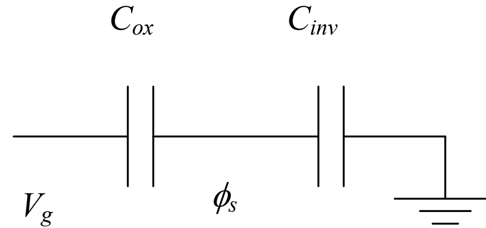


Fig.1.2: *Equivalent circuit of the MOS capacitor at the inversion condition.* C_{ox} *is the oxide capacitance and* C_{inv} *the inversion layer capacitance.*

The displacement of the carrier distribution from the surface due to quantum mechanical effects, which suggests the finite thickness of the inversion/accumulation layer, will result in a decrease of the device total capacitance. Figure 1.2 is the equivalent circuit of the MOS structure at inversion. The inversion layer capacitance is in series with the oxide capacitance. In the presence of an inversion layer capacitance, the total capacitance of the MOS structure will be reduced. Unfortunately, the inversion layer capacitance is physically inherent to the MOS structure and cannot be eliminated by any methods. The degradation to the total capacitance from carrier quantization is equivalent to an increase of the effective oxide thickness, which is estimated to be 2-4 Å [9]. When the gate oxide is thinner, the difference between the total capacitance and the oxide capacitance becomes more significant [9, 10]. Therefore, the impact of inversion layer capacitance on the device characteristics becomes more important as MOSFET scales down.

This capacitive reduction due to substrate quantization has effects of degrading the transconductance and the saturation driving current of MOSFETs [10]. The inversion layer capacitance inherent to MOS structure also limits the scaling of the

supply voltage [10]. From Fig.1.2, an additional voltage ϕ_s is dropped across the inversion layer, which increases as the electric field increases. This additional voltage drop in the inversion layer may significantly affect the operation of deep submicron devices and also make the scaling of supply voltage very difficult.

1.3.3 Threshold Voltage Shift due to Quantum Mechanical Effect

In the quantum mechanical treatment, carriers in the inversion layer are not only distributed away from the surface, but also occupy discrete subband energy levels. Since the lowest subband lies at a finite energy above the bottom of the bulk band, more band bending or a larger surface potential is required to populate the inversion layer than that in classical case. This has the effect of shifting the threshold voltage to a higher value, particularly for deep submicron MOSFETs with heavily doped substrates. This shift in threshold voltage can be as large as 0.1 V when the substrate doping concentration reaches 10^{18} cm^{-3} and it is crucial for the design of sub-0.1 μm devices with operating voltage of less than 1 V.

The threshold voltage shift due to quantum mechanical effects depends on the substrate doping concentration and the gate oxide thickness. In the following, we will express its definition explicitly. The surface electric field F_s can be determined by the total charge Q_t at the substrate:

$$F_s = -\frac{Q_t}{\varepsilon_{Si}} = -\frac{(Q_s + Q_{dep})}{\varepsilon_{Si}} \tag{1.2}$$

where Q_s and Q_{depl} are the mobile and depletion charge, respectively. At threshold voltage, the contribution from the inversion charge is negligible because its value is much smaller than that of the depletion charge. Under the depletion approximation, surface potential ϕ_s and surface electric field F_s of the substrate with doping density N_{sub} have the following relation [3]:

$$F_s = \sqrt{\frac{2qN_{sub}}{\varepsilon_0 \varepsilon_{Si}} \phi_s} \tag{1.3}$$

At threshold region, Maxwell-Boltzmann statistical function can be used and the classical inversion charge sheet density at surface potential ϕ_s is given by:

$$N_s^{CL} = \int_0^\infty N_{C(V)} e^{(E_f - q\phi(z))/kT} dz \tag{1.4}$$

where $N_{C(V)}$ is the effective density of states (DOS) of Si conduction (valence) band.

The classical definition of threshold voltage is the gate voltage when surface potential $\phi_T^{CL} = 2\phi_B$, with $\phi_B = (kT/q)ln(N_{sub}/n_i)$, where n_i is the intrinsic carrier concentration [3]. The 2-D threshold voltage is determined by the gate voltage to populate the 2-D inversion layer to the same inversion charge sheet density as the classical one, i.e. $N_s^{CL}(\phi_T^{CL})$ [11]. In quantum mechanical scheme, if the subband dispersion E_n and DOS $D_n(E)$ are determined, the inversion charge sheet density is:

$$N_s^{QM} = \sum_n \int f(E_n) D_n(E) dE \tag{1.5}$$

By equating $N_s^{QM}(\phi_T^{QM})$ in Eq. (1.5) to $N_s^{CLs}(\phi_T^{CL})$ in Eq. (1.4), we can obtain ϕ_T^{QM}, the surface potential at 2-D threshold voltage. Finally, the threshold voltage shift due to quantum mechanical effects can be expressed as [11]:

$$\Delta V_T = (\phi_T^{QM} - \phi_T^{CL}) + T_{OX} \cdot \frac{\varepsilon_{Si}}{\varepsilon_{OX}} \sqrt{\frac{2qN_{sub}}{\varepsilon_{Si}\varepsilon_0}} \left(\sqrt{\phi_T^{QM}} - \sqrt{\phi_T^{CL}} \right) \tag{1.6}$$

1.3.4. Models for Carrier Quantization in MOS Devices

In literature, 2-D carriers have been studied extensively [8]. Many methods have been proposed to study the carrier quantization. In this section, we will give a brief review on these models to study the quantum mechanical effects in MOS devices.

Self-Consistent Model

So far, the most accurate model for carrier quantization in MOS devices is to solve the Poisson and effective mass Schrödinger equations self-consistently [12-14].

At a semiconductor surface, the band bending can be characterized by an electrostatic potential $\phi(z)$. Based on the effective mass approximation, the wave function can be expressed as: $\Psi_i(x,y,z) = \xi_i(z)e^{ik_x x + ik_y y}$, where $\xi_i(z)$ is the solution of the simplified 1-D Schrödinger equation:

$$\frac{d^2\xi_i(z)}{dz^2} + \frac{2m_z}{\hbar^2}[E_i + q\phi(z)]\xi_i(z) = 0 \tag{1.7}$$

where E_i is the subband energy and m_z the effective mass perpendicular to the surface. The potential $\phi(z)$ is the solution of the Poisson equation:

$$\frac{d^2\phi}{dz^2} = -[\rho_{depl} - \sum_i N_i \xi_i^2(z)] / \varepsilon_{Si} \tag{1.8}$$

where ρ_{depl} is the depletion charge density, ε_{Si} the dielectric constant of silicon and N_i the carrier concentration in the ith subband. For 2-D carriers, the DOS $\frac{g_i m_{di}}{\pi\hbar^2}$ is independent of energy, where m_{di} and g_i are the DOS effective mass of the bulk Si and the degenerate factor of the ith subband, respectively. Then N_i can be expressed as:

$$N_i = \left(\frac{kT}{\pi\hbar^2}\right) g_i m_{d\,i} \ln\left(1 + \exp\left(\frac{E_f - E_i}{kT}\right)\right) \tag{1.9}$$

where the E_f and the E_i are the Fermi and the ith subband energy, respectively.

The self-consistent method starts from an initial estimate for the potential $\phi(z)$ and then solves Eqs. (1.7) and (1.8) successively until the output potential from Eq. (1.8) agrees with the input potential in Eq.(1.7) within a specified tolerant limit.

Triangular Well Approximation Model

In self-consistent model, numerical method must be used to solve the coupled Schrӧdinger and Poisson equations. Generally it demands much computational effort. Triangular well approximation is one of the most widely used simplified methods for carrier quantization because it leads to an analytical formula [12, 13].

In this model, the potential $\phi(z)$ in substrate is approximated by a triangular well, $\phi(z) = F_s \cdot z$, in Eq.(1.7), and by an infinite barrier for z<0 in oxide. Under triangular well approximation, Eq.(1.7) can be solved analytically in form of Airy functions. The *i*th subband energy is explicitly expressed as:

$$E_i = \left(\frac{\hbar^2}{2m_z} \right)^{1/3} \left[\frac{3}{2} \pi q F_s (i - \frac{1}{4}) \right]^{2/3} \tag{1.10}$$

where m_z is the energy quantization effective mass of the bulk Si.

The triangular well approximation has been proved to be reasonable in depletion or weak inversion region [13], it is thus effective in evaluating the carrier quantization at threshold [15-17]. Under traditional one-band effective mass approximation, as widely used for electrons, the derivation of threshold voltage shift can be done in a straightforward manner. From Eq.(1.10), the threshold voltage shift due to quantum effects can be calculated analytically [15-17]. An efficient capacitance-voltage (C-V) simulator taking carrier quantization into account in triangular well approximation has also been developed [18].

Band-Gap Widening Model

Another simple model is to treat quantum mechanical effects associated with the confinement of the carriers as an effective ***band-gap widening*** [19]. In this quasi-classical method, the energy splitting can be incorporated into a widening of silicon band gap.

$$E_g^{QM} = E_g^{Class} + \Delta E + F_s \Delta z \tag{1.11}$$

where ΔE is the energy gap between the bottom of the lowest subband and the bulk band. The effect of the displaced carrier distribution is incorporated through the third term in Eq.(1.11). F_s is the electric field perpendicular to the interface and Δz is the increase in the carrier average distance to the interface compared to the classical one. Finally, the band-gap widening results effectively in a modification of the intrinsic carrier concentration n_i. Combined with the triangular well approximation, it is concluded:

$$\Delta E_g = \beta(\frac{\varepsilon_{Si}}{4qkT})^{1/3}(F_s)^{2/3} \tag{1.12}$$

$$n_i^{QM} = n_i^{Class}\exp(-\frac{\Delta E_g}{2kT}) \tag{1.13}$$

The empirical parameter β can be obtained by comparing to results of the rigorous self-consistent method, which has typical values of 5.92×10^{-8} eV and 6.10×10^{-8} eV for electron and hole, respectively [19, 20]. This model can be easily implemented into the classical device simulators by the introduction of an empirical field dependent intrinsic carrier concentration. Electron and hole quantization have been investigated by such a method and empirical values of β at inversion and accumulation were documented in [19-21]. *C-V* and threshold voltage shift due to quantum mechanical effects have also been studied using this empirical model [19-22].

1.4. Direct Tunneling through Ultrathin Gate Dielectrics

1.4.1. Basics of Direct Tunneling

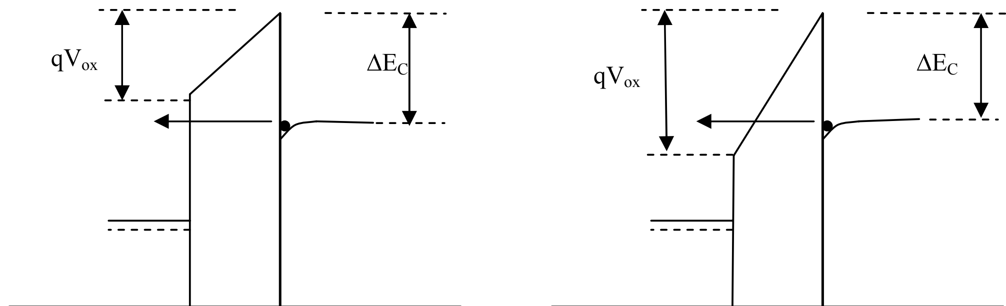


Fig.1.3: *Illustrations of direct (left) and Fowler-Nordheim (F-N) (right) tunneling in a nMOS structure.* V_{ox} *is the oxide voltage drop and* ΔE_C *the conduction band offset of* SiO_2/Si.

The study of tunneling through a classically forbidden energy barrier has a very long history and its basic mechanism has been known for several decades [23]. The tunneling phenomenon in a MOS structure can be schematically shown in Fig.1.3. From quantum mechanical physics, there is still a probability for carrier to tunnel through a classically prohibited barrier. At high gate voltage, when $qV_{ox} > \Delta E_C$, electrons tunnel through a barrier of triangular shape into the conduction band of the oxide layer, which is the well-known Fowler-Nordheim (F-N) tunneling. Instead of tunneling into the conduction band of the SiO_2 layer, when $qV_{ox} < \Delta E_C$, electrons can tunnel directly through the forbidden energy gap of the SiO_2 layer, which is the so-called "direct tunneling". It is projected that the operating voltage will be reduced to

1.0 V or less within this decade, and modern MOSFETs will thus operate in direct tunneling regime. Different from F-N tunneling, in direct tunneling through a trapezoidal barrier, the tunneling distance does not vary with the oxide field. As a result, direct tunneling current shows a distinct characteristic of much less dependence on oxide field than F-N tunneling.

The downsizing of MOSFETs is accomplished in a large part by decreasing the gate oxide thickness. As the thickness of the oxide layer decreases, the tunneling current increases approximately in an exponential manner. In deep submicron MOSFETs, the gate oxide has been scaled to below 2 nm. For such ultrathin oxide, a significant direct tunneling current is observed even at normal operating voltage. The most prominent impact of direct tunneling current is to greatly increase the power consumption of a chip. It thereby serves as one of the limiting factors for CMOS scaling. The direct tunneling current also adversely impacts the MOS device performance when oxide thickness is so thin that gate tunneling current is comparable to drain current [24]. Therefore, a study of direct tunneling through ultrathin gate oxide is valuable for the development of modern MOSFETs.

1.4.2. A Review of the Models for Tunneling Current

Since the invention of MOSFETs, tunneling through gate oxides in MOS structures has received much attention [25-38]. Many theoretical methods have been proposed to study the tunneling current. Here a brief review is given on the direct tunneling models.

Wentzel-Kramers-Brilliouin (WKB) Approximation

The most simple and well-known method to study tunneling through a barrier is the Wentzel-Kramers-Brilliouin (WKB) approximation, where the transmission probability of the barrier can be expressed as [23]:

$$T_{WKB}(E) = e^{-2\int \kappa(z)dz} \tag{1.14}$$

where κ(z) is the particle imaginary wave number inside the barrier. The transmission probability can be obtained if the dispersion relationship within the barrier, *κ(E)*, is known.

Classical Tunneling Model

Classical model on tunneling current treats the carriers available for tunneling as extended states (3-D) [27, 28]. In 3-D case, transmission probability is a well-defined concept and has a value equal to the ratio of transmission and incident flux. The tunneling current is determined directly by weighting the electron distribution function by the carrier transmission probability. If the tunneling occurs from electrode ***s*** to ***g***,

$$\begin{aligned} J_{sg}(E_z) &= \frac{qm_d(kT)}{2\pi^2\hbar^3}T(E_z)\left[f_s(1-f_g)-f_g(1-f_s)\right]dE_z \\ &= \frac{qm_d(kT)}{2\pi^2\hbar^3}T(E_z)(f_s-f_g)dE_z \end{aligned} \tag{1.15}$$

where m_d is the density of states, *T(E)* the transmission probability, f_s and f_g the carrier occupations at electrode ***s*** and ***g***. The total current is the integration over the energy E_z.

$$J_{sg} = \int_{Ec}^{\infty} J_{sg}(E_z) dE_z \tag{1.16}$$

Full Quantum Mechanical Model

In classical model described in the previous section, the 2-D quantum effects are neglected and the transmission probability represents the ability of free carriers hitting the oxide barrier to cross the oxide potential barrier by tunneling. In real MOS structures, however, the carriers are 2-D in nature and occupy discrete subbands, as discussed in section 1.3. For confined carriers in a potential well, such a concept of transmission probability is no longer meaningful. The tunneling current from such quasi-bound states can only be evaluated from the life-time τ of these quasi-bound states [29-31]:

$$J = \sum_{n} N_n / \tau_n(E_n) \tag{1.17}$$

where N_n is the carrier density of *n*th subband.

On the other hand, due to tunneling current, the substrate region cannot be regarded as being uncoupled from the gate. Consequently, the gate, the oxide and the substrate should be treated together for predicting charge distribution in the channel region and the tunneling current flowing between the gate and the substrate. By solving the coupled schrödinger and Poisson equations self-consistently in a quantum box of gate/oxide/substrate, the discrete subband states are emerging as quasi-bound

states, showing as resonant peaks in the energy spectrum. The penetration of the subband wave function into the oxide region determines the width of the resonance peaks. The width of the bound states resonance is inversely proportional to the life-time of the quasi-bound state, which provides the calculation of direct tunneling current in a full quantum mechanical scheme [29-31].

Quasi-classical Quantum Model

Although the life-time of quasi-bound states can be evaluated by the width of the quasi-bound states resonance, tremendous numerical effort is demanded. Particularly, this method is difficult or impossible for thick oxide and high-K materials because the width of the resonance is too small to be evaluated by any numerical method. Therefore, an efficient evaluation of the life-time of quasi-bound states is necessary and this has been done by a quasi-classical approach [32-35]. In this method, the lifetime of the *n*th quasi-bound state is approximately given by:

$$\frac{1}{\tau_n(E)} = \frac{T(E)}{\int_0^{z_n} \sqrt{2m_{zn}^* / \left[E_n - E_C(z)\right]} dz} \tag{1.18}$$

where E_n is the subband energy of the *n*th quasi-bound state, $E_C(z)$ is the edge of the Si conduction band, and z_n is the classical turning point for the *n*th bound state. $T(E)$ is the transmission probability of a particle through the barrier.

A comparison of the quasi-classical form of Eq.(1.18) to the full quantum numerical calculation has been conducted [36, 37]. A good agreement justifies the applicability of this quasi-classical treatment to the modeling of direct tunneling current in MOS structures.

Microscopic Model

All models above, either classical or quantum mechanical, are based on the effective mass approximation. The influence of the microscopic structure, composition of oxide and its interface with Si has not been considered. In their formalism, SiO_2 is assumed to be composed of single ellipsoidal band. In addition, the ultra-thin oxide with only several atomic layers may deviate significantly from the bulk properties. Recently, M. Stadele et al [38] presented a microscopic model in a tight-binding scheme to calculate the band structure of ultra-thin oxide and tunneling current. The results provided a physical insight into the fundamental issues relevant to oxide tunneling. Most importantly, it is demonstrated that the transmission through oxide can be qualitatively described within a bulk band structure scheme down to 1.0 nm thickness. Tunneling is chiefly determined by the dispersion of a single imaginary band, which explains the applicability of the traditional effective mass method.

1.5. Alternative High Permittivity (High-K) Gate dielectrics

1.5.1. Scaling limit of SiO_2

In the past decades, the development of the CMOS technology heavily relies on the material properties of SiO_2, a key element as the gate dielectric. The amorphous, thermally grown SiO_2 offers several advantages compatible with CMOS processing. They include the high quality interface with Si, superior electrical isolation properties associated with the large energy gap, as wells as the desired reliability. However, the continued scaling of SiO_2 in future CMOS technology meets several limits.

Theoretical simulations have demonstrated that the full SiO_2 band gap, which is crucial for the effective isolation, can be maintained for ultra-thin SiO_2 down to 7-8 Å [39]. On the other hand, further scaling of the dielectric thickness, which is propelled by the rapid shrinking of the transistor feature size, is also limited by the inherent gate leakage current from direct tunneling [40-42]. The presence of the direct tunneling current limits the scaling of SiO_2 by increasing the power consumption of the circuit. The defect generation and reliability issues may serve as another factor limiting the oxide thickness scaling [43].

In order to decrease the direct tunneling current while maintaining the total capacitance, alternative insulators with higher permittivity than SiO_2 are necessary. In recent years, alternative dielectric materials have been of intense research interest. Several kinds of dielectric materials have been exploited as candidates for gate dielectrics, including Si_3N_4, Al_2O_3 as well as Group IVB metal oxides, such as HfO_2 and ZrO_2.

1.5.2. High-K Gate Dielectrics

As shown in Eq.(1.1), it can be seen that high drive current can be achieved by increasing the capacitance at inversion. In the case of the gate capacitance, considering a parallel plate capacitor,

$$C = \frac{K\varepsilon_0 A}{t} \tag{1.19}$$

where K is the gate dielectric constant, ε_0 is the permittivity of free space, A is the area of the capacitor, and t is the thickness of the dielectric. It can be seen, in order to maintain the same capacitance, the physical thickness of the dielectric can be larger if the used material has higher dielectric constant. In practical application, the expression of C can be rewritten in terms of an equivalent oxide thickness (EOT), t_{ox}, which is equivalent to the theoretical thickness of SiO_2 required to achieve the same capacitance density. Thus, if an alternative dielectric with the physical thickness t_{High-k} is used, the equivalent oxide thickness can be obtained from [4]

$$t_{ox} = \frac{K_{ox}}{K_{High-k}} t_{High-k} \tag{1.20}$$

where K_{High-K} and K_{OX} are the dielectric constants of alternative high-K and SiO_2, respectively. The increased physical thickness will result in much lower direct tunneling current. High-K materials generally have lower band offset values than that

of SiO_2, which gives rise to an increase of the tunneling current, however, the increased physical thickness is the dominant factor.

The fundamental requirements for alternative gate dielectrics are listed here [4, 44]:

(1) The dielectric should be thermodynamically stable on Si substrate with respect to formation of uncontrolled SiO_2 or silicates at the Si/high-K interface during the deposition or post deposition annealing (PDA).

(2) The dielectric should remain amorphous after device integration to prevent the deleterious effects of mass or electrical transport along grain boundaries. Therefore, immunity to crystallization upon high temperature annealing is pertinent.

(3) To achieve low leakage current with minimum EOT, it should have sufficient high value of K and large band offsets with respect to the conduction and valence bands of Si and the gate. For good isolation, it should also have a large band gap.

(4) Interface states, fixed charge and trapped charge densities in the film are required to be low enough to avoid the capacitance-voltage (C-V) hysteresis, flat-band shift, and the degradation of the device performance.

(5) If poly-Si is used as the gate electrode, stability in contact with poly-Si is required. For p+ poly-Si pMOSFET, there should also be immunity to Boron penetration

(6) Minimum degradation to channel mobility is important for sustaining the high device drive current.

To date, several kinds of dielectric materials have been explored as candidates for gate dielectrics [4, 44]. Among them, Si_3N_4 and Al_2O_3 are the most widely studied because of their favourable properties, such as large band gap and superior thermal

stability in contact with silicon [4, 44]. Furthermore, they are also compatible with poly-Si gate processing. However, Si_3N_4 (K~7) and Al_2O_3 (K~10) have slightly higher K values than SiO_2 and thus provide only a near-term solution for CMOS scaling [41, 42]. Al_2O_3 also suffers from high charge density in the film. For long term solution, other alternative candidates with higher K values than Si_3N_4 and Al_2O_3 are required.

In recent years, a substantial amount of investigation has gone into the group IVB metal oxides, specifically Hafnium and Zirconium oxides. Research results on HfO_2 and ZrO_2 systems obtained before 2002 have been summarized in [4, 44]. For gate dielectric application, they possess excellent material properties, i.e., high dielectric constant (> 20), large band gap (5-6 eV) with band offsets to Si of > 1.5 eV. Compared to ZrO_2, HfO_2 is more stable when it is in contact with Si, Moreover, HfO_2 is compatible with poly-Si gate processing, while degradation of chemical properties and transistor performance is reported for ZrO_2 due to the interaction of the poly-Si gate electrode with the ZrO_2. Hafnium oxide family thus emerges as the most promising high-K candidate and becomes the main focus of recent studies.

MOS devices with HfO_2 gate dielectric have been fabricated using PVD (reactive sputtering), chemical vapour deposition (CVD), atomic layer deposition (ALD) and excellent electrical properties have been obtained. The capability of EOT scaling to below 1 nm is also demonstrated for HfO_2. The preliminary studies on the reliability of HfO_2 also provide encouraging results. The extrapolated voltage from TDDB data is well above the operating voltage of 1V required in future CMOS application.

Despite the above encouraging results, there are still a few unfavourable properties associated with HfO_2, which impede the replacement of the conventional

SiO_2. One drawback of HfO_2 is the low crystallization temperature. HfO_2 cannot remain amorphous when temperature is higher than about 400-500 ^oC [45, 46]. It is also a poor barrier against oxygen diffusion, which leads to the interface layer growth after PDA [46]. Because the interfacial layer, generally SiO_2 rich silicate, has much lower K value than bulk HfO_2, the presence of the interface layer growth makes it very difficult to scale the EOT of HfO_2 to below 1 nm. Although some Si surface treatment techniques, such as NH_3 passivation, have been proposed to minimize the interface layer growth, however, it degrades the channel carrier mobility so significantly that it cannot serve as a practical solution for CMOS application.

To eliminate the undesirable properties of poor thermal stability for HfO_2, it is possible to combine HfO_2 with another thermal stable oxide to form pseudobinary alloys. The effect of adding SiO_2 and Al_2O_3 to HfO_2 is to produce an amorphous film that is more thermodynamically stable on Si. Recent works on Hafnium Silicates (HfSiO) and Aluminates (HfAlO) indicate that such alloy materials exhibit encouraging gate dielectric properties [45-47]. For HfAlO, the incorporation of Al has been verified to increase the crystallization temperature significantly. The film remains amorphous after 900 ^oC annealing for HfAlO with Al_2O_3 mole fraction of 30% [45, 48]. The Al in HfAlO also has the effect to block the oxygen diffusion and in turn to reduce the interfacial layer growth remarkably [46]. The overall permittivity of the alloy dielectric is inevitably lower than that of the pure metal oxide, therefore a trade off has to be made between the improved thermal stability and degradation of permittivity [41].

The above HfO_2 system also has the problem of boron penetration in p+ poly-Si PMOS devices. In order to improve the immunity to Boron diffusion, effects of nitrogen incorporation have been investigated by various nitridation techniques. The

effective suppression of Boron diffusion after high temperatures processing has been demonstrated for HfON [48, 49], HfAlON [50] and HfSiON [51]. The nitrogen also blocks against oxygen diffusion during annealing and thus better control of the interface layer growth is obtained. Besides, HfON and HfSiON show increased crystallization temperature than HfO_2 and HfSiO. Excellent structural stability of HfSiON against phase separation is also reported. For HfAlO, the addition of nitrogen has an additional effect of reducing the C-V hysteresis and thus results in transconductance (Gm) improvement.

Although tremendous efforts have been made on the development of high-K dielectric materials, difficult challenges still remain for the implementation into current CMOS processing technology. One of them is the channel mobility degradation compared to SiO_2, which may be associated with the high densities of fixed and trapped charges, the poor interface quality, or enhanced phonon scattering in MOSFETs built with high-K gate dielectrics [4]. Another difficulty is the integration with poly-Si gate processing. Using HfO_2-based gate dielectrics, high threshold voltage is reported for p+ poly-Si gate pMOSFET. This phenomenon was attributed to the Fermi level pinning at the poly-Si/HfO_2 interface, which is induced by the dipoles created by the interfacial Hf-Si bonds [52]. In the presence of Fermi level pinning, it is difficult to achieve appropriate threshold voltage for PMOS using p+ poly/HfO_2 gate stack. Therefore, for practical applications, more efforts will be required in order to identify a suitable high-K material to replace SiO_2.

The predominant requirement of a high-K dielectric is the low gate leakage current. The gate leakage is one of the important phenomena to characterize and assess the high-K films. However, the high defect density associated with the high-K film may make the carrier conduction through gate dielectric more complicated

because of the presence of trap assisted conduction, which is directly associated with the quality of the dielectric film. Schottky emission, Frenkel-Poole and trap-assisted tunneling have been identified as the conduction mechanism in high-K materials [45]. The leakage current magnitude associated with defects can be reduced by the optimization of the high-K deposition or post deposition annealing conditions. In high quality dielectric films, the observed gate leakage currents have been identified as due to direct tunneling, especially for ultrathin high-K film. The tunneling currents through high-K materials have been studied theoretically by a number of researchers, most of which were made on silicon nitride family [4]. The incorporation of stacked structures was also considered for potential gate stack materials [53]. From WKB scheme, the direct tunneling current, which is determined by the dielectric thickness, band offset and tunneling effective mass, can be regarded as an intrinsic characteristic associated with a dielectric material and represents the minimum achievable gate leakage current. In terms of gate leakage from direct tunneling, the assessment of the scalability of potential high-K materials in CMOS technology has been made [40-42].

1.6. Metal Gate Technology

1.6.1. Polysilicon Gate Depletion Effect

The poly-Si depletion is an effect that exists when a MOS device is biased into depletion or inversion. Its origin can be illustrated from a n+ poly-Si nMOSFET, as shown in Fig.1.4. When a positive bias is applied between the gate and the substrate, a depletion layer with finite thickness is formed in the poly-Si gate at the poly-Si/oxide interface, which is indicated by the non-negligible band bending in the poly-Si gate.

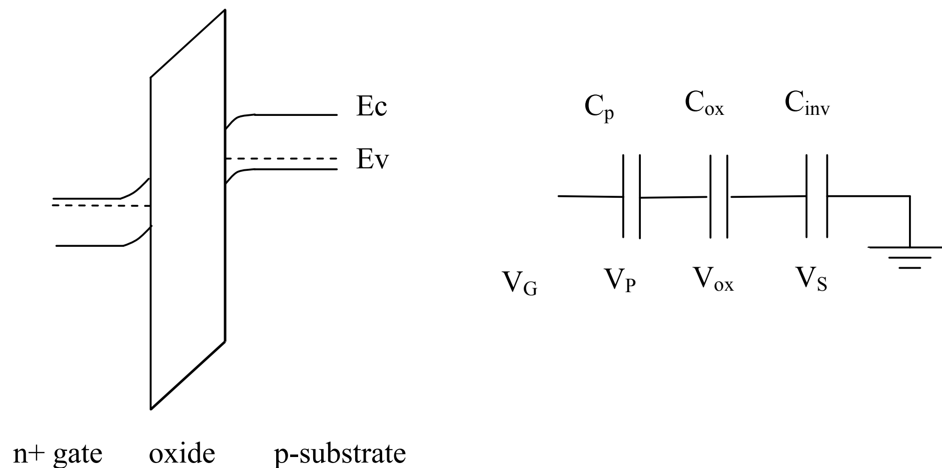


Fig.1.4: *Illustration of poly-Si gate depletion effect in nMOSFET. Cp, Cox and Cinv represent the capacitance from the poly depletion layer, gate oxide and substrate inversion layer, respectively.*

The poly-Si depletion has an effect of reducing the total capacitance. From an equivalent circuit diagram shown in Fig.1.4, the depletion capacitance C_P due to the finite depletion layer in poly-Si gate is in series with the gate oxide capacitance. Thus

the total capacitance is reduced due to the contribution of the C_P, which is equivalent to an increase of the effective oxide thickness. The existence of a voltage drop in the poly-Si gate means the effective voltage to invert the channel is smaller than that without poly-Si depletion. This will lead to a smaller inversion charge density and a resultant reduction of the drive current of the MOSFET.

Although C_P remains constant at the same poly-Si doping and electric field, the poly-Si depletion effect becomes more significant as gate oxide gets thinner [9]. It can be qualitatively explained by the circuit diagram in Fig.1.4. The total capacitance is determined largely by the smaller capacitor. For thick oxide, the gate oxide capacitance is very small and makes Cp negligible. However, when the gate oxide is thinner, the gate capacitance is increased and the impact of Cp on the total capacitance becomes more noticeable. Typically, the increase of effective oxide thickness due to poly-Si depletion is 5-6 Å, which is significant compared to the dielectric EOT below 1 nm required in nanometer scale CMOS. Hence, its effect cannot be ignored and the degradation to MOSFET performance due to poly-Si depletion becomes a major issue.

1.6.2. Metal Gate Technology

The poly-Si gate depletion can be minimized by increasing the poly-Si doping density. However, it appears that it is very difficult to get electrically active doping densities significantly above 10^{20} cm^{-3}, especially for p+ poly-Si doped with boron [54]. Other issues also include the high gate resistance and boron penetration. A metal gate material not only eliminates the gate depletion and greatly reduces the gate sheet

resistance, but also provides potential tunable gate work function. Several metal gate electrodes have been studied in literatures, such as Ta, Mo, TiN, TaN, TaSiN, and Ni or Co silicides [55-59].

The major challenge for metal gate technology is to find metals with suitable work functions. To maintain the performance advantage, metal gates with work functions near the conduction band and valence band edges of Si are desired for the optimal design of bulk n- and p-MOSFETs, respectively [54]. One simple way is to use two metal gate electrodes with distinct work functions. However, such dual metal approach is difficult and costly. Its implementation also introduces process integration complexity. Therefore, a single metal alternative is desirable. One scheme is based on the molybdemum (Mo) metal gate. Its work function has been successfully modified over a wide range of 4.5-4.9 eV by implantation of nitrogen [58]. Another promising solution is the full silicidation of poly-Si gates [59]. The silicide work function can be engineered by the type of the dopants and the doping concentrations of poly-Si prior to silicidation. Yeo et al [60] also found that the metal work function depends on the underlying gate dielectric and the dependence is explained by Fermi level pinning due to interface dipoles. Therefore, the work function engineering should be performed in close conjunction with the selection of gate dielectrics.

The integration of metal gate into a CMOS process also faces some difficulties. The metal gate is required to be thermally stable on ultra-thin gate dielectric without interface reaction and contamination of the gate dielectric beneath. Although gate-last processing utilizing damascene/replacement technique shows its potential in avoiding the high temperature budget, contamination and metal etching [61], it requires complex process.

1.7 Novel Device Architectures on SOI Technology

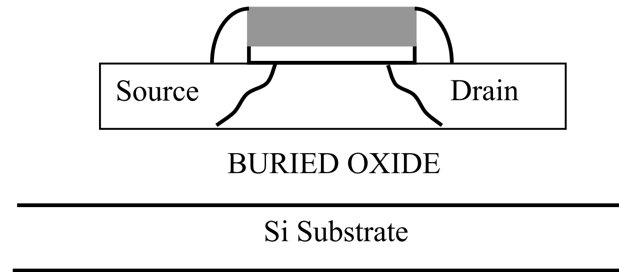


Fig.1.5: *Cross-section schematic of a MOSFET fabricated on an SOI Wafer.*

Bulk CMOS has been the mainstream VLSI technology for the past three decades. Below 100 nm gate length, however, the scaling of bulk CMOS is severely constrained by several fundamental issues [62]. As discussed in Section 1.2, the scaling principles for bulk-Si CMOS require a reduction in junction depth and an increase in doping level, which adversely affect the junction capacitance and carrier mobility. Doping fluctuation is also a problem, which induces a variation of the threshold voltage and will probably terminate the scaling of bulk CMOS. As an alternative, new device structure on silicon on insulator (SOI) technology provides a possible solution beyond the bulk CMOS [63].

In SOI technology, a buried oxide underneath the active area for device fabrication (Si Body) is used to provide the vertical isolation. The SOI devices can be categorized as partially-depleted (PD) and fully-depleted (FD) ones, which are defined according to the Si body thickness relative to the depth of the depletion region in the channel of the transistor. Here we focus on the FD-SOI, in which the depletion of the channel extends entirely through the body of the SOI structure, because the

operation of PD-SOI is quite similar to bulk one and only FD-SOI represents superior scalability. The benefits offered by SOI devices over bulk ones have been summarized in [63].

The scaling capability of CMOS can be analysed by 2-D effects inside the FET, which is done numerically using 2-D simulation tools [62], but the recent analytical analysis [64] reveals the primary dependencies on the silicon body thickness, t_{Si}, and gate oxide thickness, t_{ox}. According to this theory, the scale length Λ1, which characterises the scalability, can be approximately solved as:

$$\Lambda 1 \cong t_{Si} + (\varepsilon_{Si} / \varepsilon_{ox}) t_{ox} - (\pi^2 / 3)(\varepsilon_{Si} / \varepsilon_{ox})(\varepsilon_{Si}^2 / \varepsilon_{ox}^2 - 1)(t_{ox} / t_{Si})^2 t_I \tag{1.21}$$

It is clear that the intrinsic length is dramatically reduced with ultrathin body. This indicates that ultra-thin body (UTB) SOI structure, even without any doping, is able to reduce or eliminate the SCEs, which limit the scaling of the bulk MOSFETs. Such a design without body doping is extremely beneficial in terms of high carrier mobility and reduced dopant fluctuation. Double-gate (DG) SOI MOSFETs have in principle two symmetric gates interconnected. The two gates exert ideal control on the potential and inversion charge, so that SCEs are highly reduced. They are now considered as the most superior candidates for device scaling. The DG MOSFETs can be planar, vertical or mixed modes. FinFET represents one promising structure in this category. As alternatives to DG FinFET [65], novel architectures based on Fin technology were also exploited in recent years, including the Omega FinFET [66] and Tri-Gates [67].

The process optimization and electrical characteristics of UTB or DG SOI MOSFETs have been investigated extensively [65-68]. The tunneling currents in UTB and DG devices were also studied [69].

1.8. Objectives of this thesis

The critical objectives of this thesis are to study the hole quantization and direct tunneling phenomena in CMOS devices.

Hole Quantization

As summarized in Section 1.3, electron quantization is widely investigated, which relies on the one-band effective mass method. Such an one-band EMA has also been used for hole quantization, in which a parabolic dispersion is assumed with the heavy and light hole effective masses determined from the dispersion of bulk Si [10, 11, 13, 16]. In such a traditional one-band EMA, the subband in the potential well is solved for different valence bands separately and the mixing between valence bands was neglected. However, such a treatment of hole quantization is unphysical due to the degeneracy of the heavy and light hole bands.

As discussed in Section 1.3.4, the numerical self-consistent model demands much computational effort. Among simplified models, the triangular well approximation is one of the most widely used because it leads to an analytical formula. One objective of this thesis is to study the hole quantization beyond the traditional one-band effective mass method. A multi-band EMA is used to capture the

essential physics of valence band mixing and additionally a triangular well approximation is used to avoid computational difficulties.

Direct Tunneling

In literature, most previous studies in this area focus on electron tunneling. Relative less attention has been paid to the modeling of hole direct tunneling, which dominates the gate leakage in *p*+ poly-Si gate *p*MOSFETs under channel inversion condition. A reliable physical model of hole direct tunneling current first requires an accurate treatment of the hole quantization effect in Si substrate, which determines the hole densities at different energies. In this thesis, an appropriate model with valence band mixing effect being taken into account will be incorporated in the modeling of hole tunneling current.

As shown in Section 1.4.2, a reliable dispersion form in the energy gap of the dielectric film is crucial in the WKB approximation. Generally a parabolic dispersion in the SiO_2 energy gap is assumed. However, the applicability of such a parabolic approximation to hole quantization has never been verified because the energy of valence holes aligns at the middle of the SiO_2 energy gap. In this thesis, the proper dispersion form to model the hole tunneling current will be explored. In addition, the applicability and limitation of a parabolic approximation to hole dispersion in the SiO_2 energy gap will also be addressed.

It is of current interest to explore a high-K dielectric with low gate leakage current. Most of theoretical studies on tunneling currents through high-K materials were made on silicon nitride family [4]. Despite the considerable efforts on material and device studies, insufficient works have been done on the simulation of tunneling currents on other high-K materials, such as Al_2O_3 and HfO_2. For HfO_2, earlier

simulations did not perform a direct comparison with the experimental data over a wide range [54]. In addition, an interface layer is always formed during the high-K deposition and post deposition annealing, which increases the tunneling leakage. As a result, the interface layer effect must be considered in order to extract the accurate tunneling parameters of the high-K materials. In this thesis, tunneling current simulations will be performed on several high-K gate stacks of current interest.

As discussed in Section 1.6, a metal gate is required in future technology in order to solve the poly depletion problem. In our knowledge, a systematic study on the tunneling characteristics in metal gated CMOS devices has never been reported. An advantage of the metal gate is the potential for work function adjustment. One of the objectives of this thesis is to investigate the effect of metal work function on the tunneling leakage and device scaling capability.

1.9. Major Achievements in this Thesis

Two physical phenomena associated with deep submicron CMOS devices were studied in this thesis, they are carrier quantization and direct tunneling through ultrathin gate dielectrics.

In Chapter 2, we studied the hole quantization in MOS devices in the six-band effective mass scheme. For hole quantization, the traditional one-band effective mass approximation is found to underestimate the subband density of states and resultantly overestimate the hole quantum mechanical effects. Therefore, valence band mixing must be taken into account in order to describe the hole quantum mechanical effect properly. Based on the numerical results from six-band calculations, an improved one-band effective mass approximation was proposed. In conjunction with an introduction of effective electric field, this simplified model demonstrates its application for simulating the hole quantization in MOS devices with computational efficiency and sufficient accuracy, indicating its potential application in routine device modeling.

The studies on direct tunneling current include three topics. Chapter 3 demonstrates an efficient physical model for direct tunneling current by the successful modeling of all terminal tunneling currents through ultrathin gate oxide in CMOS devices. For hole tunneling current, using a Freeman-Dahlke dispersion form, which takes the difference of conduction and valence band effective masses into account, the agreement with the experimental data is significantly improved over a wide range of oxide thickness and gate voltage. It is also found that the widely used parabolic dispersion remains applicable to hole tunneling only when oxide is thinner than 2 nm.

In Chapter 4, the modeling of tunneling currents through high-K gate stacks was performed using the physical model. Excellent agreements between simulation results and measured tunneling currents have been achieved over Si_3N_4, Al_2O_3 and HfO_2. The model was also used to analyse the scalability of high-K dielectrics in term of gate leakage. It is found that, high-K material is expected to be first required in low power application. HfO_2 or HfAlO is demonstrated to be a viable dielectric replacing SiO_2 to the end of the roadmap. The interfacial layer between high-K and Si substrate is also included in the model, and the simulations show that this low-K interfacial layer affects significantly the gate leakage of the high-K gate stacks. Guidelines for interface layer engineering were also provided.

Chapter 5 presents a systematic study of tunneling leakage current in metal gate MOSFETs. Physical model used for simulations was corroborated by experimental results from SiO_2 and HfO_2 gate dielectrics with TaN electrodes. Due to the elimination of poly-Si depletion, metal gate has the advantage of an appreciable gate leakage reduction over poly-Si gate, when at the same CET (capacitance equivalent oxide thickness at inversion). The use of mid-gap metal gate results in significant reduction of gate to source/drain extension tunneling in both n- and p-MOSFETs. As a result, metal double gate MOSFET has much lower off-state leakage than the bulk one, especially when high-K gate dielectric is used, which demonstrates the superiority of metal double gate structure in device scaling.

Chapter 2

Hole Quantization in MOS Devices

2.1. Introduction

With the miniaturization of CMOS devices, the operation of MOSFETs is now entering a regime in which quantum mechanical effects become important and classical physics is no longer sufficient for describing accurately the device characteristics at either operating or threshold [70]. Thus, a good understanding of the quantum mechanical effects is crucial to the development of CMOS technology.

In the past, the characteristics of 2-D carriers have been studied extensively [8]. Most of the previous studies were concentrated on electron quantization in MOS devices and they heavily relied on the one-band effective mass approximation (EMA) with parabolic dispersion [8-10, 12-15, 19]. In literature, similar method is also applied to study the hole quantizaiton and the heavy and light hole effective masses were determined from the dispersion of bulk Si [10, 11, 13, 16]. However, such a simplified one-band effective mass approximation with a parabolic dispersion is inadequate for hole quantization due to the mixing between different valence bands in strong electric field [71]. From the quantum mechanical point of view, when there is a strong electric field perpendicular to the (100) interface, the symmetry of the point group of the Hamiltonian is reduced from O_h to C_{2v} and splitting as well as mixing between different bulk valence bands are expected [71]. A few works have been published on the hole quantization taking the complicated valence band structure of Si into

consideration [72-75]. They can be classified as the pseudopotential [72] and effective mass methods [73-75]. However, these multi-band models in self-consistent method are prohibitive to routine device simulation because of their computational difficulties. In this chapter, simplified models will be explored for hole quantization studies.

This chapter consists of three major parts. Section 2.2 is an introduction to the multi-band effective mass method in the treatment of valence bands of semiconductors. In Section 2.3, hole quantization and valence band mixing effect are investigated using a simple model in six-band effective mass approximation. As a further approximation, an improved one-band effective mass approximation for hole quantization is proposed in Section 2.4, which is efficient and easy to implement for conventional device simulation.

2.2. Multi-band Effective Mass Approximation Model

In this section, we will introduce the multi-band effective mass model by describing its general formalism.

From the effective mass theory, the Hamiltonian of the system can be obtained by the unperturbed bulk band Hamiltonian adding a diagonal electric potential energy term V [71]. In a biased MOS device fabricated on (100) silicon substrate, V is only a function of z, where z is perpendicular to the (100) plane. If J bulk bands are included, the multi-band effective mass equation can be generally written as J (i = *1,2, to J*) coupled equations:

$$\sum_{j=1}^{J}(D_{ij}(-i\nabla)+qV(z)\delta_{ij})\xi_j^n = E_i\xi_j^n \qquad i = 1 \text{ to J} \tag{2.1}$$

here, $\{\xi_j{}^{n}\}$ is the envelope function of nth subband, which is a combination of J *(j=1,2, to J)* bulk bands and thus has J components. $D(-i\nabla)$ is a $J\times J$ matrix of the bulk Hamiltonian and its form can be explicitly obtained by k·p method [71].

For valence bands of Si, there are three nearly degenerated bands: the heavy hole (***hh***), light hole (***lh***) and spin-orbit split-off (***so***) bands. The spin-orbit split-off energy of Si is 44 meV, which is much smaller than the Si band gap (1.12 eV). Therefore, ***so*** band cannot be neglected and a 6×6 Hamiltonian is necessary to include ***hh***, ***lh***, ***so*** holes and their respective spin degenerate bands. It is given by [73-76]:

$$H = \begin{bmatrix} P+V(z) & \alpha & \beta & 0 & i\alpha/\sqrt{2} & -i\sqrt{2}\beta \\ \alpha^* & L+V(z) & 0 & \beta & -iD/\sqrt{2} & i\sqrt{3/2}\alpha \\ \beta^* & 0 & L+V(z) & -\alpha & -i\sqrt{3/2}\alpha^* & -iD/\sqrt{2} \\ 0 & \beta^* & -\alpha^* & P+V(z) & -i\sqrt{2}\beta^* & -i\alpha^*/\sqrt{2} \\ -i\alpha^*/\sqrt{2} & iD/\sqrt{2} & i\sqrt{3/2}\alpha & i\sqrt{2}\beta & S+V(z) & 0 \\ i\sqrt{2}\beta^* & -i\sqrt{3/2}\alpha^* & iD/\sqrt{2} & i\alpha/\sqrt{2} & 0 & S+V(z) \end{bmatrix} \tag{2.2}$$

where

$$P = \frac{\hbar^2}{2m_0}\left[\left(k_x^2 + k_y^2\right)\left(\gamma_1 + \gamma_2\right) + k_z^2\left(\gamma_1 - 2\gamma_2\right)\right]$$

$$L = \frac{\hbar^2}{2m_0}\left[\left(k_x^2 + k_y^2\right)\left(\gamma_1 - \gamma_2\right) + k_z^2\left(\gamma_1 + 2\gamma_2\right)\right]$$

$$\alpha = \frac{\hbar^2}{2m_0} 2\sqrt{3}\left[k_z(ik_y - k_x)\gamma_3\right]$$

$$\beta = \frac{\hbar^2}{2m_0}\sqrt{3}\left[2ik_x k_y \gamma_3 - (k_x^2 - k_y^2)\gamma_2\right]$$

$$D = \frac{\hbar^2}{2m_0}\left[2(k_x^2 + k_y^2)\gamma_2 - 4k_z^2\gamma_2\right]$$

$$S = \frac{\hbar^2}{2m_0}\left[(k_x^2 + k_y^2 + k_z^2)\gamma_1\right] + \Delta_0 \tag{2.3}$$

The Luttinger parameters [77] are: $\gamma_1 = 4.22$, $\gamma_2 = 0.39$, $\gamma_3=1.44$ [78]. $\Delta_0 = 44$ meV is the spin-orbit splitting energy of bulk Si [78].

The electric potential energy $qV(z)$ in the Si substrate can be obtained by solving Eq.(1.7) and the coupled Poisson equation (1.8) self-consistently [73-75].

2.3. A New Simple Model for Hole Quantization by Six-band Effective Mass Approximation

Due to the complexity of the valence band of silicon, the physical validity of traditional one-band effective mass approximation (EMA) is questionable. For an accurate treatment, multi-band EMA must be used. However, the present multi-band models [72-75] in self-consistent method are prohibitive to routine device simulation because of their tremendous computational efforts needed. In this section, we will present a new method to study the hole quantization. This method in the frame of multi-band effective mass approximation captures the essential physics of valence band mixing, while it has the advantage of computational efficiency due to the employment of the triangular well approximation.

2.3.1. The Algorithm of the Model

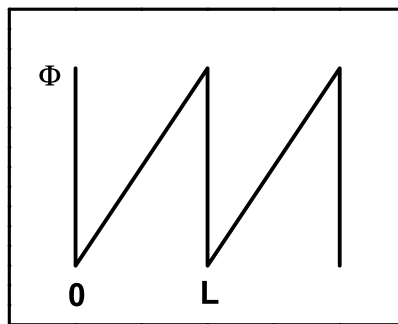


Fig.2.1: *The schematic of the multiple quantum wells with zigzag potential energy profile used in our model.* $\Phi(z) = qF_s \cdot L$.

In our model [79, 80], a triangular well approximation is applied to describe the potential in the substrate. This approximation has been widely used in studying electron quantization as discussed in Section 1.3.4. In order to simplify the mathematical calculation, the single triangular well approximation is further replaced by a periodic multiple triangular well, i.e. a zigzag potential as schematically illustrated in Fig.2.1. Then the external electric potential energy $qV(z)$ in Eq.(2.2) can be expressed as:

$$V(z) = qF_s \cdot z', \quad 0 < z' < L, \quad z = z' + nL, \quad n = 0,\ 1,\ 2...... \tag{2.4}$$

where F_s is the surface electric field in the Si substrate, q is the charge of the hole, and L is the periodic length of the zigzag potential. The advantage of using a zigzag potential lies on the fact that the Hamiltonian can be reduced to a superlattice Hamiltonian with period L, so that the well-known technique and criterion already developed in the superlattice theory can be used [74]. The wave function is expanded in plane waves and the number of plane waves, M, can be greatly reduced because only those plane waves with wave vector $k_z+2\pi m/L$, where m is integer, are used [76]. Following [76], we expand the nth subband envelope function $\xi^n{}_{j,k}$ at wave vector $k = (k_x, k_y, k_z)$ into plane waves:

$$\xi^n{}_{j,k} = \exp\left[i(k_x x + k_y y)\right] \sum_{m=1}^{M} a^n{}_{j,k,m} \frac{1}{\sqrt{L}} \exp\left[i(k_z + m\frac{2\pi}{L})z \right] \tag{2.5}$$

Combining equations (2.4) and (2.5), the matrix elements of $qV(z)$ in the plane wave representation are:

$$qV_{m,m'} = \begin{cases} \dfrac{qF_s \cdot L}{2} & m = m' \\ \dfrac{qF_s \cdot L}{i \cdot 2\pi(m-m')} & m \neq m' \end{cases} \tag{2.6}$$

The legitimacy of using a zigzag potential in replacing a single triangular well one is supported by superlattice theory [71]. When L is large enough, the coupling of lowest few energy states between the neighbouring wells disappears and each well in the zigzag potential can be considered as an isolated single triangular well. The coupling between neighbouring wells can be tested by k_z dispersion of the subband [71]. We will justify this point by using the one-band EMA and the zigzag potential in the electron quantization case in Section 2.3.2.

To assess the band mixing effect quantitatively, the quantum mechanical projection (or probability) functions are introduced:

$$P_j^n(k) = \sum_{m=1}^{M} a_{j,k,m}^{n*} a_{j,k,m}^{n} \tag{2.7}$$

and the following sum role holds,

$$\sum_{j=1}^{N} P_j^n(k) = 1 \tag{2.8}$$

$P_j^n(k)$ represents the component of the jth bulk band in the nth subband wave function $\xi^n(z)$.

If the subband dispersion and density of states (DOS) are known, the characteristics of MOS system at inversion can be determined by the following self-consistency of parameters [81]: the surface electric field F_s, the inversion charge sheet density N_s, depletion charge sheet density N_{depl}, the surface potential ϕ_s, the band bending due to depletion charge ϕ_{depl}.

$$
\begin{aligned}
F_S &= q(N_s + N_{depl}) / \varepsilon_{Si}\varepsilon_0 \\
\phi_s &= \phi_{depl} - qN_s \bar{z}_{QM} / \varepsilon_{Si}\varepsilon_0 - kT/q \\
N_{depl} &= \left[2\varepsilon_{Si}\varepsilon_0\phi_{depl}\left(N_A - N_D\right)/q\right]^{1/2} \\
\bar{z}_{QM} &= \sum_n z_n N_n / N_s
\end{aligned}
\tag{2.9}
$$

Here, the inversion charge distribution width or centroid z_n is defined as:

$$
z_n = \int z\phi_n^*(z)\phi_n(z)dz \tag{2.10}
$$

N_n is the charge sheet density of nth subband, it is related to the subband DOS $D_n(E)$:

$$
N_n = \int_{E_{n0}}^{\infty} D_n(E) f(E) dE \tag{2.11}
$$

here we use Fermi-Dirac distribution function. By iteration method to achieve the self-consistency of Eq.(2.9), the electrostatics of MOS structure can be quickly determined. Additionally, the carrier occupation factor of nth subband is defined by:

$$OF_n = \frac{N_n}{\sum_n N_n} \tag{2.12}$$

2.3.2. Application to Electron Quantization

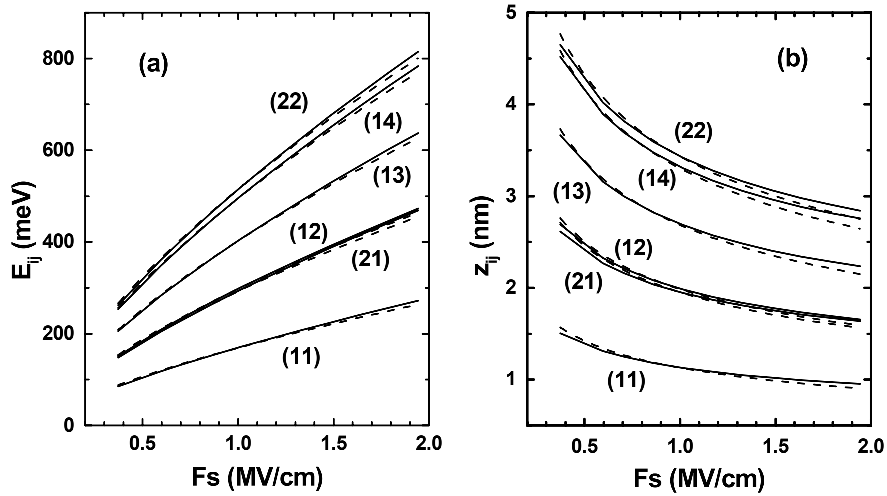


Fig.2.2: *The calculated (a) subband energies E_{ij} and (b) electron centroid z_{ij} in electron inversion layer of nMOSFET. The dotted curves are from the method of infinite triangular well. Results of the model using a periodic multiple quantum wells structure are show as solid lines. For subband index (ij), i = 1, 2 is the longitudinal and transverse valleys, respectively, while j represents the ladder number.*

First, we will apply our model to electron inversion layer. Because the electrons can be well described by one-band EMA with parabolic dispersion, comparison with the rigorous analytical results will be used to justify the approximation of our zigzag potential. By comparing with the experiments, an insight of the triangular well approximation can also be further addressed.

In Fig.2.2, the dotted curves show the subband energies E_{ij} and the electron charge centroid z_{ij} for *n*MOS device at inversion as a function of surface electric field F_s determined by the method of [81]. In the calculations, we assume a uniform channel doping of 5×10^{17} cm^{-3}, typically used in state-of-art CMOS devices. i = 1, 2 represents the longitudinal (m_z = 0.916 m_0) and transverse (m_z = 0.19 m_0) valleys, respectively. The energy of the subband minimum is given by Eq.(1.10) and the carrier centroid is:

$$z_{ij} = \frac{2E_{ij}}{3qF_s} \tag{2.13}$$

The solid lines are the numerical results by our model with a zigzag potential. An overall agreement is observed from depletion to strong inversion. The energy dispersion along k_z is less than 0.5 meV, indicating negligible coupling between neighbouring wells. Theoretically our multiple quantum wells model will converge to the above infinite triangular well approximation if L and M are large enough. However, a trade-off must be made between the computing accuracy and efficiency. In the following calculations of hole quantization, we will first compare the results of our model with Eq.(1.10) in one-band EMA. The criterion to choose L and M for convergence is that the errors in the subband energies are all within 5%. In our studied range of surface electric field, L is typically 50 –100 nm in order to achieve a good accuracy.

In Fig.2.3, we show the calculated inversion layer capacitance and threshold voltage shifts in *n*MOS devices. The experimental data are from [10] and [82], respectively. Good agreements between the simulations and experimental data are

achieved. It suggests that, comparing to the rigorous self-consistent method, the simple triangular well approximation maintains a satisfactory accuracy in evaluating the electrical characteristics of MOS structures.

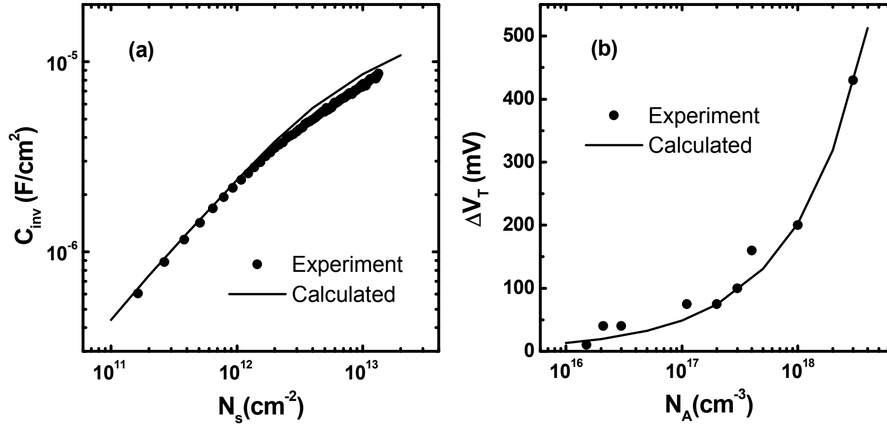


***Fig.2.3**: The comparison between the calculated and experimental (a) capacitance of electron inversion layer C_{inv} versus surface charge density N_s and (b) threshold voltage shift ΔV_T as a function of channel doping concentration N_A. The experimental data are from [10] and [82], respectively.*

2.3.3. Application to Hole Quantization

In this section, our proposed model is used to study hole quantization in pMOSFET at inversion condition [79, 80]. For valence bands, the multi-band effective mass method depicted in Section 2.2 is used.

(1). Valence Band Mixing

Figure 2.4 shows the in-plane dispersions of the six lowest subbands at surface electric fields F_s = 0.5 MV/cm and 2.0 MV/cm, respectively. The subband dispersions

are found to be anisotropic, far from parabolic, and electric field dependent. The degeneracy of the ***hh*** and ***lh*** bands at Γ point (k=0) is lifted by the electric field and their separation depends on the electric field. In addition, there are reversed camel back structures (two reversed peaks) [83] with negative hole effective mass near k = 0 for n = 2, 3, 5 and 6 subbands.

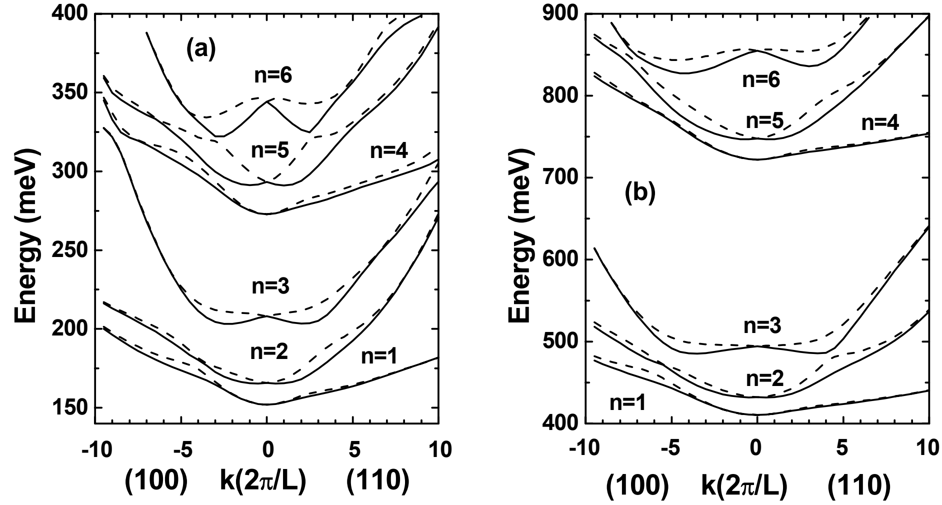


Fig.2.4*: The in-plane dispersion of the first six subbands calculated by our model in hole inversion layer. The surface electric field is (a) F_s = 0.5 MV/cm and (b) F_s = 2 MV/cm. Both the dispersions along (100) and (110) in the k_z = 0 plane are shown. The dashed lines are dispersions of corresponding spin degenerate holes and the split is induced by external field.*

Figure 2.5 shows the projection functions P^n_{hh}, P^n_{lh} and P^n_{so} of the first three subbands (n=1, 2, 3). At k = 0, the n=1 subband is purely ***hh*** and n = 2 and 3 subbands are mainly ***lh*** and ***so*** with some band mixing between ***lh*** and ***so*** bands. As the electric field is increased, the magnitude of band mixing also increases. When k ≠ 0, there are strong band mixing in all subbands.

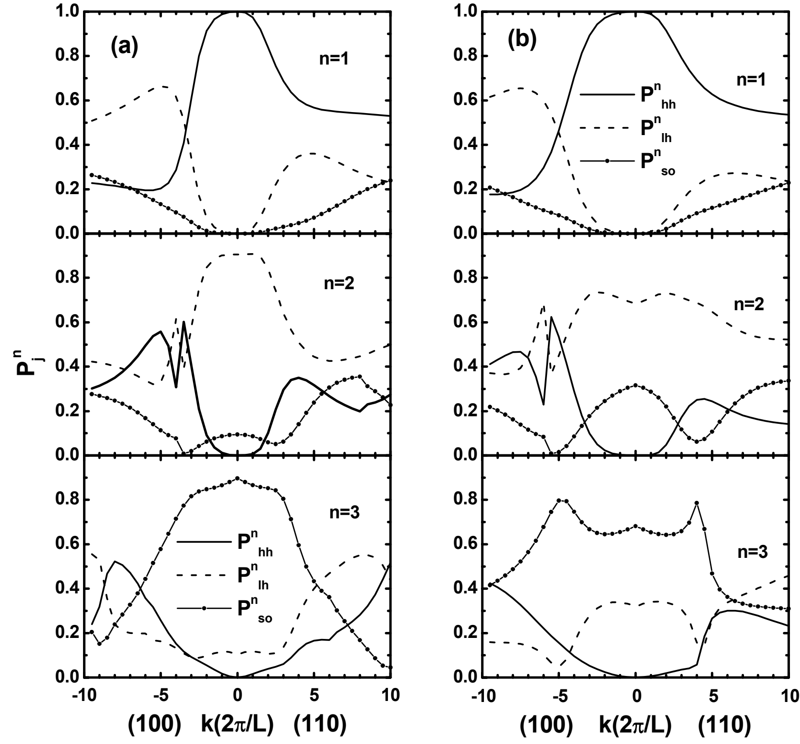


Fig.2.5*: The variation of the heavy hole (hh), light hole (lh) and spin orbit split-off (so) hole components in the three lowest subbands (n=1, 2, 3) versus in-plane wave vector k. The surface electric field is (a) Fs = 0.5 MV/cm and (b) Fs = 2 MV/cm.* P^n_j *is the projection of the* n^{th} *(n=1, 2, 3) subband wave function to the j (hh, lh or so) component defined by (2.7).*

(2). Density of States (DOS)

Figure 2.6 illustrates the simulated results of DOS's of the three lowest subbands. The DOS profiles deviate from the step-like function as predicted by the traditional one-band EMA. In particular, near the band minimum of the n = 3 subband, there are two peaks. They are caused by the camel back structure [83] at the band minimum as observed in Fig.2.4. The shapes of DOS's are also electric field dependent. The DOS's obtained from traditional one-band EMA are also shown in Fig.2.6, which overall underestimates the DOS's.

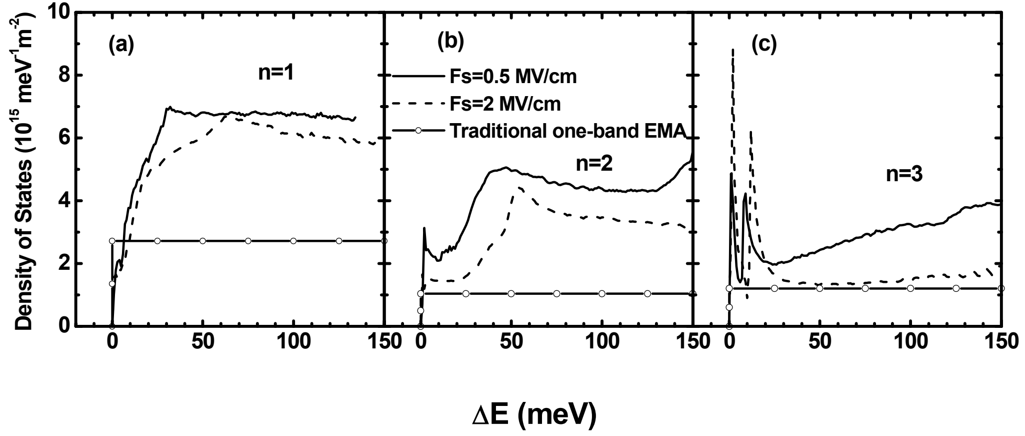


Fig.2.6*: The obtained density of states of the three lowest subbands in hole inversion layer. The relative energy ΔE is the subband energy referenced from the subband edge. The solid and dashed curves are for surface electric field F_s = 0.5 and 2 MV/cm respectively. The solid lines with open circles are the results from the traditional one-band effective mass approximation.*

(3). Subband Energies and Carrier Occupations

From the above hole quantization model along with the parametric self-consistency method [81], the subband energy levels of the first six subbands and the occupation factors on the three lowest subbands are shown in Fig.2.7 by solid lines. The corresponding data calculated by traditional one-band EMA are also shown with dashed lines. In Fig.2.7 (a), the energies of n = 1 subband obtained by two methods are in good agreement. This is expected because this subband is purely ***hh*** at the band minimum as shown in Fig.2.5. However, the results are quite different for n=2 or n=3 subbands due to the band mixing. Furthermore, from traditional one band EMA, there is a crossing between the ***lh*** and ***so*** subbands at about 1.5 MV/cm and the n = 2 subband will change from ***lh*** to ***so*** hole characteristic at electric field higher than 1.5 MV/cm. However, our calculation does not show such a crossing up to 3 MV/cm.

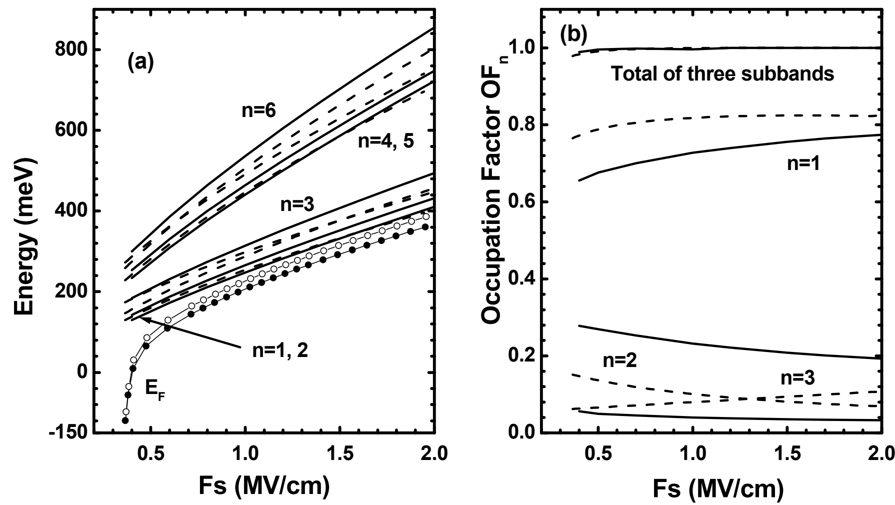


Fig.2.7: *The calculated (a) subband energies of the first six subbands, and (b) occupation factors of the three lowest subbands for hole inversion layer in pMOS device at various surface electric field. The substrate doping is* 5×10^{17} cm^{-3}*. The dashed curves are from the traditional one band effective mass approximation. The results of our six band model are shown as solid lines. Fermi energy is also added in (a) for reference (solid circles for our model, open circles for traditional one band effective mass approximation).*

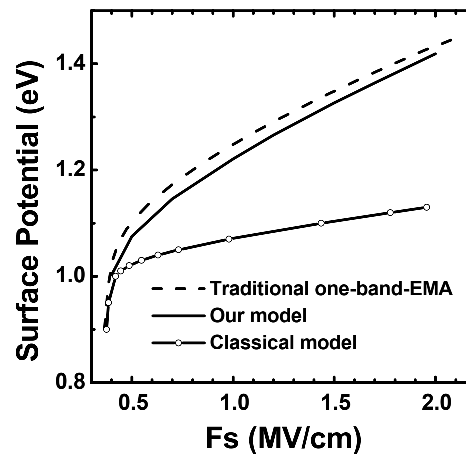


Fig.2.8: *The calculated surface potential (band bending) of pMOS structure at inversion. The substrate doping is* 5×10^{17} cm^{-3}*. The dashed curve is from the traditional one band effective mass approximation. The results of our six band model are shown as solid line. The solid line with open circles is that from classical calculation with Fermi-Dirac statistics.*

For the occupation factors OF_n in Fig.2.7 (b), our calculation shows an overall lower occupation on the n=1 subband and higher occupation on the n = 2 subband, compared to the traditional one-band EMA calculation. Both calculations predict that more than 95% of the carriers are distributed within the three lowest subbands. For subbands of higher order, the accuracy of the energy levels determined in triangular potential approximation compared to the self-consistent method is reduced, because the actual potential profile deviates from being linear as the energy increases. Fortunately, from Fig.2.7 (b), more than 95% of holes occupy the lowest three subbands and the inaccuracy of higher energy levels does not lead to significant error for prediction of device parameters, especially for state-of-art CMOS devices with high substrate doping concentrations.

The surface potential (band bending) of the pMOSFET is shown in Fig.2.8. Due to the QM effects, the surface potential is not pinned as in the classical case. It is obvious that our calculation (solid lines) leads to lower band bending than the traditional one-band EMA (dashed lines). It is because the traditional one band EMA underestimates the subband DOS as indicated in Fig.2.6. In order to achieve the same inversion charge density, the band needs to bend more in traditional one-band EMA and it leads to an overestimation of the band bending.

(4). Hole Inversion Capacitance and Threshold Voltage Shift due to Hole Quantization in PMOS Devices

Now we simulate the threshold voltage shifts and inversion layer capacitance due to quantum effects. The results are shown in Figs.2.9-2.11. Figure 2.9 shows that the simulated threshold voltage shift conducted by our method is in agreement with

the full-band self-consistent model [72]. Figures 2.10 (a) and (b) show the comparisons between the experimental data and our simulations. Our model achieves a good agreement with the experiments. It has been found that the hole inversion layer capacitance is not very sensitive to the complicated valence band structure [10]. We actually found no considerable difference between our model and the traditional one-band EMA. However, the situation is different for the threshold voltage shift (ΔV_T). ΔV_T is due to the lower DOS of 2-D carriers than the 3-D case and this parameter is expected to be more sensitive to the subband DOS. A significant discrepancy between the ΔV_T calculated by our model and the traditional one-band EMA is found, as shown in Fig.2.10. The traditional one-band EMA gives larger ΔV_T values than those by our model due to its underestimation of the subband DOS. Our model shows a better overall agreement with the experiments.

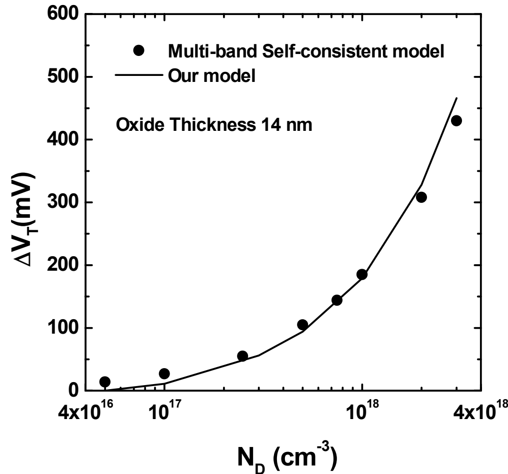


Fig.2.9*: The threshold voltage shifts due to hole quantum effects, ΔV_T, at different channel doping concentrations N_D. Solid and lines are results from our model and the multi-band* pseudopotential method *[72] respectively, with oxide thickness 14 nm.*

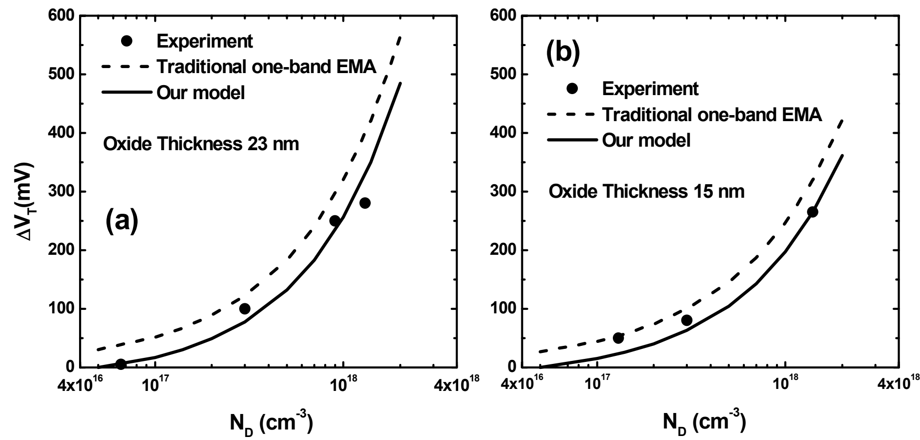


Fig.2.10: *The threshold voltage shifts due to hole quantum effects, ΔV_T, at different channel doping concentrations N_D. (a) and (b) are comparisons between our model (solid lines), the traditional one-band EMA (dashed lines), and the experimental data [82](solid circles) for oxide thickness 23 and 15 nm, respectively.*

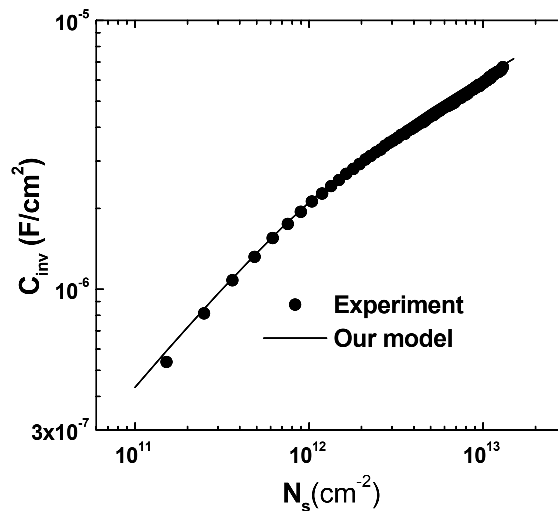


Fig.2.11: *The comparison of calculated and experimental capacitance of hole inversion layer. The solid line is our model simulation while the solid circles are experimental data from [10].*

2.4. Improved One-band Effective Mass Approximation

As discussed above, the DOS profiles of hole subbands deviate from those of traditional one-band EMA. Its shape is not regular and also depends on the magnitude of the surface electric field. Therefore, the hole subband structure cannot be described by a simple analytical model for accurate characterization. However, in some cases, we may not need to know the physical details of the hole subband structure, but only some macroscopic electrical parameters are of interest. In these cases, an improvement to the traditional one-band EMA assuming a parabolic dispersion is still valuable. In such a treatment, the analytical formulas in one-band EMA are still valid, the effective mass values are not derived from the bulk Si, but from the numerical results of the six-band EMA in an empirical way [79].

2.4.1. Empirical Effective Masses

From the numerical results of the six-band EMA, empirical effective masses can be derived for the improved one-band EMA. The results are plotted in Fig.2.12. Attention is first paid to empirical energy quantization effective mass m_z, as shown in Fig.2.12 (a). They are determined inversely by Eq.(1.10) from the energies of the subband minima obtained by six-band EMA. For n=1, the empirical energy quantization effective mass is 0.29 m_0 and is independent of F_s because of the purity of heavy hole. For n=2 and n=3 subbands, the empirical effective masses display an electric field dependent behaviour due to the field dependence of the band mixing. The empirical effective mass value of n=2 subband increases as F_s increases. Near k =

0, this band is mainly light hole (bulk mass 0.20 m_0). As electric field increases, more and more split-off hole (bulk mass 0.29 m_0) is mixed into this band, manifesting itself as an increase of the empirical effective mass. On the other hand, the decrease of the empirical effective mass of n = 3 subband with increasing electric field corresponds to the increased mixing of light hole into this primarily split-off hole subband near k = 0. From the inversion charge density occupying the respective subband calculated by six-band EMA, the obtained empirical DOS effective masses determined from Eq.(1.9) are shown in Fig.2.12 (b).

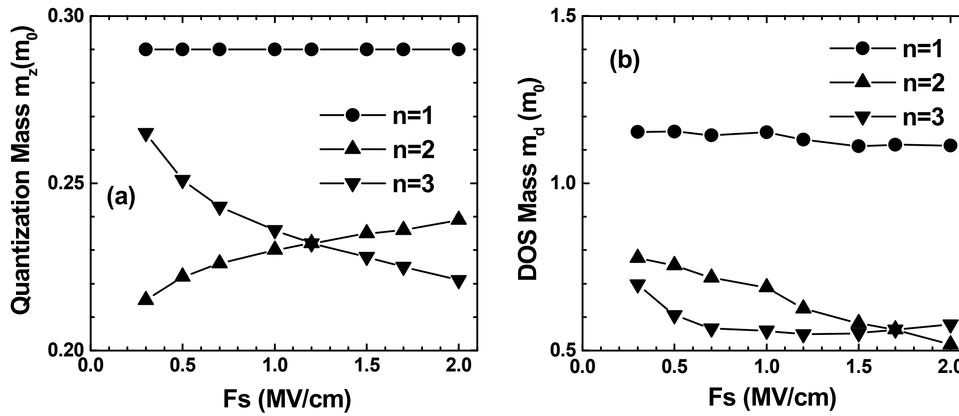


Fig.2.12: *The electric field F_s dependence of (a) empirical energy quantization effective mass m_z and (b) empirical DOS effective mass m_d, determined from the numerical results of six-band EMA calculations.*

Although some of these empirical effective mass values in Fig.2.12 are electric field dependent, this dependence can be neglected in the first order approximation. The reason is that at room temperature most of inversion holes are occupied on the n=1 subband (over 70%) as shown in Fig.2.7 (b). From Fig.2.12, both the energy quantization and DOS masses of n = 1 subband have weak electric field dependencies.

This leads us to propose a set of constant empirical effective mass values for an improved one-band EMA. They are found to be 0.29/1.16, 0.23/0.70, 0.23/0.60 m_0 for the first three subbands. As shown in Fig.2.13, such an improved one-band EMA can achieve consistent results on the subband energy levels, carrier occupations and surface potentials in comparison with the numerical results of six-band effective mass approximation.

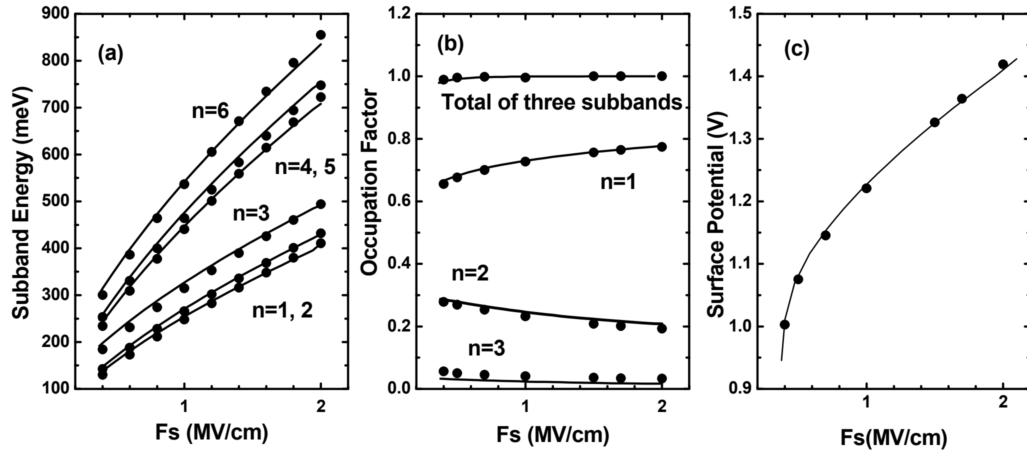


Fig.2.13: *Comparisons of (a) subband energies, (b) occupation factors and (c) surface potentials in the hole inversion layer calculated by the improved one-band effective mass approximation (solid lines) and the six-band effective mass theory (solid circles). The substrate doping concentration in the calculation is* 5×10^{17} cm^{-3}.

2.4.2. Effective Field Triangular Well Approximation

Triangular well approximation is often used because it yields an analytic solution to the Schrodinger equation. But it has been shown that triangular well approximation is invalid for strong inversion if the surface electric field is used as the electric field in

triangular well [13]. In order to extend the triangular well approximation into strong inversion, instead of the surface electric field, an effective field is introduced. This electric field can be regarded as the mean field in the inversion layer. It can be expressed as

$$F_s = F_{Eff} = q(\eta N_s + N_d)/\varepsilon_{si} \tag{2.14}$$

where N_s and N_d are inversion carrier sheet density and depletion charge sheet density respectively. η is a weighting coefficient to be properly chosen. In the universal electron mobility of inversion layer, η is taken to be 0.5 [3]. However, in the calculation of carrier sheet density, η is calculated to be 0.75 for inversion (electron) and 0.8 for accumulation layer (hole) in nMOS using one-band EMA [84]. With the subband energy levels, inversion and depletion charge densities determined from the six-band self-consistent model [75], we are able to obtain the value of η for hole quantization.

Based on the numerical data, η is calculated to be ~ 0.78 and 0.75 for n=1 and 2 hole subbands respectively at substrate doping of 10^{17} and 10^{18} cm^{-3} as shown in Fig.2.14. Apparently, η for n=3 hole subband is sensitive to surface field. Its strong dependency on electric field is a characteristic of poor triangular well fitting for higher energy levels. Lower energy levels are able to achieve a relatively constant η. The weighting coefficient η is also weakly dependent of the substrate doping over a range of electric fields as included in Fig.2.14.

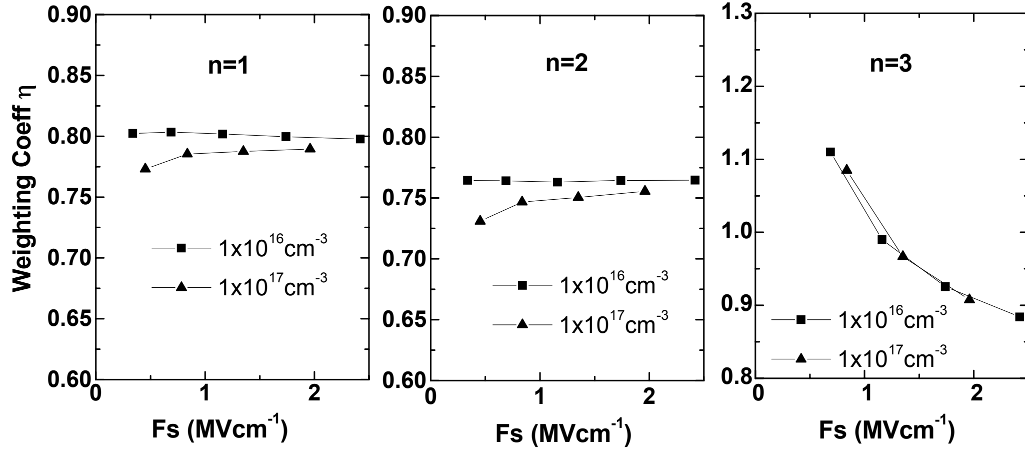


Fig. 2.14: *Weighting coefficient η as a function of surface electric field for n=1,2,3 hole subbands respectively for different doping concentration.*

Using the parameterized self-consistency method for the triangular model [81], we are able to obtain the MOS electrostatics efficiently. Using the constant effective masses in previous section, assuming a weighting coefficient of 0.77 (an average of n=1, and n=2), good fit for the subband energies, occupation factor and surface potential can be obtained as shown in Fig. 2.15. However, for triangular well model, the hole centroid falls short of the full band result at strong inversion although the results are consistent in depletion or weak inversion. This indicates that the triangular well approximation fails to describe the hole distribution in the inversion layer. Generally the triangular well approximation gives a smaller hole centroid compared to the full band self-consistent method. This point is similar to what found in electron quantization in nMOS [84].

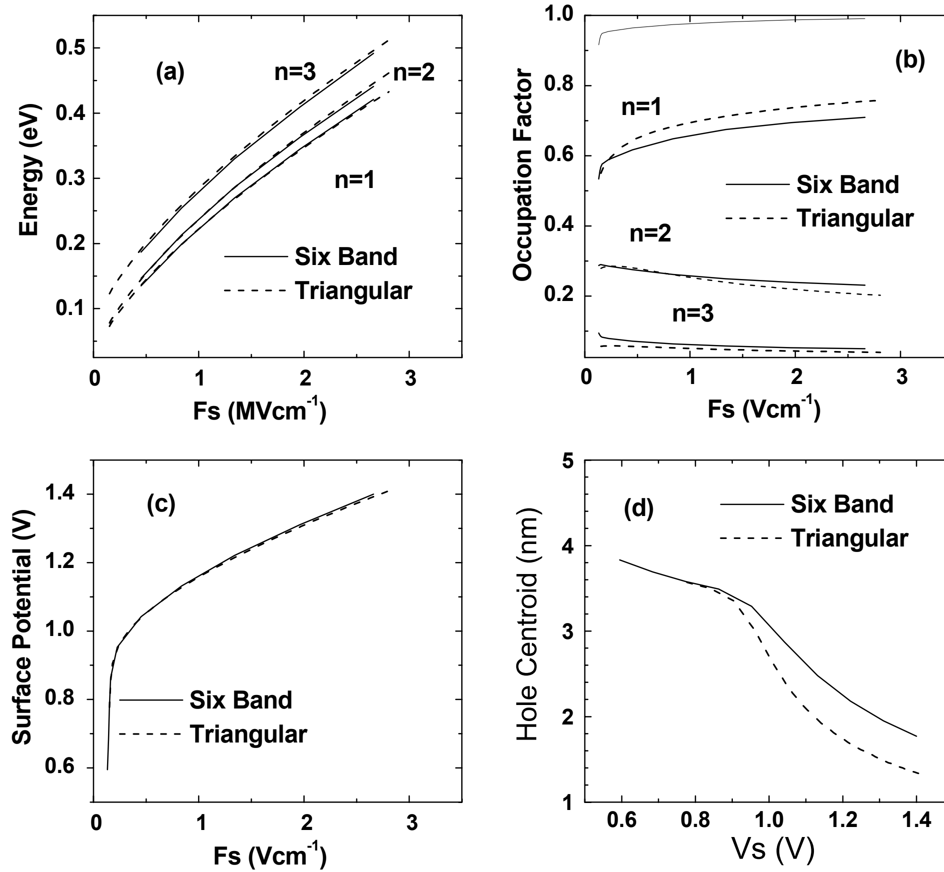


Fig. 2.15. *Comparisons of the results from improved one-band triangular approximation using an effective field (η=0.77) and six-band self-consistent EMA. They are (a) subband energies, (b) occupation factor and (c) surface potential versus surface electric field, and (d) the hole carrier centroid at different surface potential* V_s.

2.4.3. Hole Quantization by Improved One-band Effective Mass Method

As discussed before, the capacitance-voltage characteristics of MOS devices are not sensitive to the actual values of effective mass. Therefore it is natural that our improved one band EMA model can be applied to model the capacitance-voltage curves of MOS devices. In Fig.2.16, a comparison of the calculated threshold voltage shift by six-band and improved one-band EMAs is given. The data justified the accuracy of the improved one-band EMA in modeling the threshold voltage shift due to hole quantization in pMOSFETs.

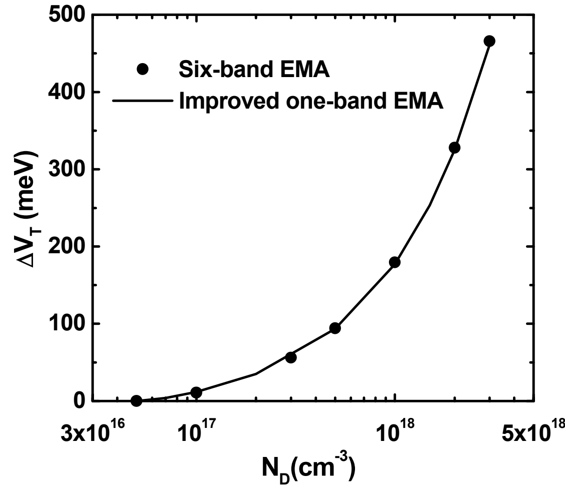


Fig.2.16. *Threshold voltage shifts ΔV_T at various substrate doping N_D . The numerical results by six-band effective mass model are shown as solid circles while solid line is those by improved one-band effective mass method using our new constant empirical effective masses.*

2.5. Conclusion

In this chapter, we presented a study of hole quantization in MOS devices based on the six-band effective mass theory and a zigzag potential well approximation. It is demonstrated that valence band mixing must be taken into account in order to describe the hole quantum mechanical effects properly. Due to the band mixing, all subband dispersions are anisotropic, far from parabolic, and electric field dependent. Correspondingly, the density of state profiles deviate substantially from the step-like functions. The traditional one-band effective mass approximation, which used the bulk effective masses, underestimates the subband density of states and therefore overestimates the hole quantum mechanical effects. Based on the numerical results of the six band effective mass approximation, an improved one-band effective mass approximation was also proposed, in which the hole quantization mass and DOS mass values were obtained from the numerical results from six-band calculations. Using this model, the calculated hole inversion capacitance and threshold voltage shifts due to hoe quantization are both in good agreement with the experiments. By introducing an effective electric field, in which the weighting coefficient was determined by comparing to the numerical results from rigorous six-band model, the triangular well approximation was successfully extended to strong inversion region. Owing to its computational efficiency and sufficient accuracy, its potential application in device modeling is demonstrated.

Chapter 3

Direct Tunneling Currents through Ultrathin Gate Oxides in CMOS Devices

3.1. Introduction

With the reduced supply voltage design for deep submicron CMOS devices, the transistor is operating in direct tunneling regime. In current manufacturing technology, the gate oxide thickness has been scaled to below 2 nm, which leads to a dramatic increase of direct tunneling current. The most obvious effect of direct tunneling is the increase of gate leakage current, the power dissipation is thus increased at both operating and stand-by modes. With the increased integrity density in the chips, the power dissipation problem becomes so severe that it limits the further scaling of gate oxide thickness, particularly for low power applications [85]. Therefore, it is crucial to thoroughly understand and accurately model the various tunneling components in CMOS devices to the design of device structure with optimized leakage current.

On the other hand, for thinner oxide, the traditional capacitance-voltage (C-V) measurement is difficult to perform due to the large direct tunneling current [86]. As an alternative, the tunneling current might be exploited to extract device physical parameters due to its sensitivity to device structures, such as oxide thickness, doping concentration, interface states, oxide charges, etc. Finally, the reliability

characterization of thin oxide involves aging with electrical stress. For thin oxide, more detailed description of the tunneling mechanism is also necessary to understand the effects of the different stressing conditions on the oxide reliability.

In this chapter, the characterization and modeling of direct tunneling current through ultrathin gate oxide in CMOS devices will be presented. First, Section 3.2 is an introduction to the conduction mechanism of dual poly-Si gate CMOS device with thin gate oxide, followed by a brief summary of the physical model used in this thesis to calculate the direct tunneling current in Section 3.3. After a description of the experiments and C-V results in Section 3.4, Section 3.5 emphasizes on the hole tunneling current in pMOSFETs, the effect of the non-parabolic hole dispersion in oxide energy gap on hole tunneling will be identified. Finally, the modeling of all terminal tunneling components in CMOS devices will be conducted in Section 3.6.

3.2. Conduction Mechanism in Dual Poly-Si Gate CMOS Transistors

Dual poly-Si gate process, in which p+ poly-Si gate is used for pMOSFET and n+ poly-Si gate for nMOSFET, is the dominant processing technology in current IC manufacturing. A complete understanding of the carrier transport processes in CMOS devices is not only important for the current-voltage (I-V) characteristics of MOSFETs at operation [87, 88], but also helpful for studying the mechanism of thin oxide reliability [89-91]. The carrier transport mechanisms in dual poly-Si gate CMOS transistors at both bias polarities will be summarized in this section.

3.2.1. Carrier Separation Measurement

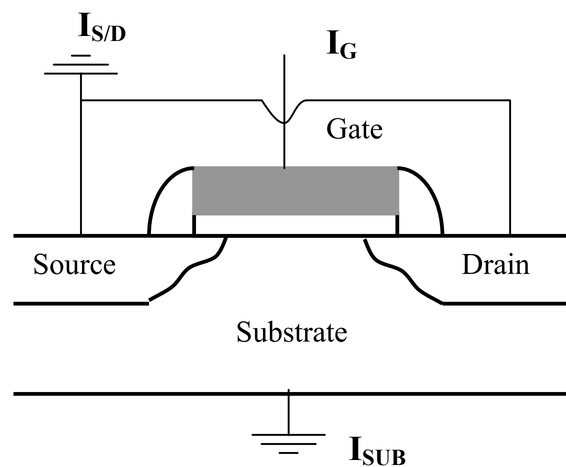


Fig.3.1: *The cross-sectional schematic of the carrier separation measurement.*

The carrier separation measurement is the most widely used experimental method in the study of the carrier conduction mechanism in CMOS transistors [87-89]. The experimental set-up of the carrier separation measurement is illustrated in Fig.3.1. During the measurement, the source and drain of the transistor were tied together, labeled as S/D terminal, which is grounded together with the substrate terminal. The carrier separation experiments were conducted to measure the gate current I_G, the sum of the source and drain currents $I_{S/D}$, and the substrate current I_{SUB} separately when a voltage V_G was applied and swept onto the gate terminal. Without the presence of other leakage channels, the summation of all terminal currents, $I_G + I_{S/D} + I_{SUB} = 0$. In the subsequent discussions, the currents flowing into (out of) the device are taken as positive (negative) signs.

The interpretation of carrier separation measurement at inversion condition is easy and direct. When MOSFET is biased in inversion mode, $I_{S/D}$ measures the current of minority carriers while I_{SUB} measures the current of majority carriers. In accumulation condition, MOSFET can be considered as a quasi-bipolar transistor. The S/D acts as the collector and $I_{S/D}$ measures the minority current in the base. The substrate acts as the base and I_{SUB} measures the majority current in the base. In the absence of electron-hole recombination, when applying negative (positive) gate voltage V_G to the nMOSFET (pMOSFET), the $I_{S/D}$ and I_{SUB} are the electron (hole) and hole (electron) currents, respectively. However, in the presence of electron-hole recombination, the situation is more complicated. When the diffusion length of minority carrier is comparable in dimension to its current path in the substrate, recombination of minority carriers with majority carriers in the substrate may occur, and consequently it may transfer to majority carrier current before reaching the S/D electrodes, resulting in smaller S/D current. In our experiments, only fresh devices

were used in order to avoid the reduction of minority carrier diffusion length arising from recombination via high density of recombination centres (interface traps, defects, etc), and the measurements were performed on short channel devices to reduce the minority carrier transport path.

3.2.2. Conduction Mechanism in n+ Poly-Si Gate NMOSFETs

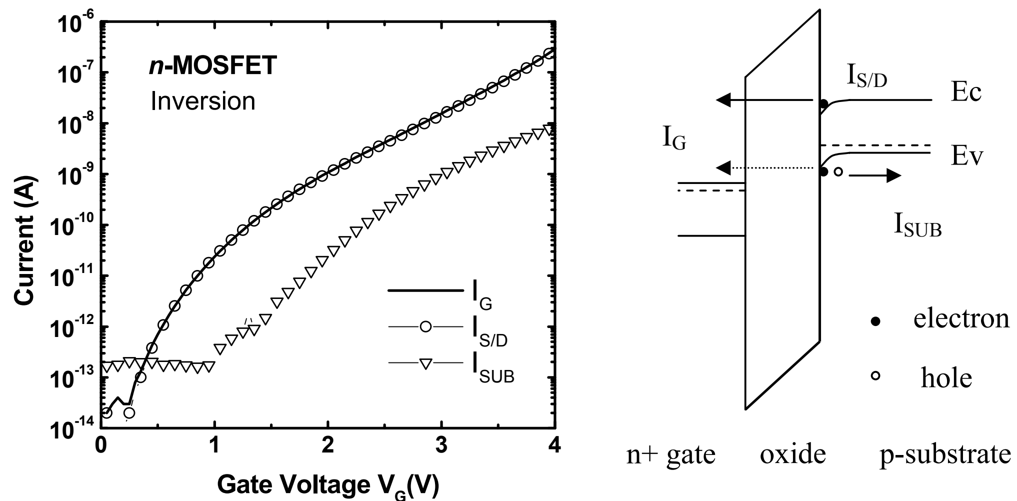


***Fig.3.2**: Current-voltage (I-V) characteristics and band diagram of a n+ poly-Si gate nMOSFET at inversion. The transistor gate length and width are 20 and 0.5 μm, respectively, and the oxide thickness is ~ 2 nm.*

Attention is first paid to inversion mode as shown in Fig.3.2. Here the $I_{S?D}$ (electron current) dominates the gate leakage, while the magnitude of I_{SUB} is lower than $I_{S/D}$ in 1-2 orders of magnitude. From the band diagram, the gate current is originated by the electron tunneling from the inversion layer in the substrate to the gate. The electron density is high at substrate surface, and steady current is

established by supplying electrons in the inversion layer from the S/D. The substrate current is interpreted as due to electron tunneling from the substrate valence band. When an electron is tunneled out, a hole is left behind, which diffuses into the substrate terminal and forms the observed I_{SUB}. The valence band electron tunneling occurs when V_G > ~1V (flat-band voltage) and its much smaller magnitude is due to its higher tunneling barrier than that of the conduction band electron by a magnitude of the Si energy gap (1.12 eV).

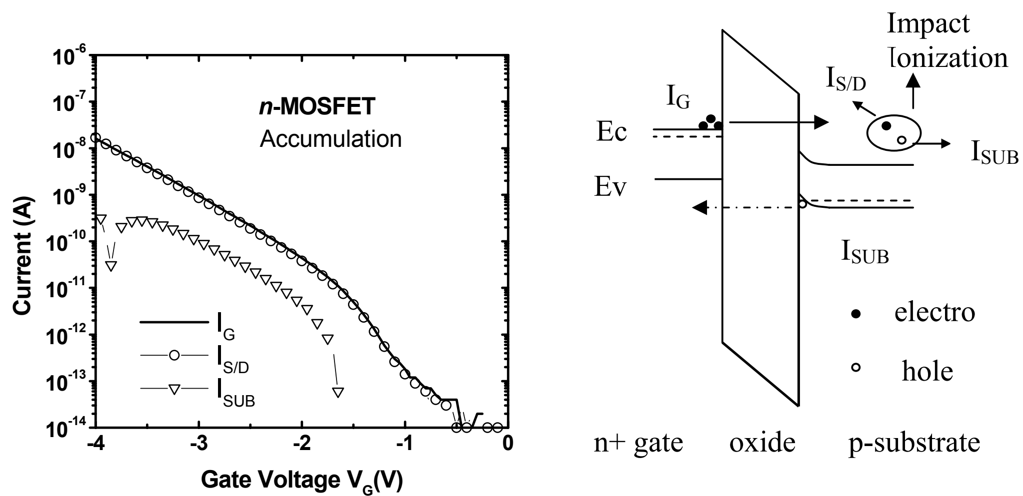


***Fig.3.3**: Current-Voltage (I-V) characteristics and band diagram of a n+ poly-Si gate nMOSFET at accumulation. The transistor gate length and width are 20 and 0.5 μm, respectively, and the oxide thickness is ~ 2 nm.*

In the case of accumulation, Fig.3.3 shows that the gate current I_G is mainly composed of the $I_{S/D}$. From the band diagram, the dominant tunneling channel is the electron tunneling from the accumulated n+ poly-Si gate to the substrate. The electrons, as minority carriers in the p-type substrate of nMOSFET, are collected by the S/D and a negative $I_{S/D}$ is thus established. The tunneling mechanism for hole

current in substrate is the hole tunneling from accumulated p-type substrate to the gate, which originates a positive I_{SUB} in Fig.3.3. For the substrate current, a much smaller magnitude and a change of sign near 4 V are observed. This phenomenon can be illustrated by the band diagram in Fig.3.3. At high gate voltage, impact ionization, induced by the high energetic electrons injected from the gate, happens and electron-hole pairs are resultantly generated in the substrate. While the generated electrons are collected by the S/D, the holes diffuse to the substrate terminal, forming a negative hole current. When the magnitude of the impact-ionization induced hole current (negative) is larger than the hole tunneling current (positive) as the gate bias increases, a change of sign in the I_{SUB} occurs.

3.2.3. Conduction Mechanism in p+ Poly-Si Gate PMOSFETs

Figure 3.4 plots the measurements on pMOSFET at inversion, which shows I-V characteristics quite different from nMOSFET. The $I_{S/D}$ (hole current) dominates the gate leakage only at about -2-0 V while the I_{SUB} starts to become the main component of gate current at $V_G < -2V$. From the corresponding band diagram, the hole current is due to the tunneling of holes in the inversion layer to the gate, resulting in a positive $I_{S/D}$. The possible channels of electron tunneling from the gate include the conduction band electrons and the valence band electrons. In normal dual poly-Si gate process, the p+ poly-Si gate is very heavily doped ($>5\times10^{19}$ cm^{-3}), which makes the inversion of the p+ poly-Si gate impossible. As a result, the electron density in the conduction band of the p+ poly-Si gate is negligible and consequently the I_{SUB} in Fig.3.4 is

primarily due to the valence band electron tunneling. A change of sign in the $I_{S/D}$ is also observed in Fig.3.4, which is similarly explained by the impact ionization. The holes generated by impact-ionization, as the minority carriers in the n-type substrate of pMOSFET, are collected by the S/D and lead to a negative Is/d.

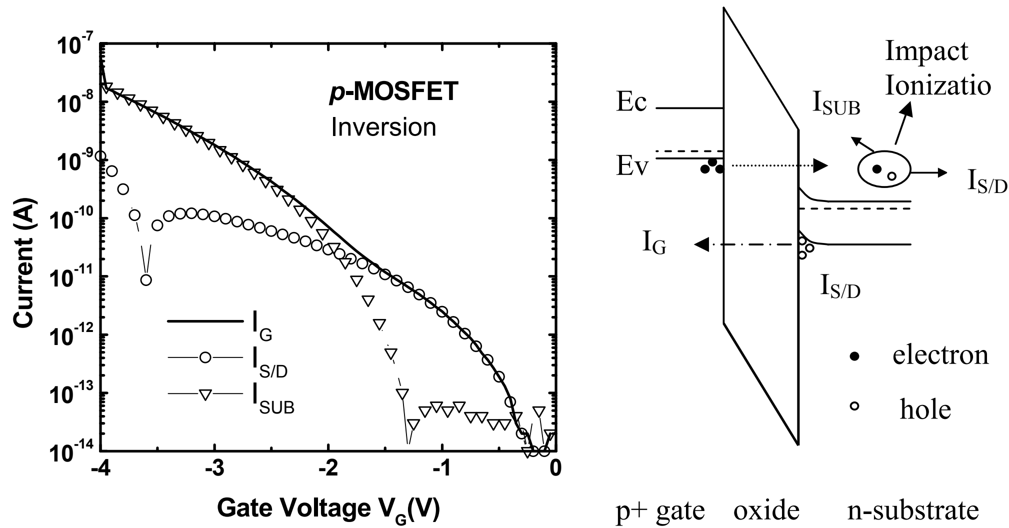


Fig.3.4: *Current-Voltage (I-V) characteristics and the band diagram of a p+ poly-Si gate pMOSFET at inversion. The transistor gate length and width are 20 and 0.5 um, respectively, and the oxide thickness is ~ 2 nm.*

In the case of accumulation mode of pMOSFET in Fig.3.5, the gate leakage is from the tunneling of electrons in the accumulated n-type substrate, which forms I_{SUB}. The interpretation of $I_{S/D}$, the hole current, is more complicated. There exist two tunneling channels: tunneling of accumulated holes in p+ gate to substrate and the diffusion of holes generated by valence band electron in the substrate tunneling out to gate. From the band diagram analysis in Fig.3.5, the valence band hole tunneling from gate occurs at V_G > ~1V (flat-band voltage V_{FB}) while the valence band electron

tunneling from the substrate is possible only when $V_G > V_{FB} + (E_G)_{Si}$ (~ 2.1 V with Si band gap = 1.12 V). It is thereby concluded that the $I_{S/D}$ below 2.1 V is due to hole tunneling, while valence band electron tunneling may possibly contribute to the $I_{S/D}$ with V_G higher than ~ 2.1 V.

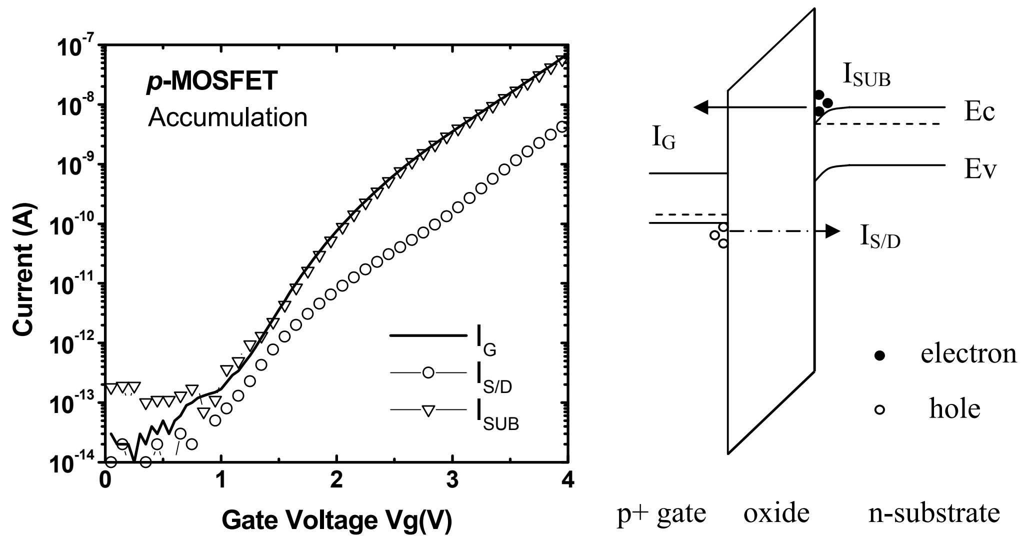


Fig.3.5: *Current-Voltage (I-V) characteristics and the band diagram of a p+ poly-Si gate pMOSFET at accumulation. The transistor gate length and width are 20 and 0.5 um, respectively, and the oxide thickness is ~ 2 nm.*

3.2.4. Conduction in Source/Drain Extension (SDE) Region

Figure 3.6 compares the current density measured from a short channel (10×0.18 μm^2) transistor and a sourced capacitor (10×10 μm^2) with ultra-thin oxide (~1.65 nm). The gate current density is of the same magnitude at inversion (not shown) or strong accumulation, however, in the gate voltage range of -1.5 to 0 V, the gate current density measured on small gate length transistor is much higher than that

from the capacitor. Because this current is not scalable with the channel area, its source is impossible to be the tunneling in the channel region. This extra current has been identified as due to the tunneling between the source/drain extension region (SDE) and the gate [90-93]. From the band diagram illustrated in Fig.3.6, electrons in accumulated n+ gate can tunnel to n+ SDE region overlapped with the gate. Due to their similar work function (V_{FB} ~ 0), this tunneling current occurs at Vg ~ 0 and dominates the gate current in gate voltage region between 0 and transistor V_{FB} because the band mis-alignment inhibits the tunneling between the gate and the channel.

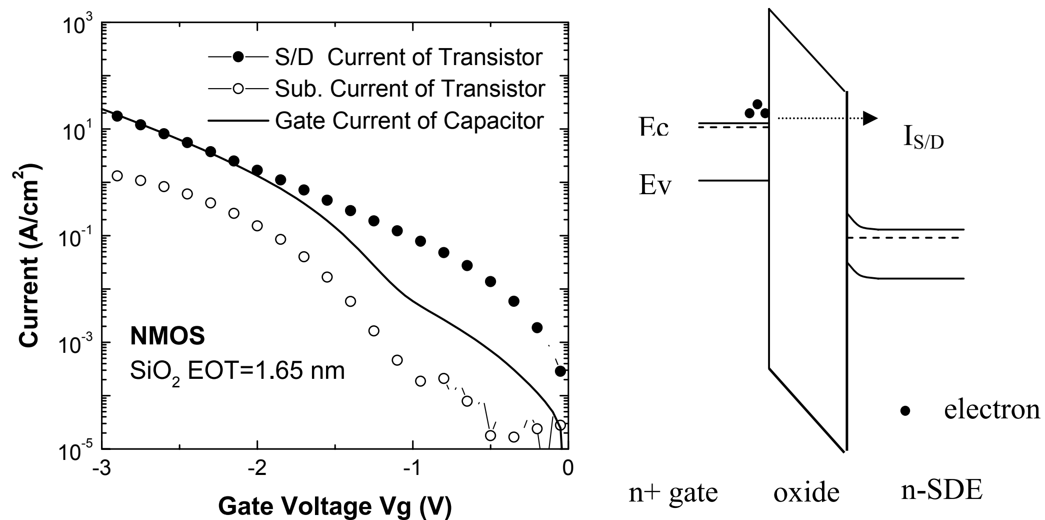


Fig.3.6: *Current-Voltage (I-V) characteristics of n+ poly-Si gate nMOS short channel transistor (10×0.18 μm² and larger area (10×10 um²) capacitor at accumulation and the band diagram at source/drain extension region overlapped with the gate. The oxide thickness is ~1.65 nm.*

The results on p+ poly-Si pMOSFET are shown in Fig.3.7. The same extra tunneling current is also observed in short channel pMOSFET. As illustrated by the

band diagram, the conduction mechanism between 0 ~ V_{FB} in pMOSFET is due to the tunneling hole current from p+ poly-Si gate to the p-doped SDE.

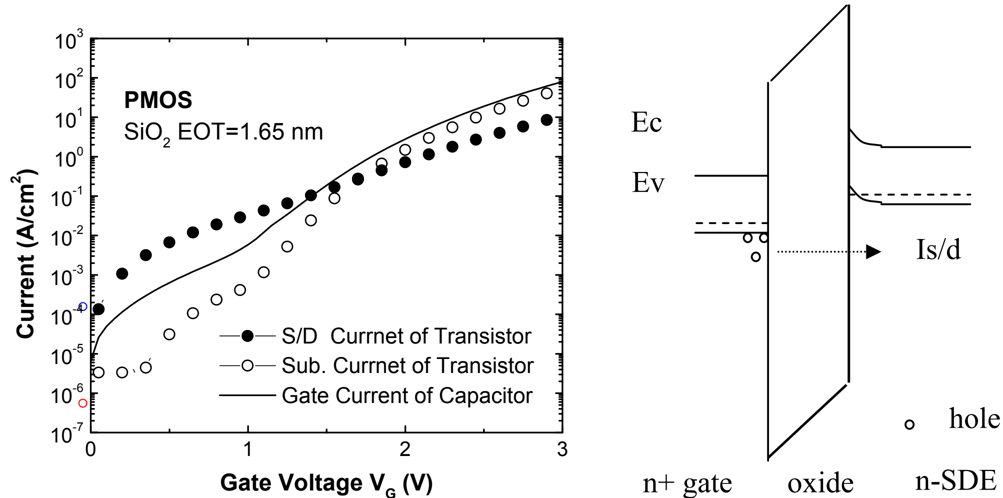


Fig.3.7: *Current-Voltage (I-V) characteristics of p+ poly-Si gate pMOS short channel transistor (10×0.18 μm^2 and larger area (10×10 um^2) capacitor at accumulation and the band diagram at source/drain extension region overlapped with the gate. The oxide thickness is ~1.65 nm.*

Because the overlap dimension of SDE with the gate is much smaller than the channel length, when compared to the tunneling in the channel area, the tunneling from this overlap region is negligible at strong accumulation. However, the dimension of the SDE overlap is not so scalable as the channel length [94], it is thus expected that tunneling in SDE overlap will play more important role in transistor with smaller gate length, particularly when the gate oxide thickness is also so scaled that the tunneling through oxide is greatly enhanced. As demonstrated from previous studies, the tunneling current in SDE overlap has significant impact on the off-state leakage in scaled transistors, as well as the ultra-thin oxide reliability [90-93].

3.3. Physical Model for Tunneling Current

In this thesis, the tunnelling current is simulated in the frame of the quasi-classical quantum model [32-35], such a method takes advantage of the computational efficiency of the classical method in obtaining the life-time of quasi-bound states, and it still preserves the main MOS quantum physics, such as the carrier quantization and tunneling nature from confined 2-D subbands. Compared to the rigorous quantum model, this efficient quasi-classical model is verified to be sufficiently accurate [36, 37].

In this model, the tunnelling current is obtained from Eqs. (1.17) and (1.18). For classical turning points, the life-time of the *n*th subband can be obtained by [34]:

$$\tau_n^{-1} = \left(\frac{n\pi\hbar}{E_n} \right)^{-1} T(E) \tag{3.1}$$

where *T(E)* is obtained in WKB scheme. However, the traditional WKB approximation is applicable only when the barrier varies slowly over a wavelength of the particle [23]. In a MOS structure, sharp boundaries exist at the interfaces between the oxide and Si substrate as well as the poly-Si gate. In the presence of a sharp boundary, wave reflection at the interface is expected, which is not taken into account in classical WKB approximation. Recently a modified WKB approximation with a correction factor accounting for such reflections at the boundaries of the oxide layer is proposed [33]:

$$T(E) = T_R(E) \cdot T_{WKB}(E) \tag{3.2}$$

where the correction factor T_R can be expressed as a function of the group velocities of the tunneling carrier:

$$T_R = \frac{4v_{Si}(E)v_{OX}(E_{OXi})}{v_{Si}^2(E)+v_{OX}^2(E_{OXi})} \times \frac{4v_{Si}(E+qV_{OX})v_{OX}(E_{OXo})}{v_{Si}^2(E+qV_{OX})+v_{OX}^2(E_{OXo})} \tag{3.3}$$

where $v_{Si}(E)$ and $v_{Si}(E+qV_{OX})$ are the group velocities of the carriers in Si incident and leaving the oxide layer, respectively. V_{OX} is the oxide voltage drop whereas $v_{ox}(E_{OXi})$ and $v_{ox}(E_{OXo})$ are the magnitudes of the imaginary group velocities of carriers tunneling in and out of the oxide layer, respectively. From the matching of wave functions and conservation of the carrier flux at the oxide boundaries, the same form of T_R as Eq.(3.3) is also derived independently by other authors [27, 28].

After obtaining N_n the carrier density occupied on nth subband from Eq. (1.9), the tunneling current from the nth subband can be expressed explicitly as:

$$J_n = \frac{qm_{dn}(kT)}{\pi^2\hbar^3}\left(\frac{E_n}{n}\right)T(E)\ln\left(1+\exp\left(\frac{E_F-E_n}{kT}\right)\right) \tag{3.4}$$

Finally, the total current is the summation over all subbands.

To determine the charge density N_n and subband energy E_n in Eq. (3.4), the calculations of carrier quantization and electrostatics of MOS structure are required, which is done in this thesis by a parametric self-consistency method in triangular well approximation as described in Chapter 2. For hole quantization, the improved one-band effective mass approximation is used to account for the valence band mixing effect. As discussed before, the accuracy to apply the triangular well approximation

beyond weak inversion may be questionable. To extend the triangular well approximation into strong inversion or accumulation, an improved triangular well method in effective electric field scheme, as presented in Section 2.4, is used to determine the MOS electrostatics and its validity has been verified by comparing to rigorous self-consistent method.

The gate voltage V_g is determined from the voltage balance equation,

$$V_g = V_{FB} + V_{ox} + V_P + \phi_s \tag{3.5}$$

where V_{FB} is the flat-band voltage, oxide voltage drop $V_{ox} = F_{ox} \cdot t_{ox}$, V_P the voltage drop in poly-Si gate and ϕ_s the substrate band bending.

3.4. Experiments and C-V Characterization

In our experiments, the MOSFETs were fabricated by a standard dual poly-Si gate CMOS processing technology. The oxide was grown by rapid thermal oxidation followed by annealing in N_2O ambient. The purpose of nitridation is to suppress the boron penetration in pMOSFET. The low concentration level of incorporated nitrogen is expected to have minor effect on the material properties of oxide. It is thus assumed that the pure SiO_2 properties, such as band offset, dielectric constant etc, are still preserved in our samples. The *C-V* measurements were performed using a HP4284A LCR meter on large area (400×60 μm^2) MOS capacitors at a frequency of 100 kHz. The current-voltage (*I-V*) characteristics were measured using the HP4156A semiconductor parameter analyzer. The tunneling currents were measured using a carrier separation method, as depicted in Section 3.2. For tunneling in channel area, devices with different areas were compared to ensure that the current density exhibits no area dependence in order to avoid the edge effect. Short channel transistors (0.18 μm gate length) were used to measure the current from gate to source/drain extension (SDE) tunneling. Tunneling currents of n- and p-MOSFET were measured on the same wafer and non-uniformity of oxide thickness within the wafer was not observed.

The device parameters, such as substrate and gate doping concentrations, gate oxide thickness and flat-band voltage, are independently determined by fitting to the measured C-V curves using a quantum C-V simulator developed by device group at UC Berkeley [95]. The results are shown in Fig.3.8. For samples with oxide thickness < 2 nm, reliable C-V is difficult to be measured due to the high gate tunneling current. For these samples, the doping concentration values determined from samples with

thick oxide, which were processed at the same conditions except the oxidation recipe, were used.

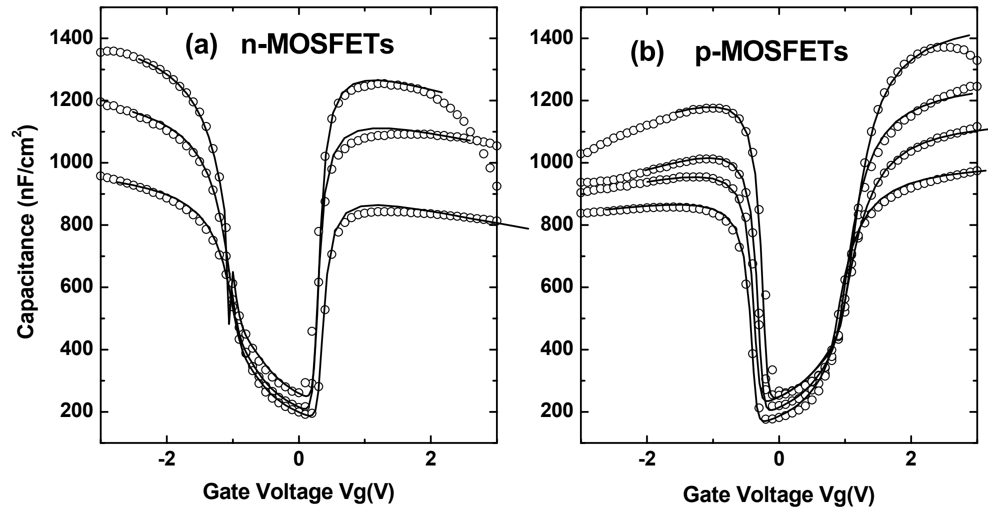


Fig.3.8. *The measured capacitance-voltage(C-V) characteristics of the MOSFETs used in the simulations. The experimental data are shown as open circles and the solid lines are the fitting results using the QM-CV model of device group at UC Berkeley. The extracted oxide thickness is 1.85, 2.07, 2.44 and 2.74 nm, respectively.*

3.5. Non-parabolic Effect in Hole Direct Tunneling Current

After obtaining the device parameters from C-V characterization, first we will perform a study on hole direct tunneling current in pMOSFET.

In literatures, most studies on tunneling currents in MOS structures are concentrated on tunneling of conduction band electrons from inversion or accumulation layers [25-38]. As for *p*+ poly-Si gate *p*MOSFETs, hole direct tunneling was found to dominate the gate current under channel inversion condition [87-89]. However, relatively less attention has been paid to the modeling of hole direct tunneling [27, 35, 96-101]. In a reliable physical model of hole direct tunneling current, two important characteristics should be present. First there should be an accurate treatment of the hole quantization effect in Si substrate, which is important for an accurate determination of the hole densities at different energies and the respective voltage drops in the oxide layer and substrate. As discussed in Chapter 2, it is done in this thesis by employing an improved one band effective mass approximation (EMA) including valence band mixing effect. Second there should be a reliable dispersion form for holes in the energy gap of the dielectric film, which is crucial in the WKB approximation. All previous models assume a parabolic dispersion in the SiO_2 energy gap during hole tunneling [27, 35, 96-99]. However, its applicability has never been verified because the energy of valence holes aligns at the middle of the SiO_2 energy gap. In this section, we focus on the simulation of hole direct tunneling current. A more appropriate Freeman-Dahlke dispersion form [102, 103] is used to calculate hole direct tunneling current. It is physically more reasonable and it also achieves a significant improvement in the matching of simulation results to the experimental data

[100, 101]. In addition, the applicability and limitation of a parabolic approximation to hole dispersion in the SiO_2 energy gap will also be addressed.

3.5.1. Dispersion Relationship in Oxide Energy Gap

From Eq.(1.14), the dispersion relationship $\kappa(E)$ in dielectric energy gap is crucial for tunneling because it appears in the exponential factor. Unfortunately, little knowledge on the exact dispersion relationship $\kappa(E)$ in dielectric is known, even for SiO_2, which has been used in MOS technology for decades.

Currently, the most widely used dispersion relationship in the energy gap of SiO_2 is the parabolic one, in which an energy independent effective mass, m_{ox}, is assumed,

$$\frac{1}{(\hbar k)^2} = \frac{1}{2m_{ox}E} \tag{3.6}$$

where the energy E is measured from the conduction (valence) band edge for electron (hole) tunneling.

For electron tunneling, a non-parabolic dispersion effect has been identified and a Franz-type dispersion, instead of the parabolic one, is also proposed [25],

$$\kappa(E) = \left(\frac{2m_{OX}}{\hbar^2}\right)^{1/2} E^{1/2}\left(1 - \frac{E}{E_{gox}}\right)^{1/2} \tag{3.7}$$

where E_{gox} is the energy gap of SiO_2. From the Franz-type dispersion, the conduction and valence band edges have the same effective mass, which has never been physically justified.

Here we introduce another form to describe the dispersion in the SiO_2 gap, which was initially proposed by Freeman and Dahlke [102]:

$$\frac{1}{(\hbar k)^2} = \frac{1}{2m_{COX}(E_{COX} - E)} + \frac{1}{2m_{VOX}(E - E_{VOX})} \tag{3.8}$$

where m_{COX} and m_{VOX} are the effective masses of conduction and valence band of SiO_2 respectively. Equation (3.8) takes into account the difference in effective masses between conduction and valence bands. When E approaches E_{COX}, the first term in the right hand side of Eq.(3.8) becomes the dominant term and $\kappa(E)$ reduces to a parabolic relationship with the conduction band effective mass m_{COX}. When E approaches E_{VOX}, the second term in the right hand side of Eq.(3.8) becomes the major term and $\kappa(E)$ reduces to a parabolic relationship with the valence band effective mass m_{VOX}. Apparently, Eq.(3.8) reduces to the Franz-type dispersion when $m_{COX} = m_{VOX} = m_{OX}$.

3.5.2. Electron Tunneling in NMOSFETs

In tunneling current simulations, the effective mass in oxide is generally used as a fitting parameter. The oxide thickness is determined by other independent methods,

such as C-V, high-resolution transmission electron microscopy (HRTEM) and optical ellipsometry. The oxide effective mass values in a range of 0.3-0.6 m_0 have been reported for electrons in literatures [25-38]. One reason for the inconsistency is most likely due to the limited accuracy of different methods in the determination of oxide thickness. In most cases, oxide thickness is determined by fitting to the experimental C-V curves using a C-V model. Unfortunately, the oxide thickness extracted from the same experimental C-V curve is quite different when using different C-V models (deviation up to 1.3 A) [104]. As pointed out in [104], it is hard to judge which model is valid or accurate at present stage. It is also reported that effective mass may depend on the oxide thickness [28, 105].

In this chapter, we determine the oxide thickness from fitting to the experimental electron tunneling currents from inversion layer in nMOSFETs by assuming a fixed value of electron effective mass in oxide, which is 0.50 m_0 for parabolic and 0.61 m_0 for Franz-type dispersion, respectively. This oxide thickness rather than that from the *C-V* method will be used in the simulation of tunneling current from other carriers, such as valance band holes, valence band electrons etc. Such a treatment should not be misinterpreted as meaning that the effective mass is always independent of oxide thickness. Considerable uncertainty still exists over whether the effective mass depends on oxide thickness. Assuming a constant effective mass ensures the consistency in the calculations of various tunneling current components. Additionally, it should not lead to significant errors if the oxide thickness is not varied in a wide range.

Figure 3.9 shows the simulations of electron direct tunneling currents from the inversion layers of nMOSFETs. In the calculations, the Si/SiO_2 conduction band offset is fixed at 3.15 eV [29, 32-34]. The solid lines are the calculated results using an

empirical Franz-type dispersion (m_{ox}=0.61m_0) [33, 34]. The best fits to the experimental results were obtained by adjusting oxide thickness T_{OX}. Similar to previous studies [33, 34], the calculated *I-V* characteristics are in good agreement with the experiments at all voltages (0–3 V) and for all oxide thicknesses (1.8–2.7 nm).

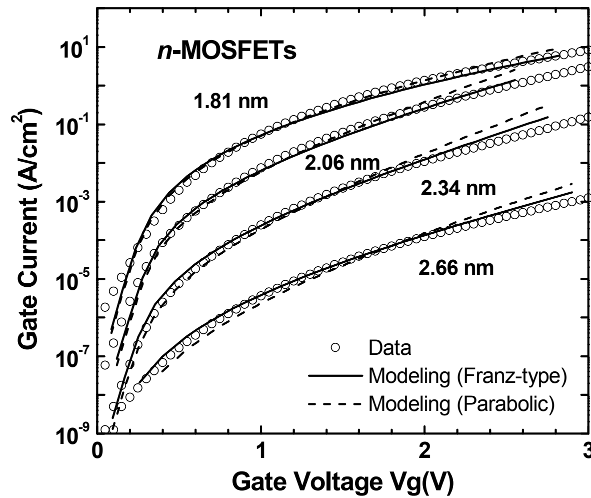


Fig.3.9: *The electron direct tunneling currents in nMOSFETs. The open circles are the measurements. The solid and dashed lines are calculations by assuming the electron dispersion in SiO_2 energy gap to be Franz-type (m_{ox}=0.61 m_0) and parabolic (m_{ox}=0.50 m_0), respectively.*

The calculated results using a simple parabolic dispersion (m_{ox} = 0.50m_0) are also displayed in Fig.3.9 as dashed lines. The fitting results using the simple parabolic dispersion are only slightly degraded for thick oxides. It indicates that the parabolic dispersion still remains to be a good approximation for modeling electron direct tunneling. The obtained effective mass values in parabolic dispersion are also in

agreement with the band structure calculation of bulk SiO_2, in which an effective mass of about 0.5 m_0 is demonstrated for the conduction band [106, 107].

The oxide thickness determined from the fitting of electron tunneling for our four samples are 1.81, 2.06, 2.34 and 2.66 nm. They are close to the values determined from the *C-V* method (1.85, 2.07, 2.44 and 2.74 nm, respectively). The maximum deviation is about 1 Å and this is within the reported limits of different experimental methods, such as *C-V*, HRTEM and optical ellipsometry [108].

3.5.3. Simulation of Hole Tunneling Current Using Freeman-Dahlke Dispersion Form

In this section, the simulations of hole tunneling current were performed on p-MOSFETs. Here, a valence band offset between Si and SiO_2 of 4.5 eV was used [96]. In the modeling of hole tunneling current, most of previous reports [27, 35, 96-99] used the simple parabolic dispersion form. However, for valence hole tunneling, the top of the Si valence band aligns at the middle of the SiO_2 energy gap. According to the more general Freeman-Dahkle form, two items in the right hand side of Eq.(3.8) are comparable and neither of them can be neglected in the condition of hole tunneling. This indicates that a parabolic approximation is not physically appropriate. The Freeman-Dahlke dispersion of Eq.(3.8) is expected to provide more accurate results on hole tunneling. Here, all the three dispersion forms will be used to study the hole tunneling and a comparison will be given.

First, we discuss the hole direct tunneling current using a parabolic dispersion in energy gap of SiO_2. The results obtained by our physical model are displayed as dashed lines in Fig. 3.10. Although a Franz-type dispersion has never been used for hole tunneling, we also show the calculations from a Franz-type dispersion ($m_{OX} = 0.55\ m_0$) as solid lines in Fig. 3.10 for a comparison. From Fig.3.10, the hole tunneling *I-V* characteristics computed from either the parabolic or Franz-type dispersion are not as close to the experimental data as those for electron tunneling. The fit for thinner oxides (< 2 nm) is better. However neither the parabolic nor the Franz-type dispersion can fit the experimental data well when the oxide thickness is larger than about 2 nm. The deviation is more significant at high gate voltage.

Assuming a parabolic dispersion, m_{OX} is found to be $0.41m_0$ in order to get the best fitting. This value is close to the previous results reported for valence band electron or hole tunneling (0.35–0.50m_0) [27, 35, 96-99]. It suggests a smaller m_{OX} value for holes than for electrons. This is in conflict with the existing results from the band structure calculations of bulk SiO_2. For SiO_2, the calculated effective mass of valence bands is typically 3–10 m_0 [106, 107], which is much larger than that of the conduction band (~0.5 m_0).

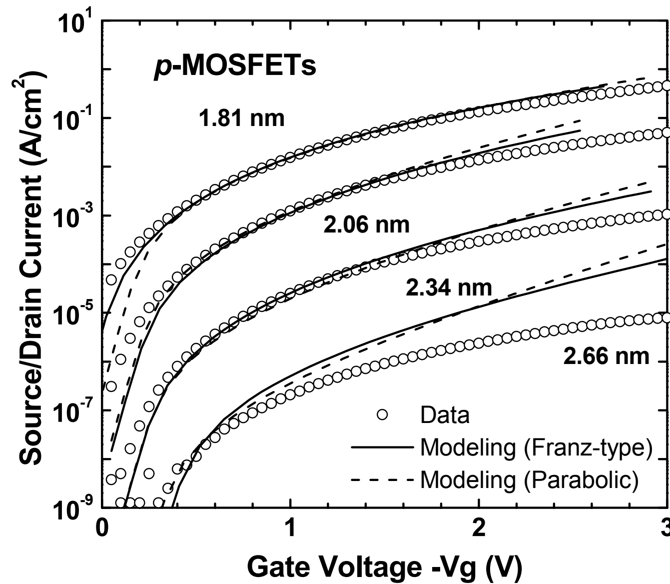


Fig.3.10: *The hole direct tunneling currents in pMOSFETs. The open circles are the measured data. The solid and dashed lines denote the calculated values by assuming the hole dispersion in SiO_2 band gap to be Franz-type ($m_{ox}=0.55m_0$) and parabolic ($m_{ox}=0.40\ m_0$), respectively.*

In the following, we use the Freeman-Dahlke form of Eq.(3.8) in SiO_2 energy gap to calculate the hole direct tunneling current. In Freeman-Dahlke form, there are two necessary effective mass values. Based on the electron tunneling and SiO_2 band structure calculation, we set the conduction band effective mass m_{COX} at $0.5m_0$, while valence band effective mass m_{VOX} is used as an adjustable parameter for best fitting to the experimental data. A T_{OX} independent value of $m_{VOX} = 0.8\ m_0$ can provide the best results and the simulations are shown in Fig.3.11 as solid lines. From Fig.3.11, it is apparent that a much better fitting of simulated hole tunneling current to the experimental data is achieved by using the Freeman-Dahlke form.

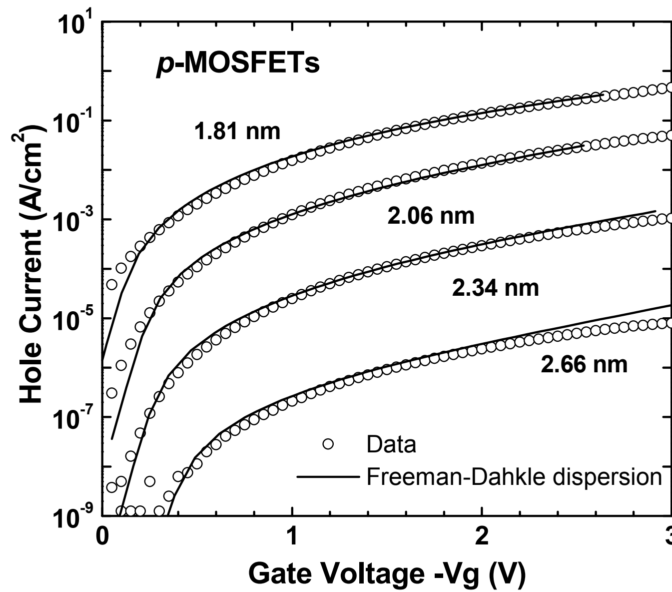


Fig.3.11: *The hole direct tunneling currents in pMOSFETs. The open circles are the measured data. The solid lines denote the calculated values by assuming a Freeman-Dahlke form dispersion in SiO_2 band gap with m_{cox}=0.50 m_0 and m_{vox}=0.80 m_0.*

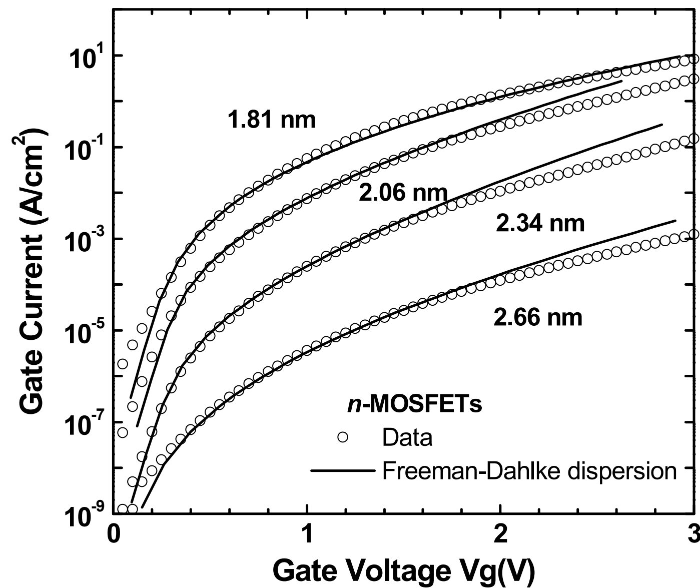


Fig.3.12: *The electron direct tunneling currents in nMOSFETs. The open circles are the measured values. The solid lines denote the calculated values by assuming a Freeman-Dahlke form dispersion in* SiO_2 *band gap with* $m_{cox}=0.50\ m_0$ *and* $m_{vox}=0.80\ m_0$.

In Fig.3.12, we also present the calculations using the Freeman-Dahlke dispersion for electron direct tunneling currents. Comparing to the measured data, this dispersion is also applicable to electron tunneling with the same effective mass parameters. As discussed before, the contribution of the second term in Eq.(3.8) is negligible for conduction electron tunneling. As a result, calculation using the Freeman-Dahlke form is very similar as that using the traditional parabolic dispersion.

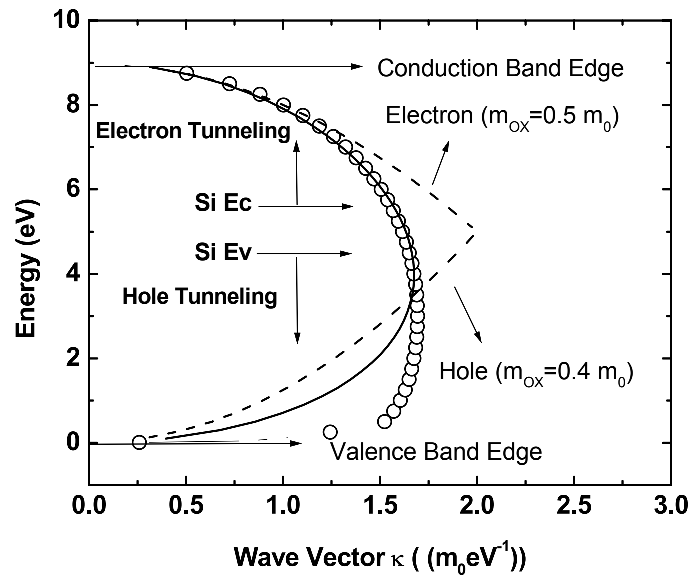


Fig.3.13: *The calculated imaginary wave vector in the energy gap of SiO_2 by Freeman-Dahkle (solid line) and microscopic model [38](open circles). The dashed lines are those determined from parabolic dispersion.*

Finally, when compared to the parabolic assumption, the Freeman-Dahlke form gives not only much better agreement with the experimental data, but also a reasonable correlation between electron and hole dispersion in the band gap of SiO_2. A further insight of the dispersion in SiO_2 energy gap can be obtained from Fig.3.13, in which we show respectively the calculated imaginary part of SiO_2 band by the Freeman-Dahkle form and the tight-binding method [38]. In our experiment, the hole energy ranges from ~ 1.5 eV to 4.5 eV measured from the oxide valence band during tunneling. In vicinity of 4.5 eV, the imaginary wave number, which is responsible for hole tunneling, does not vary strongly in this energy region. This explains why the tunneling effective mass for holes is similar to that of electrons. Comparing the shapes of the dispersion in Fig.3.13, it is easily concluded that the dispersion from the

Freeman-Dahlke form is much more close to that by the microscopic model in the whole band gap (0-9 eV) than that from the parabolic approximation. Although hole tunneling may involve several imaginary bands with a rather complicated structure and it is difficult to be described in the simple effective mass scheme, the analytical Freeman-Dahlke form is demonstrated to be a reasonable approximation, which can provide hole tunneling current in qualitative consistency with experiments.

From the dispersion in Fig. 3.13, the non-parabolic effect observed in Fig.3.10 can be tentatively explained. At a fixed gate voltage V_G, a thicker oxide layer has a smaller oxide field, and hence smaller voltage drops in both substrate and poly-Si gate. According to Eq.(3.5), the oxide voltage drop is larger. Correspondingly, there is a larger energy range to be integrated in Eq.(1.14). As a result, a simple parabolic dispersion, which deviates from the real dispersion, will lead to a larger integration error in Eq.(1.14). A similar argument is also applicable to the case of high gate voltage. This explains the deviation between the experimental and simulated results in Fig.3.10.

3.6. Simulations of All Terminal Direct Tunneling Currents in CMOSFETs

As discussed in previous section, the direct tunneling current can be simulated with a proper consideration of the dispersion in oxide energy gap. Despite its limitation in describing hole tunneling, the parabolic dispersion has the advantages of simplicity in formalism and efficiency in computation. When oxide thickness is less than 2 nm, the use of the parabolic dispersion still preserves a fair accuracy in the modeling of hole direct tunneling current. In current CMOS processing technology, the gate oxide thickness has been scaled to below 2 nm. Therefore, a parabolic dispersion will remain to be the most effective and important dispersion form used in the modeling of tunneling current. In this section, the simulations of all terminal direct tunneling currents in CMOS transistors with ultra-thin oxide (< ~2 nm) were performed by using the proposed physical model.

3.6.1. Conduction Band Electron Tunneling Current

So far, conduction band electron tunneling is the most widely studied tunneling components. As discussed in Section 3.2, the tunneling from conduction band electrons can occur in both biases of nMOSFET and in accumulation bias of pMOSFET.

Figure 3.14 shows the results of conduction band electron direct tunneling from the inversion layers of nMOSFETs with oxide thickness down to 1.45 nm. The solid

lines are the calculated results using parabolic dispersion (m_{ox}=0.50 m_0). Similar to previous section, the best fits to the experimental results were obtained by adjusting oxide thickness T_{OX}. The T_{OX} values are shown in the figure and will be used in the subsequent calculations of other tunneling current components. When oxide thickness is below 2 nm, the C-V measurement is difficult to perform because of the very high gate leakage, especially at strong accumulation from which oxide thickness is extracted. From comparisons between oxide thickness values extracted from C-V and I-V measurements for thick oxide, as discussed in Section 3.5.2, direct tunneling provides an alternative method to measure the thickness of ultra-thin oxide layer. The uncertainty in the extracted oxide thickness value is estimated to be ~ 0.02 nm from 10% variation of tunneling current.

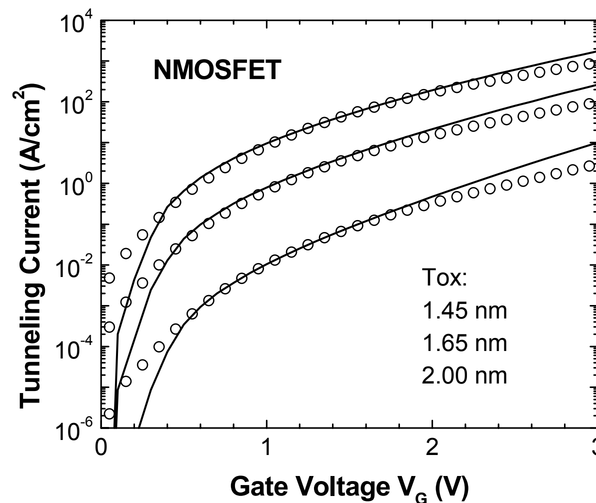


Fig.3.14: *The electron direct tunneling currents in nMOSFETs. The open circles are the measurements. The solid lines are the calculations by assuming the electron dispersion in SiO_2 band gap to be parabolic (m_{ox}=0.50 m_0).*

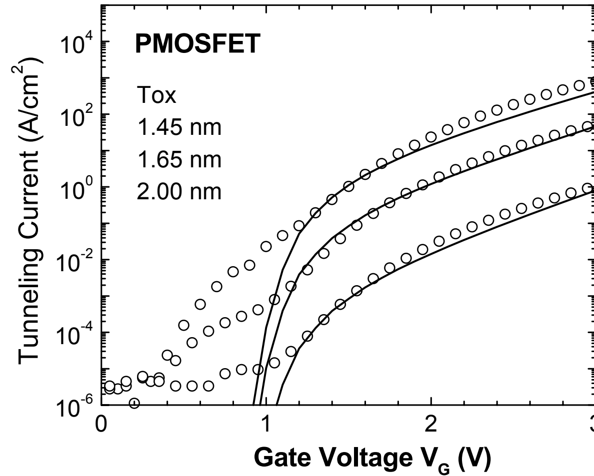


Fig.3.15: *The electron direct tunneling currents in pMOSFETs. The open circles are the measurements. The lines are the calculations by assuming the electron dispersion in* SiO_2 *band gap to be parabolic* ($m_{ox}=0.50\ m_0$).

The substrate currents measured in accumulation mode on pMOSFETs are shown in Fig.3.15, which is identified as the conduction band electron tunneling from the accumulated n-type substrate. Although some authors used different effective mass values for inversion and accumulation electrons in order to achieve the best fitting to the experimental data [9], here we used a constant m_{OX} value for conduction band electrons. The solid lines are the calculated results and the agreement with experimental data is fairly good without any adjusting parameters.

3.6.2. Valence Band Hole Tunneling Current

From Section 3.5, when using 4.5 eV as the valance band offset between SiO_2 and Si, the hole effective mass in oxide is determined to be 0.41 m_0 in a parabolic dispersion. The tunneling from valence band holes can be measured by carrier separation method at $V_G < 0$ bias. They are identified as the S/D current in pMOSFETs and substrate current in nMOSFETs, which are shown in Fig.3.16 and Fig.17, respectively. It is found that the experimental data follow closely the theoretical calculations without any fitting parameters for tunneling from both inversion or accumulated holes. The deviation in high V_G for 2 nm oxide is attributed to the non-parabolic dispersion effect, as discussed in previous section. The results demonstrate that a simple parabolic dispersion is applicable to calculate the hole tunneling through ultra-thin oxides.

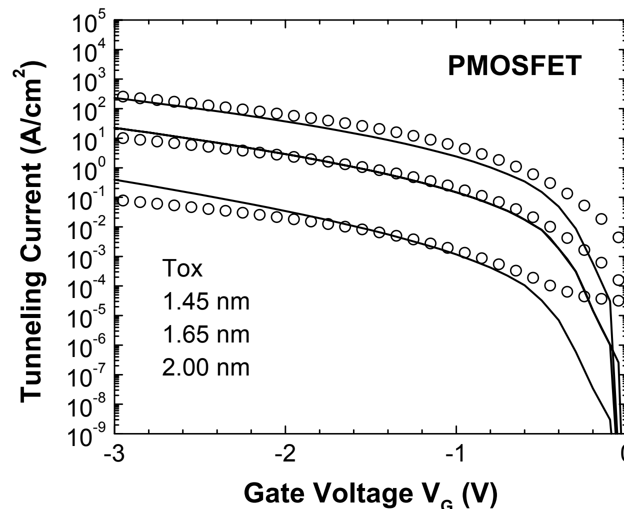


Fig.3.16: *The hole direct tunneling currents in pMOSFETs. The open circles are the measurements. The solid lines are the calculations by assuming the hole dispersion in SiO_2 band gap to be parabolic (m_{ox}=0.41 m_0).*

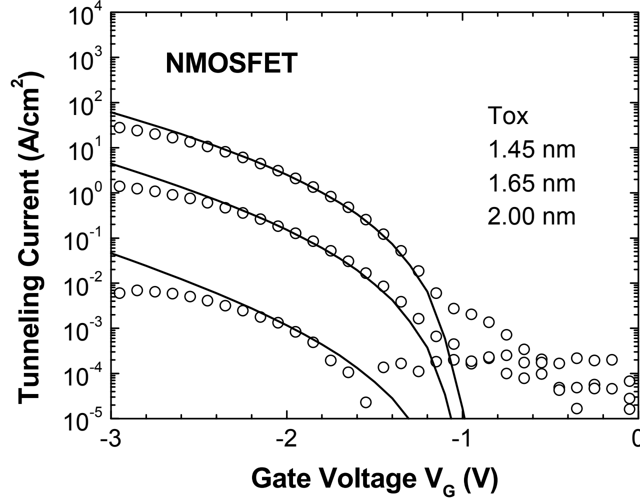


Fig.3.17: *The hole direct tunneling currents in nMOSFETs. The open circles are the measurements. The solid lines are the calculations by assuming the hole dispersion in* SiO_2 *band gap to be parabolic (*m_{ox}*=0.41* m_0*). The noise level at low voltage is due to the measurement limit of the analyzer.*

3.6.3. Valence Band Electron Tunneling Current

The treatment of valence band electrons (VBEs) in MOS structure can follow the classical model because there is no quantization of valence band electrons. From the classical scheme, the tunneling current from valence band (electrode ***s***) to conduction band (electrode ***g***) with energy E_z (referenced from Ev=0) can be written as [27]:

$$J_{VE}(E_z) = \frac{qm_d}{2\pi^2\hbar^3} T(E_z)(E_{vs} - E_{cg} - E_z)dE_z \qquad (3.9)$$

Here, it is assumed that the valence band is fully filled with electrons and *T(E)* is

independent of lateral momentum k_x and k_y. In the current model, the VBE tunneling occurs only when Ev (***s***) > Ec (***g***) when tunneling occurs from ***s*** to ***g*** electrode. The total current is the integration of E_z from valence band edge *Ev* (***s***) to conduction band edge Ec (***g***):

$$J_{VE} = \int_0^{qVox-Eg} J_{VE}(E_z)dE_z \quad (3.10)$$

In CMOS transistors, VBE tunneling can be measured by substrate current in nMOSFET at $V_G > 0$ and pMOSFET at $V_G<0$. We show here only the former case in Fig.3.18 because the I-V characteristics of VBE tunneling observed in n- and pMOSFETs are the same [109]. In the calculations of VBE tunneling, the effective mass in oxide is assumed to be the same as holes (0.41 m_0) because the tunneling barrier for VBE (4.27 eV) is in similar magnitude as hole barrier (4.5 eV). As shown by the solid lines, the calculated VBE tunneling current is consistent with measurements at high V_G. However, higher VBE tunneling currents than the calculated one is measured at low V_G, which is also observed by other authors [96, 109]. This might be due to the limitation of the current method in treating VBE tunneling. In the current model, only electrons with energy below the surface valence band edge are taken into account. However, from the band diagram shown in Fig.3.18, valence band electrons above surface valence band edge probably contribute to the observed currents by tunneling through the very thin inversion layer and subsequently the oxide layer. This can qualitatively explain the earlier occurring of VBE tunneling, however, further quantitative study is necessary.

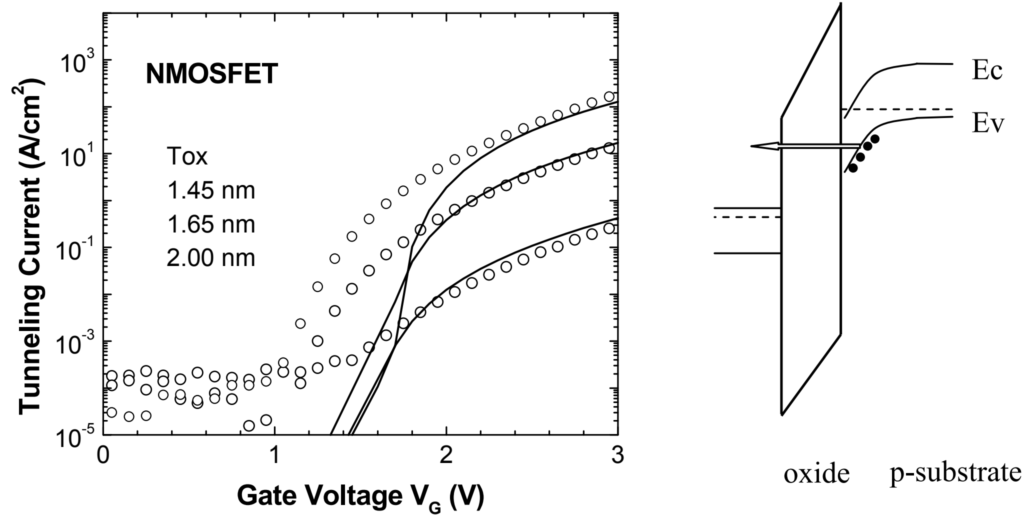


Fig.3.18: *The direct tunneling currents from valence band electrons in nMOSFETs. The open circles are the measurements. The solid lines are the calculations by assuming the electron dispersion in SiO_2 band gap to be parabolic ($m_{ox}=0.41\ m_0$). The left band diagram illustrates the possible tunneling from valence band electrons with energy above the surface valence band edge, which has not been included in the present calculations.*

3.6.4. Tunneling in Source/Drain Extension Overlap Region

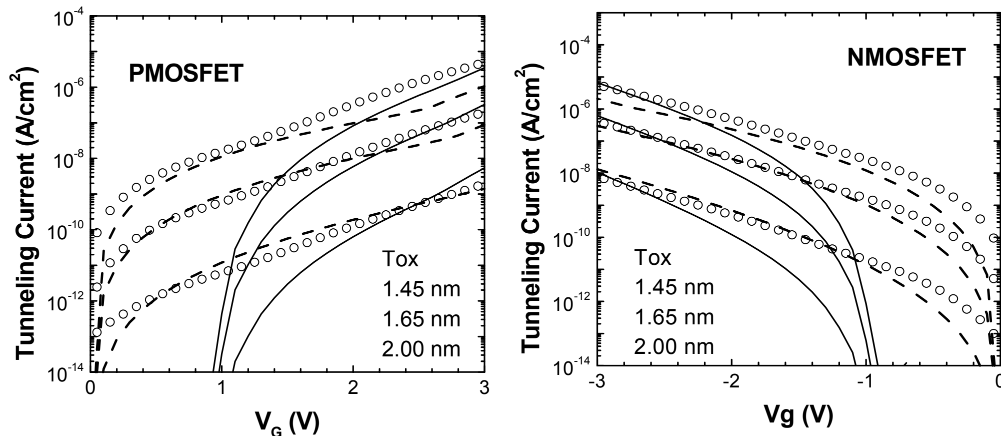


Fig.3.19: *The gate to source/drain extension (SDE) tunneling currents in CMOSFETs. The open circles are the measurements. The solid and dashed lines are the calculated tunneling currents in channel and SDE area, respectively.*

As presented in Section 3.2, the gate current of a short channel transistor measured between 0 and V_{FB} is dominated by the tunneling in SDE region. In nMOSFET (pMOSFET), it is originated from the conduction band electron (valence band hole) tunneling from the accumulated poly-Si gate to the n-doped (p-doped) SDE region at depletion. As a result, the calculations of tunneling in SDE can follow the same method by considering poly-Si as same as Si (100). The calculated tunneling currents in SDE region (dashed lines) are shown in Fig.3.19. The calculated tunneling currents from poly-Si gate to channel area are also shown as solid lines. By comparing to the measurements, the overlapped dimension of the SDE with the gate is estimated to be ~ 5 nm for nMOSFET and 6 nm for pMOSFET, which is consistent with the reported values by other researchers [92, 93]. Based on the gate to SDE tunneling, an effective method to extract the channel length in short channel transistors has also been proposed [110].

Finally, the tunneling currents measured and simulated in CMOS transistors with 1.65 nm oxide are summarized in Fig.3.20. The physical model can produce all tunneling components of CMOS transistors in agreements with the experiments.

Although the simulations show good agreements with the measurements at high gate voltage, for some components, some difference between the simulations and measurements can be seen. For electron or hole tunneling from accumulated substrate (Fig.15 and Fig,17) shown as substrate currents at $V_G>0$ in pMOSFET and $V_G<0$ in nMOSFET in Fig.3.20, when V_G is between 0 and V_{FB}, tunneling cannot happen because of the band misalignment. However, a measurable current can be observed in this region, especially for thinner oxide, this current has been identified as tunneling into the interface states [147]. For valence band tunneling currents, shown as substrate currents at $V_G>0$ in nMOSFET and $V_G<0$ in pMOSFET in Fig.3.20, the possible

reason have been given in Section 3.6.3. At $V_G \sim 0$, the measured current is always found to be higher than simulation, which may be caused by the interface states and trapped charges [34].

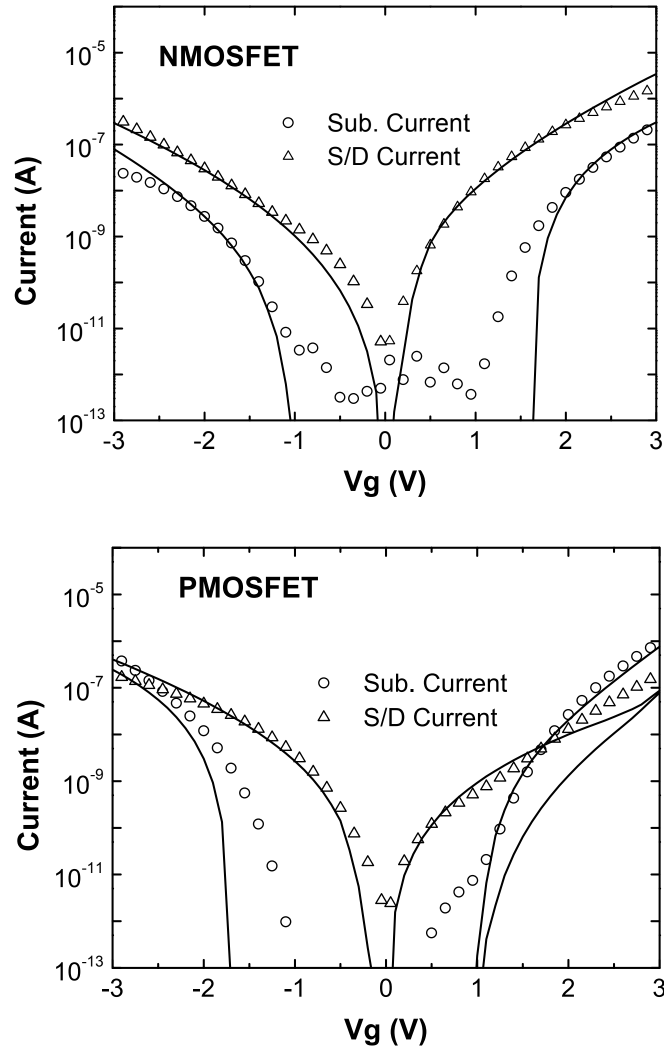


Fig.3.20: *The simulations of all terminal direct tunneling currents in CMOSFETs. The open symbols are the measurements on 10×0.18 μm² transistor by the carrier separation method. The solid lines are the calculations using the physical model. The oxide thickness is ~1.65 nm .*

3.7. Conclusion

An efficient physical model was proposed to calculate the direct tunneling current in CMOS devices. This model takes carrier quantization into account. In particular, a new improved one-band effective mass approximation, including the valence band mixing, was employed to compute the hole quantization effect. A modified Wentzel-Kramers-Brilliouin (WKB) approximation accounting for the reflections at oxide interfaces was used for the tunneling probability calculation. The validity of the model is verified by simulated results consistent with experiments for electron tunneling by using either parabolic or Franz-type dispersion relationship in oxide energy gap. Hole tunneling current was studied more intensively, and a Freeman-Dahlke dispersion form is found more appropriate for describing the hole dispersion in oxide energy gap. After taking the difference of conduction and valence band effective masses into account by the Freeman-Dahlke form, the agreement of the simulated hole tunneling currents with the experimental data is significantly improved over a wide range of oxide thickness and gate voltage. The applicability of the widely used parabolic approximation to hole tunneling is also addressed. It is demonstrated that, when oxide is thinner than 2 nm, the parabolic dispersion remains quantitatively accurate in the modeling of hole direct tunneling current. Using the simple parabolic dispersion, the successful modeling of all terminal tunneling currents in CMOS transistors by the proposed model was also presented.

Chapter 4

Tunneling Current and Scalability of High-K Gate Dielectrics in CMOS Technology

4.1. Introduction

As discussed in Section 1.5, the practical limit for reducing SiO_2 thickness is at about 10-12 Å, which corresponds to the 65 nm technology node to be deployed in 2007 as projected by ITRS 2001. With the continued scaling of the equivalent oxide thickness (EOT) required by performance improvement, gate dielectric materials with higher permittivity (high-K) than SiO_2 are demanded in order to suppress the direct tunneling current. In recently years, alternative high-K materials are of immense research interest. Several kinds of candidate dielectrics have been exploited, including Si_3N_4, Al_2O_3, HfO_2 and ZrO_2 [4]. On the selection of gate dielectrics, the direct tunneling current inherent to the dielectric materials serves as one of the most critical criterions. It is thus technologically important to have a deep understanding of the direct tunneling characteristics of these high-K dielectrics. Further, technological requirements may vary significantly from one application to another, including the performance, off-state power and supplied voltage. This makes it also necessary to evaluate the dielectric materials based on these application dependent requirements.

In this Chapter, we investigate the tunneling current through several popular gate dielectric materials, including Si_3N_4 and Al_2O_3 in Section 4.2 and HfO_2 system in Section 4.3. With the band offsets and dielectric constants obtained by other independent experimental methods, the effective mass values inherent to these materials, which determine the magnitude of the direct tunneling current, are extracted by comparing the simulations with experiments. Finally, the scalability of these dielectric materials in future CMOS technology with different applications is assessed in terms of gate leakage, which is included in Section 4.4.

4.2. Direct Tunneling through Si_3N_4 and Al_2O_3 Gate Dielectric Stacks

In this section, we will first study the direct tunneling characteristics of Si_3N_4 and Al_2O_3. They are among the most sufficiently investigated dielectric materials to date. Both of them show promising thermal properties, which ensure their compatibility with current CMOS processing. In addition, they often act as basic elements in the engineering of alternative dielectrics. For example, nitrided SiO_2 or oxynitride (SiON) still remains to be the predominant gate dielectric material employed in current CMOS manufacturing. Due to the improved thermal stability, it is also of intense interest to alloy HfO_2 with Al_2O_3 to form HfAlO.

4.2.1 Tunneling Currents through Si_3N_4, Oxynitride Gate Stacks

Among various dielectric materials other than SiO_2, silicon nitride (Si_3N_4) system, including oxynitride and oxide/nitride stack, has attracted the most considerable attention and has been investigated intensively in the past [111-123]. In addition to its slightly higher K value than SiO_2 (pure Si_3N_4 ~ 7.8) for reduced leakage current, the addition of nitrogen to SiO_2 also greatly reduces boron penetration, which is a significant concern in p+ poly-Si pMOSFET with ultra-thin gate oxide. In current CMOS technology, in which the nitridation of gate oxide is usually based on N_2O or NO technique, nitrogen with very low concentration (< ~3 at.%) is incorporated into oxide. Despite the improved immunity of gate oxide to

boron penetration, the nitridation has negligible impact on the gate leakage [117]. Recently, new nitridation techniques, such as remote plasma nitridation (RPN) [118] and decoupled plasma nitridation (DPN) [119], were employed to increase the nitrogen concentration in oxide with appreciable reduction in the gate leakage. Oxynitride or oxide/nitride stack with high nitrogen content was also obtained by other deposition techniques, such as chemical vapour deposition (CVD) [117], jet vapour deposition (JVD) [111-113] and atomic layer deposition (ALD). Pure Si_3N_4 films with ultrathin thickness were also reported [114, 115, 120]. Leakage current lower ~100× than that of SiO_2 was demonstrated by the best results.

The tunneling current through oxynitride or oxide/nitride stack has been studied by simulation tools [112-114, 116, 121-122]. Basic characteristics on tunneling were obtained. However, relatively less modeling work was presented with recognized parameters justified by reliable experiments. This is due to the high defect and trap density in the deposited films, which enhances the trap-related current conduction. In Ref. [111-115], high quality oxynitride or Si_3N_4 films with low leakage current were obtained by JVD technique, making it possible to model the tunneling current accurately. The electron tunneling currents were calculated in [112-113] on relative thick film, in which F-N tunneling is the main topic. In Ref. [114], the simulation on direct tunneling through pure Si_3N_4 film was performed using an empirical model. Based on these reliable experimental data, here we present our calculations of direct tunneling current through ultrathin Si_3N_4 gate dielectric.

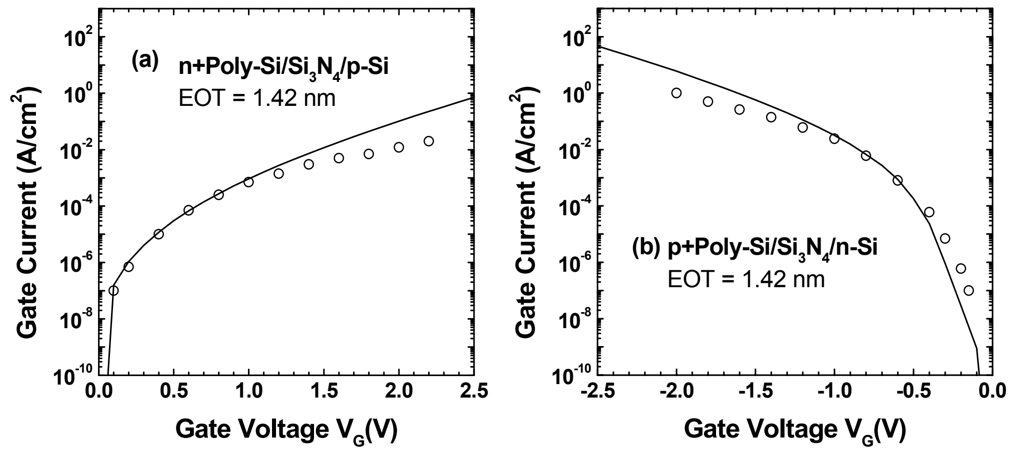


Fig.4.1: *Calculated (solid lines) (a) electron and (b) hole direct tunneling currents through a Si_3N_4 film with EOT of 1.42 nm from inversion layer of n - and pMOSFETs. The measured data (open circles) are from [114].*

Figure 4.1 displays the electron and hole direct tunneling currents through 1.42 nm JVD Si_3N_4, the measured data are from [114]. The device parameters, such as the channel doping concentration, poly doping concentration, and flat-band voltage were determined from the capacitance-voltage (C-V) measurements [115]. In our calculations, the conduction and valence band offsets between Si_3N_4 and Si were taken from literatures and their values are 2.1 eV and 1.9 eV, respectively [112-114]. The calculated results are shown as solid lines in Fig.4.1, and the effective mass in Si_3N_4 was used as the single fitting parameter. To fit the experiments best, the tunneling effective mass values for electron and hole were found to be ~0.50 m_0 and 0.41 m_0, respectively, in a parabolic dispersion, which are consistent with the reported results [112-114]. For nitride deposition, it is challenging to suppress the oxide interfacial layer

growth during the deposition. In JVD technique, the high impact energy provides the energy for the deposition process and therefore high quality film has been formed at room temperature. The room temperature deposition process provides the best control of interfacial layer [115].

As seen in Fig.4.1, an important feature of tunneling current characteristics of Si_3N_4 is that, the hole tunneling current is larger in magnitude than the electron current. As a result, in CMOSFETs with Si_3N_4 gate dielectric, pMOSFET exhibits larger tunneling current than nMOSFET, suggesting that the leakage current limit will first be reached by pMOSFET. This is contrary to the case of SiO_2 and will be critical in the assessment of the scalability of Si_3N_4 dielectric in CMOS applications [41,121, 122].

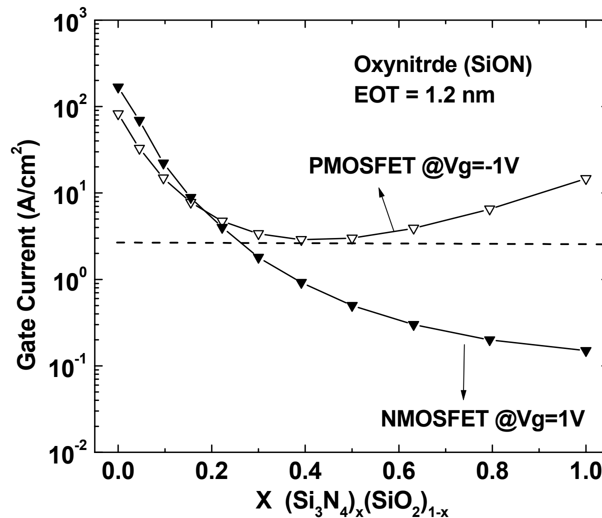


Fig.4.2*: Calculated tunneling currents in CMOSFETs with oxynitride (SiON) versus the nitrogen composition.*

From the simulation results, it is also found that the electron and hole effective mass values in Si_3N_4 are very close to those of SiO_2. It is thereby reasonable to

assume these effective mass values are independent of the nitrogen content in oxynitride film. Given the band offset and dielectric constant values linearly extrapolated between those of SiO_2 and Si_3N_4 [113], the tunneling current through oxynitride film can be readily obtained. The comparisons between the simulated tunneling currents to the experimental ones were presented in [121, 122], in which good agreements have been achieved with experiments for oxynitride/oxide stacks.

Figure 4.2 shows the simulated electron and hole tunneling currents, as a function of film composition, at inversion bias of $|V_G| = 1$ V for an oxynitride film with EOT of 1.2 nm. The simulations clearly reveal that the electron tunneling current decreases monotonically with an increase in nitrogen concentration, which is due to the higher dielectric constant and thereby larger physical thickness at same EOT. However, as the nitrogen content is increased, the hole tunneling current, after an initial reduction, exhibits an increase when more nitrogen is incorporated. As a result, hole tunneling current is lowest at nitride mole fraction of x~0.4 (nitrogen concentration of 30 at.%). This can be explained by a compromise between the increase of physical thickness and the significant decrease of valence band offset ΔE_V, when nitrogen content in oxynitrides is increased. The different trend in the electron tunneling is probably due to the small decrease of conduction band offset (ΔE_C) at oxynitride/Si interface. More importantly, Fig.4.2 also demonstrates that hole tunneling current becomes larger than electron current when x > ~0.2 (15 at.%), implying that the evaluation on gate leakage of oxynitride dielectric with high nitrogen content should be made on hole tunneling current in pMOSFET. Assessed from both electron and hole tunneling currents, it is clearly seen that the optimal nitrogen concentration is ~ 30 at.% in oxynitride gate dielectric for obtaining the lowest gate leakage in CMOS devices. The same conclusion is also true for

oxynitride/oxide stack structures [121, 122]. The above phenomenon has been experimentally confirmed in a recent study [123]. In subsequent discussions, this optimal nitrogen content will be used for assessing the scalability of oxynitride dielectric.

4.2.2. Tunneling Current through Al_2O_3 Stacks

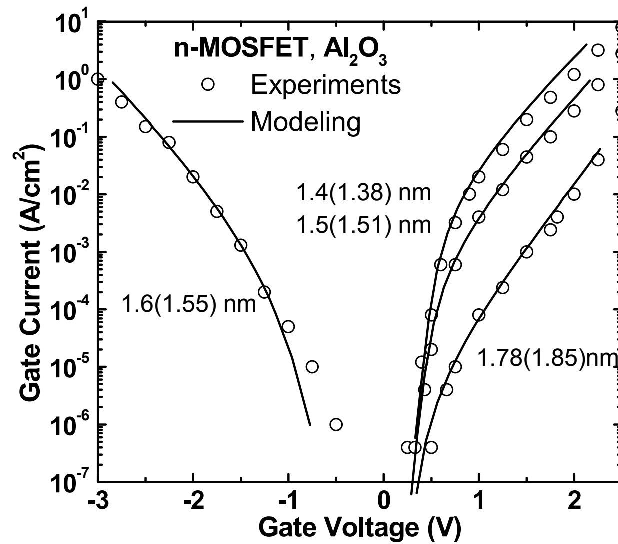


***Fig.4.3**: Simulated electron tunneling currents of nMOSFETs with Al_2O_3 gate dielectric. The experimental data are from [125, 126]. The tunneling effective mass is found to be 0.28 m_0 from overall fitting of all the data. The thickness values from best fitting to the measured data match well with those in [125, 126] from C-V method (in parenthesis).*

As an alternative gate dielectric, Al_2O_3 has many favourable properties, including a large band gap (~ 8 eV), thermodynamic stability on Si and resistance to crystallization at high temperatures. Good Electrical results, including low leakage current, have been reported for Al_2O_3 dielectric [124-126]. The scalability of EOT down

to below 1 nm was also achieved using Al_2O_3 [124]. Short channel transistors with 80 nm physical gate length using ALD Al_2O_3 have been fabricated by the standard CMOS processing technology and promising electrical results obtained [125, 126].

Despite the considerable efforts on material and device studies, few works were done on the simulation of tunneling currents through Al_2O_3 gate dielectric [127]. In this section, we present the simulation results and the comparisons with experimental data of ALD Al_2O_3/SiO_2 stacks. In the simulations, the conduction band offset between Al_2O_3 and Si was taken as 2.24 eV, measured from high resolution X-ray photoelectron spectroscopy (XPS) method [128-130]. A SiO_2 interfacial layer of 7.5 Å, which is determined from experiments [125, 126], is also included in the model. As seen in Fig.4.3, a single value of electron effective mass in Al_2O_3 m_{AlO} (0.28 m_0) is able to give the best fitting to the four experimental curves with different EOTs and bias polarities. This effective mass value is comparable to that determined from independent band structure calculations of α–Al_2O_3 (0.35 m_0) [131].

4.3. Direct Tunneling through HfO_2 and HfAlO Gate Stacks

In recent years, HfO_2 attracted much research attention because of its superior electrical characteristics, including high dielectric constant (>20), large band gap (> 5 eV), reasonable high band offset (>1.5 eV) [132]. Its compatibility to poly-Si gate processing has also been demonstrated. More encouragingly, scaling of EOT down to below 1 nm has been achieved for HfO_2 by several groups [133-135]. It is, therefore, extremely important to study tunneling currents through HfO_2 as well as their scalability in CMOS technology. However, physical modeling of tunneling current through HfO_2 gate stack is far from sufficient. Earlier simulations on tunneling through high-K stacks lack of direct comparison with the experimental data over a wide range [53, 136]. Particularly for HfO_2, the most reliable electrical data are available only in recent years. Q. Lu et al [137] and Y.C. Yeo [127] reported the modeling of tunneling currents through HfO_2 using an empirical model, however, the critically important interfacial layer between high-K and Si substrate is not considered in their model. As a result, in their modeling, HfO_2 stack is regarded as an equivalent single dielectric layer. In fact, the formation of an interface layer is inevitable during the high-K deposition. To aggravate the problem, high temperature post deposition annealing, which is generally necessary for obtaining a high quality film, always leads to the further growth of the interfacial layer [138]. It is commonly observed that the interfacial layer has much lower dielectric constant than the high-K film [138], which, in turn, has an effect of increasing the tunneling leakage [139]. As a result, the interface layer effect must be considered in order to extract accurately the tunneling parameters of high-K materials. In this section, we apply the physical model to the simulations of

tunneling currents through HfO_2 gate stack. Comparisons to experimental electrical data were made and the effective mass values in HfO_2 determined.

The simulation results were compared with experimental data measured from devices with ultra-thin HfO_2 gate dielectrics deposited by chemical vapour deposition (CVD) or physical vapour deposition (PVD) technique. All samples were with poly-Si gate. More details on the processing and device characteristics can be found in the corresponding references. Only current-voltage (*I-V*) data from ultrathin films were used, where tunneling dominates the current conduction. The assumed tunneling mechanism is also supported by the observed very small temperature dependence of the gate current. The EOT and other device parameters, such as channel doping concentrations and flat-band voltages, were extracted from C-V measurements.

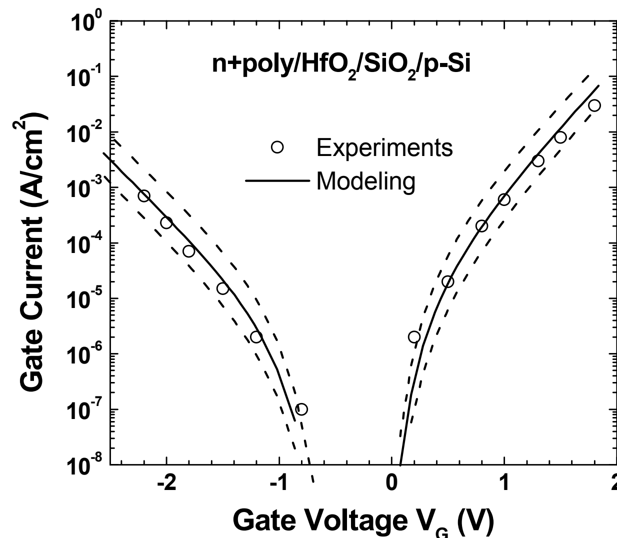


Fig.4.4: *Simulated gate current of a n+ poly/HfO_2 /SiO_2/p-Si device. The measured data are from [141], the physical thickness are HfO_2(38Å)/IL(6Å) from HRTEM and interface layer (IL is likely SiO_2 from XPS [141]. The fitted tunneling mass m_{HfO} is 0.18 m_0. When the uncertainty of HRTEM is 1 Å, the resulted m_{HfO} error is ±0.02 m_0. The dashed lines are simulations with m_{HfO} = 0.20 and 0.16 m_0.*

In the simulations, the band offset values were determined independently by high resolution XPS measurements. From XPS, valence band offsets ΔE_V between dielectric and Si is determined from the valence band XPS and the energy gap E_G by the O 1s energy loss spectra [128]. With the known Si energy gap 1.12 eV, the conduction band offsets ΔE_C can be readily obtained. The ΔE_C thus obtained is 1.9 eV for HfO_2, which is consistent with that determined from photon emission measurements (2.0 eV) [140]. This band offset value will be used for HfO_2 in the following simulations.

First, we focus on electron tunneling currents in nMOSFETs. The contributions to the gate current from valence band electron and hole tunneling are neglected due to their larger tunneling barriers compared to the conduction band electron barrier for HfO_2/Si [128, 132 140]. Figure 4.4 shows the comparison between simulations and experimental data of PVD HfO_2 on p-Si from [141]. The physical thicknesses of PVD HfO_2 (38 Å) and interfacial layer (IL) (6 Å) used in the simulations were determined by HRTEM [141]. The IL is found to be SiO_2-like by XPS [141, 142]. At $V_G > 0$, the gate current mainly consists of electron tunneling from channel inversion layer to gate. Electrons in n+ poly-Si gate accumulation layer tunneling to channel form the gate current when $V_G < 0$. The only fitting parameter in our simulations is the electron tunneling effective mass in HfO_2, m_{HfO}, and a value of 0.18 m_0 fits the experiments well in both the inversion and accumulation polarities. Considering the uncertainty of HRTEM as 1 Å in the measured thickness values of both HfO_2 and IL, the resulting error of m_{HfO} is ±0.02 m_0. This m_{HfO} value (0.18 m_0) is larger than that reported by other authors (0.1 m_0) [137], where the IL was neglected in the modeling.

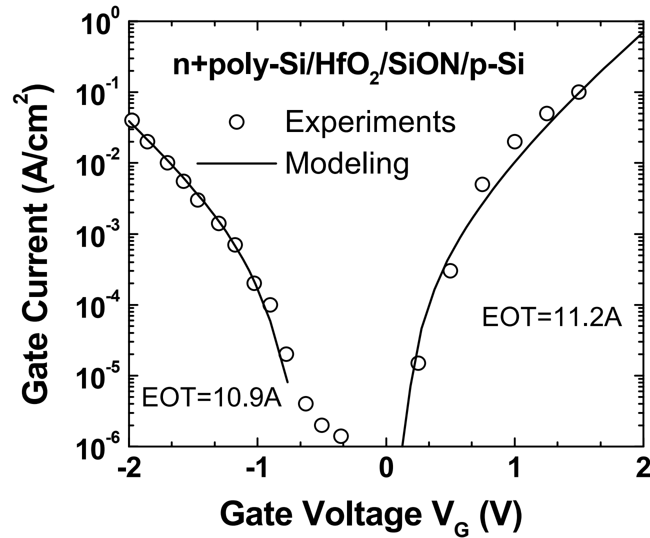


Fig.4.5: *Simulated tunneling currents of n+ poly-Si/HfO_2 on NH_3 nitrided p-Si. The $V_G > 0$ data is from [137]. The interfacial layer (IL) is assumed as $(SiO_2)_{0.5}(Si_3N_4)_{0.5}$. Using m_{HfO}=0.18 m_0 and K=22 for HfO_2, the effective IL physical thickness can be determined and their values are 6.5 Å and 9 Å, respectively.*

Using the same electron tunneling effective mass of HfO_2, we further compared the simulations with the experimental data for CVD and PVD HfO_2 on HN_3 nitrided p-Si in Fig.4.5. The IL is oxynitride (SiON) with corresponding barrier, dielectric constant and electron effective mass (0.50 m_0) parameters from [121]. The IL physical thickness is adjusted to fit the simulation curves to the experimental data. The obtained interfacial layer thicknesses of 6.5 Å and 9 Å are consistent with the HRTEM results, 6-7 Å [133] and 8.5 Å [137]. It is worth noting that, it is difficult or impossible to determine accurately the nitrogen concentration in the ultrathin interface layer. For simplicity, the mole fraction of Si_3N_4 in the interfacial oxynitride layer is assumed at 50% in present calculations.

As for hole tunneling in HfO_2, there are less experimental data reported in literature. It has been experimentally observed that hole tunneling is generally lower

in magnitude than electron tunneling for HfO_2 [137, 143]. In Fig.4.6, the results on pMOSFET fabricated on NH_3 nitrided n-Si are shown. Using the valence band offset value of 2.22 eV from XPS [128], the hole effective mass in HfO_2 is determined to be 0.18 m_0 by fitting the simulated hole current to the measurement. The similar value of hole effective mass as the electron one in HfO_2 obtained here is in correspondence with their similar band offset values.

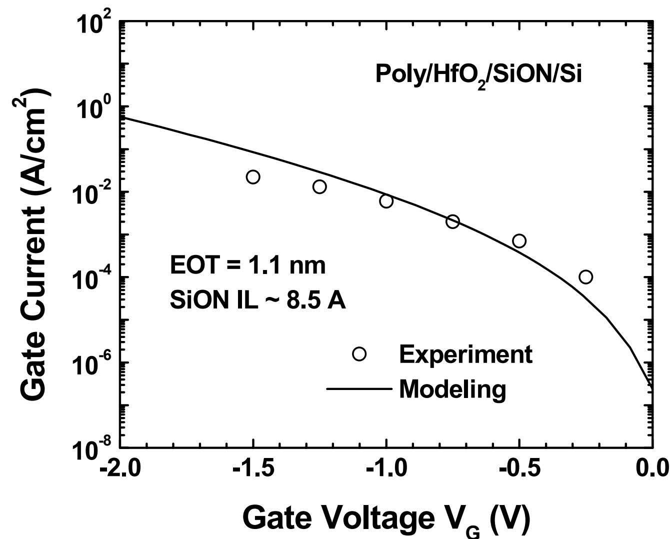


Fig.4.6: *Hole tunnelling current simulation of p+ poly-Si gate pMOSFET with HfO_2 stack. An oxynitride interface layer is also concluded in the modeling. The experimental data is from [137]. The gate stack structure is same as the nMOSFET in Fig.4.5.*

As discussed in Section 1.5, HfO_2 tends to crystallize at ~ 400 ^{0}C, which is not compatible with the high temperature processing in CMOS fabrication. As a potential solution, Al was added into HfO_2 to form HfAlO in order to improve its immunity to crystallization and oxygen diffusion [45, 46]. In order to assess the impact of Al incorporation on the gate leakage current, the tunneling current through HfAlO is also studied.

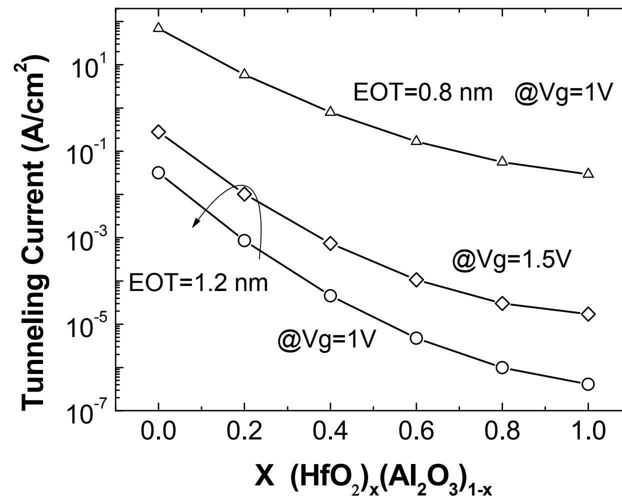


***Fig.4.7**: Calculated tunneling currents of $(HfO_2)_x(Al_2O_3)_{1-x}$ for various Hf compositions.*

For HfAlO, the band offset data were also determined from the XPS experiments [128]. Their dependences on the Hf composition are demonstrated to be in a linear relationship. Using the electron effective mass and the dielectric constant values for HfAlO linearly interpolated between those of HfO_2 and Al_2O_3, the tunneling currents through $(HfO_2)_x(Al_2O_3)_{1-x}$ is calculated and shown in Fig.4.7. It is seen from Fig.4.7 that Al incorporation increases the leakage current of HfAlO monotonically. Hence, the increase of tunneling current after Al incorporation suggests a trade-off between the increase of crystallization temperature and the degradation of gate leakage for $(HfO_2)_x(Al_2O_3)_{1-x}$. On the other hand, X-ray diffraction results indicate that a 30% mole fraction of Al incorporation into HfO_2 can raise the crystallization temperature to ~900°C [45, 46]. Therefore, Al_2O_3 mole fraction of 30% is regarded as the optimized value for HfAlO in the subsequent analysis of scalability of high-K gate stack.

4.4. Scalability of Gate Dielectrics in CMOS Technology

The scaling of CMOS device gate length requires the according scaling of other device parameters, such as gate oxide thickness, doping profile, junction depth, as well as the operating voltage. The International Technology Roadmaps for Semiconductors (ITRS) provides a consensus scenario of how these device parameters will scale for each technology generations ranging from today's 0.13 um technology to devices as small as 22 nm node in the year 2016 [1]. It is projected that EOT down to a few angstroms will be required in order to minimize the short channel effects and to obtain high device drive current. There are many requirements on such ultrathin gate dielectrics. One of them is the need to reduce the gate leakage current, since it contributes to the leakage currents and power dissipation at both operation and off states. The gate leakage requirements for the future generation CMOS are also documented in ITRS 2001, which indicates that the gate dielectric must be physically thick so that the direct tunneling is minimized. In particular, it is challenging to keep the stringent low gate leakage in low power application. It is, therefore, useful to assess the gate leakage of the potential gate dielectric candidates in future technologies. In this section, we perform the simulations on the gate leakage of several viable dielectric materials using the proposed direct tunneling model. Based on the simulated results, their scalability in future generation CMOS devices is quantitatively analyzed in terms of gate leakage.

Before the analysis of the scalability, the tunneling parameters used in the simulations of direct tunneling current are summarized in Table 4.1 for various dielectric materials. In the table, hafnium silicate ($HfSiO_4$) is also included with the parameters obtained by averaging those of HfO_2 and SiO_2, which is used for a

qualitative study. Using these parameters, the tunneling currents as a function of EOT are displayed in Fig.4.8 for the listed dielectrics.

Table 4.1: Material parameters used in tunneling simulations. $\Delta E_C/\Delta E_V$ (eV): conduction/valence band offsets with Si; m_e/m_h (m_0): electron/hole effective mass in dielectric, and K dielectric constant.

	K	**electron conduction band**		**hole valence band**	
		ΔE_C (eV)	**m_e (m_0)**	**ΔE_V (eV)**	**m_h (m_0)**
SiO_2	3.9	3.15	0.50	4.5	0.41
Si_3N_4	7.8	2.10	0.50	1.90	0.41
SiON	5.08	2.63	0.50	3.2	0.41
Al_2O_3	11	2.24	0.28	--	---
$HfSiO_4$	13	2.0	0.34	--	--
HfO_2	22	2.0	0.18	2.24	0.18

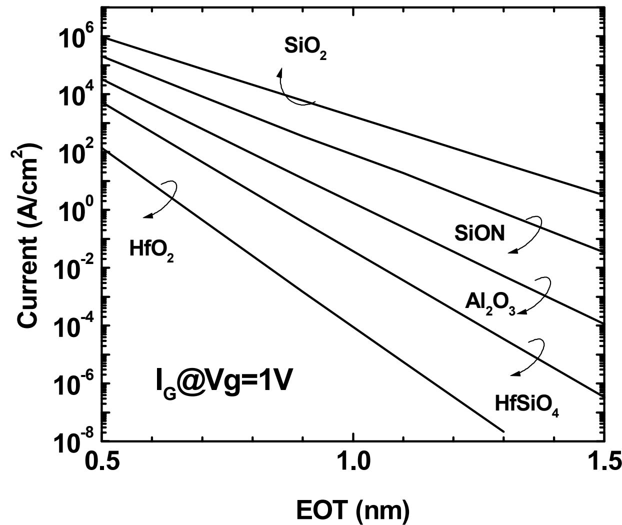


Fig.4.8: *Simulated tunneling current of MOSFET versus EOT for various gate dielectrics. The substrate doping is 10^{18} cm^{-3}. Al_2O_3 mole fraction is 30% for HfAlO and Si_3N_4 mole fraction 40% for optimized SiON. For SiON, hole tunneling in pMOSFET, which determines the scalability, is shown. Electron tunneling in MOSFET is shown for other dielectrics.*

In the following, the gate leakage is calculated for each CMOS technology generation according to the ITRS 2001. For each generation, the gate leakage is estimated by the gate current value at $V_G = V_{DD}$ with corresponding EOT and operating voltage V_{DD} values taken from ITRS2001.

4.4.1. Scalability of Gate Dielectrics in High Performance Application

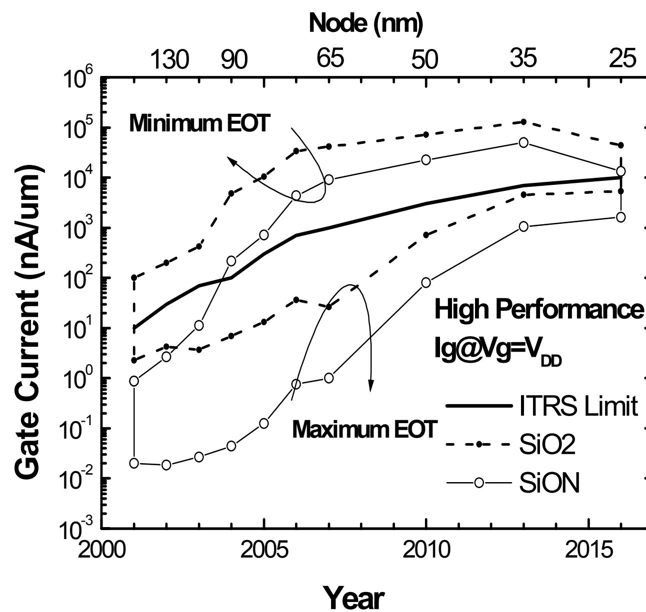


***Fig.4.9**: The calculated gate leakage of high performance CMOS. The calculated high (low) gate leakage for each generation corresponds to the minimum (maximum) EOT proposed in ITRS 2001. For oxynitride(SiON), gate leakage is from hole tunnelling in pMOSFET and the Si_3N_4 mole fraction is 40%.*

Attention is first paid to high performance CMOS and the results are shown in Fig.4.9. The calculated high (low) gate leakage for each generation corresponds to the

minimum (maximum) EOT proposed in the roadmap. The simulations indicate that, if the EOT is engineered at the maximum side of the ITRS projections, the gate leakage using even pure oxide is acceptable for all technology nodes in high performance CMOS, suggesting the applicability of conventional oxide to high performance application in terms of gate leakage. However, it is still uncertain whether oxide down to 4-5 Å in thickness is still applicable in manufacturing industry. Nevertheless, the above simulations suggest that the aggressive scaling of oxide will probably continue in the future until its physical limit, especially for high performance application. Same conclusion is also made by other authors [40].

In fact, most leading edge devices still employ oxynitride or oxynitride/oxide stack, which takes advantages of the high resistance to boron penetration. On the other hand, the nitridation also allows for a slightly thicker dielectric at the same EOT and a resultant reduction of gate leakage. To assess the scalability of oxynitride, the results for optimized SiON (x=0.4) is also shown in Fig.4.9, in which the hole current in pMOSFET is used as the scaling criterion. It shows that the optimized oxynitride can reduce the gate leakage by about 1-2 orders of magnitude, which provides more flexibility in dielectric thickness engineering for device integration. The above simulations indicate that it is not difficult to meet the gate leakage requirement in high performance application and a dielectric material with only slightly higher K value than SiO_2 is applicable. Therefore, the most difficult challenges to select a viable dielectric for high performance application seem likely to be the preservation of high channel mobility and compatibility with CMOS manufacturing. Although success of scaling alternative dielectric materials, such as Si_3N_4, Al_2O_3 and HfO_2, to below 1nm has been achieved, from manufacturing point of view, their applications in high performance application are still limited by the severe mobility degradation associated

with the poor interface to Si, high density of charges, or the incompatibility to current CMOS processing. Since SiON is probably able to meet the ITRS target from leakage current viewpoint, this gate dielectric system is expected to remain predominant in main stream manufacturing technology in foreseeable future and its aggressive scaling would be continued, particularly for high performance application [144].

4.4.2 Scalability of Gate Dielectrics in Low Power Application

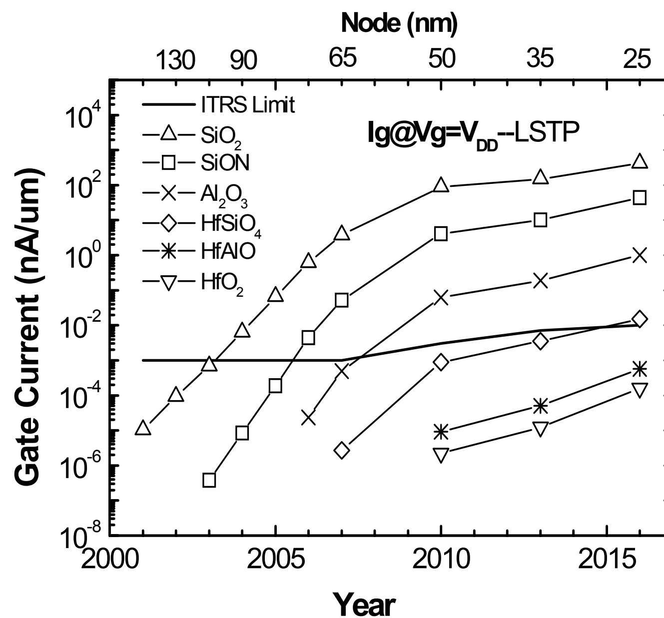


Fig.4.10: *The calculated gate leakage for low standby power (LSTP) application. Here, an average value of the proposed maximum and minimum EOT from ITRS 2001 is used for each generation.* Al_2O_3 *mole fraction is 30% for HfAlO and* Si_3N_4 *mole fraction 40% for optimized SiON.*

The difficult challenge for reducing the gate leakage is expected to be in low power application, in which low standby power (LSTP) CMOS imposes the most

stringent requirement on gate leakage. The calculated gate leakage in LSTP CMOS is displayed in Fig.4.10. It is shown that, the gate leakage of SiO_2 exceeds the gate leakage specification at as early as 90 nm node (2003). The use of oxynitride, which allows for a reduction of the gate leakage around 1-2 orders of magnitude, can only extend the scaling by about 1-2 generations. The simulations also suggest that, some medium-K dielectrics, such as Al_2O_3, can only act as a short-term solution and they are not qualified as alternative dielectrics in long term. Thus, the low power applications present the earliest driving need for high-K dielectrics.

Also shown in the figure are the simulated results on HfO_2, which is regarded as the most promising high-K candidate. The viability of HfO_2 as a long-term solution for alternative gate dielectric, which has lower gate leakage than the oxide counterpart by 4-5 orders of magnitude, is demonstrated. When Al is added into HfO_2 to solve the problem of thermal instability, despite the slightly degradation of gate leakage characteristics (< 1 order of magnitude for Al_2O_3 concentration at typical 30%), the gate leakage still remains low enough to meet the requirements of the roadmap. From the simulations, it is thus concluded that HfO_2 system is a viable candidate as a solution to the gate dielectrics scaling for future low power CMOS in long term, assessed from gate leakage prospective. From Fig.4.10, $HfSiO_4$ seems to be another choice for alternative gate dielectrics. It is thermally stable on Si and high interfacial properties with Si can be obtained. However, as a long term solution, its relatively low dielectric constant makes it crucial to eliminate the low-K interfacial layer in order to meet the stringent gate leakage requirement for low stand-by power application. Additionally, phase separation of silicate system is generally not perceived as being acceptable.

4.4.2 Interface Engineering on Gate Leakage of High-K Gate Stacks

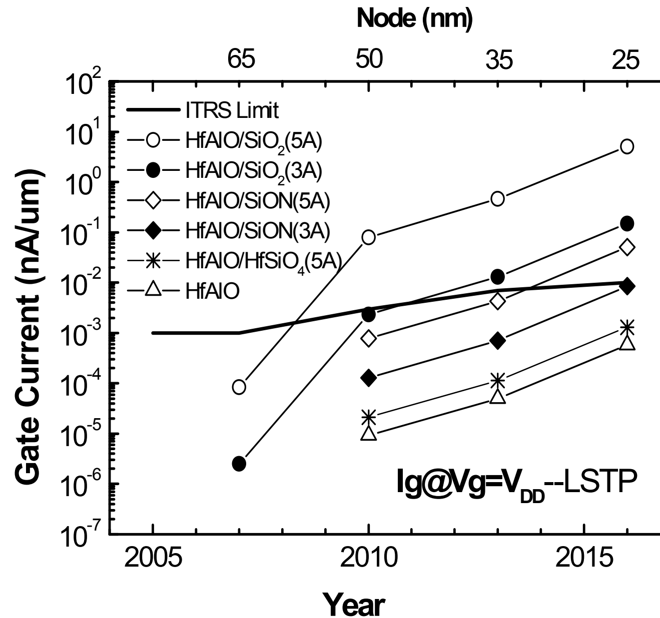


Fig.4.11*: The calculated gate leakage for low standby power applications of HfAlO dielectric stacks with different interface layers (ILs). Physical 5 Å IL of SiO_2, optimized SiON, $HfSiO_4$ were presented. A minimum 3 Å SiO_2 and SiON ILs are also shown to demonstrate the limit of SiO_2-based dielectrics as an IL layer.*

As discussed earlier, it is required to preserve high channel mobility for any alternative gate dielectric to replace SiO_2. Unfortunately, materials having a high dielectric constant always exhibit high density of charges due to the high ionic bonding. In particular, when comparing to SiO_2, the plasmon and phonon scattering in high-K are likely to be more significant due to the much lower phonon frequencies. Combined with the degraded surface roughness, the channel mobility is generally severely reduced in devices with high-K gate dielectrics [4]. One scenario to meet the

mobility requirement is to use an interfacial layer. Here, we studied the impact of several interface layer materials including SiO_2, oxynitride and hafnium silicates, on the gate leakage of high-K gate stack. They are among the available materials to date, which can form high quality interface on Si with atomic layer flatness and low density of interface states. Considering the thinnest physical thickness of the interfacial layer down to 3 Å (a single molecular layer), Fig.4.11 shows the calculated gate leakage for LSTP application using HfAlO gate stacks. If the interface layer is SiO_2, the gate leakage of the gate stack is higher by more than 2 orders of magnitude than the single HfAlO film, which makes it difficult to meet the gate leakage requirements of the roadmap. This indicates the necessity to increase the dielectric constant of the interface layer. From our simulations, either oxynitride or silicates is justified to be the candidates for such interface engineering. Considering the easiness of oxynitride in manufacturing, HfO_2/SiON stack structure is likely to be the most promising candidate as alternative gate dielectrics scalable to 22 nm node LSTP applications by 2016.

4.5. Conclusion

We reported the modeling of tunneling currents through high-K gate stacks using a physically based model and its application for analyzing the scalability of high-K materials in future CMOS technology. In the modeling, the ultrathin interfacial layer between high-K and Si substrate is included. The energy band offsets of high-K to Si were determined by high resolution X-ray photoelectron spectroscopy. Excellent agreements between simulation results and measured tunneling currents have been achieved over several high-K dielectric materials, such as Si_3N_4, Al_2O_3 and HfO_2. The effective mass values in these high-K dielectrics were determined. Subsequently, the model was used to predict when high-K dielectrics are to replace SiO_2 according to 2001 ITRS for different CMOS device applications. For high performance application with high tolerance of gate leakage, the continue applicability of traditional oxynitrde dielectric is justified. It is also found that, high-K material is expected to be first required in low power application because of the stringent requirement on leakage current. In terms of gate leakage, HfO_2 or HfAlO is demonstrated to be a viable candidate for long term solution. The impact of the interfacial layer on the gate leakage of high-K stacks was further analyzed. The simulations show that the low-K interfacial layer gives rise to a significant increase of the gate leakage, making interface layer engineering critical to the scaling of high-K gate stacks. Recommendations for interface layer materials were also made for high-K gate stacks that will meet ITRS roadmap.

Chapter 5:

Metal Gate Engineering on Gate Leakage Characteristics of MOSFETs

5.1. Introduction

Aggressive scaling of gate length and gate oxide thickness in CMOS transistors aggravates the problems of poly-Si gate depletion, high gate resistance and boron penetration from the p+-doped poly-Si gate into the channel region. The poly-Si depletion reduces the gate capacitance in the inversion regime and hence the inversion charge density, or leads to a lower gate over drive, thus degrades the device performance. As a result, metal gate technology is recently exploited. The replacement of poly-Si gate with metal not only eliminates the gate depletion and dopant penetration, but also greatly reduces the gate sheet resistance. For bulk CMOS, metal gates with work functions corresponding to the conduction band and valence band edges of Si are preferred for the optimal design of n- and p-MOSFETs, respectively [54]. On the other hand, novel device architectures, such as ultrathin body (UTB) or double gate (DG) structure fabricated on silicon-on-insulator (SOI) wafers may be utilized in future CMOS technology due to their excellent scaling capability [62]. In such SOI devices, in order to improve channel carrier mobility and eliminate threshold voltage (V_T) instability induced by dopants fluctuation, low doping in body is generally desirable and metal gate with work function near mid-gap of Si is thus

required to obtain appropriate V_T [145, 146]. At the same time, the introduction of an alternative high-K material is also underway in order to suppress gate leakage [4]. Therefore, an understanding of metal gate engineering on the tunneling characteristics of MOSFETs, especially with high-K dielectric stack, is technically important and timely.

In this chapter, we present a systematic investigation on the tunneling phenomena in metal gate MOSFETs with oxide and high-K dielectrics. First, the simulated tunneling currents will be compared to measurements on TaN metal gated devices with SiO_2 and HfO_2 dielectrics in Section 5.2. Subsequently, in Section 5.3, the advantage of the metal electrode over poly-Si on the gate leakage is demonstrated in bulk CMOS due to the elimination of poly-Si depletion. It is followed by a systematic study on how the change of metal gate work function affects the various tunneling components in CMOS transistors, as presented in Section 5.4. Finally, in Section 5.5, the better scalability of SOI devices over bulk one in terms of tunneling leakage current is demonstrated for future CMOS technology.

5.2. Tunneling Currents in Metal Gate CMOS Devices

The tunneling currents through SiO_2 and high-K gate dielectric in poly-Si gated CMOS devices have been studied in Chapter 3 and 4. Relevant material parameters with tunneling, including the band offset and effective mass values, were extracted for various gate dielectrics. These parameters and the same simulation method will be used in the following calculations of tunneling currents in metal gate devices. In the simulations, free electron gas is assumed in the metal gate. The equivalent oxide thickness (EOT) is extracted from capacitance-voltage (C-V) measurements and the work function of the metal electrode is determined from the flat-band voltage (V_{FB}).

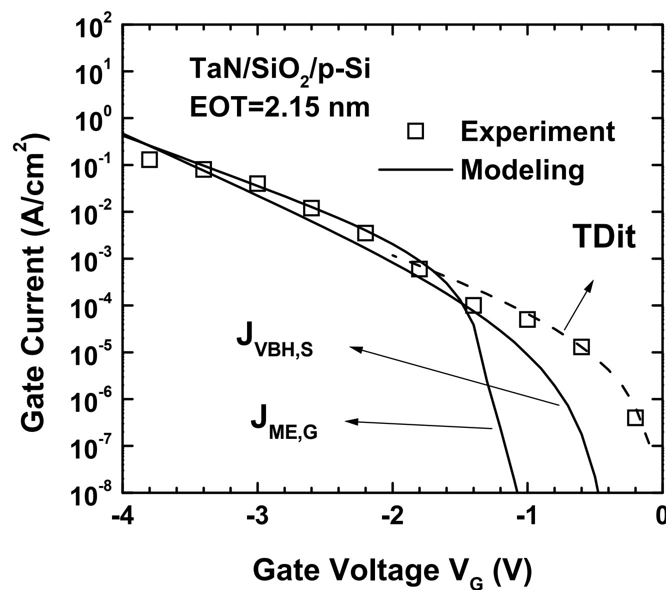


Fig.5.1*: Tunneling currents in a $TaN/SiO_2/p$-Si capacitor. The experimental data is from [56]. $J_{VBH,S}$ and $J_{ME,G}$ represents the valence band hole tunneling from substrate and metal electron tunneling from the gate, respectively.*

Figure 5.1 shows simulations on a metal-gated TaN/SiO_2 capacitor. From the calculations, for SiO_2 with the specific EOT (2.15 nm) [56], the gate current at accumulation is mainly comprised of the electron tunneling from the metal gate to the substrate. However the hole tunneling from the accumulated p-Si substrate to the gate also contributes to the gate current appreciably and it even dominates over the electron tunneling current in low voltage regime. This is different from poly-Si gated SiO_2 devices, in which hole tunneling is always negligible compared to electron tunneling at the accumulation bias of nMOS as shown in Fig.3.3. Good agreements between simulations and experiments have been obtained at strong accumulation. In gate voltage V_G region between V_{FB} and 0 V, the much higher gate leakage experimentally observed is attributed to the electron tunneling through interface state (TDit) [147], which is shown by the dashed line.

The simulations of gate currents in CMOS transistors with HfO_2 stacks are displayed in Fig.5.2. Solid lines are simulations assuming parabolic dispersion in HfO_2 with effective mass values obtained in Chapter 4. Using one fitting parameter of 8 Å oxide interfacial layer (IL), overall good agreements between simulations and experiments [148] were obtained for eight current-voltage (I-V) curves measured on both n- and p-MOSFETs with different EOTs and V_G polarities. Higher gate current values obtained from the simulations than the measurements are found at high V_G, which might arise partly from the non-parabolic effects, as discussed in Chapter 3. The dashed lines are simulations using Freeman-Dahlke dispersion for HfO_2 [101], in which second adjusting parameter of hole effective mass ($0.27m_0$) is introduced. Better agreement with experimental data at high V_G is obtained.

From Fig.5.2, higher current is observed than simulated one at low gate voltage. This excess current flow may be caused by charging and discharging of interface states and trapped charges, which is generally much higher in density in HfO_2.

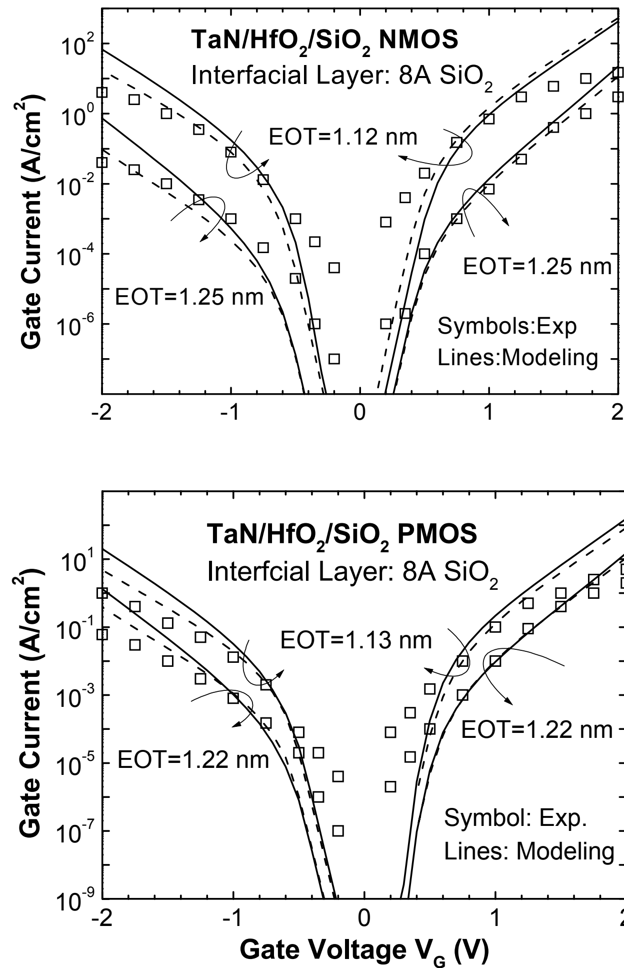


Fig.5.2: *Tunneling currents in TaN gated MOSFETs with HfO_2 stacks. Solid lines are simulations assuming parabolic dispersion in HfO_2 with effective mass values listed in Table I. Using one fitting parameter of 8 Å oxide interfacial layer (IL), overall good agreements between simulations and experiments are obtained. Better agreement with experimental data at high V_G can be obtained by using Freeman-Dahlke dispersion for HfO_2 (dashed lines).*

5.3. Reduction of Gate Leakage by Metal Gate

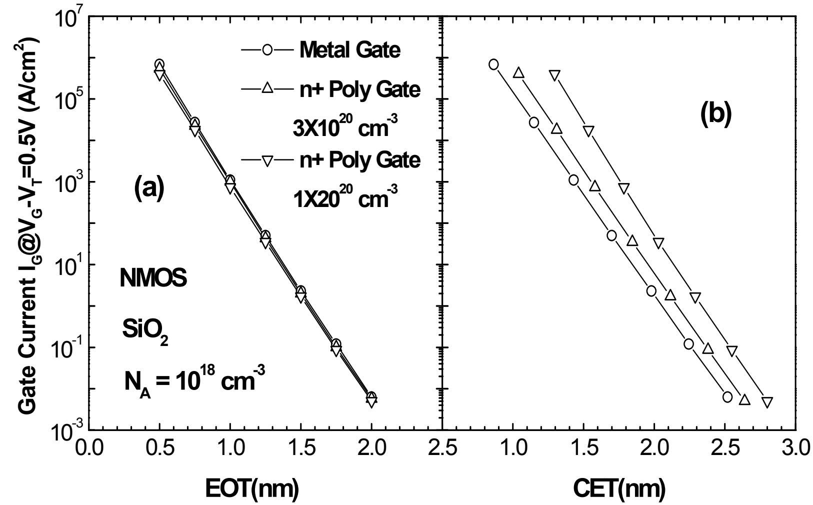


Fig.5.3: *Gate leakage of metal and poly-Si gate nMOSFET with SiO_2 gate dielectric. The metal gate work function is assumed at Si conduction band edge. The EOT is the equivalent oxide thickness, and CET the capacitance equivalent thickness at inversion (V_G-V_T = 0.5V).*

As discussed previously, for bulk CMOS, metal gates with work functions corresponding to poly-Si are desirable. In this section, we first compared the gate leakage between metal and poly-Si gated devices with the same gate work functions and the same substrate doping concentrations. Typical results in nMOSFET with SiO_2 are shown in Fig. 5.3, in which the gate leakage at the same gate overdrive (i.e. oxide field), V_G-V_T, of 0.5 V is presented. V_T is defined by the gate voltage to induce inversion charge of 10^{11} cm^{-2}. In Fig. 5.3(a), when having the same EOT, V_T and thereby V_G is higher in poly device than that in metal device due to the additional voltage drop in poly depletion layer. From the calculations, they show similar gate leakage in magnitude due to the same oxide field and oxide thickness. In Fig.5.3(b),

the gate leakage was compared at the same capacitance equivalent thickness (CET) at inversion. The same oxide field and inversion capacitance means the same gate terminal voltage. From Fig.5.3(b), the use of metal gate results in lower gate leakage than poly-Si gate. This can be explained by the elimination of gate depletion by using metal gate. In poly-Si device, poly-Si depletion leads to capacitance attenuation, and for compensation, EOT smaller than that in metal gate device is thereby required to achieve the same CET, which in turn results in higher gate tunneling current.

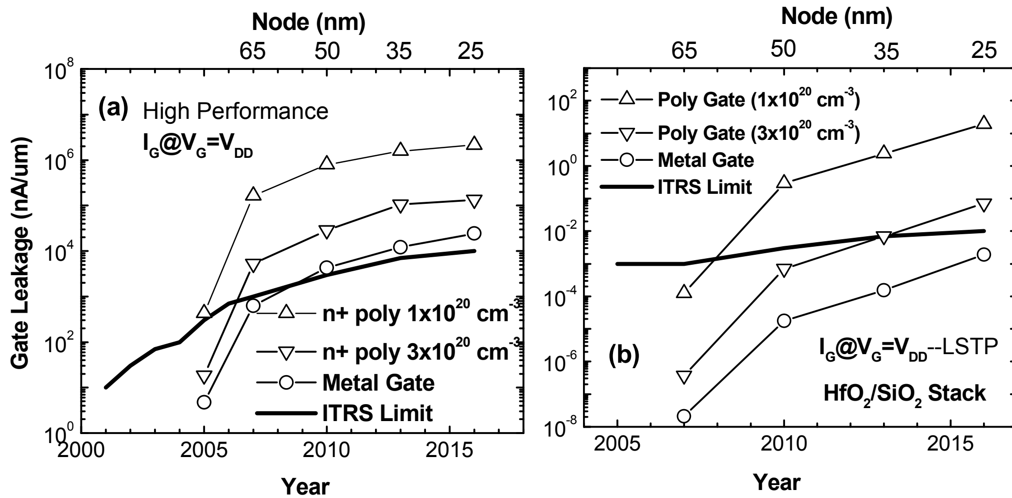


Fig.5.4: *NMOS gate leakage in future generation CMOS for (a) high performance application using SiO_2 and (b) low stand-by power application using HfO_2/SiO_2 stack with SiO_2 interfacial layer of 3Å. In the calculations, the EOTs of gate dielectrics were selected to meet the required CET by ITRS 2001.*

The advantage of metal gate over poly-Si in terms of gate leakage is also elucidated in Fig.5.4, in which the gate leakage in future CMOS is predicted by choosing the dielectric physical thickness to meet the requirements of CET from ITRS

2001 [1]. It is projected that metal gate devices can lower the gate leakage by 1-2 orders of magnitude over poly-Si ones with typical poly-Si doping density in current processing technology of ~10^{20} cm^{-3}. This reduction of gate leakage suggests the capability to scale gate dielectrics more aggressively (by additional ~2-3 Å) by employing metal gate in bulk CMOS. This benefit of suppressing gate leakage by metal gate technology might be lost with extremely high poly-Si gate doping. However, it presents great challenge since the active dopant density at poly-Si/dielectric interface saturates at ~10^{20} cm^{-3} and it is difficult to dope the poly-Si gate more than 10^{20} cm^{-3}, especially for p+ poly-Si gate [54].

5.4. Metal Gate Work Function Engineering on Tunneling Characteristics of MOSFETs

For CMOS with SOI structure, metal gate electrodes with work function deviated from poly-Si gates are possible and near mid gap metal gates are required for SOI devices with low doping body. In bulk CMOS, there also exists possibility to engineer the work functions of metal gates in some extent around the poly-Si work functions [61]. In this section, we focus on the impact of the metal gate work function on the tunneling currents in metal gated MOSFETs. For such studies, a double gate (DG) structure is selected as a typical example. The body thickness is assumed at 20 nm and body doping concentration 10^{16} cm^{-3}. Although some minor difference may exist, the following results obtained on DG structure are also applicable to other device architectures, such as bulk CMOS or ultra-thin body SOI devices.

5.4.1 Gate to Channel Tunneling

The tunneling in channel area is first studied, and the tunneling mechanisms are illustrated by the band diagrams in Fig.5.5. Throughout this chapter, the tunneling components are labelled by the type of the tunneling carrier (CBE: conduction band electron; VBE: valence band electron; VBH: valence band hole; ME: metal gate electron) and the electrode supplying the carriers (G: gate and S: substrate electrode).

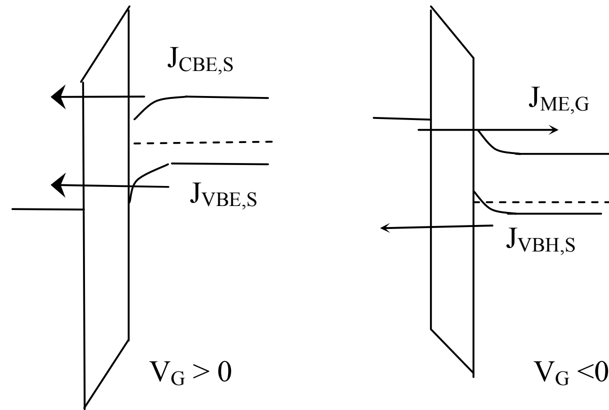


Fig.5.5: *Band diagram schematics of tunneling in channel area of nMOSFET. It is similar for pMOSFET except the substrate Fermi energy. Labels: CBE: conduction band electron; VBE: valence band electron; VBH: valence band hole; ME: metal gate electron; G: gate and S: substrate electrode.*

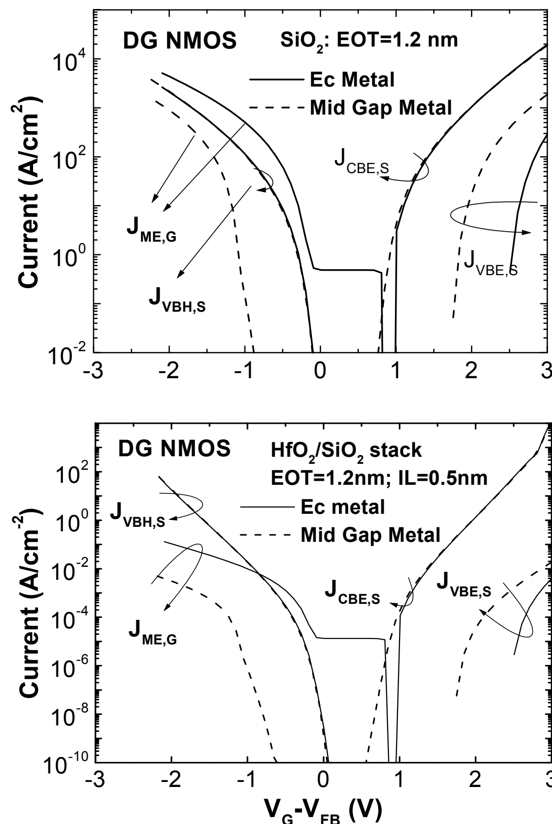


Fig.5.6: *Tunneling currents of metal double gate (DG) nMOSFET with (a) SiO_2 and (b) HfO_2 stack as a function of V_G-V_{FB}. Solid lines are of metal work function Φ_B at Si conduction band edge (E_C metal), while dashed lines are those with mid-gap metal. The band diagrams for various tunneling components are shown in Fig.5.5.*

Fig.5.6 plots the simulated tunneling currents in channel area of metal gated nMOSFETs with metal work function at Si conduction band edge (E_C metal) and mid-gap of Si. From the comparisons, the gate current at inversion, which is dominated by $J_{CBE,S}$ and $J_{VBH,S}$ in n- and p-MOSFETs, respectively, is independent of metal work function (Φ_B) after deducting the flat-band voltage V_{FB} shift caused by Φ_B variation. However, the increase of Φ_B will reduce metal electron tunneling $J_{ME,G}$ due to the decrease of electron tunneling barrier. The most significant effect is observed in nMOSFET when $V_G - V_{FB} < 0$. As shown in Fig.5.6, when metal Fermi energy changes from Si conduction band (E_C) to mid-gap, $J_{ME,G}$ changes from larger to smaller than $J_{VBH,S}$, the hole tunneling current from substrate to the metal gate. As a result, the total gate current, as elucidated in Fig.5.7, is first decreased due to the reduced $J_{ME,G}$, then becomes independent of Φ_B after $J_{VBH,S}$ dominates over $J_{ME,G}$. It is also found that the gate current in nMOSFET with high-K dielectric has a less dependency on Φ_B than that with SiO_2.

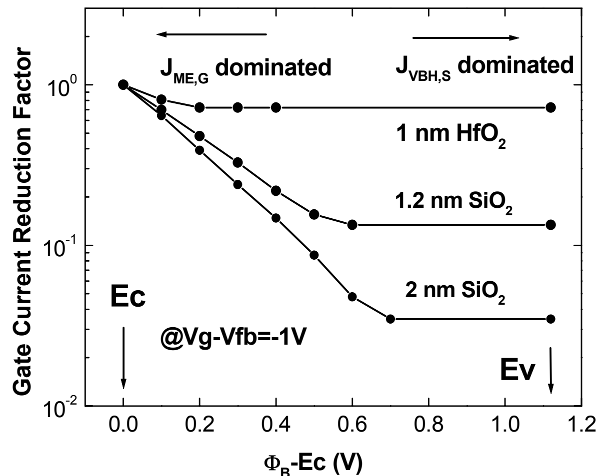


***Fig.5.7**: Gate current of nMOSFET at V_G-V_{FB}=-1V as a function of metal work function Φ_B.*

The results in pMOSFETs are shown in Fig.5.8. Effects of Φ_B variation on various tunneling components are similar as those observed in nMOSFETs. One noticeable characteristic is that the valence band electron tunneling $J_{VBE,S}$ at accumulation bias, which usually contributes negligibly to the gate current in p+ poly-Si gate pMOSFET, becomes comparable in magnitude with the gate electron tunneling for SiO_2 gate dielectric when metal work function is at Si valence band edge (E_V metal) [149].

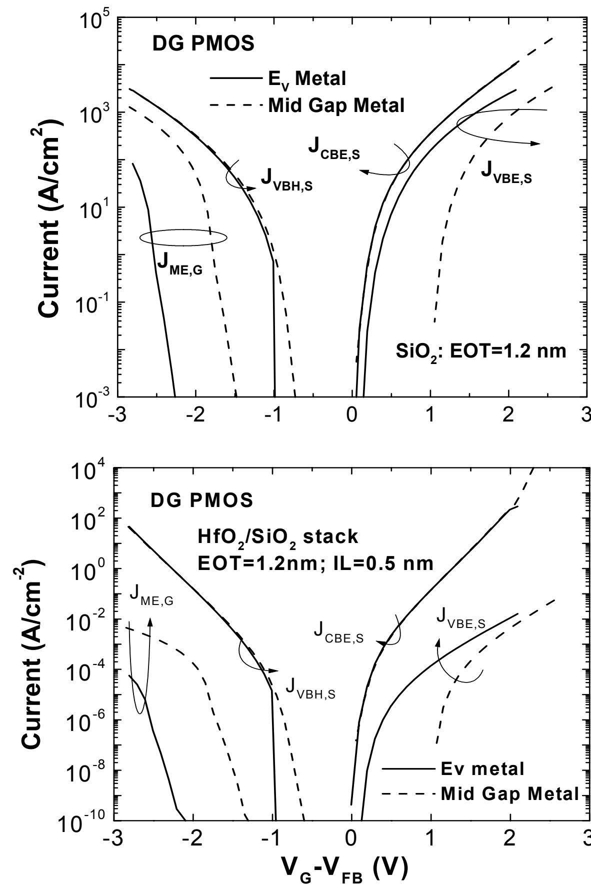


Fig.5.8: *Same as Fig.5.7 but for metal double gate (DG) pMOSFET. E_V metal means metal work function Φ_B is at Si valence band edge.*

5.4.2 Gate to Source/Drain Extension (SDE) Tunneling

In the following, we will study the tunneling between the gate and the SDE overlap region (J_{SDE}). As discussed in Section 3.2.4, SDE dimension is not so scalable as channel and the contribution of gate to SDE overlap tunneling (J_{SDE}) to circuit leakage current becomes increasingly important as channel gate length is scaled down. Tunneling in SDE has been verified even to be the dominant source of off-state leakage in MOSFETs with ultrathin gate dielectric [90-93, 150].

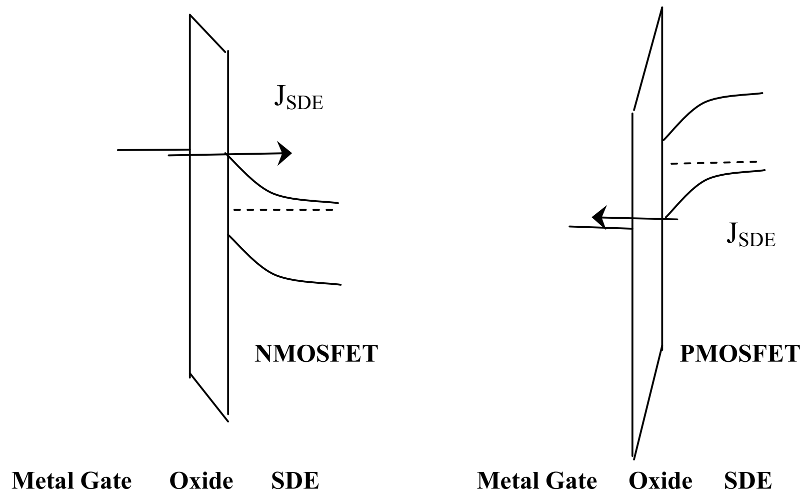


Fig.5.9: *Schematics of band diagrams of gate to source-drain extension (SDE) tunneling (J_{SDE}) at accumulation bias. In SDE region, the metal electrons tunnel to n+ SDE in nMOSFET and valence band electrons tunnel from p+ SDE to metal gate in pMOSFET.*

The band diagrams of gate to SDE tunneling are illustrated in Fig.5.9 for metal gate CMOSFETs. Compared with those in poly-Si gate devices, gate to SDE tunneling in metal gate devices shows different mechanism for pMOSFET. In metal gate pMOSFET, tunneling in SDE comes from valence band electron tunneling from

p+ SDE to metal gate. The hole tunneling current, which dominates the SDE tunneling in p+ poly-Si gated p-MOSFET, cannot occur in the presence of metal gate electrode. In nMOSFET, metal electron tunneling to n+ SDE forms the gate to SDE tunneling.

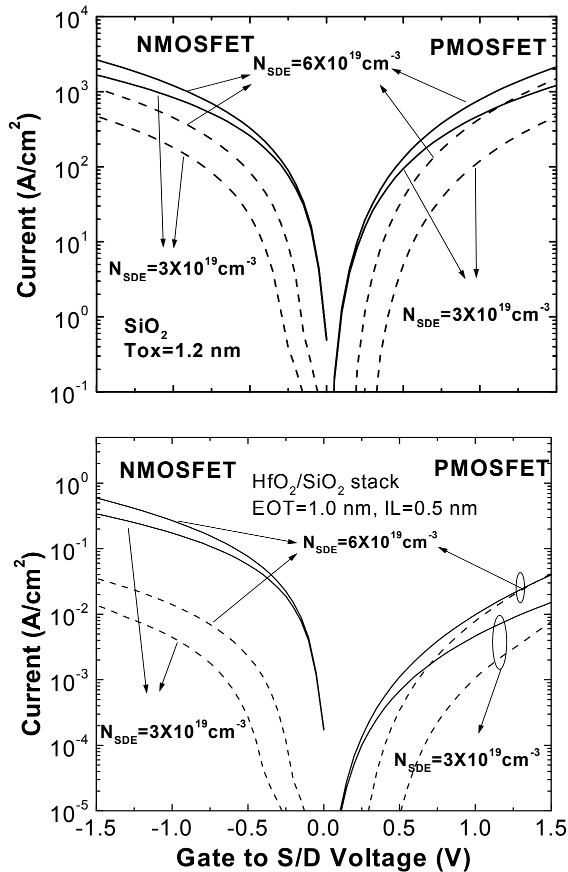


Fig.5.10: *Gate to source-drain extension (SDE) tunneling in metal double gate (DG) MOSFETs with SiO_2 and HfO_2 stack. Solid lines are those with E_C metal (nMOS) and E_V metal (pMOS) gates, while dashed lines are those using mid-gap metal gates.*

Figure 5.10 compares the gate to SDE tunneling with different metal gate work functions. It is found that J_{SDE} is sensitive to Φ_B and the use of mid gap metal gates reduces the J_{SDE} in both n- and p-MOSFETs. J_{SDE} is also found to have a dependence

of SDE doping concentration. Higher doping in SDE leads to higher magnitude of J_{SDE} because of the higher oxide field resulting from less voltage drop in SDE. This indicates that a trade-off between leakage current and SDE resistance should be made for device optimization.

The dependence of J_{SDE} on Φ_B is summarized in Fig.5.11. J_{SDE} in nMOSFET is always higher than that in pMOSFET in the whole rang of Φ_B, indicating that the leakage current limit is first reached by nMOSFET. It is also found that increasing Φ_B reduces J_{SDE} significantly and the reduction is further enhanced when using high-K dielectric. This decreased J_{SDE} is expected to have an effect of reducing the MOSFET off-sate leakage current.

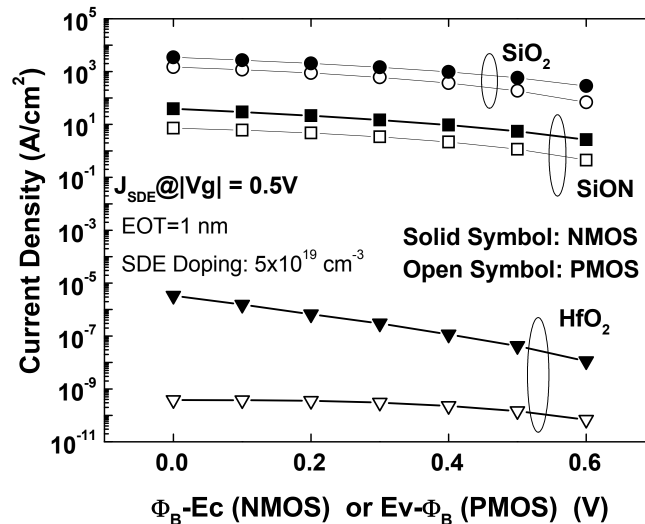


Fig.5.11: *The effect of metal gate work function Φ_B on the gate to source-drain extension (SDE) tunneling J_{SDE} for various gate dielectrics*

5.4.3. Advantage of Metal Double Gate MOSFETs on Leakage Current

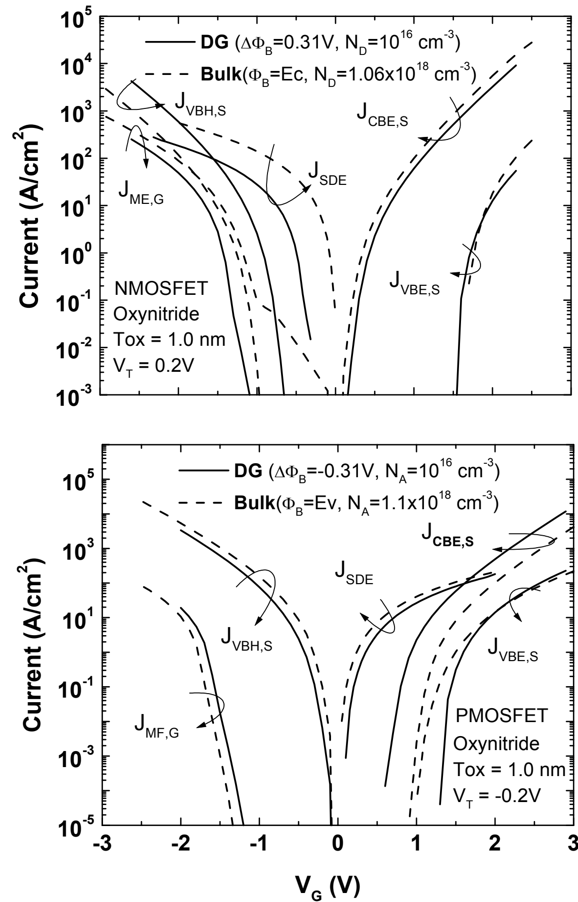


Fig.5.12*: Comparison of metal double gate (DG) and bulk MOSFETs with oxynitride gate dielectrics at the same threshold voltage V_T (defined by inversion charge of 10^{11} cm^{-2}). In DG MOSFET, low body doping is assumed and V_T is adjusted by metal gate work function Φ_B while in bulk MOSFET, V_T is tuned by channel doping.*

Figure 5.12 compares tunneling currents in DG and bulk MOSFETs with the same threshold voltage V_T of 0.2 V. In DG device, the ultra-thin body provides the well control for short channel effects and thereby low body doping is assumed. The

V_T is achieved by metal gate Φ_B. However, in bulk devices, the substrate cannot be undoped due to short channel effects and work function must locate at Si conduction band edge for optimal design. Therefore, V_T adjustment can be only be achieved by channel doping in bulk device. For MOSFET on state, DG shows slightly lower gate leakage, which is explained by the electric field lowering due to low body doping [69]. Compared to its bulk counterpart, DG MOSFET exhibits significant advantage of the reduced J_{SDE}, which is due to the adjustment of metal Φ_B to near the mid-gap of Si. For nMOSFET, which is the limiting case as demonstrated in Fig.5.11, J_{SDE} is ~1 order of magnitude lower for DG than for bulk MOSFET, demonstrating the advantage of DG structure on suppressing the off –state leakage induced by gate to SDE tunneling.

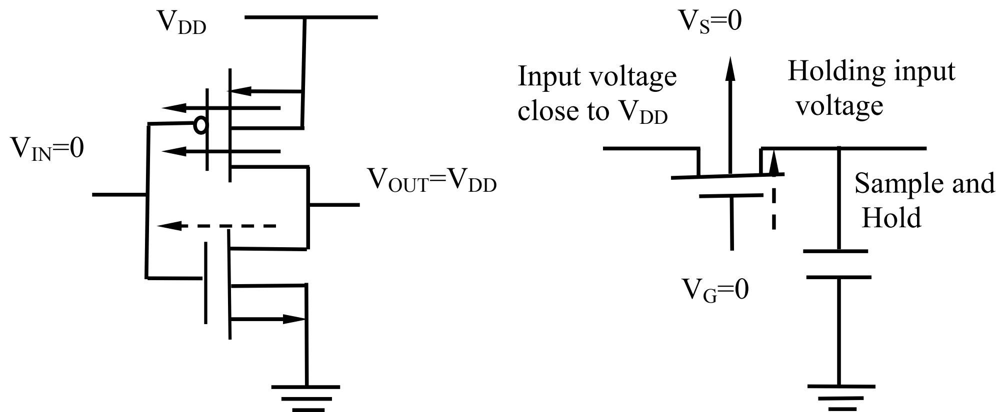


Fig.5.13*: Circuits of an inverter (left) and a sample/hold with an nMOSFET switch (right).*

The leakage current impacts adversely on circuit performance [151]. Examples of an inverter in digital circuit and a sample/hold in analog circuit are illustrated in

Fig.5.13. For the inverter at high output state, the input gate current is comprised of pMOSFET gate-channel leakage at inversion and off-state leakage current of nMOSFET. When the device is scaled down, J_{SDE} may dominate this off-state leakage [90-93, 150]. J_{SDE} can be changed by changing Φ_B of the nMOSFET metal gate. For a sample/hold, when nMOSFET is off to hold a high level signal close to supply voltage V_{DD}, both $J_{ME,G}$ and J_{SDE} will charge or discharge the hold capacitor, degrading the accuracy of the sample/hold. Similar discussion can be applied to pMOSFET in inverter and sampling/hold. The reduced tunneling in DG SOI devices at both on and off states may benefit to lower the standby power of an inverter in digital circuit and improve the accuracy of sample/hold in analog circuit.

5.5. Scalability of Metal Gate Advanced MOSFETs

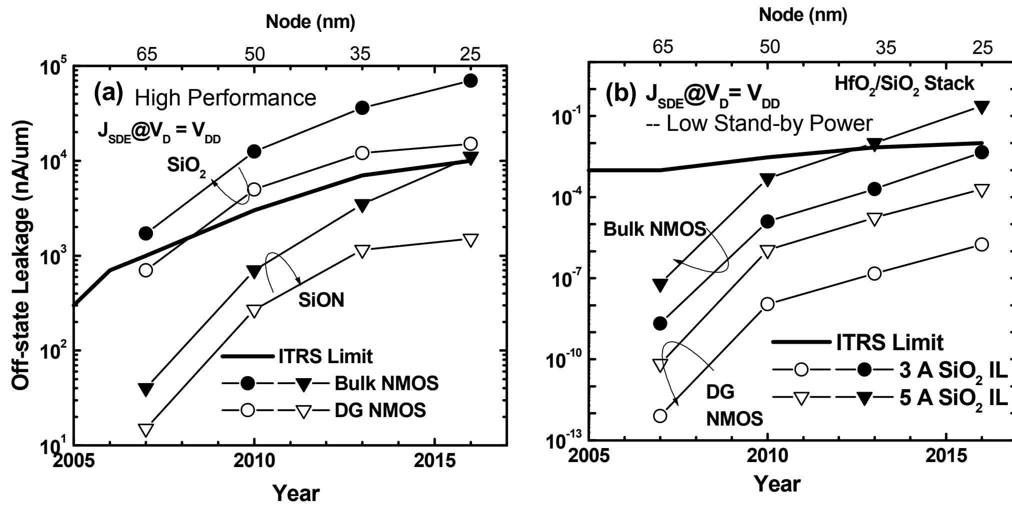


Fig.5.14*: Off-state leakage of metal gate nMOSFET contributed by gate to source-drain extension (SDE) tunneling estimated using 5 nm SDE dimension for (a) high performance application using SiO_2 and SiON, (b) low power application using HfO_2/SiO_2 stack.*

To study the effects of metal gate work function engineering on the device scalability, the off-state leakage (I_{OFF}) contributed by J_{SDE} in future CMOS technology is calculated and presented in Fig.5.14 for both bulk and DG SOI nMOSFETs with metal gates. In the calculations, values of EOT, threshold voltage V_T and operating voltage V_{DD} from ITRS2001 were used for each generation. In DG, low body doping is assumed and metal gate work function is determined by V_T, while, in bulk MOSFET, metal work function is at Si conduction band edge and uniform channel doping is used for V_T adjustment. From the simulated I_{OFF}, metal DG SOI MOSFET demonstrates its potential in suppressing J_{SDE}. This reduction of J_{SDE} is expected to be as much as 2-3 orders of magnitude in low power application when high-K is employed, suggesting the superior scaling capability from leakage perspective by utilizing metal gate DG SOI structure.

5.6. Conclusion

In this chapter, we presented a systematic study of tunneling leakage current in metal gate MOSFETs and how it is affected by the work functions of the metal gate electrodes. The physical model used for simulations was corroborated by experimental results from SiO_2 and HfO_2 gate dielectrics with TaN electrodes. In bulk MOSFET, results show that, at the same CET (capacitance equivalent oxide thickness at inversion), replacing poly-Si gate by metal can reduce gate leakage appreciably by 1-2 orders of magnitude due to the elimination of poly-Si gate depletion. It is also found that, the work function of metal gate affects the various tunneling components in MOSFETs. Specifically, increase of metal work function reduces gate to channel leakage in off-biased nMOSFET and the use of mid-gap metal gate results in significant reduction of gate to source/drain extension tunneling in both n- and p-MOSFETs. Comparing to bulk MOSFET, SOI MOSFET has much lower off-state leakage due to the smaller gate to source/drain extension tunneling. The reduction of off-state leakage can be as much as three orders of magnitude when high-K gate dielectric is used. Finally, the benefits of employing metal gate SOI structure in circuit applications and in future CMOS scaling were demonstrated.

Chapter 6

Conclusions and Recommendations

6.1. Conclusions

6.1.1. Hole Quantization in CMOS Devices

In this thesis, we demonstrated the importance of valence band mixing to the hole quantization in CMOS devices. Compared with the calculations by using the six-band effective mass approximation, the traditional one-band effective mass approximation, which used the bulk Si effective masses, underestimates the subband density of states and thereby overestimates the threshold voltage shifts in pMOSFET caused by hole quantization.

Based on the numerical results of the six-band effective mass approximation, we also demonstrated an improved one-band effective mass approximation method for hole quantization in pMOSFET. In this method, the hole quantization mass and DOS mass values were derived from the numerical results from six-band calculations. Further, the triangular well approximation was successfully extended to strong inversion or accumulation region by introducing an effective electric field with a weighting coefficient, which is determined from the numerical results of rigorous six-band self-consistent model. This method shows significant advantages in modeling the hole quantization:

(1) This improved one-band effective mass method is same in formalism as the conventional one, so it preserves the efficiency in computation. It is also easy to be

implemented into routine device simulators by just replacing the corresponding effective mass values.

(2) The accuracy of this method is ensured by extracting the new effective masses from numerical results of rigorous six-band approach. Using this model, the calculated hole inversion capacitance and threshold voltage shifts are in good agreements with both the six-band approach and the experiments.

(3) After establishing the effective electric field, this method in triangular well approximation is applicable for obtaining the MOS electrostatics at both inversion and accumulation biases with sufficient accuracy.

6.1.2. Direct Tunneling Currents through Ultra-thin Gate Dielectrics

In this thesis, studies of direct tunneling currents through ultrathin gate dielectrics consist of three major parts:

(1) Direct tunneling current through gate oxide in CMOS devices;

(2) Direct tunneling and scalability studies of high-K gate dielectrics;

(3) Direct tunneling in metal gated MOSFETs and the impact of metal work function on the gate leakage characteristics in advanced MOSFETs.

From these studies, the proposed direct tunneling model is demonstrated for SiO_2, Si_3N_4, Al_2O_3 and HfO_2. Using tunneling current simulations, the scalability of these gate dielectric materials in future CMOS technology was analyzed. Guidelines for the selection of high-K dielectrics were also provided.

(I) Direct Tunneling Currents through Ultra-thin Gate Oxide in CMOS Devices

The direct tunneling currents through gate oxide in state-of-art CMOS devices were simulated using an efficient physical model. In the direct tunneling model, the carrier quantization is properly treated, especially for hole quantization, in which the improved one-band effective mass approximation was employed to account for the important valence band mixing effect. The tunneling probability was calculated by a modified Wentzel-Kramers-Brilliouin (WKB) approximation with the reflections at oxide/Si interfaces being taken into account.

(a) Non-parabolic dispersion effect in hole tunneling current

Hole tunneling currents were studied intensively in pMOSFETs with gate oxide thickness ranging from 1.35 to 2.7 nm. Significant deviations from the experiments were observed by using the traditional parabolic dispersion in oxide energy gap when oxide thickness is thicker than 2 nm, which has been explained by the non-parabolic hole dispersion. A Freeman-Dahlke dispersion form is found more appropriate for describing the hole dispersion in oxide energy gap. After taking the difference of conduction and valence band effective masses into account by using the Freeman-Dahlke form, the agreement of the simulated hole tunneling currents with the experimental data is significantly improved over a wide range of oxide thickness and gate voltage.

(b) Modeling of all terminal tunneling currents in CMOS transistors

We demonstrated the proposed direct tunneling model by the successful modeling of all terminal tunneling currents in CMOS transistors. The validity of the

model is verified by simulated results consistent with experiments for electron tunneling by using parabolic dispersion relationship in the oxide energy gap. When oxide is thinner than 2 nm, the applicability of the parabolic approximation to hole tunneling is also demonstrated. Using the simple parabolic dispersion, the successful modeling of all terminal tunneling currents through gate oxide, including conduction band electrons, valance band electrons and valence band holes in CMOS devices, were made by the proposed model.

(II) Tunneling Current and Scalability of High-K Gate Dielectrics in CMOS Technology

The modeling of tunneling current through alternative high-K gate stack using the physically based model has been done for several high-K candidates, such as Si_3N_4, Al_2O_3 and HfO_2. For high-K materials, an interfacial layer is always formed during deposition or post deposition annealing. The ultrathin interfacial layer between high-K and Si substrate was considered properly in our calculations, which ensures the reliability of the modeling results. Excellent agreements between simulation results and measured tunneling currents were achieved. Based on the simulation results on gate leakage and the 2001 ITRS, the scalability of these gate dielectrics was analyzed for different CMOS device applications.

(a) The traditional nitrided oxide (oxynitride) used in current manufacturing shows high magnitude of gate leakage, however, it still can meet the leakage current requirement of the recent roadmap for high performance application. Due to the high

tolerance of gate leakage in high performance CMOS devices, it is expected that the continued aggressive scaling of the oxynitride dielectric is demanded by manufacturing industry.

(b) In terms of gate leakage current, an alternative high-K material is expected to be first required in low power application. For low stand-by power application, which represents the most stringent requirement on leakage current, Si_3N_4 and Al_2O_3 with medium K values can act only as a near term solution. The newly studied HfO_2, or more thermally stable HfAlO, exhibits low gate leakage and demonstrates to be a viable high-K gate dielectric in long term until the year of 2016.

(c) The impact of the interfacial layer on the gate leakage of high-K stacks was analyzed. The interfacial layer generally has lower K value than the high-K film. The simulations show that this low-K interfacial layer gives rise to a significant increase of the gate leakage. The interface layer engineering on the scalability of high-K gate stacks was studied. Recommendations for interface layer materials were made for high-K gate stacks that will meet ITRS roadmap. Specifically, for HfAlO gate stacks, it is found that the existence of an oxide interfacial layer degrades the gate leakage so much that oxynitride or silicate should be exploited for interface engineering in the high-K gate stacks for low stand-by power application.

(III) Metal Gate Engineering on Gate Leakage of MOSFETs

In metal gated CMOS devices, a systematic study of tunneling leakage currents has been made and it is to study how the gate leakage is affected by the work function of the metal gate electrode. Physical model used for simulations was corroborated by experimental results from SiO_2 and HfO_2 gate dielectrics with TaN electrodes.

(a) In bulk CMOS, in which metal gates with work functions at conduction and valence band edges of Si are required for optimal performance, the gate leakage currents simulated in poly-Si and metal gated MOSFETs were compared. The use of metal gate can reduce gate leakage appreciably by 1-2 orders of magnitude at the same CET (capacitance equivalent oxide thickness at inversion). Due to the elimination of poly-Si gate depletion, metal gate material exhibits superior capability in gate dielectric scaling over the conventional poly-Si gate.

(b) The effects of metal gate work function on various tunneling components in MOSFETs have been investigated. In SOI MOSFETs with un-doped body, the use of mid-gap metal gate results in significant reduction of gate to source/drain extension (SDE) tunneling in both n- and p-MOSFETs. As a result, SOI MOSFETs exhibit much lower off-state leakage current contributed by gate to SDE tunneling and this reduction of off-state leakage current can be as much as three orders of magnitude when high-K gate stack is used.

6.2. Recommendations for Future Works

(I) Quantization in ultrathin body SOI devices

As discussed in Section 1.7, the ultrathin body or double gate (DG) SOI demonstrates superior scalability in device scaling. When the body thickness is scaled to nanometer scale, the carrier confinement due to the body, which has been known as the volume inversion [152], can occur in conjunction with the confinement by the inversion layer. In current fabrication technology, the body thickness has been scaled to below 5 nm, in which body confinement plays a critical role in determining the MOS electrostatics, especially at threshold. Therefore, as an extension of the hole quantization study in this thesis, it is proposed to study the hole quantization using six-band effective mass model and to establish the improved one-band effective mass method in ultrathin body or DG SOI devices. Although electron quantization in such devices have been analyzed extensively, hole quantization [153], especially including the valence band mixing, is less investigated. Such a study should be helpful for determining the accurate pMOSFET electrostatics, and also be beneficial for the study of hole transport in such scaled devices

(II) Carrier Quantization in other Surface of Si.

Because the Si (100) surface widely used in current processing has the disadvantage of low hole mobility, other Si surfaces, such as (110), is being explored because of the high hole mobility [154]. As discussed in Section 1.7, the FinFET, demonstrated by recent studies, is one of the most promising device architecture in device scaling [65]. In Fin technology, the channel surface is generally on (110) Si surface. Therefore, it is also interesting to study the carrier quantization in channel

surface other than the conventional (100). Recently, such a study on hole quantization using six-band self-consistent method has been reported [155], however, the exploratory of simple simulation methods beyond the six-band self-consistent method is extremely valuable.

(III) Tunneling in High-K dielectrics

As discussed in Chapter 4, most of studies on tunneling currents in high-K stacks were concentrated on oxynitride system. Modeling studies on HfO_2 system have been performed only in recent years. However, it is still difficult to have HfO_2 film with quality as high as the SiO_2. The existence of the interfacial layer also makes the situation complicated. The thickness and composition of the interfacial layer may depend very much on processing details, such as the surface cleaning and treatment, deposition recipe and technique, and annealing conditions. At present, the analysis on the interfacial layer is quite limited. This may partly account for the difference in the extracted parameters of HfO_2 from different authors [41, 45, 127, 137]. Therefore, a further study is still necessary in conjunction with the future developments on HfO_2 dielectric.

As presented in Section 4.4, HfAlO and HfSiO have been demonstrated as promising high-K candidates. However, there still are less studies on tunneling in this category of alloy dielectric materials [41, 156]. In this thesis, although the gate leakage is estimated for these materials, comparisons to experiments have not been made because less reliable experimental data is available to date. The accurate determination of the alloy composition is still difficult at present stage. It is also unknown how the relevant parameters, such as dielectric constant, band offset and tunneling effective mass, vary with the film composition. A further study of tunneling

in such alloy dielectrics will be meaningful for the optimization of these alloy high-K materials.

Another proposed study is the interfacial layer effects on tunneling currents. In this thesis, the interfacial layer is assumed to be uniform in dielectric properties due to the lack of detailed information on the interface layer. In practical case, the interfacial layer may not be uniform and it is more appropriate to consider the interface layer as a transition layer [138]. The study of such a structural non-uniform transition layer on the gate leakage as well as its impact on the high-K scalability is another interesting topic in future [157].

(IV) Tunneling in metal gate CMOS

The Fermi level pinning at the metal/dielectric interface plays an important role in the determination of the effective work function of metal on dielectric [52, 60]. As we have demonstrated in Chapter 5, metal gate work function impacts significantly the direct tunneling currents in CMOS devices. In Chapter 5, only general guidelines were provided, the properties of the specific metal/dielectric interface were not considered. Here, we propose the future studies on tunneling in metal gate CMOS with high-K dielectric. With the Fermi level pinning properly considered at the metal/High-K interface, it will be valuable to evaluate tunneling current performance for some metal gates currently being explored, such as TiN, TaN, or silicide.

References:

[1] International Technology Roadmap for Semiconductors, 2001. http://public.itrs.net/Files/2001ITRS/.

[2] S. Wolf, Silicon Processing for the VLSI Era, Vol.3: The Submicron MOSFET, Chap. 5, pp.205-320, Sunset Beach, CA: Lattice Press, 1995.

[3] Y. Taur and T. H. Ning, Fundamentals of modern VLSI devices, Cambridge: Cambridge University Press, 1998.

[4] G. D. Wilk, R. M. Wallace, J. M. Anthony, "High-K gate dielectrics: current status and materials properties considerations", *J. Appl. Phys*., vol.**89**, pp.5243-5275, 2001.

[5] S. Thompson, P. Packan, and M. Bohr, "MOS scaling: Transistor challenges for the 21st century", *Intel Technology Journal*, Q**3**, p.1-19, 1998.

[6] S. Wolf, Silicon Processing for the VLSI Era, Vol.4: Deep Submicron Processing Technology, Chap. 5, pp.181-226, Sunset Beach, CA : Lattice Press, 2002.

[7] J. L. Hoyt, H. M. Mayfeh, S. Eguchi, I. Aberg, G. Xia, T. Drake, E. A. Fitzgerald, and D. A. Antoniadis, "Strained silicon MOSFET technology," in *IEDM Tech. Dig*., pp.23-26, 2002.

[8] T. Ando, A. Fowler, and S. Stern, "Electronic properties of two-dimensional systems," *Rev. Mod. Phys*., vol.**54**, pp.437-672, 1982.

[9] S.-H. Lo, D. A. Buchanan, and Y. Taur, "Modeling and characterization of quantization, polysilicon depletion, and direct tunneling currents in MOSFETs with ultrathin oxides", *IBM J. Res. Develop*., vol.**43**, pp.327-337, 1999.

[10] S. Takagi, M. T. Takagi, and A. Toriumi, "Accurate characterization of electron and hole inversion-layer capacitance and its impact on low voltage operation of scaled MOSFETs," in *IEDM Tech. Dig*., pp.619-622, 1998.

[11] C.-Y. Hu, S. Banerjee, K. Sadra, B. G. Streetman, and R. Sivan, "Quantization effects in inversion layers of pMOSFET's on Si (100) substrate," *IEEE Electron Device Lett*. vol. **17**, pp.276-278, 1996.

[12] F. Stern, "Self-consistent results for n-type Si inversion layer," *Phys. Rev. B*, vol.**5**, pp.4891-4899, 1972.

[13] C. Moglestue, "Self-consistent calculation of electron and hole inversion charges at silicon-silicon dioxide interfaces," *J. Appl. Phys.*, vol. **59**, pp.3175-3183, 1986.

[14] J. Sune, P. Olivo, and B. Ricco, "Self-consistent solution of the Poisson and Schrodinger equations in accumulated semiconductor-insulator interfaces," *J. Appl. Phys*. vol. **70**, pp.337-345, 1991.

[15] Y. Ma, Z. Li, L. Liu, L. Tian, and Z. Yu, and, "Effective density of states approach to QM correction in MOS structures", *Solid Stat. Electron*., vol. **44**, pp.401-407, 2000.

[16] Y. Ma, L. Liu, Z. Yu, and Z. Li, "Characterization and modeling of threshold voltage shift due to quantum mechanical effects in pMOSFETs", *Solid Stat. Electron*. vol. **44**, pp.1335-1339, 2000.

[17] Y. T. Hou and M. F. Li, "Hole quantization effects and threshold voltage shift in pMOSFET -assessed by improved one-band effective mass approximation," *IEEE Tran. Electron Devices*, vol. **48**, pp.1188-1193, 2001.

[18] J. R. Hauser and K. Ahmed, "Characterization of ultrathin oxides using electrical C-V and I-V measurements," in *Int. Conf. on Characterization and Metrology for ULSI Technology*, pp.235-239, 1998.

[19] M. J. van Dort, P. H. Woerlee, and A. J. Walker, "A simple model for quantization effects in heavily-doped silicon MOSFETs at inversion conditions," *Solid Stat. Electron.*, vol. **37**, pp.411-414, 1994.

[20] S. A. Hareland, S. Jallepalli, G. Chindalore, W. K. Shih, A. F. Tasch, and C. M. Maziar, “A simple model for quantum mechanical effects in hole inversion layers in silicon pMOS devices”, *IEEE Trans. Electron Devices*, vol.**44**, pp.1172-1173, 1997.

[21] S. A. Hareland, M. Manassian, W. K. Shih, S. Jallepalli, H. Wang, G. L. Chindalore, A. F. Tasch, and C. M. Maziar, “Computationally efficient models for quantization effects in MOS electron and hole accumulation layers”, *IEEE Trans. Electron Devices*, vol.**45**, pp.1487-1492, 1998.

[22] C. H. Choi, J. S. Goo, T. Y. Oh, Z. Yu, R. W. Dutton, A. Bayoumi, M. Cao, P. V. Voorde, D. Vook, and C. H. Diaz, “MOS C-V characterization of ultrathin gate oxide thickness (1.3-1.8 nm)”, *IEEE Electron Lett*. vol. **20**, pp.292-294, 1999.

[23] E. Burstein and S. Lundqvist, *Tunneling Phenomena in Solids*, Plenum Press: New York, 1969.

[24] M. S. Krishnan, L. Chang, T. J. King, J. Bokor, and C. Hu, “MOSFETs with 9 to 13 A thick gate oxides,” in *IEDM Tech. Dig*., pp.241-244, 1999.

[25] J. Maserjian, “Tunneling in thin MOS structures,” *J. Vac. Sci. Technol*., vol.**11**, pp.996-1003, 1974.

[26] Z. A. Weinberg, “On tunneling in metal-oxide-silicon structures,” *J. Appl. Phys*., vol.**53**, pp.5052-5056, 1982.

[27] J. Cai and C. T. Sah, “Gate tunneling currents in ultrathin oxide metal-oxide-silicon transistors,” *J. App. Phys*., vol. **89**, pp.2272-2286, 2001.

[28] Khairurrijal, W. Mizubayashi, S. Miyazaki, and M. Hirose, “Analytic model of direct tunnel current through ultrathin gate oxides,” *J. Appl. Phys*., vol.**87**, pp.3000-3005, 2000.

[29] S.-H. Lo, D. A. Buchanan, Y. Taur and W. Wang, "Quantum-mechanical modeling of electron tunneling current from the inversion layer of ultra-thin-oxide nMOSFETs," *IEEE Electron Lett*., vol. **18**, pp.209-211, 1997.

[30] E. Cassan, "On the reduction of direct tunneling leakage through ultrathin gate oxides by a one-dimensional Schrodinger-Poisson solver," *J. Appl. Phys*., vol. **87**, pp.7931-7939, 2000.

[31] W. Magnus and W. Schoenmaker, "Full quantum mechanical model for the charge distribution and the leakage currents in ultrathin metal-insulator-semiconductor capacitors," *J. Appl .Phys*., vol. **88**, pp.5833-5842, 2000.

[32] F. Rana, S. Tiwari, and D. A. Buchanan, "Self-consistent modeling of accumulation layers and tunneling currents through very thin oxides," *Appl. Phys. Lett.*, vol. **69**, pp.1104-1106, 1996.

[33] L. F. Register, E. Rosenbaum, and K. Yang, "Analytic model for direct tunneling current in polycrystalline-silicon-gate metal-oxide-semiconductor devices," *Appl. Phys. Lett*., vol. **74**, pp.457-459, 1999.

[34] N. Yang, W. K. Henson, J. R. Hauser, and J. J. Wortman, "Modeling study of ultrathin gate oxides using direct tunneling current and capacitance-voltage measurements in MOS devices," *IEEE Trans. Electron Device*, vol. **46**, pp.1464-1471, 1999.

[35] A. Ghetti, C. T. Liu, M. Mastrapasqua, and E. Sangiorgi, "Characterization of tunneling current in ultrathin gate oxide," *Solid Stat. Electron*., vol.**44**, pp,1523-1531, 2000.

[36] W. K. Shin, E. X. Wang, S. Jallepalli, F. Leon, C. M. Maziar, and A. F. Tasch, "Modeling gate leakage current in nMOS structures due to tunneling through an ultrathin oxide," *Solid-State Electron*., vol.**42**, pp.997-1006, 1998.

[37] A. Dalla Serra, A. Abramo, P. Palestri, L. Selmi, F. Widdershoven, "Closed- and open-boundary models for gate-current calculation in n-MOSFETs**,"** *IEEE Trans. Electron Device,* vol.**48**, pp.1811-1815, 2001.

[38] M. Stadele, B. R. Tuttle, and K. Hess, "Tunneling through ultrathin SiO_2 gate oxides from microscopic models," *J. Appl. Phys.,* vol. **89**, pp.348-363, 2001.

[39] D. A. Muller, T. Sorsch, S. Moccio, F. H. Baumann, K. Evans-Letterodt, and G. Timp, "The electronic structure at the atomic scale of ultrathin gate oxides", *Nature*, vol. **399**, pp.758-761, 1999.

[40] C. M. Osburn, I. Kim, S. K. Han, I. De, K. F. Yee, S. Gannavaram, S. J. Lee, C. H. Lee, Z. J. Luo, W. Zhu, J. R. Hauser, D. L. Kwong, G. Lucovsky, T. P. Ma, M. C. Ozturk, "Vertically scaled MOSFET gate stacks and junctions: how far are we likely to go?" *IBM J. Res. & Dev*., vol. **46**, pp.299-313, 2002.

[41] Y. T. Hou, M. F. Li, H. Y. Yu, Y. Jin and D. L. Kwong, "Quantum tunneling and scalability of HfO_2 and HfAlO gate stacks," in *IEDM Tech. Dig*., pp.731-734, 2002.

[42] Y. C. Yeo, T. J. King, and C. Hu, "MOSFET gate leakage modeling and selection guide for alternative gate dielectrics based on leakage considerations," *IEEE Trans. Electron Device,* vol.**50**, pp.1027-1035, 2003.

[43] J. H. Stathis, "Reliability limits for the gate insulator in CMOS technology," *IBM J. Res. Develop*., vol. **46**, pp.265-285, 2002

[44] S. Wolf, Silicon Processing for the VLSI Era, Vol.4: Deep Submicron Technology, Chap. 4, pp.145-180, Sunset Beach, CA : Lattice Press, 2002.

[45] W. Zhu, T.P. Ma, T. Tamagawa, Y. Di, J. Kim, R. Carruthers, M. Gibson, and T. Furukawa, "HfO_2 and HfAlO for CMOS: thermal stability and current transport," in *IEDM Tech. Dig*., pp.463-466, 2001

[46] H. Y. Yu, N. Wu, M. F. Li, C. X. Zhu, B. J. Cho, D. L. Kwong, C. H. Tung, J. S. Pan, J. W. Chai, W. D. Wang, D. Z. Chi, C. H. Ang, J. Z. Zheng, S. Ramanathan, "Thermal stability of $(HfO_2)x(Al_2O_3)_{1-x}$ on Si," *Appl. Phys. Lett.*, vol. **81**, pp.3618-3620, 2002.

[47] G. D. Wilk, R. M. Wallace, and J. M. Anthony, "Hafnium and zirconium silicates for advanced gate dielectrcis," *J. Appl. Phys*., vol.**87**, pp.484-492, 2000.

[48] C. H. Choi, S. J. Rhee, T. S. Jeon, N. Lu, J. H. Sim, R. Clark, M. Niwa, and D. L. Kwong, "Thermally stable CVD HfOxNy advanced gate dielectrics with poly-Si gate electrodes," in *IEDM Tech. Dig*., pp.857-860, 2002.

[49] C. S. Kang, H. J. Cho, K. Onishi, R. Choi, Y. H. Kim, R. Nieh, J. Han, S. Krishnam, A. Shahriar, and J. C. Lee, "Nitrogen concentration effects and performance improvement of MOSFETs using thermally stable HfOxNy gate dielectrics," in *IEDM Tech. Dig*., pp.865-868, 2002.

[50] H. S. Jung, Y. S. Kim, J. P. Kim, J. H. Lee, J. H. Lee, N. I. Lee, H. K. Kang, K. P. Suh, H. J. Ryu, C. B. Oh, Y. W. Kim, K. H. Cho, H. S. Baik, Y. S. Chung, H. S. Chang, and D. W. Moon, "Improved current performance of CMOSFETs with nitrogen incorporated HfO_2-Al_2O_3 laminate gate dielectric," in *IEDM Tech. Dig*., pp.853-856, 2002.

[51] M. Koyama, A. Kaneko, T. Ino, M. Koike, Y. Kamara, R. Iijima, Y. Kamimuta, A. Takashima, M. Suzuki, C. Hongo, S. Imumiya, M. Takayanagi, and A. Nishiyama, "Effects of nitrogen in HfSiON gate dielectric on the electrical and thermal characteristics," in *IEDM Tech. Dig*., pp.849-852, 2002.

[52] C. Hobbs, L. Fonseca, V. Dhandapani, S. Samavedam, B. Taylor, J. Grant, L. Dip, D. Triyoso, R. Hegde, D. Gilmer, R. Garcia, D. Roan, L. Lovejoy, R. Rai, L.

Heber, H. Tseng, B. White, and P. Tobin, "Fermi level pinning at the polySi/metal oxide interface," in *Symp. VLSI Tech.*, pp.9-10. 2003.

[53] E. M. Vogel, K. Z. Ahmed, B. Hornung, W. K. Henson, P. K. McLarty, G. Lucovsky, J. R. Hauser, and J. J. Wortman, "Modeled tunneling currents for high dielectric constant dielectrics," *IEEE Trans. Electron Devices*, vol.**45**, pp.1350-1355, 1998.

[54] E. Josse and T. Skotnicki, "Polysilicon gate with depletion or metallic gate with buried channel: weir will worse ?",in *IEDM Tech. Dig*., pp. 661-664, 1999.

[55] D. G. Park, K. Y. Lim, H. J. Cho, T. H. Cha, J. J. Kim, J. K. Ko, I. S. Yeo and J. W. Park, "Novel damage free direct metal gate process using atomic layer deposition," in *Symp. VLSI Tech.*, pp.65-66. 2001.

[56] Y. H. Kim , C. H. Lee, T. S. Jeon, W. P. Bai, C. H. Choi, S. J. Lee, L. Xinjian, R. Clarks, D. Roberts, and D. L. Kwong, "High quality CVD TaN gate electrode for sub-100 nm MOS devices, " in *IEDM Tech. Dig*., pp.667-670, 2001.

[57] H. Zhong. S.H. Hong, Y. S. Suh, H. Lazar, G. Heuss, and V. Misra, "Properties of Ru-Ta alloys as gate electrodes for NMOS and PMOS silicon devices," in *IEDM Tech. Dig*., pp.467-470, 2001.

[58] Q. Lu, R. Lin, P. Panade, T. J. King and C. Hu, "Metal gate work function adjustment for future CMOS Technology," in *Symp. VLSI Tech.*, pp.45-46. 2001

[59] J. Kedzierski, E. Nowark, T. Kanarsky, Y. Zhang, D. Boyd, R. Carruthers, C. Cabral, R. Amos, C. Lavioe, R. Roy, J. Newbury, E. Sullivan, J. Benedict, P. Saunders, K. Wong, D. Canaperi, M. Krishnam, K. L. Lee, B. A. Rainey, D. Fried, P. Cottrell, H. S. P. Wong, M. Leong, and W. Haensch, "Metal-gate FinFET and fully-depleted SOI devices using total silicidation," in *IEDM Tech. Dig*., pp.247-250, 2002.

[60] Y. C. Yeo, T. J. King, and C. Hu, "Metal-dielectric band alignment and its implications for metal gate complementary metal-oxide-semiconductor technology," *J. Appl. Phys*., vol.**92**, pp.7266-7271, 2002.

[61] W. P. Maszara, Z. Krivokapic, P. King, J. S. Goo, and M. R. Lin, "Transistors with dual work function metal gates by single full silicidation (FUSI) of polysilicon gates," in *IEDM Tech Dig*., pp.367-370, 2002.

[62] D. J. Frank, R. H. Dennard, E. Nowak, P. M. Solomon, Y. Taur, and H. P. Wong, "Device scaling limits of Si MOSFETs and their application dependencies," *Proc. IEEE*, vol.**89**, p.259, 2001.

[63] S. Wolf, Silicon Processing for the VLSI Era, Vol.4: Deep Submicron Technology, Sunset Beach, CA : Lattice Press, 2002. Chap. 11, pp.501-572.

[64] D. J. Frank, Y. Taur, and H.-S. P. Wong, "Generalized scale length for two-dimensional effects in MOSFETs," *IEEE Electron Device Lett.*, Vol. **19**, pp.385-387, 1998.

[65] B. Yu, L. Chang, S. Ahmed, H. Wang, S. Bell, C.Y. Yang, C. Tabery, C. Ho, Q. Xiang, T.J. King, J. Bokor, C. Hu, M.R. Lin, and D. Kyser, "FinFET scaling to 10 nm gate length," in *IEDM Tech. Dig*., pp.251-254, 2002.

[66] F. L. Yang, H. Y. Chen, F. C. Chen, C. C. Huang, C. Y. Chang, H. K. Chiu, C. C. Lee, C. C. Chen, H. T. Huang, C. J. Chen, H. J. Tao, Y. C. Yeo, M. S. Liang, and C. Hu, "25 nm CMOS Omega FETs," in *IEDM Tech. Dig*., pp.255-258, 2002.

[67] 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," in *Symp. VLSI Tech.*, pp.133-134. 2003.

[68] K. Uchida, H. Watanabe, A. Kinoshita, J. Koga, T. Numata, and S. Takagi, "Experimental study on carrier transport mechanism in ultrathin-body SOI n- and

p-MOSFETs with SOI thickness less than 5 nm," in *IEDM Tech Dig*., pp.47-50, 2002.

[69] L. Chang, K. J. Yang, Y. C. Yeo, Y. K. Choi, T. J. King, and C. Hu, " Reduction of direct tunneling gate leakage current in double-gate and ultra-thin body MOSFETs," in *IEDM Tech Dig*., pp.99-102, 2001.

[70] Y. Taur, D.A. Buchanan, W. Chen, D.J. Frank, K.E. Ismail, S.H. Lo, G.A. Sai-Halasz, R.G. Viswanathan, H.J.C. Wann, S.J. Wind, and H.S. Wong, "CMOS scaling into the nanometer regime", *Proc. IEEE*, vol.**85**, pp. 486-504, 1997.

[71] Ming-Fu Li, *Modern Semiconductor Quantum Physics*, chap.3, p.243 and chap.5, p.441, World Scientific, Singapore, 1994;

[72] S. Jallepalli, J. Bude, W.-K. Shih, M. R. Pinto, C. M. Maziar, and A. F. Tasch, "Electron and hole quantization and their impact on deep submicron silicon p- and n-MOSFET characteristics," *IEEE Tran. Electron Devices*, vol. **44**, pp.297-303, 1997.

[73] S. Rodriguez , J. A. Lopez-Villaneuva , I. Melchor , and J. E. Carceller , "Hole confinement and energy subbands in a silicon inversion layer using the effective mass theory," *J. Appl. Phys*., vol. **86**, pp.438-444, 1999.

[74] A. Pirovano, A. L. Lacaita, G. Zandler, and R. Oberhuber, "Explaining the dependence of the hole and electron mobilities in Si inversion layers", *IEEE Trans. Electron Devices*, vol.**47**, pp.718-724, 2000.

[75] T. Low, Y.T. Hou, and M. F. Li, "Improved One-band Self-consistent Effective Mass Methods for Hole Quantization in p-MOSFET," *IEEE Trans. Electron Devices*, vol.**50**, pp.1284-1289, 2003.

[76] W. J. Fan, M. F. Li, T. C. Chong, and J. B. Xia, "Valence hole subbands and optical gain spectra of $GaN/Ga_{1-x}Al_xN$ strained quantum wells," *J. Appl. Phys*. vol.**80**, pp.3471-3478, 1996.

[77] J. M. Luttinger , "Quantum Theory of Cyclotron Resonance in Semiconductors: General Theory ", *Phys. Rev*., vol. **102**, p.1030 and p.1956.

[78] Landolt-Bornstein, *Numerical Data and Functional Relationships in Science and Technology*, Vol.17a, edited by O. Madelung (Springer-Verlag , Berlin , 1987).

[79] Y. T. Hou and M. F. Li, "A Simple and efficient Model for Quantization Effects of Hole Inversion Layers in MOS Devices," *IEEE Tran. Electron Devices*, vol. **48**, pp.2893-2898, 2001.

[80] Y. T. Hou and M. F. Li, "A novel simulation algorithm for Si valence hole quantization of inversion layer in metal-oxide-semiconductor devices," *Jpn. J. Appl. Phys*. part 2, vol. **40**, pp.144-146, 2001.

[81] H. H. Mueller, and M. J. Schulz, "Simplified method to calculate the band bending and the subband energies in MOS capacitors," *IEEE Tran. Electron Devices*, vol. **44**, pp.1539-1543, 1997.

[82] G. Chindalore, S. A. Hareland, S. Jallepalli, A. F. Tasch, C. M. Maziar, V. K. F. Chia, and S. Smith, "Experimental determination of threshold voltage shifts due to quantum mechanical effects in MOS electron and hole inversion layers," *IEEE Electron Device Lett.*, Vol. **18**, pp.206-208, 1997.

[83] Landolt-Bornstein, *Numerical Data and Functional Relationships in Science and Technology*, Vol.22a, edited by O. Madelung (Springer-Verlag, Berlin, 1987), p.72 and p.304.

[84] Y. Ma, L. Liu, Z. Yu, and Z. Li, "Validity and applicability of triangular potential well approximation in modeling of MOS structure inversion and accumulation layer," *IEEE Tran. Electron Devices*, vol. **47**, pp.1764-1767, 2000.

[85] D. J. Frank, "Power-constrained CMOS scaling limits," *IBM J. Res. & Dev.*, vol.**46**, pp.234-244, 2002.

[86] K. Ahmed, E. Ibok, G.C.F. Yeap, Q. Xiang, B. Ogle, J.J. Wortman, and J.R. Hauser, "Impact of tunneling currents and channel resistance on the characterization of channel inversion layer charge and polysilicon-gate depletion of sub-20 A gate oxide MOSFETs," *IEEE Trans. Electron Devices*, vol.**46**, pp.1650-1655, 1998.

[87] Y. Shi, T. P. Ma, S. Prasad, and S. Dhanda, "Polarity-dependent tunneling current and oxide breakdown in dual-gate CMOSFETs," *IEEE Electron Lett.*, vol.**19**, pp.391-393, 1998.

[88] W. C. Lee, T. J. King, and C. Hu, "Evidence of hole direct tunneling through ultrathin gate oxide using p+ poly-SiGe gate," *IEEE Electron Lett.*, vol.**20**, pp.268-270, 1999.

[89] H. Guan, M. F. Li, Y. D. He, B. J. Cho, and Z. Dong, "A through study of quasi-breakdown phenomenon of thin gate oxide in dual-gate CMOSFETs", *IEEE Trans. Electron Devices,* vol.**47**, pp.1608-1616, 2000.

[90] N. Yang, W.K. Hension, and J.J. Wortman, "Analysis of tunneling currents and reliability of NMOSFETs with sub-2 nm gate oxides," in *IEDM Tech Dig.*, pp.453-456, 1999.

[91] N. Yang, W. K. Henson, and J. J. Wortman, "A comparative study of gate direct tunneling and drain leakage currents in n-MOSFETs with sub-2 nm gate oxides," IEEE *Trans. Electron Devices*, vol. **47**, pp.1636-1644, 2000.

[92] K. N. Yang, H. T. Huang, M. J. Chen, Y. M. Lin, M. C. Yu, S. M. Jang, D. C. H. Yu, and M. S. Yang, "Characterization and modeling of edge direct tunneling (EDT) leakage in ultrathin gate oxide MOSFETs," *IEEE Trans. Electron Devices*, vol.**48**, pp.1159-1162, 2000.

[93] K. N. Yang, H. T. Huang, M. J. Chen, Y. M. Lin, M. C. Yu, S. M. Jang, C. H. Yu, and M. S. Liang, "Edge hole direct tunneling in off-state ultrathin gate oxide p-channel MOSFETs," in *IEDM Tech Dig*., pp.679-682, 2000.

[94] S. Thompson, P. Packan, T. Ghani, M. Steller, M. Alavi, I. Post, S. Tyagi, S. Ahmed, S. Yang, and M. Bohr, "Source/drain extension scaling for 0.1 um and below channel length MOSFETs," in *Symp. VLSI Tech.*, pp.132-133, 1998.

[95] Berkeley Device Group, [online] www.device.berkeley.eecs.edu/qmcv.

[96] W. C. Lee and C. Hu, "Modeling gate and substrate currents due to conduction- and valence-band electron and hole tunneling," in *Symp. VLSI Tech.*, pp.198-199, 2000.

[97] 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," *IEEE Trans. Electron Devices*, vol.**47**, pp.2161-2166, 2000.

[98] H. Iwata, "Fully quantum-mechanical calculation of hole direct tunneling current in ultrathin gate oxide p-channel metal-oxide-semiconductor devices," *Jpn. J. Appl. Phys*., vol.**41**, part 1, pp.552-556, 2002.

[99] A. Haque, and K. Alam, "Accurate modeling of direct tunneling hole current in p-metal-oxide-semiconductor devices," *Appl. Phys. Lett*., vol.**81**, pp.667-669, 2002.

[100] Y. T. Hou, M. F. Li, W. H. Lai, and Y. Jin, "Modeling and characterization of direct tunneling hole current in *p*-metal-oxide-semiconductor field-effect transistors," *Appl. Phys. Lett*., vol.**78**, pp.4034-4036, 2001.

[101] Y. T. Hou, M. F.Li, Y. Jin, and W. H. Lai, Direct tunneling hole current through ultrathin gate oxides in metal-oxide-semiconductor devices, *J. Appl. Phys.*, vol.**91**, pp.258-264, 2002.

[102] L. B. Freeman and W. E. Dahlke, "Theory of tunneling into interface states," Solid-State Electron., vol.**13**, p.1483-1503, 1970.

[103] B. Majkusiak and M. H. Badri, "Semiconductor thickness and back-gate voltage effects on the gate tunnel current in themOS/SOI system with an ultrathin oxide," *IEEE Trans. Electron Devices,* vol. **47**, pp.2347-2351, 2000.

[104] C. A. Richter, A. R. Hefner and E. M. Vogel, "A comparison of quantum-mechanical capacitance-voltage simulators," *IEEE Device Lett*., vol.**22**, pp.35-37, 2001.

[105] R. Ludeke and A. Schenk, "Energy-dependent conduction band mass of SiO_2 determined by ballistic electron emission microscopy," *J. Vac. Sci. Technol. B*, vol.**17**, pp.1823-1830, 1999.

[106] P. M. Schneider and W. B. Fowler, "Band structure and optical properties of silicon dioxide," *Phys. Rev. Lett*., vol.**36**, pp.425-428, 1976.

[107] E. Gnani, S. Reggiani, R. Colle, and M. Rudan, "Band structure calculations of SiO_2 by means of Hartree-Fock and density functional techniques," *IEEE Trans. Electron Devices,* vol. **47**, pp.1795-1803, 2000.

[108] K. Ahmed, E. Ibok, G. Bains, D. Chi, B. Ogle, J. J. Wortman, and J. R. Hauser, "Comparative physical and electrical metrology of ultrathin oxides in the 6-1.5 nm regime," *IEEE Trans. Electron Devices*, vol.**47**, pp.1349-1354, 2000.

[109] A. Shanware, J.P. Shiely, H.Z. Massoud, E. Vogel, K. Henson, A. Srivastava, C. Osburn, J.R. Hauser, and J.J. Wortman, "Extraction of the gate oxide thickness of N- and P-channel MOSFETs below 20 A from the substrate current resulting from valence-band electron tunneloing," in *IEDM Tech. Dig*., pp.815-818, 1999.

[110] S. Hong, Y. Zhang, Y. Luo, T. Suligoj, S.D. Kim, J.C.S. Woo, R. Li, B.W. Min, B. Hradsky, A. Vandooren, B.Y. Nguyen, and K. L. Wang, "Novel direct tunneling current (DTC) method for channel length extraction beyond sub-50 nm gate CMOS," in *IEDM Tech Dig*., pp.297-300, 2001.

[111] T. P. Ma, "Making silicon nitride film a viable gate dielectric," *IEEE Trans. Electron Devices*, vol.**45**, pp.680-690, 1998.

[112] Y. Shi, X. Wang, and T. P. Ma, "Electrical properties of high-quality ultrathin nitride/oxide stack dielectrics," *IEEE Tran. Electron Device*, vol.**46**, pp.362-368, 1998.

[113] X. Guo and T.P. Ma, "Tunneling leakage current in oxynitride: dependence on oxygen/nitrogen content," *IEEE Electron Device Lett*. vol.**19**, pp.207-209, 1998.

[114] Y. C. Yeo, Q. Lu, W. C. Lee, T. J. King, C. Hu, X. Wang, X. Guo, and T. P. Ma, "Direct tunneling gate leakage current in transistors with ultrathin silicon nitride gate dielectric", *IEEE Electron Device Lett*. vol.**21**, pp.540-542, 2000.

[115] Q. Lu, Y. C. Yeo, K. J. Yang, R. Lin, I. Polishchuk, T. J. King, C. Hu, S. C. Song, H. F. Luan, D. L. Kwong, X. Guo, Z, Luo, X. Wang, and T. P. Ma, "Two silicon nitride technologies for post-SiO_2 MOSFET gate dielectric," *IEEE Electron Device Lett.,* vol.**22**, pp.324-326, 2001.

[116] G. Lucovsky, Y. Wu, H. Niimi, H. Yang, J. Keister, and J.E. Rowe, "Separate and independent reductions in direct tunneling in oxide/nitride stacks with

monolayer interface nitridation associated with the (i) interface nitridation and (ii) increased physical thickness" *J. Vac. Sci. Tachnol. A*, vol.**18**, pp.1163-1168, 2000.

[117] S. C. Song, H. F. Luan, C. H. Lee, A. Y. Mao, S. J. Lee, J. Gelpey, S. Marcus, and D. L. Kwong, "Ultra thin high quality stack nitride/oxide gate dielectrics prepared by in-situ rapid thermal N_2O oxidation of NH_3-nitrided Si," in *Symp. VLSI Tech.*, pp.137-138. 1999

[118] C. H. Chen, Y. K. Fang, C. W. Yang, S. F. Ting, Y. S. Tsair, M. C. Yu, T. H. Hou, M. F. Wang, S. C. Chen, C. H. Yu, and M. S. Liang, "Thermally-enhanced remote plasma nitrided ultrathin (1.65 nm) gate oxide with excellent performances in reduction of leakage current and boron penetration," *IEEE Electron Device Lett.*, vol.**22**, pp.378-380, 2001.

[119] H. H. Tseng, Y. Jeon, P. Abramowitz, T. Y. Luo, L. Hebert, J. J. Lee, J. Jiang, P. J. Tobin, G. C. F. Yeap, M. Moosa, J. Alvis, S. G. H. Anderson, N. Cave, T. C. Chua, A. Hegedus, G. Miner, J. Jeon, and A. Sultan, "Ultra-thin decoupled plasma nitridation (DPN) oxynitride gate dielectric for 80 nm advanced technology," *IEEE Electron Device Lett.*, vol.**23**, pp.704-706, 2002.

[120] A. Nakajima, Q. D. M. Khosru, T. Yoshimoto, T. Kidera, and S. Yokoyama, "NH_3 annealed atomic layer deposited silicon nitride as a high-k gate dielectric with high reliability," *Appl. Phys. Lett.*, vol.**80**, pp.1252-1254, 2002.

[121] H. Y. Yu, Y. T. Hou, M. F. Li, and D. L. Kwong, "Investigation of hole tunneling current through ultrathin oxynitride/oxide stack gate dielectrics in p-MOSFETs, " *IEEE Trans. Electron Devices*, vol.**49**, pp.1158-1164, 2002

[122] H. Y. Yu, Y. T. Hou, M. F. Li, and D. L. Kwong, "Hole tunneling current through oxynitride/oxide stack and the stack optimization for p-MOSFETs," *IEEE Electron Device Lett.*, vol.**23**, pp.285-287, 2002.

[123] K. Muraoka, K. Kurihara, N. Yasuda, and H. Satake, "Optimum structure of deposited ultrathin silicon oxynitride film to minimize leakage current," *J. App. Phys.,* Vol.**94**, pp.2038-2045, 2003.

[124] A. Chin, Y. H. Wu, S. B. Chen, C. C. Liao, and W. J. Chen, "High quality La_2O_3 and Al_2O_3 gate dielectrics with equivalent oxide thickness 5-10 Å," in *Symp. VLSI Tech.*, pp.16-17. 2000.

[125] D. A. Buchanan, E. P. Gusev, E. Cartier, H. Okorn-Schmidt, K. Rim, M. A. Gribelyuk, A. Mocuta, A. Ajmera, M. Copel, S. Guha, N. Bojarczuk, A. Callegari, C. D'Emic, P. Kozlowski, K. Chan, R. J. Fleming, P. C. Jamison, J. Brown, and R. Arndt, "80 nm poly-silicon gated n-FETs with ultrathin Al_2O_3 gate dielectric for ULSI applications," in *IEDM Tech. Dig*., pp.223-226, 2000.

[126] E.P. Gusev, D.A. Buchanan, E. Cartier, A. Kumar, D. DiMaria, S. Guha, A. Callegari, S. Zafar, P.C. Jamison, D.A. Neumayer, M. Copel, M.A. Gribelyuk, H. Okorn-Schmidt, C. D'Emic, P. Kozlowski, K. Chan, N. Bojarczuk, L.A. Ragnarsson, P. Ronshein, K. Rim, R.J. Fleming, A. Mocuta, and A. Ajmera, "Ultrathin high-K gate stacks for advanced CMOS devices," in *IEDM Tech. Dig*., pp.451-454, 2001.

[127] Y. C. Yeo, T. J. King, and C. Hu, "Direct tunneling leakage current and scalability of alternative gate dielectrics," *Appl. Phys. Lett*., vol.**81**, pp.2091-2093, 2002.

[128] H. Y. Yu, M. F. Li, B. J. Cho, C. C. Yeo, M. S. Joo, D.-L. Kwong, J. S. Pan, C. H. Ang, J. Z. Zheng, and S. Ramanathan, "Energy Gap and Band Alignment for $(HfO_2)_x(Al_2O_3)_{1-x}$ on (100) Si", *Appl. Phys. Lett.*, vol.**81**, pp.376-378, 2002

[129] V. V. Afanas'ev, M. Houssa, A. Stesmans, and M. M. Heyns, "Electron energy barriers between (100) Si and ultrathin stacks of SiO_2, Al_2O_3, and ZrO_2 insulators," *Appl. Phys. Lett.*, vol.**78**, pp.3073-3075, 2001.

[130] S. Miyazaki, "Photoemission study of energy-band alignments and gap state density distributions for high-K gate dielectrics," *J. Vac. Sci. Technol. B*, vol.**19**, pp.2212-2216, 2001.

[131] Y. N. Xu and W. Y. Ching, "Self-consistent band structures, charge distributions, and optical-absorption spectra in MgO, α-Al_2O_3, and $MgAl_2O_4$", *Phys.Rev.* B, vol.**43**, pp.4461-4472, 1991.

[132] J. Robertson, "Band offsets of wide band gap oxides and implications for future electronic devices," *J. Vac. Sci. Technol. B*, vol.**18**, pp.1785-1791, 2000.

[133] S. J. Lee, H. F. Luan, C. H. Lee, T. S. Jeon, W. P. Bai, Y. Senzaki, D. Roberts, and D. L. Kwong, "Performance and reliability of ultra thin CVD HfO_2 gate dielectrics with dual poly-Si gate electrodes," in *Symp. VLSI Tech.*, pp.133-134, 2001.

[134] B. H. Lee, R. Choi, L. Kang, S. Gopalan, R. Nieh, K. Onishi, Y. Jeon, W.J. Qi, C. Kang, and J.C. Lee, "Characteristics of TaN gate MOSFET with ultrathin hafnium oxide (8-12 Å)," in *IEDM Tech. Dig.*, pp.39-42, 2000

[135] H. Harris, K. Choi, N. Mehta, A. Chandolu, N. Biswas, G. Kipshidze, S. Nikishin, S. Gangopadhyay, and H. Temkin, "HfO_2 gate dielectric with 0.5 nm equivalent oxide thickness," *Appl. Phys. Lett.*, vol.**81**, pp.1065-1067, 2002.

[136] Y. Y. Fan, S. Mudanai, W. Qi, J. C. Lee, A. F. Tasch, L. F. Register, and S. K. Banerjee, "Modeling high K gate tunneling current from p-type Si inversion layers," in *Proceedings of 58th Device Research Conference*, pp. 63-64, 2000.

[137] Q. Lu, R. Lin, H. Takeuchi, T. J. King, C. Hu, K. Onishi, R. Choi, C. S. Kang, and J. C. Lee, "Deep-submicron CMOS process integration of HfO_2 gate dielectric with poly-Si gate," in *Proceedings of Semiconductor Research Symposium*, pp.377-380, 2001.

[138] Y. Harada, M. Niwa, S. J. Lee, and D. L. Kwong, "Specific structural factors influencing on reliability of CVD HfO_2," in *Symp. VLSI Tech.*, pp.26-27, 2002.

[139] Y. T. Hou, M. F. Li, H. Y. Yu, and D. L. Kwong, "Modeling of tunneling currents through HfO_2 and $(HfO_2)_x(Al_2O3)_x$ gate stacks," *IEEE Electron Device Lett.*, vol. **24**, pp.96-98, 2003.

[140] V. V. Afanas'ev, A. Stesmans, F. Chen, X. Shi, and S. A. Campbell, "Internal photoemission of electrons and holes from (100) Si into HfO_2," *Appl. Phys. Lett.*, vol.**81**, pp.1053-1055, 2002.

[141] L. Kang, K. Onishi, Y. Jeon, B. H. Lee, C. Kang, W. J. Qi, R. Nieh, S. Gopalan, R. Choi, and J. C. Lee, "MOSFET devices with polysilicon on single layer HfO_2 high-K dielectrics," in *IEDM Tech. Dig*., pp.35-38, 2000

[142] P. D. Kirsch, C. S. Kang, J. Lozano, J. C. Lee, J. G. Ekerdt, "Electrical and spectroscopic comparison of HfO_2/Si interfaces on nitrided and un-nitrided Si (100)," *J.Appl. Phys*., vol.**91**, pp.4353-4363, 2002.

[143] B. Guillaumot, X. Garros, F. Lime, K. Oshima, B. Tavel, J.A. Chroboczek, P. Masson, R. Truche, A. M. Papon, F. Martin, J. F. Damlencourt, S. Maitrejean, M. Rivoire, C. Leroux, S. Cristoloveanu, G. Ghibaudo, J. L. Autran, T. Skotnicki, and S. Deleonibus, "75 nm damascene metal gate and high-k integration for advanced CMOS devices," in *IEDM Tech. Dig*., pp.355-358, 2002.

[144] M. Fujiwara, M. Takayanagi, T. Shimizu, and Y. Toyoshima, “Extending gate dielectric scaling limit by NO oxynitride: design and process issues for sub-100 nm technology,” in *IEDM Tech Dig*., pp.227-230, 2000.

[145] L. Chang, S. Tang, T. J. King, J. Bokor, and C. Hu, “Gate length scaling and threshold voltage control of double-gate MOSFETs,” in *IEDM Tech. Dig*., pp.719-722, 2000.

[146] Y. Taur, “An analytical solution to a double-gate MOSFET with undoped body,” *IEEE Electron Device Lett*., vol.**21**, pp.245-247, 2000.

[147] A. Ghetti, E. Sangiogi, J. Bude, T.W. Sorsch, and G. Weber, “Tunneling into interface states as relaibility monitor for ultrathin oxides”, *IEEE Trans. Electron Devices*, vol.**47**, pp.2358-2365, 2000.

[148] C. H. Lee, J. J. Lee, W. P. Bai, S. H. Bae, J. H. Sim, X. Lei, R. D. Clark, Y. Harada, M. Niwa, and D. L. Kwong, “Self-aligned ultra thin HfO_2 CMOS transistors with high quality CVD TaN gate electrode, ” in *Symp. VLSI Tech.*, pp.82-83, 2002.

[149] C. Bowen, C. L. Fernando, G. Klimeck, A. Chatterjee, D. Blanks, R. Lake, J. Hu, J. Davis, M. Kulkarni, S. Hattangady, and I.C. Chen, “Physical oxide thickness extraction and verification using quantum mechanical simulation,” in *IEDM Tech*. Dig., pp.869-872 1997.

[150] S. Song, H. J. Kim, J. Y. Yoo, J. H. Yi, W. S. Kim, N. I. Lee, K. Fujihara, H. K. Kang, and J. T. Moon, “On the gate oxide scaling of high performance CMOS transistors,” in *IEDM Tech Dig*., pp.55-58, 2001.

[151] C. H. Choi, K. Y. Nam, Z. Yu, and R. Dutton, “ Impact of gate direct tunneling current on circuit performance: A simulation study,” *IEEE Electron Device Lett*. vol.**48**, pp.2823-2829, 2001.

[152] Y. Omura, S. Horiguchi, M. Tabe, and K. Kishi, "Quantum-mechanical effects on the threshold voltage of ultrathin-SOI nMOSFETs", IEEE Electron Device Lett., vol.**14**. , pp.569-571, 1993.

[153] W. K. Choi, D. Ha, T. J. King, C. Hu, "Ultra-thin body PMOSFETs with selectively deposited Ge source/drain," in *Symp. VLSI Tech.*, pp.19-20, 2001.

[154] T. Mizuno, N. Sugiyama, T. Tezuka, Y. Moriyama, S. Nakaharai, and S. Takagi, "(110)-surface strained-SOI CMOS devices with higher carrier mobility," in *Symp. VLSI Tech.*, pp.97-98. 2003

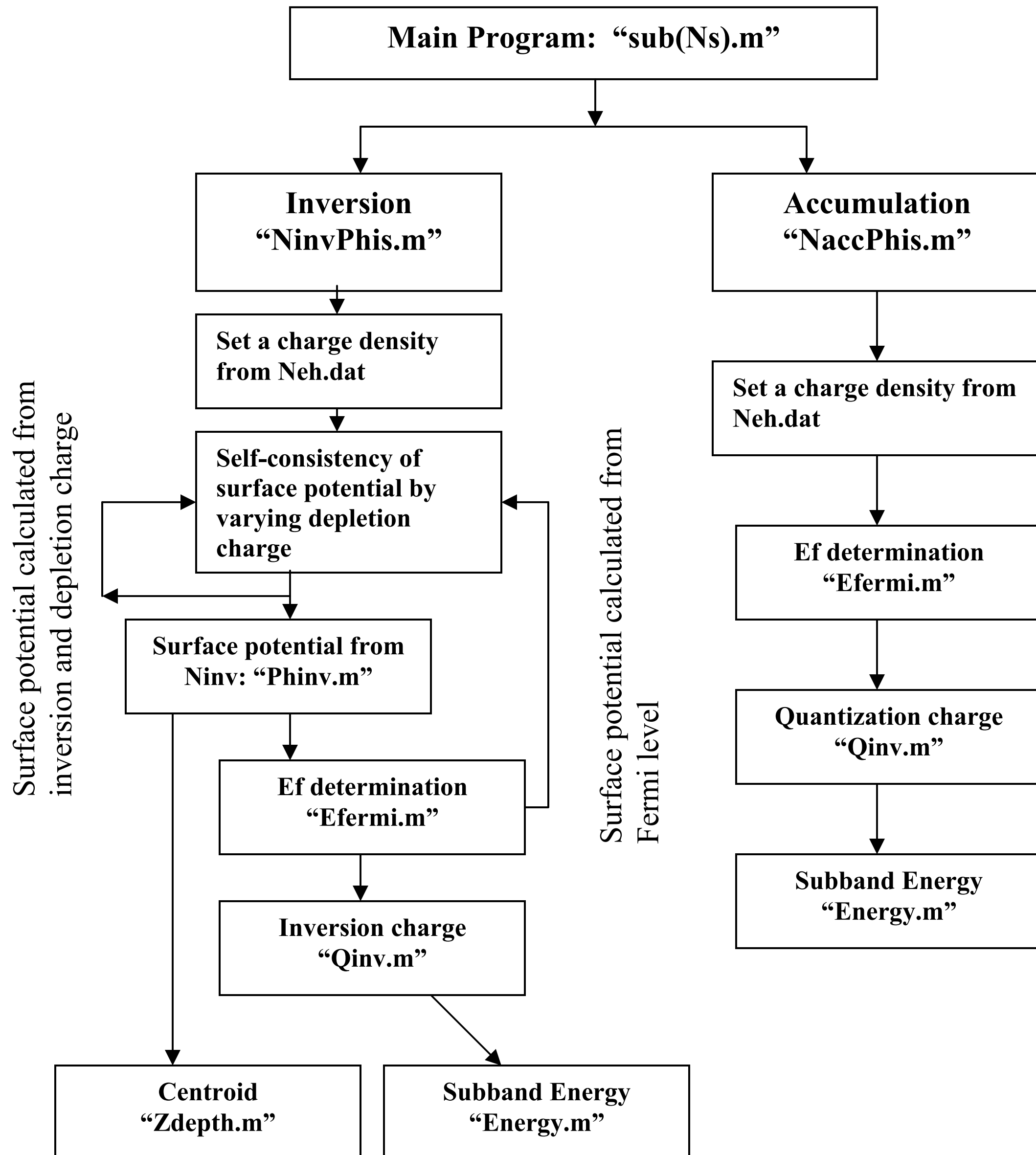
[155] M. V. Fischetti, Z. Ren, P. M. Solomon, M. Yang, and K. Rim, "Six-band k.p calculation of the hole mobility in silicon inversion layers: dependence on surface orientation, strain , and silicon thickness," *J. Appl. Phys.,* vol.**94**, pp.1079-1095, 2003.

[156] H. Takeuch and T.J. King, "Scaling limits of hafnium-silicate films for gate-dielectric applications," *Appl. Phys. Lettt.,* vol.**83**, pp.788-790, 2003.

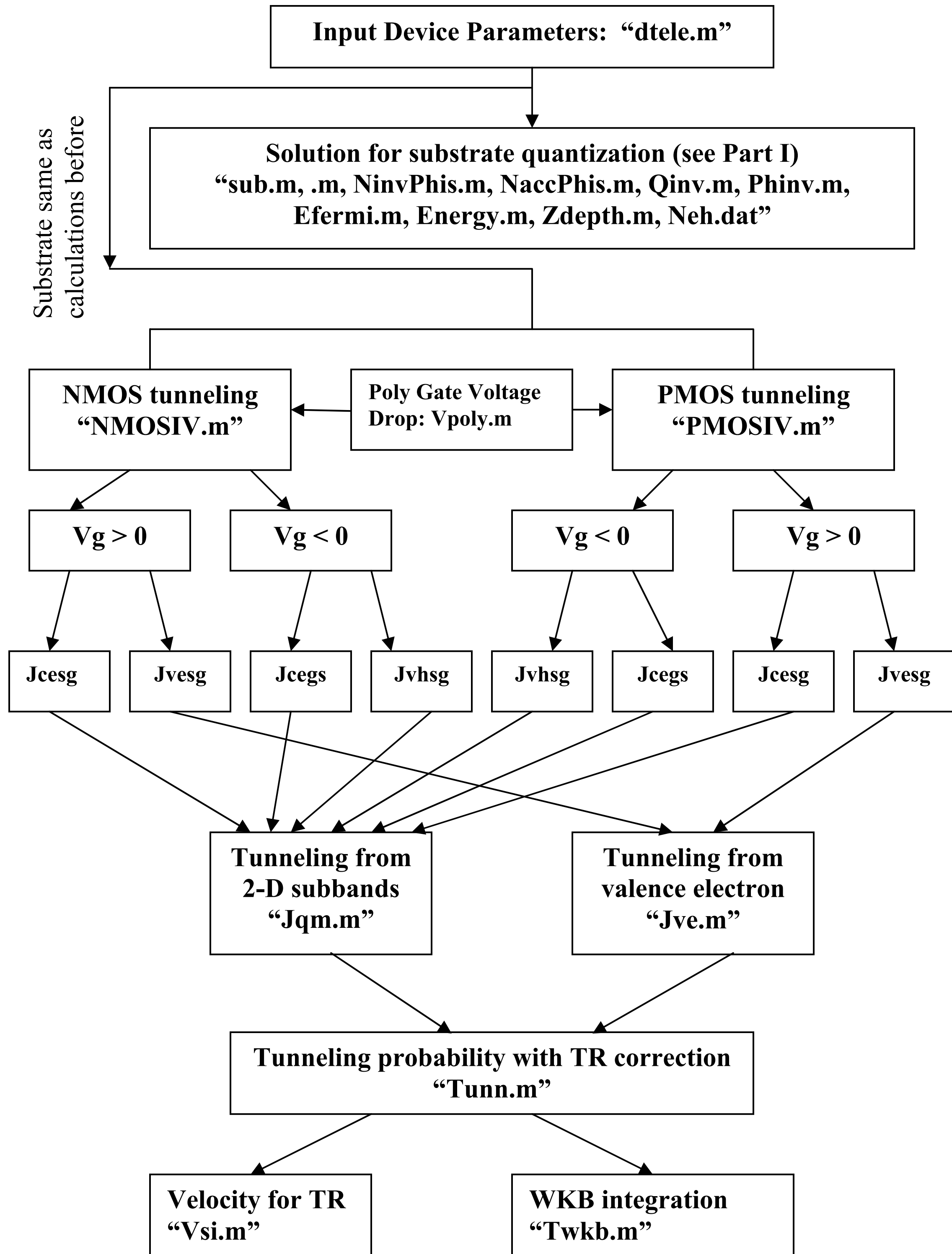
[157] Y. Y. Fan, Q. Xiang, J. An, L. F. Register, and S. Banerjee, "Impact of interfacial layer and transition region on gate current performance for high-K gate dielectric stack: its tradeoff with gate capacitance," *IEEE Trans. Electron Devices,* vol.**50**, pp.433-439, 2003.

Brief Descriptions of Simulation Programs

I. Program Flow Chart for Quantization



II. Program Flow Chart for Direct Tunneling



Jabcd: a-band (conduction/valence), b-carrier (electron/hole); c-injecting, d-outgoing electrodes.

List of Publications

Journals:

1. **Y. T. Hou** and M. F. Li, "A novel simulation algorithm for Si valence hole quantization of inversion layer in metal-oxide-semiconductor devices," *Jpn. J. Appl. Phys. Part 2*, **40**, L144 (2001).
2. **Y. T. Hou** and M. F. Li, "Hole quantization effects and threshold voltage shift in pMOSFET -- assessed by improved one-band effective mass approximation," *IEEE Tran. Electron Devices*, **48**, 1188 (2001)
3. **Y. T. Hou** and M. F. Li , "A Simple and Efficient Model for Quantization Effects of Hole Inversion Layers in MOS Devices," *IEEE Tran. Electron Devices*, **48**, 2893 (2001).
4. **Y. T. Hou**, M. F. Li, W. H. Lai, and Y. Jin, "Modeling and characterization of direct tunneling hole current in p-MOSFETs," *Appl. Phys. Lett.*, **78**, 4034 (2001).
5. **Y. T. Hou**, M. F. Li, Y. Jin, and W. H. Lai, "Direct tunneling hole current through ultrathin gate oxides in metal-oxide-semiconductor devices," *J. Appl. Phys.*, **91**, 258 (2002).
6. H. Y. Yu, **Y. T. Hou**, M. F. Li, and D. L. Kwong, "Hole Tunneling Current through Oxynitride /Oxide Stack and the Stack Optimization for p-MOSFET's," *IEEE Electron Device Lett.*, **23**, 285 (2002).
7. H. Y. Yu, **Y. T. Hou**, M. F. Li, and D. L. Kwong, "Investigation of Hole Tunneling Current through Ultrathin Oxynitride/Oxide Stack Gate Dielectrics for p-MOSFET's," *IEEE Trans. Electron Devices*, **49**, 1158 (2002)

8. **Y. T. Hou**, M. F. Li, D. L. Kwong, "Modeling of Tunneling Currents Through HfO_2 and $(HfO_2)_x(Al_2O_3)_{1-x}$ Gate Stacks," *IEEE Electron Device Lett.*, **24**, 96 (2003).

9. T. Low, **Y. T. Hou**, M. F. Li, "Improved One-band Self-consistent Effective Mass Methods for Hole Quantization in p-MOSFET," *IEEE Trans. Electron Devices*, **50**, 1284 (2003).

10. **Y. T. Hou**, M. F. Li, T. Low, and D. L. Kwong, "Metal Gate Work Function Engineering on Gate Leakage of MOSFETs," submitted to *IEEE Trans. Electron Devices.*

Conferences:

1. **Y. T. Hou**, M. F. Li, and Y. Jin, "Hole quantization and hole direct tunneling in deep submicron p-MOSFETs," (**invited paper**) *International Conference on Semiconductor Integrated Circuit Technology*, Shanghai, China, p.895, 2001.

2. **Y. T. Hou**, M. F. Li, W. H. Lai, and Y. Jin, "A Physical Model for Hole Direct Tunneling Currents Through Ultrathin Gate Dielectrics in Advanced CMOS Devices," in the extended abstract of the International Conference on Solid State Devices and Materials (SSDM), Tokyo, Japan, p.144, 2001.

3. **Y. T. Hou**, M. F. Li, H. Y. Yu, Y. Jin, and D. L. Kwong, "Quantum tunneling and scalability of HfO_2 and $HfAlO_2$ gate stacks," International Electron Device Meeting (*IEDM*), Francisco, USA, pp.731-734. 2002.

4. T. Low, **Y. T. Hou**, M. F. Li, C. Zhu , D. L. Kwong, and A. Chin, "Germanium MOS: An Evaluation from Carrier Quantization and Tunneling Current," Symposium on VLSI Technology, Kyoto, Japan, p.177-178, 2003.

5. Y. Jin, W. Y. Teo, **Y. T. Hou**, F. H. Gn, H. F. Lim, Z. Y. Han, and M. F. Li, “Enhanced Plasma Charging Damage due to AC Charging Effect, International Reliability Physics Symposium (IRPS),” Dallas, USA, pp.359-365, 2002.
6. T. Low, **Y. T. Hou**, M. F. Li, C. Zhu, A. Chin, G. Samudra and D. L. Kwong, “Investigation of Performance Limits of Germanium Double-Gated MOSFETs, International Electron Device Meeting (IEDM), Washington, USA, pp.691-694, 2003.