NANO-SCALED LOGIC AND MEMORY DEVICES: MODELING AND FABRICATION

by

Peiqi Xuan

Memorandum No. UCB/ERL M04/13

11 December 2003

NANO-SCALED LOGIC AND MEMORY DEVICES: MODELING AND FABRICATION

by

Peiqi Xuan

Memorandum No. UCB/ERL M04/13

11 December 2003

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

Nano-Scaled Logic and Memory Devices: Modeling and Fabrication

by

Peiqi Xuan

B.S. (Peking University, China) 1996
M.S. (University of California, Berkeley) 2000

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering – Electrical Engineering
and Computer Sciences

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor Jeffrey Bokor, Chair
Professor Tsu-Jae King
Professor Oscar Dubon

Fall 2003

The dissertation of Peiqi Xuan is approved:

12/3/03

Chair Date

12/3/03

Date

12/3/03

Date

University of California, Berkeley

Fall 2003

Nano-Scaled Logic and Memory Devices: Modeling and Fabrication

by

Peiqi Xuan

Abstract

Nano-Scaled Logic and Memory Devices: Modeling and Fabrication

by

Peiqi Xuan

Doctor of Philosophy in Engineering –
Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Jeffrey Bokor, Chair

This dissertation investigates both the modeling and fabrication of ultra-thin-body (UTB) and double gate (DG) MOSFETs, which are proposed to suppress short channel effects (SCE) in nano-scaled MOSFETs. An analytic model is developed to evaluate the effectiveness of the structures. The minimum channel length with certain performance criteria can be derived from the physical dimensions of the transistor. The 2D effects in both the body and the high κ gate dielectric are included. The influences of high body doping and pocket implants on SCE are also modeled. The results of the analytical model form the basis of the subsequent discussion of device design and fabrication.

Lateral solid-phase-epitaxy (SPE) is a practical approach to realizing the UTB structure with good uniformity and controllability of the thin Si channel film. SPEFETs are fabricated, and the quality of the SPE films is investigated. Within a short SPE range (≤60nm), the resulting film has good quality close to that of a perfect Si film, and good device performance has been achieved. The easy integration of SPEFET with bulk MOSFET makes it suitable for sub-50nm device generations.

A correct threshold voltage (V_t) can be achieved only by gate workfunction engineering in sub-30nm transistors. NiSi is proposed as a single gate material for multiple V_t CMOS applications because the workfunction of NiSi can be continuously adjusted over a large range by dopants implanted into the silicon film before silicidation. Furthermore, the NiSi gate has excellent compatibility with the current CMOS process because it causes no degradation of the resulting MOSFET performance. After all, nickel silicide is highly advantageous as a single gate material for future CMOS technologies.

The fully depleted structure is also applied to flash memory, and the resulting FinFET SONOS can be successfully scaled to sub-40nm. The large V_t windows and high current ratio between programmed/erased states enable multi-bit storage for even higher storage density. Good program/erase speeds, endurance and retention are demonstrated in FinFET SONOS memory devices. Devices fabricated on (100) sidewall surfaces show more resistance to electrical stress than do (110) devices. The FinFET SONOS device is a promising candidate for sub-100nm embedded flash memories.

The dissertation abstract of Peiqi Xuan is approved:

12/4/03

Professor Jeffrey Bokor Date

Committee Chair

To my family

For their love, support and encouragement

Table of Contents

Chapter 1 Introduction 1
1.1 MOSFET scaling 1
1.2 Ultra-thin-body (UTB) and double-gate (DG) devices 4
1.3 Silicide gate electrodes 7
1.4 FinFET SONOS flash memory 8
1.5 References 9

Chapter 2 Subthreshold Model for Fully Depleted SOI Transistors 15
2.1 Introduction 15
2.2 Performance characteristics 16
2.3 The Poisson equation in the body 18
2.4 General solution and its simplification 20
2.4.1 General solution 20
2.4.2 2nd order approximation 21
2.4.3 Subthreshold swing and Vt definition 22
2.5 Uniformly doped body 23
2.5.1 Potential profile and short channel effects 23
2.5.2 Critical doping and leakage depth 26
2.6 Channel with pocket doping 30
2.6.1 Long channel model 30
2.6.2 Short channel model 31
2.6.3 Model verification 33
2.7 4th order approximation 35
2.8 2D effects in the gate dielectric 36
2.9 Conclusion 40
2.10 References 41
Appendix 2A Simplification of DV 45
Appendix 2B Calculation of the barrier and V_t with uniform doping 46

Chapter 3 Solid Phase Epitaxy for Ultra-Thin-Body MOSFETs 47
3.1 Introduction 47
3.2 Fabrication process 49
3.2.1 Lateral solid-phase-epitaxy 49
3.2.2 SPEFET fabrication 52
3.3 Results and discussion 54
3.3.1 60nm CMOS performance 54
3.3.2 Defects and channel length dependence 55
3.3.3. SPE Twin boundary 57
3.3.4 SPE film quality from activation energy 57
3.3.5 Film thickness dependence 60
3.3.6 Trench size dependence 61
3.4 Conclusion 62
3.5 References 63
Appendix 3A Process flow for SPEFETs 66

Chapter 4 Silicide Gates for Workfunction Engineering 68
4.1 Introduction 68
4.2 Motivation for silicide gates 69
4.2.1 Metal gates 69
4.2.2 Silicide gates 71
4.3 NiSi and TiSi workfunction extraction 72
4.3.1 Fabrication process for MOSCAPs 72
4.3.2 Workfunction extraction 75
4.4 CMOS with silicide gates 78
4.4.1 Fabrication process for MOSFETs 78
4.4.2 Bulk PMOS results with NiSi gate 79
4.4.3 SOI PMOS results with NiSi gate 82
4.4.4 Bulk NMOS results with TiSi gate 84
4.5 Conclusion 87
4.6 References 88
Appendix 4A Process flow for bulk silicide gate MOSFETs 92
Appendix 4B Process flow for SOI silicide gate MOSFETs 94

Chapter 5 FinFET SONOS Flash Memory 95
5.1 Introduction 95
5.1.1 Operation of flash memory 95
5.1.2 Flash memory scaling and SONOS memory 97
5.1.3 FinFET SONOS memory 99
5.2 Fabrication of FinFET SONOS memory 100
5.3 Performance characteristics 103
5.3.1 P/E speeds of FinFET SONOS memory 103
5.3.2 Reliability of FinFET SONOS memory 105
5.3.3 Retention and reading operation 108
5.3.4 Scalability of FinFET SONOS memory 110
5.4 Conclusion 111
5.5 References 112
Appendix 5A: Process flow for FinFET SONOS memory 117

Chapter 6 Conclusion 119
6.1 Summary 119
6.2 Suggestion for future works 121
6.2.1 Integration of NiSi gate with FinFET and UTBFET 121
6.2.2 Tunneling FET for sub 60mV/dec subthreshold swing 122
6.3 References 124

Acknowledgments

First, I would like to express my thanks to and admiration for my advisor, Professor Jeffrey Bokor, whose unfailing support, encouragement and supervision has made this journey possible. His openness and flexibility have given me the independence to do my research, while his strictness with respect to English has driven me to improve my writing skills. From the very beginning, he has encouraged me to pursue my own ideas, but he is always available for guidance and suggestions with his keen insights into solid-state physics.

I humbly thank Professor Chenming Hu and Tsu-Jae King, not only for serving as chair for my qualifying exam and member of my dissertation committee, but also for their insightful advice over the last six years, which has been crucial to the success of my projects. I would like to thank Professor Jan Rabaey of EECS Department and Professor Oscar Dubon of MSE Department for serving on my qualifying exam and/or dissertation committee. Their guidance and suggestions have benefited me greatly. I would also like to thank Vivek Subramanian, whose broad knowledge of device physics, material science and fabrication techniques has always been my last resort for help.

My sincere thanks go to my senior colleague Jakub Kedzierski. It has always been instructive talking with him. The idea of the silicide gate was first conceived in our discussion right before his graduation. His passion for and experience in microfabrication were a great help in my overcoming the challenges when I joined the group.

To Min She and Qiang Lu, who cooperated with me on the SONOS memory and metal gate projects, respectively. Without their deep knowledge of and help in fabrication, these research projects would not have been so fruitful. I am also very grateful to Hideki

Takeuchi for his invaluable help and discussion on process issues.

I am also indebted to Mrs. Palma Lower, who has proofread my Ph.D. dissertation and my Master report as well. Her professional help with my written English has given this dissertation a sound English style and helped it meet the requirements of my advisor.

I pay special tribute to Eric Anderson, Alex Liddle and Bruce Harteneck from the Lawrence Berkeley National Laboratory. Without their help with the electron beam lithography, this work would not have been possible. Their knowledge about and skills in generating nano-scaled features have enabled me to research on the frontier of device scaling.

My appreciation also goes to the UC Berkeley Microlab staff for all their help. Jimmy Chang has always been available to answer my process questions, and David Lo constantly asked to switch the Heatpulse1 chamber for me. Marilyn Kushnar has helped generate dozens of masks, and Hongbin Liu has made the time I spent in the Microlab pleasant. Without their help, none of my processes could have been realized.

During my stay in Berkeley, many people have assisted me and offered their support. Here I would like to acknowledge my former and present colleagues: Yee-Chia Yeo, Yan Wang, Stephen Tang, Nick Lindert, Yang Kyu Choi, Donggun Park, Wen-Chin Lee, Ya-Chin King, Mark Cao, Xiying Xiong, Qing Ji, Leland Chang, Cathy Huang, Kevin Cao, Daewon Ha, Pushkar Ranado, Hui Wan, Jane Xi, Gang Liu, Liuyung Wong, Kyoungsub Shin, Yu-Chih Tseng, Pin Su, and Jin He.

I would like to sincerely thank my parents and sister for their support, sacrifices and encouragement in the past 30 years. Finally, my deepest gratitude to my wife, Fan Si, who has always stood beside me in the peaks and valleys of my life.

Chapter 1

Introduction

1.1 MOSFET scaling

In the past few decades, the world semiconductor market has grown explosively mainly due to the steady improvement in circuit performance made possible by the scaling of MOSFETs (Metal-Oxide-Semiconductor Field Effect Transistors). Since the 1960s, transistor dimensions have been shrinking 30% every 3 years, as predicted by Moore's law [1,2]. This reduction of the device gate length has improved both circuit speed and density (Fig. 1.1).

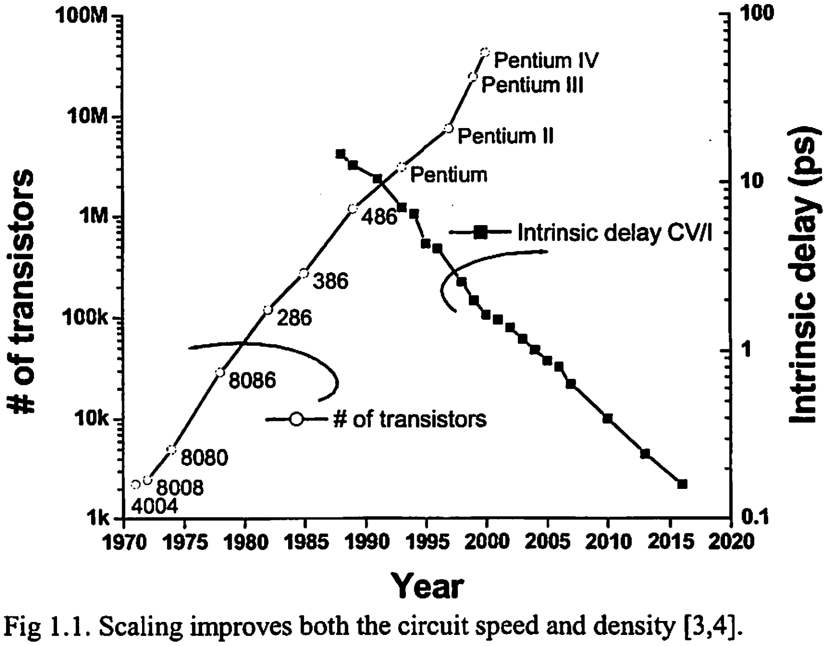
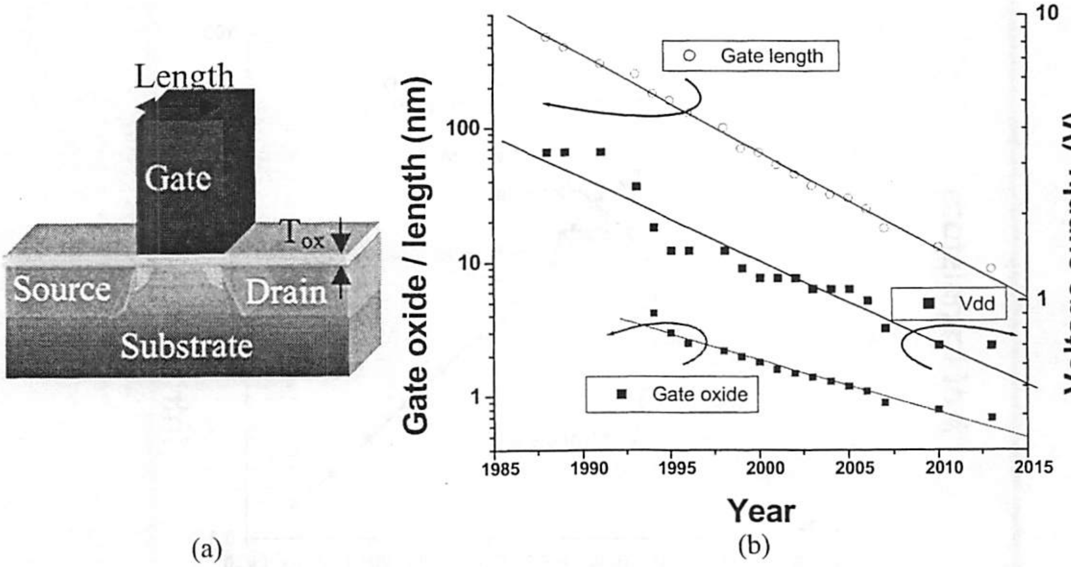


Fig 1.1. Scaling improves both the circuit speed and density [3,4].

Currently, the challenges that prevent devices from being scaled down are process difficulties and short channel effects (SCE), such as current leakage, subthreshold swing

(S), drain induced barrier lowering (DIBL) and threshold voltage (V_t) roll-off [5]. Although this scaling will not continue indefinitely, the present device sizes are still far from the fundamental limits of physics, such as atomic size and quantum effect limits. Simulation shows that devices with a gate length of less than 10nm are feasible and still not dominated by tunneling phenomenon [6]. For each technological generation, many innovations have to be made to overcome the process difficulties and SCE [7]. In the sub-micron region, constant field scaling is the major method for scaling down the transistor to maintain its reliability. In this approach, the supply voltage and vertical dimensions are scaled together with the gate length (Fig. 1.2). Currently, the state-of-the-art production technology is at a gate length of around 70nm, while sub-10nm devices are under investigation in laboratories [8].



(a) (b)

Fig 1.2. (a) Schematic of a conventional bulk MOSFET. (b) Scaling the supply voltage and oxide thickness together with the gate length.

In an ideal MOSFET, the gate electrode has complete control of the body potential; therefore, ideal performance can be achieved, i.e. S=kT•ln(10)=60mV/dec at room temperature. When the channel length becomes shorter, the two-dimensional (2D) distribution of the potential in the channel becomes important. The source/drain and body compete with the gate in controlling the channel potential. This drives the scaling of the vertical dimensions together with the gate length to minimize the 2D effect. However, further scaling imposes great technological challenges in the manufacturing process.

The gate oxide thickness (T_{ox}), source/drain junction (X_j) and depletion depth (W_{dep}) have been the three major vertical dimensions being scaled aggressively in the past. However, they are all approaching their limits. Gate oxide thickness is limited by the tunneling current from the gate electrode to the channel [9]. High κ materials are proposed to suppress the tunneling current while maintaining the same effective oxide thickness (EOT) [10,11]. Although the gate leakage tolerance continues to increase from generation to generation, gate oxide scaling is slowing down and will eventually become saturated. The scaling of the junction depth is prevented by the high sheet resistance of the source/drain, which reduces the drive current and device speed [12], and process issues. Although high channel doping can achieve narrow depletion depth, it also degrades the carrier mobility in the inversion layer and thus the drive current [13]. The *pn* junction current leakage might also become the limiting factor of scaling when high body doping is used [14].

Table 1.1 is an excerpt from the International Technology Roadmap for Semiconductors (ITRS) of 2002 [4], which predicts the performance, structure, and dimensions of future integrated circuit (IC) technologies. As pointed out in the table, many innovations have to be made to meet the technological requirements.

	Near term							*Long term*		
Year	*2001*	*2002*	*2003*	*2004*	*2005*	*2006*	*2007*	*2010*	*2013*	*2016*
Technology node(nm)	*130*	*115*	*100*	*90*	*80*	*70*	*65*	*45*	*32*	*22*
Gate length (nm)	*65*	*53*	*45*	*37*	*32*	*28*	*25*	*18*	*13*	*9*
Gate oxide (Å)	*13–16*	*12–15*	*11–14*	*9–14*	*8–13*	*7–12*	*6–11*	*5–8*	*4–6*	*4–5*
Drain extent X_j (nm)	*27-45*	*22-36*	*19-31*	*15-25*	*13-22*	*12-19*	*10-17*	*7-12*	*5-9*	*4-6*
Channel doping for $W_{dep}<L_{eff}/4$ (cm^{-3})	*4.0E18*	*6.0E18*	*8.0E18*	*1.1E 19*	*1.4E 19*	*1.6E19*	*2.3E19*	*1.0E19*	*1.3E20*	*5.0E20*
Channel doping for V_t =0.4V (cm^{-3})	*0.8–1.5 E18*	*0.8–1.5 E18*	*1.5–2.5 E18*	*1.5–2.5 E18*	*1.5–2.5 E18*	*2.0–4.0 E18*	*2.5–5.0 E18*	*5.0–9.0 E18*	*0.9–1.5 E19*	*1.5–3.0 E19*
Poly doping for 25% depletion (cm^{-3})	*9.2E19*	*9.2E19*	*1.14E20*	*1.50E20*	*1.66E20*	*1.66E20*	*1.87E20*	*2.58E20*	*2.99E20*	*3.74E20*
Gate sheet Rs (Ω/sq)	*5*	*5*	*5*	*5*	*5*	*5*	*5*	*5*	*6*	*7*

□ Solutions exist ▨ Solutions being pursued ■ No known solutions

Table 1.1 Relevant entries from ITRS 2002 update for front-end processes.

1.2 Ultra-thin-body (UTB) and double-gate (DG) devices

As indicated in the above table, significant challenges have to be overcome to extend Moore's law into the future, and new structures and devices have been active research topics for many years [15,16]. Because modern circuitry consists of millions of transistors, a minor modification in the manufacturing process can result in significant degradation in the yield. A process with minimum deviation from the current technology would be highly appreciated for easy migration. Ultra-thin-body devices, based on the maturing SOI (silicon-on-insulator) technology [17], overcome most of the scaling challenges discussed previously and are the most promising structures for nano-scaled devices.

MOSFETs fabricated on SOI wafers have become part of the semiconductor industry standard [18]. Currently, SOI technology uses a thick silicon layer of 50-100nm, which is partially depleted. Both the fabrication and characteristics of the devices are similar to those of bulk MOS. The advantages of SOI devices are that they eliminate the source/drain junction capacitance, which slows down the circuit, and the junction leakage, which consumes power even at idle. However, the devices suffer from the floating body effect due to the existence of an isolated neutral region in the body [19].

The UTB structure (Fig 1.3a), on the other hand, uses an ultra-thin, fully depleted silicon film as the channel [20]. It provides the following advantages over the conventional bulk MOS or partially depleted SOI devices. First, it doesn't have the floating body effect because there is no neutral region inside the body. Second, in this device structure, both the junction depth and depletion depth are determined by the thin silicon film thickness, which can be precisely controlled by the process. Ultra shallow junctions can be fabricated by making the channel film thinner than the junction depth possibly made by bulk MOS technology. Moreover, the current leakage path at a low gate bias is typically not at the surface, but deep in the depletion region, which is far away from and less effectively controlled by the gate terminal. If the body thickness is reduced, the leakage path is forced to be closer to the surface, which ensures a stronger gate control and less leakage.

UTB devices can be further optimized to double-gate devices (Fig 1.3b). In this structure, the channel potential is controlled by two connected gates, which ensures better gate-control and SCE. To the first order, a DG MOSFET is just two UTB FETs standing back to back. Therefore, twice the body film thickness of a UTB device can be tolerated, while maintaining the same SCE. Actually, it provides even better SCE due to the elimination of the buried oxide, which is also a medium for the penetrating fields from the source/drain to the channel [21]. This relaxed requirement for the body film thickness is highly advantageous since the formation of a uniform ultra-thin film is a major challenge in this device's fabrication. Simulation results show that a DG MOSFET has the best scalability and can be successfully scaled down to sub-10nm gate length devices [6]. The detailed mechanism of UTB and DG transistor suppression of SCE will be modeled in Chapter 2.

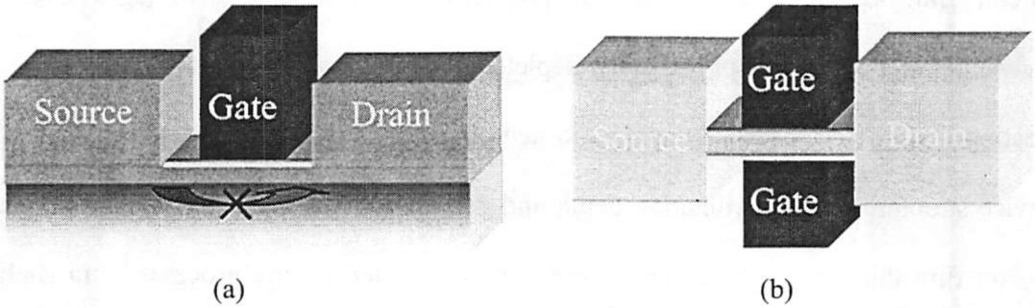


Fig. 1.3 Schematic of (a) a UTB MOSFET (b) a double-gate MOSFET.

Moreover, UTB and DG devices provide further benefits due to the fact that no high body doping is required to minimize the depletion depth. Actually, a lightly doped or even an undoped thin film can be used as the channel, which improves the device in the following ways:

a. Low doping in the silicon film results in less impurity scattering and thus higher carrier mobility in the inversion layer. Therefore, better performance of the transistor can be obtained.

b. Without the depletion charge in the body, a lower vertical field can be realized with the same driving force. Therefore, carrier mobility can be further enhanced. [22]

c. In a bulk MOS, the body doping determines the threshold voltage. In an ultra-scaled device, the number of doping atoms in the channel decreases due to the reduced volume, and the doping fluctuation causes V_t fluctuation [23,24]. Since body doping is not crucial in UTB devices, it does not cause V_t fluctuation.

UTB devices offer a way of creating ultra shallow depletion and junction depths, but they still have the disadvantage of the high resistance resulting from the ultra thin film. Raised source and drain are also introduced to improve the external resistance of the device and make the later metallization scheme feasible [25]. The formation of an ultra-thin but still uniform Si film with high crystalline quality has become the major challenge in the fabrication of UTB devices. For example, the channel film thickness has to be scaled down to less than 5nm when the channel length is scaled to 20nm. Many approaches have been proposed, and a promising one, solid-phase-epitaxy (SPE), will be addressed in Chapter 3.

1.3 Silicide gates

Table 1.1 indicates that polysilicon is no longer a good gate material in deeply scaled technologies. The poly depletion effect adds about 0.5nm to the physical gate oxide thickness, making the already difficult oxide scaling even more difficult [26]. The high resistivity of polysilicon degrades the circuit speed at high frequencies [27]. A metallic gate is required to improve the gate conductance and eliminate the gate depletion layer.

As mentioned, a UTB MOSFET does not require channel doping for its controlling of SCE. Actually, the transistor performance does not depend on either the body doping type or the concentration as long as the concentration is within a quite large range ($<1\times10^{18}$cm^{-3}). Although this property provides UTB MOSFETs with immunity to the dopant fluctuations, it also rules out the possibility of adjusting V_t through body doping, which has been used successfully in the conventional bulk MOSFETs for decades. Other methods must be developed to achieve appropriate threshold voltages in UTB devices.

Gate workfunction engineering seems the most feasible way to manipulate V_t in a UTB or DG MOSFET [6]. Two gate materials with the correct workfunctions, for NMOS and PMOS, respectively, are required for the implementation of these structures, and polysilicon no longer meets the workfunction requirement. To meet the roadmap specification, new metallic gate materials have to be introduced into the process, which significantly complicates the IC manufacture.

Nickel silicides formed from silicon films with various doping levels are promising candidates for future gate electrodes. First, NiSi is metallic and has a resistivity of 6μΩ•cm, much lower than that of polysilicon. Second, the workfunction of NiSi depends on the type and dose of ion implanted into the silicon film before the formation of the silicide, and its range covers the desired workfunctions for both NMOS and PMOS [28]. More detailed results will be presented in Chapter 4.

1.4 FinFET SONOS flash memory

Since mobile electronics, such as cellular phones, digital cameras, personal digital assistants, and global positioning systems, are widely used, nonvolatile memory (NVM) devices have become an indispensable semiconductor electronics component because they provide 10 years of retention time even without a power supply. Among all NVM devices, SONOS (silicon-oxide-nitride-oxide-silicon) flash has shown the best scalability and the size of the flash cell has been scaled as quickly as the logic devices to achieve the ultra high capacity of memory chips [29]. A SONOS flash memory cell is simply a

MOSFET with an extra silicon nitride film sandwiched between the tunnel oxide and the inter-poly oxide to form a charge storage layer [30]. Electrons tunnel into and out of the Si_3N_4 layer in programming and erasing, respectively. The charge in the Si_3N_4 layer alters the V_t of the transistor, through which the stored information can be determined.

A SONOS gate stack shows significantly advantages over a floating gate stack by storing charges in trap states inside the sandwiched nitride layer. Since the traps are isolated from each other, even if a defect path forms in the tunnel oxide, most charge will remain in the nitride and the information can be still retained [31]. Because the stored charge can leak out of the nitride layer through the thin oxide even when the device is at idle, the 10 years retention time sets the minimum tunnel oxide thickness in a SONOS device. Significant V_t window closure after 10 years retention can be observed when the tunnel oxide is below 2nm [32]. After all, the minimum EOT of the SONOS gate stack is around 7nm, and scaling the flash memory beyond 100nm is very challenging.

On the other hand, double-gate MOSFETs have demonstrated the best scalability by offering an alternative way of scaling: the thinning of the body. Because the fabrication of a self-aligned bottom gate imposes tremendous process challenges, many novel structures and processes have been proposed. Among all DG structures proposed so far, the FinFET (Fig 1.4a) is the most manufacturable because it eliminates the need for the bottom gate by putting channels on the two sidewalls of the silicon fin [33]. In Chapter 5, a process combining the FinFET and SONOS technology is described. Fig. 1.4b shows the cross section of the gate stack in a SONOS FinFET cell. The excellent performance and scalability of such a device will be demonstrated.

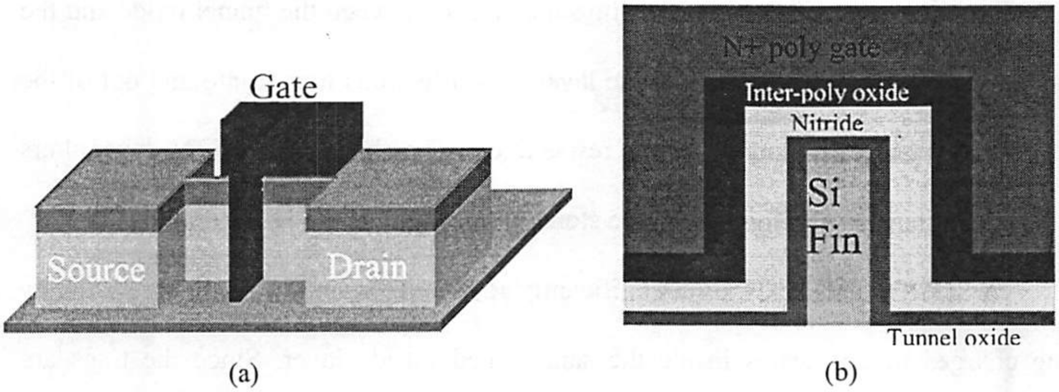


Fig. 1.4 (a) Schematic of a FinFET (b) Cross-section of the SONOS cell gate stack

1.5 References

[1] M.T. Bohr, "Nanotechnology goals and challenges for electronic applications," *IEEE, Trans. On Nanotechnology, Vol. 1, No. 1, pp. 56-62, March 2002*

[2] P.A. Packan: "Device physics: Pushing the limits". *Science. 285(5436), pp. 2079-2081, Sep. 1999*

[3] Intel technology and manufacturing research: silicon showcase of Moore's law, *http://www.intel.com/research/silicon/mooreslaw.htm*

[4] International Technology Roadmap for Semicondutors, Semicondutor Industry Association, 2001. *http://public.itrs.net/Files/2002Update/2001ITRS/Home.htm*

[5] D.J. Frank and Y. Taur, "Design considerations for CMOS near the limits of scaling," *Elsevier. Solid-State Electronics, Vol. 46, No. 3, pp. 315-20, March 2002*

[6] L. Chang, S. Tang, T-J. King, J. Bokor and C. Hu, "Gate length scaling and threshold voltage control of double-gate MOSFETs," *Proceedings of the International Electron Device Meeting, pp. 719-722, Dec. 2000*

[7] H-SP. Wong, D.J. Frank, P.M. Solomon, C.H.J. Wann and J.J. Welser, "Nanoscale

CMOS." *Proceedings of the IEEE, Vol. 87, No. 4, pp. 537-570, April 1999*

[8] 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 10nm gate length", *Proceedings of the International Electron Device Meeting, pp. 251-254, Dec. 2002*

[9] N. Yang, W.K. Henson and J.J. Wortman, "Analysis of tunneling currents and reliability of NMOSFETs with sub-2 nm gate oxides," *IEEE. Proceedings of International Electron Devices Meeting, pp. 453 -456. Dec. 1999*

[10] S.A. Campbell, D.C. Gilmer, X. Wang, M. Hsieh, H-S. Kim, W.L. Gladfelter and J. Yan, "MOSFET Transistor Fabricated with High Permitivity TiO_2 Dielectrics". *IEEE Trans. Elec. Dev. Vol. 44, No. 1, pp. 104-109, Jan. 1997*

[11] C.H. Choi, S.J. Rhee, T.S. Jeon, N.Lu, J.H. Sime, R. Clark, M. Niwa and D.L. Kwong, "Thermally stable CVD HfO_xN_y advanced gate dielectric with poly-Si gate electrode", *Proceedings of the International Electron Device Meeting, pp. 857-860, Dec, 2002*

[12] P. Keys, H.J. Gossmann, K.K. Ng and C.S. Rafferty, "Series resistance limits for 0.05um MOSFETs," *Supperlattices and Microstructure, Vol. 27, No. 2-3, pp. 125-136, 2000*

[13] S. Takagi, A. Toriumi, M. Iwase and H. Tango, "On the universality of inversion layer mobility in Si MOSFET's: part I – effects of substrate impurity concentration," *IEEE, Trans. on Elec. Devices, Vol. 41, No. 12, pp. 2357-2362. Dec. 1994*

[14] T. Ghani, K. Mistry, P. Packan, S. Thompson, M. Stettler, S. Tyagi and M. Bohr,

"Scaling challenges and device design requirements for high performance sub-50nm gate length planar CMOS transistors," *Proceeding of the 2000 symposium on VLSI technology, pp. 174-175, June 2000*

[15] D.P. DiVincenzo, "Prospects for quantum computing," *Proceedings of the International Electron Device Meeting 2000, pp. 12-15, Dec. 2000*

[16] S.J. Wind, J. Appenzeller, R. Martel, V. Derycke and P. Avouris, "Vertical scaling of carbon nanotube field-effect transistors using top gate electrodes" *Applied Physics Letters, Vol. 80, No. 20, pp. 3817-3819, May, 2002*

[17] M. Yoshimi. "MOS scaling crisis and SOI technology." *IEEE. Proceedings of 6th International Conference on Solid-State and Integrated Circuit Technology, pp.637-642, Vol. 1, 2001*

[18] M.M. Pelella, W. Maszara, S. Sundararajan, S. Sinha, A. Wei, D. Ju, W. En, S. Krishnan, D. Chan, S. Chan, P. Yeh, M. Lee, D. Wu, M. Fuselier, R. VanBentum, G. Burbach, C. Lee, G. Hill, D. Greenlaw, C. Riccobenc, O. Karlsson, D. Wristers and N. Kepler, "Advantages and challenges of high performance CMOS on SOI." *Proceedings of IEEE International SOI Conference, pp. 1-4, 2001*

[19] S. Fung, N. Zamdmer, P.J. Oldiges, J. Sleight, A. Mocuta, M. Sherony, S-H. Lo, R. Joshi, C.T. Chuang, I. Yang, S. Crowder, T.C. Chen, F. Assaderaghi and G. Shahidi, "Controlling floating-body effects for 0.13um and 0.10um SOI CMOS," *Proceedings of the International Electron Device Meeting, pp. 231-234, Dec. 2000*

[20] B. Yu, Y-J. Tung, S. Tang, E. Hui, T-J. King and C. Hu, "Ultra-thin-body silicon on insulator MOSFETs for terabit-scale integration," *Proceedings of the International Semiconductor Device Research Society (ISDRS), pp. 623-628, 1997*

[21] B. Van Meer and K. De Meyer, "Threshold voltage model for deep-submicron fully depleted SOI CMOS transistors including the effect of source/drain fringing fields into the buried oxide." *Solid-State Electronics, Vol. 45, No. 4, pp. 593-598, April 2001*

[22] D.A. Antoniadis. "MOSFET scalability limits and "new frontier" devices." *IEEE. Symposium on VLSI Technology. Digest of Technical Papers, pp. 2-5, 2002*

[23] P.A. Stolk and D.B.M. Klaassen, "The effect of statistical dopant fluctuations on MOS device performance," *IEEE. Proceedings of International Electron Devices Meeting, pp. 627–630, Dec. 1996*

[24] D.J. Frank, Y. Taur, M. Ieong and H-SP. Wong, "Monte Carlo modeling of threshold variation due to dopant fluctuations," *Symposium on VLSI Technology, pp. 169–170, June 1999*

[25] C. Yin, V.W.C. Chan and P.C.H. Chan, "Low S/D resistance FDSOI MOSFETs using polysilicon and CMP," *IEEE. Proceedings of Hong Kong Electron Devices Meeting, pp. 89–92, June 2001*

[26] B. Cheng, M. Cao, R. Rao, A. Inani, P. Vande Voorde, W.M. Greene, J.M.C. Stork, Z. Yu, P.M. Zeitzoff and J.C.S. Woo, "The impact of high-κ gate dielectrics and metal gate electrodes on sub-100 nm MOSFETs," *IEEE. Transactions on Electron Devices, Vol. 46, No. 7, pp. 1537 –1544, July 1999*

[27] T. Hirose, Y. Momiyama, M. Kosugi, H. Kano, Y. Watanabe and T. Sugii, "A 185 GHz f_{max} SOI DTMOS with a new metallic overlay-gate for low-power RF applications." *IEEE. Proceedings of International Electron Devices Meeting. pp. 943-945, Dec. 2001*

[28] J. Kedzierski, E. Nowak, T. Kanarsky, Y. Zhang, D. Boyd, R. Carruthers, C. Cabral, R. Amos, C. Lavoie, R. Roy, J. Newbury, E. Sullivan, J. Benedict, P. Saunders, K. Wong, D. Canaperi, M. Krishnan, K-L. Lee, B.A. Rainey, D. Fried, P. Cottrell, H-SP. Wong, M. Ieong and W. Haensch, "Metal-gate FinFET and fully-depleted SOI devices using total gate silicidation." *IEEE. Proceedings of International Electron Devices Meeting, pp. 247-250, Dec. 2002*

[29] R. Bez, E. Camerlenghi, A. Modelli and A. Visconti, "Introduction to Flash memory," *Proceedings of the IEEE, Vol. 91, No. 4, pp. 489-502, April 2000*

[30] Y-K. Lee, S-K. Sung, J-S. Sim, C-J. Lee, T-H. Kim, S-H. Lee, J-D. Lee, B-G. Park, D-H. Lee and Y-W. Kim, "Multi-level vertical channel SONOS nonvolatile memory on SOI," *Symposium on VLSI Technology, pp. 208 -209, June 2002*

[31] V-Y. Aaron and J-P. Leburton, "Flash memory: towards single-electronics," *IEEE. Potentials, Vol. 21, No. 4, pp. 35 –41, Oct. 2002*

[32] J. Bu and M.H. White, "Retention reliability enhanced SONOS NVSM with scaled programming voltage." *IEEE. Proceedings of Aerospace Conference, Vol. 5, pp. 5.2383-5.2390, 2002*

[33] X. Huang, W-C. Lee, D. Hisamoto, L. Chang, J. Kedzierski, E. Anderson, Y-K. Choi, V. Subramanian T-J. King, J. Bokor and C. Hu, "Sub 50-nm FinFET: PMOS," *Proceedings of the International Electron Device Meeting, pp. 67-70, Dec. 1999*

Chapter 2

Subthreshold Model for Fully Depleted SOI Transistors

2.1 Introduction

Analytical modeling has played an important role in the evolution of semiconductor industry. The understanding and accurate modeling of the device physics have been guiding device design. As the channel length is scaled to its operational limits, understanding two-dimensional (2D) short channel effects (SCE) becomes increasingly crucial for device scaling. Since the conventional quasi-2D approach [1] is not adequate to capture the 2D characteristics of the potential profile in the channel, a better model with improved accuracy and an extended valid region is required. Moreover, as novel structures and materials, such as pocket (also called halo) implants [2], high κ gate dielectrics [3] and ultra-thin-body (UTB) devices [4], are proposed to suppress SCE, new models including their influences on device performance are in demand.

In this chapter, a true 2D model is developed for the single gate (SG) fully depleted (FD) and double gate (DG) MOSFET. It includes the influence of pocket doping, which has become an industry standard in suppressing SCE, and the 2D effect in the gate dielectric, which is significant with the use of thick high κ gate dielectric materials. Compared with Frank's model [5], which gives only an implicit equation for the scale length, our model results in an explicit analytical scale length and 2D potential profiles in the entire body.

2.2 Performance characteristics

In modern circuitry, power, delay and density are the three dominant metrics for a given technology. The density indicates both the amount of functionality that can be integrated on a chip and its cost. The delay refers to the speed at which the function is fulfilled, and it has been the major driving force of scaling for decades. Power shows both the energy consumption and, more importantly, the heat generation. The latest integrated circuit (IC) chips utilize a power density as high as 20W/cm^2, which approaches that of an electrical oven. Self-heating of chips has become a decisive issue in circuit design [6].

Although a MOSFET is often modeled as an ideal switch in many digital designs, in reality, it consumes both a finite delay and finite power. For simplicity, the delay used in this report will be the intrinsic delay, defined as t_p=CV/I. The power consists of the switching power and standby power. The switching power is CV^2f_{switch}, and the standby power is $I_{off}V$. To maintain a reasonable level of power consumption, the current leakage I_{off} must be kept below a certain limit, which depends on the application type and is set by the ITRS roadmap [7]. This power is particularly crucial in the ULSI era because there are billions of transistors consuming the standby power even at idle [8].

A MOSFET has two distinct operational modes separated by the threshold voltage V_t: on and off. Fig 2.1 shows a typical drain current, in both logarithmic scale and linear scale, versus the gate voltage. The threshold voltage can be defined in a number of ways. In this chapter, we will choose it as the gate voltage at a drain current of I_o=40nA/μm.

In the off region, the current is limited by the number of carriers that can thermionically overcome the barrier in the channel, which, in turn, is controlled by the gate voltage. Therefore, the current depends exponentially on the gate bias, i.e. $I_d = I_o \bullet 10^{(V_g - V_T)/S}$. The subthreshold swing S, which is defined as the V_g change for each decade of I_d change in the subthreshold region, indicates the effectiveness of the gate control over the channel barrier; therefore, $V_t = S \log I_o / I_{off}$. Since the application sets the acceptable level of I_{off}, the minimum V_t is proportional to S and does not scale with device dimensions [9].

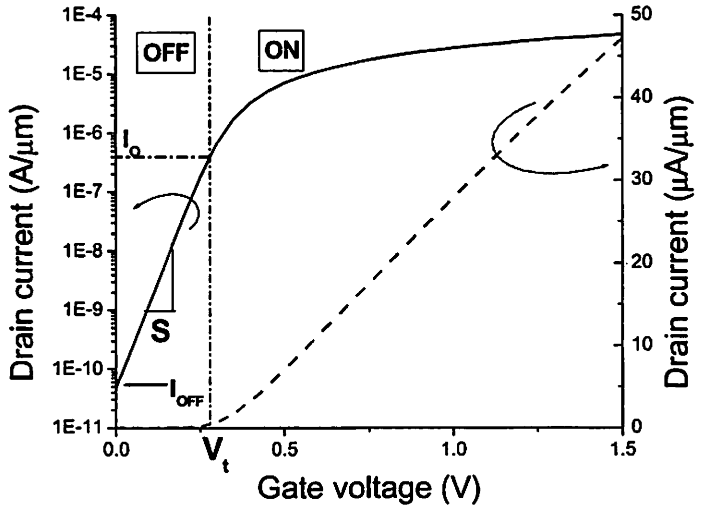


Fig. 2.1. Typical I_d vs. V_g curve of a transistor. The threshold voltage separates its on and off operation modes.

When the device is on, the current is limited by the number and velocity of charge carriers. $I_{on} = nev \approx vC_{ox}W(V_g - V_t)$. At a high drain bias, the device is in saturation, and the carrier velocity is v_{sat}, a material constant of around 10^7cm/s for electrons and $8*10^6$cm/s for holes in silicon at room temperature [10]. At a low drain bias, the carrier velocity is $v=\mu E=\mu V_{ds}/L$, where μ is referred to as the carrier mobility.

More important is the intrinsic delay t_p, which is the index of the circuit speed.

$$t_p \equiv \frac{CV}{I} = \frac{C_{ox}WLV_{dd}}{I_{on}} = \frac{L}{v}\frac{V_{dd}}{V_{dd} - V_t} \qquad (1)$$

The above equation clearly shows that scaling L is the major method for speeding up the MOSFETs because it reduces gate capacitance as well as increases the drive current. Although higher V_{dd} results in faster circuit speed, as shown in equation (1), unfortunately, V_{dd} actually is scaled down to reduce the power consumption. This reduction is a compromise between maximizing performance and minimizing power dissipation. As a matter of fact, V_{dd} is scaled as quickly as the device dimensions are to maintain a constant electric field inside the transistor for its reliability. Therefore it is crucial to keep V_t as low as possible, especially when V_{dd} is scaled to be comparable to V_t. Improving the mobility μ, is another way to increase I_{on} and reduce T_p [11,12].

In summary, S, μ and V_t values are the essentials indices for device performance with a given L and voltage supply. The smaller S, the lower V_t can be for a given level of I_{off}. High mobility results in a high drive-current and fast circuit. Here, an appropriate V_t is assumed to be achievable with gate workfunction engineering, which will be discussed in Chapter 4.

2.3 The Poisson equation in the body

Since the major challenge of scaling is the SCE, this model focuses on the subthreshold behavior of a fully depleted transistor. Unless explicitly stated, NMOS will be assumed throughout this chapter, and similar results can be derived for PMOS. For convenience, the origin of the x-axis is set at the bottom of an SG transistor or the middle of a DG transistor (Fig. 2.2). The label d refers to the body thickness of an SG device, or one half of that in a DG MOSFET. Also, the source potential is used as the reference point, i.e. V_s=0.

The following assumptions and approximations are made to set up the problem.

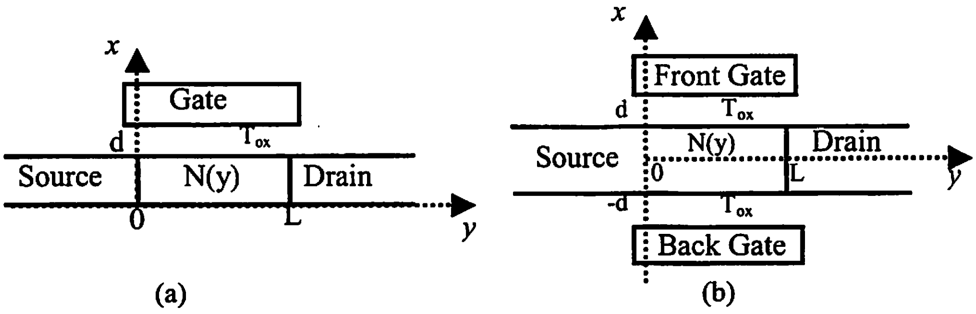


Fig. 2.2. Schematic cross sections of (a) SG thin-body and (b) DG MOSFETs

1. Since only the subthreshold behavior of the MOSFET is involved, the full depletion approximation is used inside the fully depleted body.

2. Ideal abrupt source/drain junctions are assumed, so that the boundaries between the source/drain and channel do not move with bias.

3. In a UTB or DG structure, the body is so thin that vertical doping engineering becomes impractical. Therefore, vertically uniform doping is assumed.

4. On the other hand, this model includes lateral doping engineering, such as pocket implants, which are widely used in suppressing SCE. The doping profile along the channel is assumed to be symmetric to the centerline at y=L/2, which applies almost exclusively to modern circuits.

5. In the case of a DG MOSFET, the devices are assumed to be symmetric to the centerline x=0. That means the front and back gates share the same gate workfunction and bias, same gate dielectric material and thickness. Therefore, only half of the device needs to be modeled. Also due to the symmetry, the vertical electric field vanishes at the middle of the device (x=0).

6. In the case of a UTB MOSFET, infinitely thick buried oxide (BOX) is assumed, and

the fringing fields through the BOX are ignored. Therefore, with no vertical field at the bottom interface, an SG MOSFET can be treated as a half of a DG device. In reality, the fringing fields through the buried oxide degrade the SCE of a UTB MOSFET [13]. Low-κ materials are desirable as the bottom insulator.

7. First, the electrical field in the gate dielectric is assumed to be strictly vertical. The 2D effect of the gate dielectric will be discussed and modeled in section 2.8.

Based on the above assumptions, the body in a fully depleted SOI MOSFET is simply a box with a fixed charge density of -qN(y) (positive N(y) for a p-type body and negative for an n-type body) . The Poisson equation in the body is:

$$\frac{d^2V}{dx^2}+\frac{d^2V}{dy^2}=-\frac{-qN(y)}{\varepsilon_{si}}\equiv\rho(y) \qquad (2)$$

The boundary conditions are:

$$V\big|_{y=0}=0,\ V\big|_{y=L}=V_{ds} \text{ and } \frac{dV}{dx}\bigg|_{x=0}=0,\ V+\frac{\varepsilon_{si}}{\varepsilon_{ox}}T_{ox}\frac{dV}{dx}\bigg|_{x=d}=V_{eff}\equiv V_{gs}-\phi_{gs} \qquad (3)$$

T_{ox} is the thickness of the silicon dioxide as the gate insulator and V_{gs} (V_{ds}) is the gate (drain) bias voltage. Here, ϕ_{gs} is defined as the workfunction difference between the gate and the source. When conventional N+/P+ poly gates are used in N+/P+ MOSFETs, ϕ_{gs} equals zero.

2.4 General solution and its simplification

2.4.1 General solution

The solution V(x,y) can be separated into two terms: V(x,y)=U(x,y)+h(y). Here h(y) satisfies $\frac{d^2h}{dy^2}=\rho(y)$ and h(0)=0, h(L)=V_{ds}, while U(x,y) satisfies $\frac{d^2U}{dx^2}+\frac{d^2U}{dy^2}=0$, and

U(x,0)=U(x,L)=0. The general solution is $U(x,y)=\sum_{k=1}^{\infty}\left[A_k\cosh\left(\frac{k\pi}{L}x\right)+B_k\sinh\left(\frac{k\pi}{L}x\right)\right]\sin\left(\frac{k\pi}{L}y\right)$. The coefficients A_k and B_k can be calculated from the boundary conditions in the x direction as: $B_k=0$ and $A_k=C_k\Big/\left[\cosh\left(\frac{k\pi}{L}d\right)+\frac{\varepsilon_{si}}{\varepsilon_{ox}}T_{ox}\frac{k\pi}{L}\sinh\left(\frac{k\pi}{L}d\right)\right]$, where C_k's are the Fourier coefficients of V_{eff}-$h(y)$.

Since the solution is a sum of $\sin\left(k\pi\frac{y}{L}\right)$ terms, it is natural to expand ρ(y) as $\rho(y)=\sum_{k=1}^{\infty}D_k\sin\left(k\pi\frac{y}{L}\right)$. Then, h(y) can be easily solved from the 1D differential equation as $h(y)=-\sum_{k=1}^{\infty}D_k\left(\frac{L}{k\pi}\right)^2\sin\left(k\pi\frac{y}{L}\right)+V_{ds}\frac{y}{L}$, and $C_k=\frac{2}{k\pi}\{[1-(-1)^k]V_{eff}+(-1)^kV_{ds}\}+D_k\left(\frac{L}{k\pi}\right)^2$.

Therefore, the final result of the potential profile is:

$$V(x,y)=\sum_{k=1}^{\infty}\frac{C_k\cosh\left(\frac{k\pi}{L}x\right)\sin\left(\frac{k\pi}{L}y\right)}{\cosh\left(\frac{k\pi}{L}d\right)+\frac{\varepsilon_{si}}{\varepsilon_{ox}}T_{ox}\frac{k\pi}{L}\sinh\left(\frac{k\pi}{L}d\right)}-\sum_{k=1}^{\infty}D_k\left(\frac{L}{k\pi}\right)^2\sin\left(k\pi\frac{y}{L}\right)+V_{ds}\frac{y}{L} \qquad (4)$$

2.4.2 2nd order approximation

Since it has been well established that d<<L is required to suppress SCE in nano-scaled MOSFETs [14], the following approximation can be made:

$$\frac{\cosh\left(\frac{k\pi}{L}x\right)}{\cosh\left(\frac{k\pi}{L}d\right)+\frac{\varepsilon_{si}}{\varepsilon_{ox}}T_{ox}\frac{k\pi}{L}\sinh\left(\frac{k\pi}{L}d\right)}\approx\frac{1}{1+\left(\frac{k\pi}{L}\right)^2\left(\frac{\varepsilon_{si}}{\varepsilon_{ox}}T_{ox}d+\frac{d^2-x^2}{2}\right)}=\frac{1}{1+\left(\frac{k\pi l}{L}\right)^2} \text{ and } l\equiv\sqrt{\frac{\varepsilon_{si}}{\varepsilon_{ox}}T_{ox}d+\frac{d^2-x^2}{2}} \qquad (5)$$

With this approximation, all of the x dependence of the potential is contained in *l*, the scale length of the MOSFET. V(x,y) can be simplified as:

$$V(x,y)\approx\sum_{k=1}^{\infty}\left\{\frac{2}{k\pi}\frac{[1-(-1)^k]V_{eff}+(-1)^kV_{ds}}{1+(k\pi l/L)^2}+\frac{D_k}{1+(k\pi l/L)^2}\left(\frac{L}{k\pi}\right)^2-D_k\left(\frac{L}{k\pi}\right)^2\right\}\sin\left(k\pi\frac{y}{L}\right)+V_{ds}\frac{y}{L}$$

$$=\sum_{k=1}^{\infty}\frac{2}{k\pi}\frac{[1-(-1)^k]V_{eff}+(-1)^kV_{ds}}{1+(k\pi l/L)^2}\sin\left(\frac{k\pi}{L}y\right)-\sum_{k=1}^{\infty}\frac{D_kl^2}{1+(k\pi l/L)^2}\sin\left(\frac{k\pi}{L}y\right)+V_{ds}\frac{y}{L} \qquad (6)$$

It is further simplified with the following identities [15] and their integral forms:

$$\sum_{k=1}^{\infty}\frac{\cos(k\pi y)}{1+(k\pi/a)^2}=\frac{a}{2}\frac{\cosh(a-ay)}{\sinh a}-\frac{1}{2} \text{ and} \sum_{k=1}^{\infty}\frac{(-1)^k\cos(k\pi y)}{1+(k\pi/a)^2}=\frac{a}{2}\frac{\cosh(ay)}{\sinh a}-\frac{1}{2} \qquad (7)$$

$$\sum_{k=1}^{\infty}\frac{2}{k\pi}\frac{\sin(k\pi y)}{1+(k\pi/a)^2}=1-y-\frac{\sinh(a-ay)}{\sinh a} \text{ and } \sum_{k=1}^{\infty}\frac{2}{k\pi}\frac{(-1)^k\sin(k\pi y)}{1+(k\pi/a)^2}=\frac{\sinh(ay)}{\sinh a}-y \qquad (8)$$

Then, finally,

$$V(x,y)=V_{eff}-V_{eff}\frac{\sinh[(L-y)/l]+\sinh(y/l)}{\sinh(L/l)}+V_{ds}\frac{\sinh(y/l)}{\sinh(L/l)}-\sum_{k=1}^{\infty}\frac{D_k l^2}{1+(k\pi l/L)^2}\sin\left(k\pi\frac{y}{L}\right)$$

Equivalently, $V(x,y)=V_{eff}\left\{1-\frac{\cosh[(L/2-y)/l]}{\cosh(L/2l)}\right\}+V_{ds}\frac{\sinh(y/l)}{\sinh(L/l)}+\Delta V(x,y)$ (9)

The last term ΔV(x,y) is the potential caused by the substrate doping and is independent of the applied bias. It can be further derived as (Appendix 2A)

$$\Delta V(x,y)=\frac{-l}{\cosh(L/2l)}\left[\cosh\frac{L/2-y}{l}\int_0^y\rho(t)\sinh\frac{t}{l}dt+\sinh\frac{y}{l}\int_y^{L/2}\rho(t)\cosh\frac{L/2-t}{l}dt\right] \qquad (10)$$

2.4.3 Subthreshold swing and V$_t$ definition

Next, we will relate the internal potential profile to the device's external electrical characteristics. In the subthreshold state, the drain current is predominantly limited by the number of carriers thermionically excited over the barrier in the channel. $N_{carrier}=N_c\exp\left(\frac{qV_{\min}}{kT}\right)$. Therefore, the current leakage depends exponentially on the barrier height V$_{min}$, i.e. $I_d\propto\exp\left(\frac{qV_{\min}}{kT}\right)$. Since the current flows along the path with a minimum barrier, the position of the barrier can be found at the minimum of the potential in the y direction, but at the maximum in the x direction. Supposing it is located at (X$_0$,Y$_0$), then $l_0\equiv\sqrt{T_{ox}d\varepsilon_{si}/\varepsilon_{ox}+(d^2-X_0^2)/2}$, and we can derive the subthreshold swing:

$$S\equiv\frac{dV_g}{d\log I_d}=\frac{kT\log 10}{q}\frac{dV_g}{dV_{\min}}=\frac{60\,mV/dec}{1-\cosh[(L/2-y)/l_0]/\cosh(L/2l_0)} \qquad (11)$$

The threshold voltage (V_t) is defined as the gate voltage at a certain level of current leakage. It corresponds to a specific barrier height $V_{min}=V_b$. The exact value of V_b depends on the actual definition of V_t, which itself is not a settled issue. Practically, $V_b \cong -0.2V$, which matches closely the widely used definition of V_t at I_d=40nA/μm. It is observed that the drain bias V_{ds} causes a reduction of the threshold voltage by modifying the potential barrier. This effect is called "drain-induced barrier lowering" (DIBL), an important index of SCE. The DIBL coefficient can be calculated as:

$$C_{DIBL} \equiv -\frac{\partial V_t}{\partial V_{ds}} = \frac{dV_{min}}{dV_{ds}} \Big/ \frac{dV_{min}}{dV_g} = \frac{\sinh(y/l_0)}{\sinh(L/l_0) - \sinh(y/l_0) - \sinh[(L-y)/l_0]} \qquad (12)$$

The minimum (optimum) swing and maximum (worst) C_{DIBL} are produced when the barrier is at the middle of the channel, i.e. y=L/2, and they are linearly related.

$$S = 60\frac{mV}{dec} \Big/ \left[1 - \frac{1}{\cosh(L/2l_0)}\right], \quad C_{DIBL} = \frac{1}{2}\frac{1}{\cosh(L/2l_0) - 1} = \frac{1}{2}\left(\frac{S}{60mV/dec} - 1\right) \qquad (13)$$

Notice that this simple relation holds as long as the barrier is at the middle of the channel. Either a finite V_{ds} or pocket doping can shift the potential barrier closer to the source side. From the formula for S and C_{DIBL}, this shift degrades the subthreshold swing, but improves the DIBL coefficient. More discussion of this subject will be presented in the next section.

2.5 Uniformly doped body

2.5.1 Potential profile and short channel effects

When ρ is constant, ΔV can be calculated as $\Delta V = -\rho l^2 \left\{1 - \frac{\cosh[(L/2 - y)/l]}{\cosh(L/2l)}\right\}$

The potential profile can be simplified to equation (14) and the channel doping merely causes a shift in V_{eff}: $V'_{eff} = V_{eff} - \rho l^2$. Fig. 2.3 demonstrates the excellent agreement throughout the entire body of the potential with numerical simulation results. The device simulation tool used throughout this chapter is ATLAS from Silvaco, which solves the potential and current based on device physics and material properties.

$$V(x,y) = V'_{eff}\left\{1 - \frac{\sinh[(L-y)/l]}{\sinh(L/l)}\right\} - (V'_{eff} - V_{ds})\frac{\sinh(y/l)}{\sinh(L/l)} \qquad (14)$$

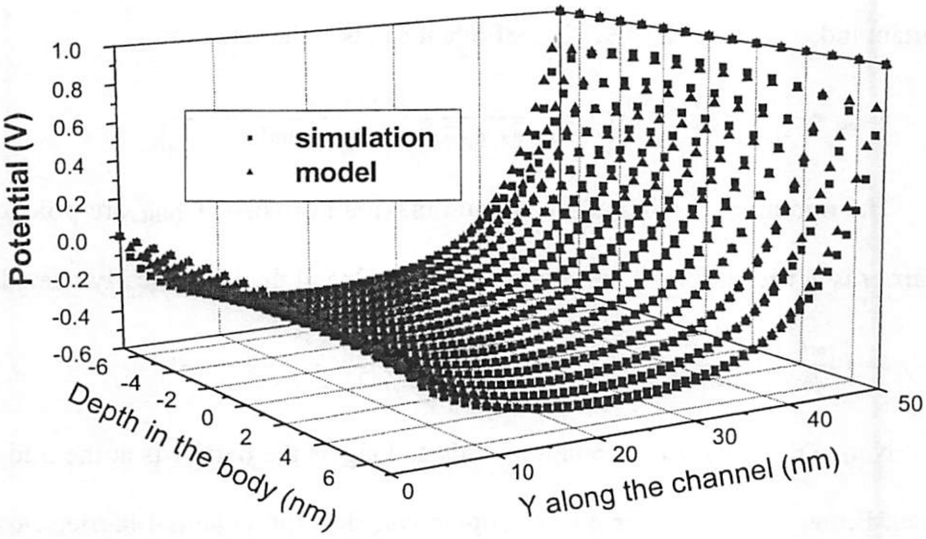


Fig. 2.3. The good agreement in the potential profile between the simulation and modeling results. (T_{ox}=2nm, d=7nm, L=50nm, V_g=0, V_{eff}= -0.48V, V_d=1V)

2D models for the potential profile in FD SOI MOSFETs with uniform doping have been previously published by Young [16] and Yan [17]. Both make virtually the same assumption that the potential profile is parabolic in the x direction [18]. The fact that equation (14) agrees with their final results verifies that their assumptions are equivalent to the 2nd order expansion of our model. The above derivation also indicates that those models are valid only for long channel devices (L>>d) since higher order terms become significant at short channel lengths.

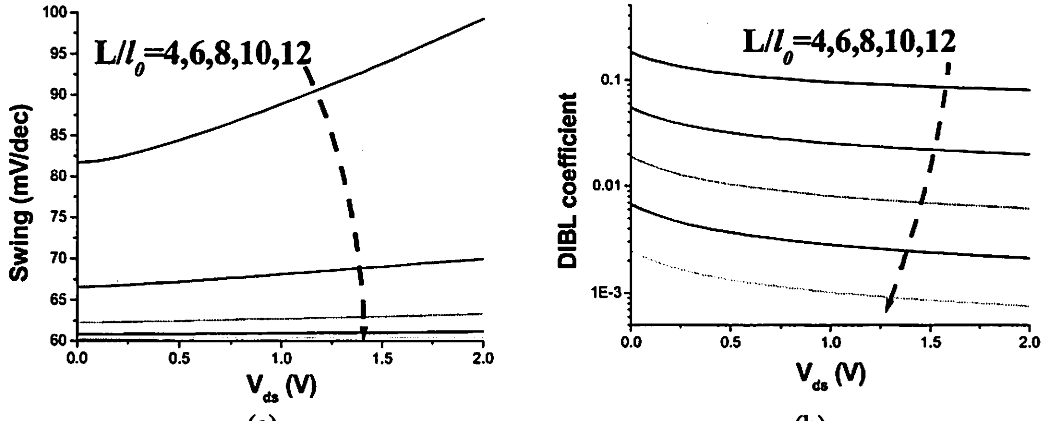
From the potential profile, the barrier height can be derived. Then V_t can be solved as the gate voltage when $V_{min}(x)=V_b$ (Appendix 2B).

$$V_{\min}(x) = V_{eff} - \rho l^2 + \frac{1}{\sinh(L/l)}\sqrt{4V'_{eff}(V'_{eff} - V_{ds})\sinh^2\frac{L}{2l} - V_{ds}^2} \tag{15}$$

$$V_t = \phi_{gs} + \rho l_0^2 + V_b + \left[V_b - \frac{V_{ds}}{2} - \sqrt{V_b(V_b - V_{ds})}\cosh\frac{L}{2l_0}\right] \Big/ \sinh^2\frac{L}{2l_0} \tag{16}$$

With a uniformly doped body, S and C_{DIBL} can be calculated at an arbitrary bias. Fig. 2.4 plots S and C_{DIBL} as functions of V_{ds}.

$$S = 60\frac{mV}{dec} \Big/ \left[1 - \frac{(V_{ds} - 2V'_{eff})\tanh(L/2l_0)}{\sqrt{4(V'_{eff} - V_{ds})V'_{eff}\sinh^2(L/2l_0) - V_{ds}^2}}\right] \text{ and } C_{DIBL} = \frac{1 + \cosh(L/2l_0) \big/ \sqrt{1 + V_{ds}/|V_b|}}{2\sinh^2(L/2l_0)} \tag{17}$$



(a) (b)

Fig. 2.4. (a) Subthreshold swing with V'_{eff}=-0.5V and (b) DIBL coefficient change with the drain bias.

It is observed that the swing is insensitive to V_{ds} and is well represented by the value at V_{ds}=0, unless the device is scaled at its SCE limit, i.e. L/l_0=4. C_{DIBL} decreases with increasing V_{ds}, reaching about 50% of its original value at 1.2V. On average, the DIBL effect can be approximated as: ΔV_t= $-\alpha C_{DIBL}(0)V_{ds}$ and α~0.8 for V_{ds}=1.2V. Therefore, the SCE of a transistor can be well represented by its swing and C_{DIBL} values at V_{ds}=0. Since the barrier is simply at the middle point of the channel at this special bias, the DIBL coefficient is still linearly related to the subthreshold swing.

$$S = 60\frac{mV}{dec}\Big/\left[1-\frac{1}{\cosh\left(L/2l_0\right)}\right],\quad C_{DIBL}(0)=\frac{1}{2}\left(\frac{S}{60mV/dec}-1\right) \tag{18}$$

Fig. 2.5a shows that L/*l* determines the SCE of a given device. It also shows that L/l_o>4 is required to achieve good SCE, which is set roughly at S=80mV/dec and $\Delta V_{t,DIBL}$<0.15V with V_{dd}=1.2V. Fig. 2.5b plots the minimum gate length for a given device geometry (T_{ox} and d), using the above criteria $L_{min}=4l_o$. This result can be used as a guide in designing deeply scaled devices. For example, a 10nm MOSFET with a 1nm T_{ox} requires a body thickness of 2nm (or 4nm in double gate structures).

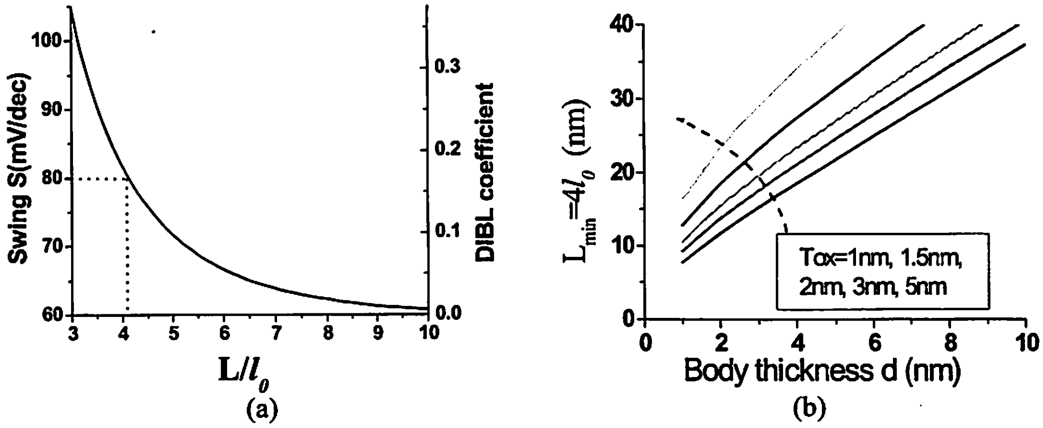


Fig. 2.5. (a) Subthreshold swing and DIBL coefficient vs. L/l_0 at V_{ds}=0. (b) The minimum channel length with S<80mV/dec as a function of device dimensions. $L_{min}=4l_0=4\sqrt{T_{ox}d\varepsilon_{si}/\varepsilon_{ox}+d^2/2}$ is used for the reason explained in Section 2.5.2.

In equation (16), the last term vanishes when $L>>l_o$, i.e. $V_t \rightarrow \phi_{gs}+\rho l_0^2+V_b$ for long channel devices. The term $\rho l_o^2=qNl_o^2/\varepsilon_{si}$ shows the extent of V_t shift caused by the body doping. Since $l_o<L_{min}/4$ is required for suppressed SCE, channel doping can't provide sufficient V_t adjustment for future device generations with L<25nm. For example, a body doping of $7\times10^{18}cm^{-3}$ is required for 0.4V V_t shift with l_o=6nm. Therefore, gate workfunction engineering is the only effective way to achieve the appropriate V_t, which will be addressed in more details in Chapter 4.

2.5.2 Critical doping and leakage depth

As shown in the previous section, the ratio L/*l* determines the SCE of a transistor. However, to calculate the scale length $l_0 \equiv \sqrt{\frac{\varepsilon_{si}}{\varepsilon_{ox}} T_{ox} d + \frac{d^2 - X_0^2}{2}}$, the leakage path depth x must be located first. Since the current leakage flows through the path with the lowest barrier, it is equivalent to find the depth with maximum $V_{min}(x)$.

Inside the body, the 2D effect bends the barrier up towards the front surface and the p-type ionized dopants do the opposite. At high doping levels (p-type), the latter dominates, and the barrier is lower at the front surface. Therefore, the mobile charge carriers are pushed towards the front surface (x=d) and the scale length is $l_0 = \sqrt{\frac{\varepsilon_{si}}{\varepsilon_{ox}} T_{ox} d}$. In this case, the scale length *l* reaches its minimum value and results in the best SCE. Provided the doping is still low enough such that the body is fully depleted, the depleted body is analogous to the depletion region in a bulk MOSFET, as shown by the similarity between this scale length and its classic counterpart $l_{bulk} = \sqrt{\frac{\varepsilon_{si}}{\varepsilon_{ox}} T_{ox} W_{dep}}$, where W_{dep} is the depletion depth in a bulk MOSFET.

When the channel is lightly doped, undoped or counter-doped, the barrier is actually lower at the back interface of a UTB transistor where the gate control is weaker. Because the leakage flows through the back interface, i.e. x=0, the scale length is modified into $l_0 = \sqrt{\frac{\varepsilon_{si}}{\varepsilon_{ox}} T_{ox} d + \frac{d^2}{2}}$, as has been pointed out in some other publications [19,20]. In this case, the scaling of body thickness is even more effective in suppressing SCE, because the shrinking of the body thickness will cut out most of the current leakage, which concentrates at the back interface.

Since the short channel characteristics are simply functions of L/l_o, devices with higher channel doping gives better SCE by pushing the leakage path to the front, and gives rise to more efficient gate control. Fig. 2.6 clearly demonstrates the improvement of SCE at the high body-doping end.

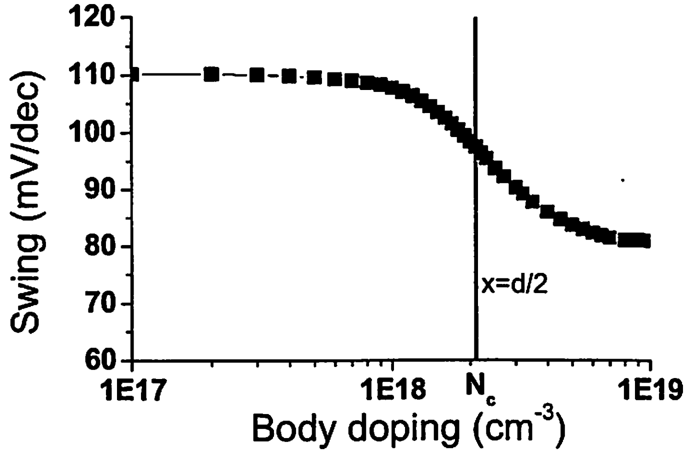


Fig. 2.6. Simulation (Atlas from Silvaco) shows that a high body doping improves the SCE by pushing the leakage to the front surface. (L=30nm, d=10nm, T_{ox}=2nm)

The depth of the current leakage can be obtained by finding the minimum barrier (or maximum potential), i.e. $\frac{dV_{\min}}{dx}=0$. Although the path moves with bias, for simplicity, the profile at V_{ds}=0 is used: $V_{\min}(x)=(V_{eff}-\rho l^2)\left[1-\frac{1}{\cosh(L/2l)}\right]$. The critical doping level N_c is defined as the doping concentration with the leakage flowing through the middle of the channel (x=d/2) at the specific bias of V_{ds}=0 and $V_{gs}=V_t$. The resulting N_c is:

$$N_c=\frac{-\varepsilon_{si}V_b L}{2ql_0^3}\frac{\sinh(L/2l_0)}{[\cosh(L/2l_0)-1]^2}=\frac{\varepsilon_{si}|V_b|L}{4ql_0^3}\frac{\cosh(L/4l_0)}{\sinh^3(L/4l_0)} \qquad (19)$$

Fig. 2.7 plots N_c as a function of the channel length with T_{ox}=2nm and different body thicknesses. For very short channel devices (L<25nm), $N_c \to 16\frac{\varepsilon_{si}|V_b|}{qL^2}$, becomes independent of the vertical dimensions. Unfortunately, N_c is typically too high to be practical in that region; therefore channel doping can not push the leakage path to the front surface for good SCE and the worst-case scale length has to be used.

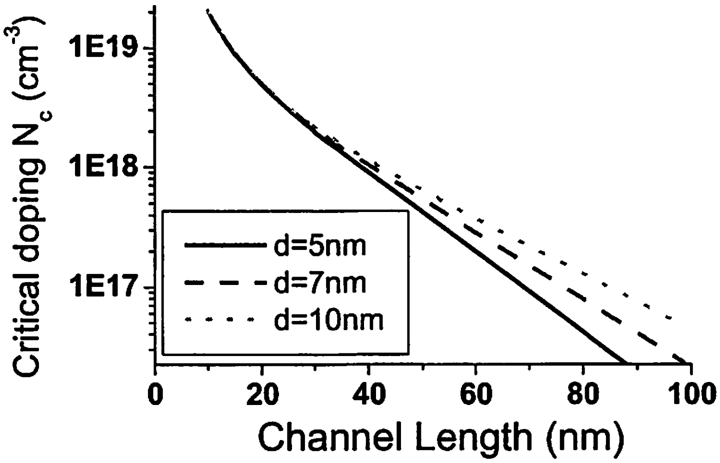


Fig. 2.7. The critical doping level N_c as a function of the channel length L.

Fig. 2.8 demonstrates the excellent agreement of the model results with numerical simulation (Silvaco). Since the channel is undoped in this case, the scale length of $l_0 = \sqrt{\frac{\varepsilon_{si}}{\varepsilon_{ox}} T_{ox} d + \frac{d^2}{2}}$ is used. The result indicates that the minimum gate length with the given dimensions is about 32nm. Devices with shorter gate lengths are possible with further reduction of the body and gate oxide thicknesses.

Our model does not address quantum effects on device performance. In the first order approximation, quantum confinement causes two modifications. First, it raises the band edge by $\frac{\hbar^2\pi^2}{2m^*d^2}$, where m* is the effective mass of the carrier [21]. This effect can be easily included as an increase in the parameter V_b. Second, the carriers are repelled from the Si/SiO$_2$ interface with a charge centroid of δ, typically at around 1nm [22]. Therefore, the depth x can not reach its extreme values, and is limited in the range of (δ, d-δ). The scale length may need to be modified to include this effect.

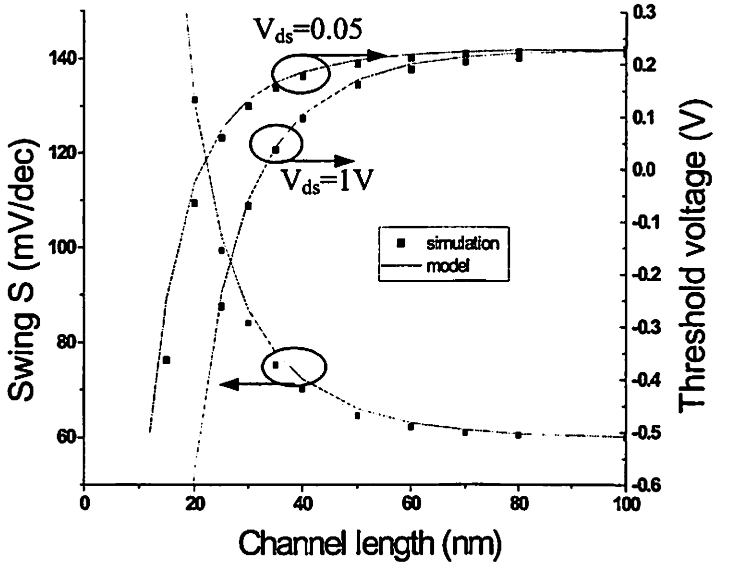


Fig. 2.8. Comparison of subthreshold swing and V_t between the 2D analytical model and numerical simulation. (T_{ox}=2nm, d=7nm, N_{body}=0. l_o=8.1nm)

2.6 Channel with pocket doping

High pocket implants (p-type for NMOS) are commonly added at the ends of the channel to suppress SCE [2]. To perform the calculation of the potential profile, the analytic form of the pocket doping profile, which is typically not available from the process, is required. Even with known pocket doping profiles, the potential and threshold voltage can be analytically solved for only in a few special cases, such as an exponential doping profile. However, some primitive results can be still achieved even with an arbitrary pocket doping.

2.6.1 Long channel model

Since pocket doping locally increases the barrier, there can be two barriers in the channel: one near the source and one near the drain. When channel length is long, these two barriers are separated, and the barrier at the source side determines the device's SCE. For long L>>l, the potential in the channel can be simplified as

$$V(x,y) \approx V_{eff}\left(1-\exp\frac{-y}{l}\right)-\frac{l}{2}\int_{0}^{L/2}\rho(t)\left[\exp\frac{-|y-t|}{l}-\exp\frac{-(y+t)}{l}\right]dt \tag{20}$$

The barrier position (X_0,Y_0) can be solved numerically from the above potential, which is independent of the channel length or drain bias. Then

$$S \approx \frac{60}{1-\exp(-Y_0/l_0)}\frac{mV}{dec} \text{ and } C_{DIBL} \approx \frac{\exp[(Y_0-L)/l_0]}{1-\exp(-Y_0/l_0)} \tag{21}$$

By putting the barrier closer to the source (y<L/2), pocket doping can significantly improve the C_{DIBL}, but will degrade the swing because now the barrier is more controlled by the source and less by the drain. The formula for the swing also indicates that the potential peak must be at least 1.4l away from the source to get an acceptable swing (S<80mV/dec). This imposes a minimum lateral displacement of the pocket doping.

2.6.2 Short channel model

For very short channel devices, which are of great interest, those two barriers merge, so that the barrier is still located at y=L/2 when V_{ds}=0. Therefore, the simple formulas for swing and C_{DIBL} still hold:

$$S = 60\frac{mV}{dec}\Big/\left[1-\frac{1}{\cosh(L/2l_0)}\right] \text{ and } C_{DIBL}(0)=\frac{1}{2}\left(\frac{S}{60mV/dec}-1\right) \tag{22}$$

Typically, the doping in the body will not change the scale length, unless it crosses the critical doping. Therefore, body-doping engineering does not help improve the subthreshold swing or DIBL. On the other hand, pocket doping in a bulk device reduces the depletion depth and, in turn, improves the scale length and SCE.

The V_t roll-off behavior involves a simple integration of the pocket profile.

$$V_t = \phi_{gs} + V_b + \left[V_b - \frac{\alpha}{2}V_{ds} + l_0 \int_0^{L/2} \rho(t)\sinh\left(\frac{t}{l_0}\right)dt\right] \Big/ \left(\cosh\frac{L}{2l_0} - 1\right) \tag{23}$$

The last term in the bracket is the SCE term, including three contributions from distinct sources. The first one is the intrinsic V_t roll-off due to the existence of the built-in voltage (V_b) between the channel and S/D. Because devices with different channel lengths have different swing values, the amount of the roll-off depends on the V_t definition or the value of V_b. If V_t is defined at a lower I_0, $|V_b|$ is higher, and more V_t roll-off will be observed (Fig. 2.9).

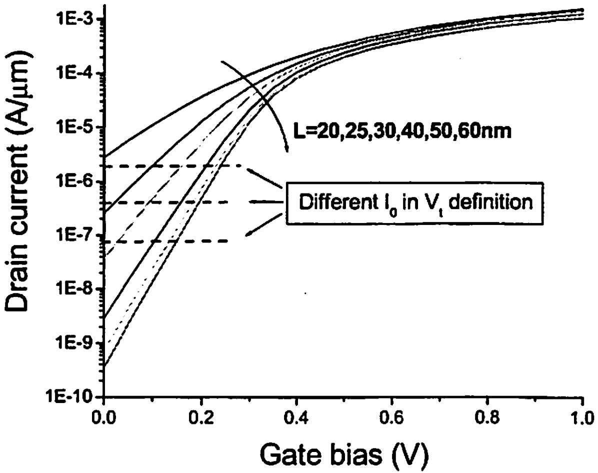


Fig. 2.9. The level of I_0, at which the threshold voltage is defined, not only changes the value of V_t, but also its L-dependence. (T_{ox}=2nm, d=7nm)

The second term is DIBL. Since V_{ds} applies only on the drain side while V_b exists on both source/drain sides, its effect on potential is half of V_b's. Moreover, because a finite V_{ds} shifts the barrier away from the drain, its voltage coupling to the channel is further reduced. Therefore, on average, the factor $\alpha/2$ is applied. It is found that α is insensitive to the exact pocket shape; therefore, α~0.8 can be still used with pockets at V_{ds}=1.2V.

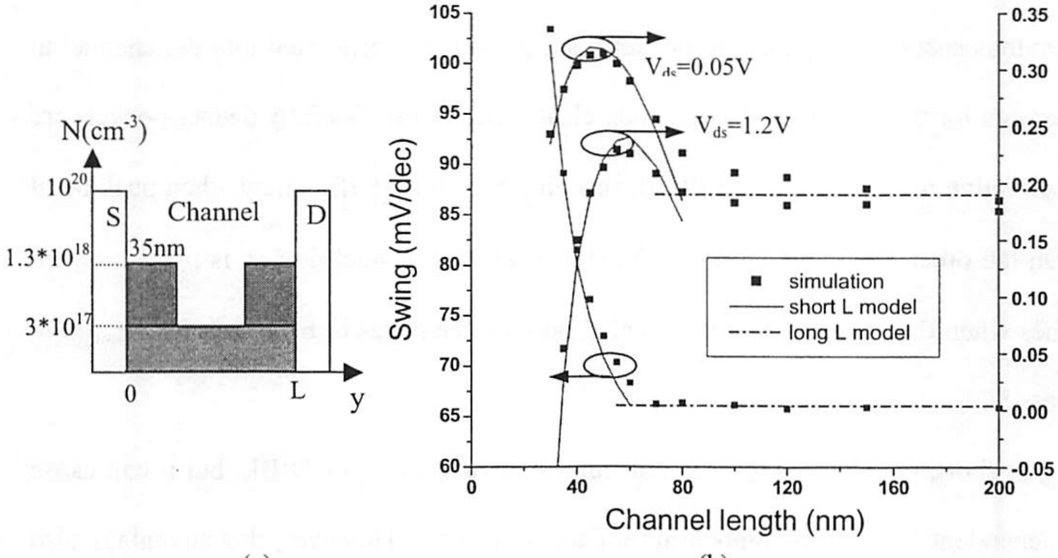
The third term is from the pocket doping and is bias independent but L dependent.

With a constant doping profile, this term reduces to a constant shift in V_t, as discussed in the previous section. Non-uniform pocket doping can be incorporated into the channel to compensate for the first item. The formula clearly shows that laterally deeper pockets are more effective in controlling V_t roll-off, reaching the highest efficiency when peaking at L/2. On the other hand, the maximum V_t shift from the channel doping is $\rho_{max}l_o^2$, which vanishes when then channel length is scaled down since l_o has to be kept below $L_{min}/4$ to suppress SCE.

In short, pocket doping does not improve either swing or DIBL, but it can cause an L-dependent V_t shift to compensate for the V_t roll-off. However, this advantage also vanishes when the channel length is scaled below 25nm.

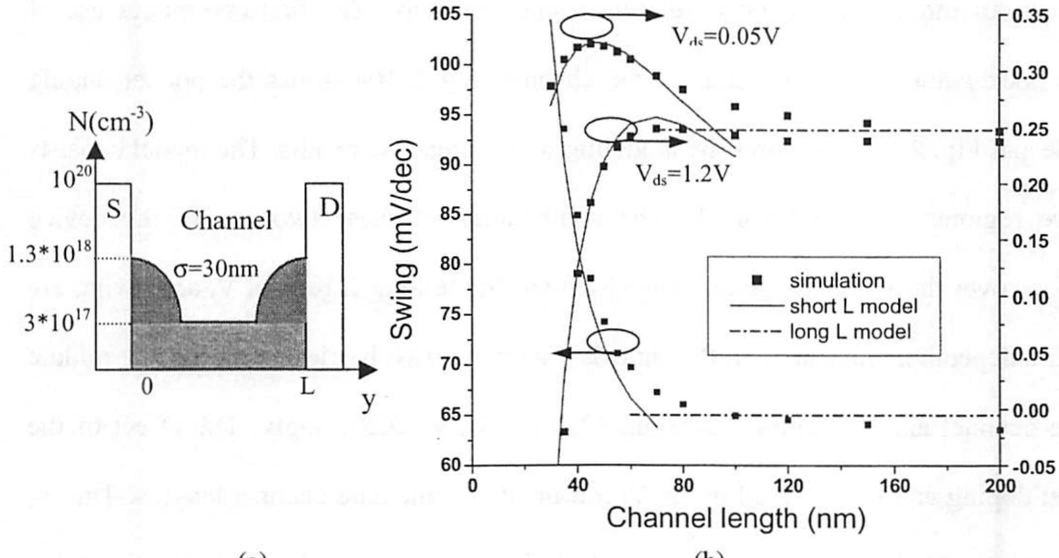
2.6.3 Model verification

Two cases are presented to examine the model, with excellent agreement achieved between our model results and simulation results (Silvaco). The first one makes use of cubic pocket doping on both ends of the channel. Fig. 2.10a shows the pocket doping profile and Fig. 2.10b compares the modeling with simulation results. The model consists of two regions: short and long L. The combination of these two can fit the device behavior over the whole range of channel length. In the long L region, V_t and swing are two L-independent constants. In the short L region, the two barriers merge at the middle of the channel and the simple equations (22) for S and DIBL apply. The effect of the pocket doping can be observed in the V_t roll-up at intermediate channel lengths. This V_t roll-up can partially balance the V_t roll-off and extend the gate length with acceptable $|\Delta V_t|$ to a lower limit. If perfectly balanced, ΔV_t would be symmetrical around the nominal value, so $|\Delta V_t|$ is half of the V_t drop due to the DIBL effect.



(a) (b)

Fig. 2.10. Model verification with cubic pocket doping. (a) The pocket profile: $N_{poc}=1\times10^{18}cm^{-3}$, with width of Δ=40nm. The two pockets merge when L<70nm. (b) The good agreement between the model and simulation. (T_{ox}=2nm and d=15nm)



(a) (b)

Fig. 2.11. Model verification with Gaussian pocket doping. (a) The pocket profile: $N_{poc}=1\times10^{18}cm^{-3}$ and σ=30nm. $N(y)=N_{poc}[\exp(-\frac{y^2}{2\sigma^2})+\exp(-\frac{(L-y)^2}{2\sigma^2})]$. (b) The good agreement between the model and simulation. (T_{ox}=2nm and d=15nm)

A Gaussian pocket profile, which is more realistic in terms of manufacturing, is investigated in the second case (Fig. 2.11). Due to the tail of the Gaussian profile, the transition between the long and short channel regions is much smoother here. On the other hand, the Gaussian pocket is less effective because it peaks at the edges of the channel, while the cubic pocket extends more towards the center.

2.7 4th order approximation

The 2nd order approximation relies heavily on the assumption of L>>d, which is typically true for devices with good SCE. Discrepancy between the model and simulation is observed near the region of L/l_o~4. Moreover, when devices are scaled down, the tolerance of SCE is also increased, and L/l_o might be extended below 4. Therefore, the prediction accuracy of the model may not be sufficient at the short channel length, which is of great interest. Higher order terms must be evaluated to achieve a better prediction capability for the model with extended SCE limits.

When the infinite sum (4) is expanded into the 4th order of d/L, two scale lengths result, but only the larger one limits the SCE. The same identities (7) and (8) can be applied to each term to convert the potential into the closed-form result as in (26). (An undoped body is assumed here for simplicity.)

$$\frac{\cosh\left(\frac{k\pi}{L}x\right)}{\cosh\left(\frac{k\pi}{L}d\right)+\frac{\varepsilon_{si}}{\varepsilon_{ox}}T_{ox}\frac{k\pi}{L}\sinh\left(\frac{k\pi}{L}d\right)}\approx\frac{1}{1+\left(\frac{k\pi}{L}\right)^2\left(\frac{\varepsilon_{si}}{\varepsilon_{ox}}T_{ox}d+\frac{d^2-x^2}{2}\right)+\left(\frac{k\pi}{L}\right)^4\left(\frac{\varepsilon_{si}}{\varepsilon_{ox}}\frac{T_{ox}d^3}{6}+\frac{d^4}{24}+\frac{5x^4}{24}-\frac{\varepsilon_{si}}{\varepsilon_{ox}}\frac{T_{ox}dx^2}{2}-\frac{d^2x^2}{4}\right)}$$

$$=\frac{1}{[1+(k\pi l_1/L)^2][1+(k\pi l_2/L)^2]}=\frac{l_1^2}{l_1^2-l_2^2}\frac{1}{[1+(k\pi l_1/L)^2]}-\frac{l_2^2}{l_1^2-l_2^2}\frac{1}{[1+(k\pi l_2/L)^2]} \qquad (24)$$

$$\text{And } l_{1,2}^2=\frac{1}{2}\left(\frac{\varepsilon_{si}}{\varepsilon_{ox}}T_{ox}d+\frac{d^2-x^2}{2}\pm\sqrt{(\frac{\varepsilon_{si}}{\varepsilon_{ox}}T_{ox}d)^2+\frac{\varepsilon_{si}}{\varepsilon_{ox}}T_{ox}d(\frac{d^2}{3}+x^2)+\frac{d^4+d^2x^2-7x^4}{12}}\right) \qquad (25)$$

$$V(x,y) \approx V_{eff} - \frac{l_1^2}{l_1^2 - l_2^2}\frac{V_{eff}\sinh[(L-y)/l_1] + (V_{eff} - V_{ds})\sinh[y/l_1]}{\sinh(L/l_1)} + \frac{l_2^2}{l_1^2 - l_2^2}\frac{V_{eff}\sinh[(L-y)/l_2] + (V_{eff} - V_{ds})\sinh(y/l_2)}{\sinh(L/l_2)} \quad (26)$$

At V_{ds}=0, the minimum potential is still at y=L/2. While analytic forms of swing and the DIBL coefficient can be still derived, the V_t formula is too complicated to be derived without further approximations, and numerical calculation is needed.

$$S = 60\frac{mV}{dec}\Big/\left[1 - \frac{l_1^2}{l_1^2 - l_2^2}\frac{1}{\cosh(L/2l_1)} + \frac{l_2^2}{l_1^2 - l_2^2}\frac{1}{\cosh(L/2l_2)}\right] \text{ and } C_{DIBL}(0) = \frac{1}{2}\left(\frac{S}{60mV/dec} - 1\right) \quad (27)$$

In Fig. 2.12, the model results are extended to an unrealistic region, where swing exceeds 500mV and V_t roll-off is more than 2V. Still, the 4th order agrees excellently with the infinite sum results (equation (4)), and the improvement of this model is significant for short channel devices down to L=10nm.

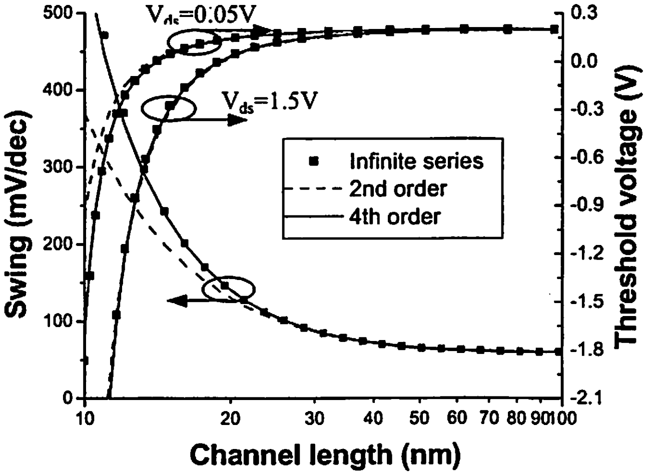


Fig. 2.12. Comparison between the 2nd and 4th order modeling results. Significant improvement is observed near the scaling limit, where severe SCE occurs. (T_{ox}=2nm, d=7nm, N_{body}=0, x=0)

2.8 2D effects in the gate dielectric

In the last few decades, the gate oxide has been dramatically scaled to achieve smaller scale lengths, but it is reaching its ultimate thickness limited by the gate tunneling current. At the 70nm technology node, the limit is about 12Å with the corresponding supply voltage and gate leakage tolerance [23]. High κ materials are proposed and investigated to suppress the gate tunneling leakage. In the expression of the scale length, only the ratio of T_{ox}/ε_{ox} is relevant because ideal front surface boundary conditions are assumed. When thick high κ materials are used, a thicker dielectric layer can achieve the same scale length with its high dielectric constant. However, the 2D effect in the gate dielectric, which was neglected in the previous discussion, becomes significant [24].

To model this phenomenon, a new approximation is made: the source and drain boundaries are extended all the way up to the gate, i.e. $V|_{y=0} = 0$, $V|_{y=L} = V_{ds}$ even in the insulator. By doing this, the 2D effect in the gate dielectric is overestimated, but an analytic solution with good fitting can result. The form of the solution in the dielectric is slightly different from that in the silicon due to the lack of ionized charges.

In the silicon: $V(x,y) = \sum_k \left[A_k \cosh\left(\frac{k\pi}{L}x\right) + B_k \sinh\left(\frac{k\pi}{L}x\right) - D_k \left(\frac{L}{k\pi}\right)^2 \right] \sin\left(\frac{k\pi}{L}y\right) + V_{ds}\frac{y}{L}$ (28)

In the gate dielectric: $V'(x,y) = \sum_k \left[E_k \cosh\left(\frac{k\pi}{L}x\right) + F_k \sinh\left(\frac{k\pi}{L}x\right) \right] \sin\left(\frac{k\pi}{L}y\right) + V_{ds}\frac{y}{L}$ (29)

At the boundaries $V(d,y) = V'(d,y)$, $\left.\frac{dV(x,y)}{dx}\right|_{x=d} = \frac{\varepsilon_I}{\varepsilon_{si}} \left.\frac{dV'(x,y)}{dx}\right|_{x=d}$ *and* $V'(d+T_I, y) = V_{eff}$

where T_I is the gate insulator thickness and ε_I is its relative dielectric constant. From the boundary conditions, the coefficients can be solved as

$$A_k = \frac{\frac{2}{k\pi}[(1-(-1)^k)V_{eff} + (-1)^k V_{ds}] + D_k \left(\frac{L}{k\pi}\right)^2 \cosh\left(\frac{k\pi}{L}T_I\right)}{\cosh\left(\frac{k\pi}{L}d\right)\cosh\left(\frac{k\pi}{L}T_I\right) + \frac{\varepsilon_{si}}{\varepsilon_I}\sinh\left(\frac{k\pi}{L}d\right)\sinh\left(\frac{k\pi}{L}T_I\right)} \text{ and } B_k = 0. \quad (30)$$

It is in agreement with equation (4) when expanded only into the first order of the T_I term. Similarly, the above result can be expanded to the 2nd order, and the identical closed-form formula for the potential profile can be achieved, except for the following modifications:

$$l=\sqrt{\frac{\varepsilon_{si}}{\varepsilon_I}T_I d+\frac{1}{2}(T_I^2+d^2-x^2)}=\sqrt{\frac{\varepsilon_{si}}{\varepsilon_{ox}}T_{eq}d+\frac{1}{2}\left[\left(\frac{\varepsilon_I}{\varepsilon_{ox}}T_{eq}\right)^2+d^2-x^2\right]} \text{ and } D_k\Rightarrow D_k\left(1-\frac{T_I^2}{2l^2}\right). \qquad (31)$$

Defined as $T_{eq}=T_I\,\varepsilon_{ox}/\varepsilon_I$, T_{eq} is the equivalent oxide thickness (EOT) when the 2D effect in the dielectric is ignored. This result implies that the 2D effect in the gate dielectric increases the scale length *l* and worsens the SCE. Since T_I does not always appear together with ε_I, the EOT alone is not sufficient in·the determination of device behaviors. When silicon oxide is used as the gate dielectric in the current technology, T_I<<d holds, and the modification is negligible. However, when thick high κ materials and ultra-thin bodies are used, T_I can be comparable to or even larger than d. This extra $T_I^2/2$ term becomes significant and its inclusion is essential for accurate modeling.

The modification of the D_k's reflects the fact that the ionized dopants do not exist in the dielectric. It can be further simplified as a scaling of the doping concentration; $N(y)\Rightarrow N(y)\left(1-\frac{T_I^2}{2l^2}\right)$.

With the use of the modified scale length and channel doping, the formulas for the potential profile, subthreshold swing, threshold voltage and DIBL coefficients remain the same. Therefore, the discussion in previous sections is still valid. L/*l* still determines the SCE of a device, but now the device dimensions must be scaled further for the same *l*.

Fig. 2.13 demonstrates the 2D effects of the high-κ dielectric on device SCE. In (a), the minimum channel length versus high-κ dielectric constant ε_I of the gate insulator is plotted. With a fixed EOT value, high-κ dielectric shows more 2D effects, and the degradation becomes obvious when ε_I >10. With high ε_I, the $T_I^2/2$ term dominates and

$L_{min} \propto \varepsilon_I$. With a fixed T_I value, the improvement from the higher dielectric constant diminishes at high ε_I. In (b), the high-κ EOT without the 2D effect in the dielectric is converted to a SiO_2 thickness for the same SCE (or the same *l*). Due to the increased 2D effect, the EOT of the high-κ dielectric must be thinner than that of SiO_2 to achieve the same scale length. For example, a gate dielectric of 0.5nm EOT with ε_I =100 is as effective as a 5nm SiO_2. Therefore, the use of the modified scale length *l*, which includes the 2D effect in the gate dielectric, is crucial to evaluate the efficiency of high-κ dielectrics in the suppression of SCE.

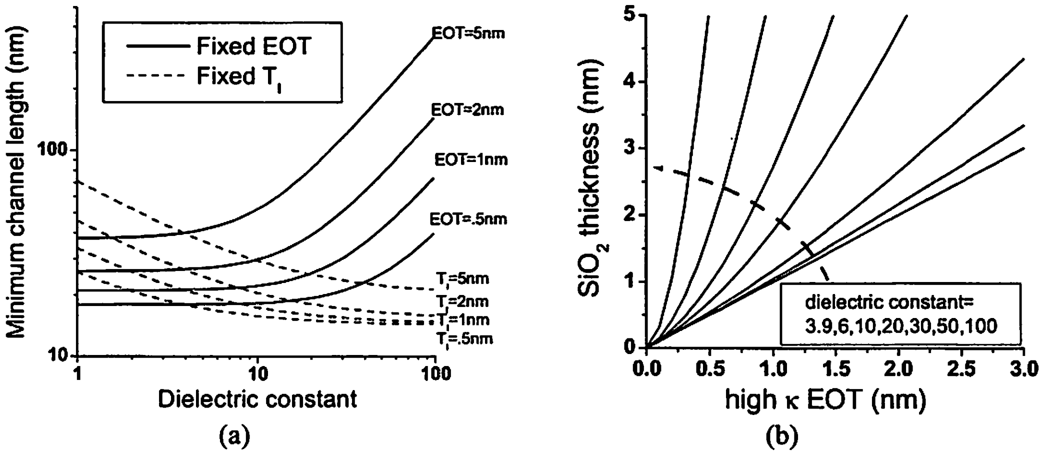


Fig. 2.13. 2D effects of the gate dielectric layer. (a) The minimum channel length as a function of the dielectric constant, either by fixing the EOT or T_I at 0.5, 1, 2 and 5nm. (d=5nm) (b) SiO_2 thickness vs. high-κ EOT for the same SCE. (d=5nm)

Equation (30) can be expanded to the 4th order too for better accuracy. With the 2D effect in the gate dielectric included, a result similar to equation (26) can be reached, except for another modification of the scale lengths.

$$l_{1,2}^2 = \frac{1}{2}\left[\frac{\varepsilon_{si}}{\varepsilon_I}T_I d + \frac{d^2+T_I^2}{2} \pm \sqrt{\frac{d^4+T_I^4-2d^2T_I^2}{12} + \left(\frac{\varepsilon_{si}}{\varepsilon_I}\right)^2 d^2T_I^2 + \frac{1}{3}\frac{\varepsilon_{si}}{\varepsilon_{ox}}T_I d(d^2+T_I^2)}\right] \quad \text{at } x=0 \qquad (32)$$

$$l_{1,2}^2 = \frac{1}{2}\left[\frac{\varepsilon_{si}}{\varepsilon_I}T_I d + \frac{T_I^2}{2} \pm \sqrt{\frac{T_I^4}{12} + \left(\frac{\varepsilon_{si}}{\varepsilon_I}\right)^2 d^2T_I^2 + \frac{1}{3}\frac{\varepsilon_{si}}{\varepsilon_{ox}}T_I d(4d^2+T_I^2)}\right] \quad \text{at } x=d \qquad (33)$$

The 2D effect in the gate dielectric is addressed through another approach by Frank [5]. There, the scaling length can be solved from an implicit equation:

$$\varepsilon_{si} / \varepsilon_I \tan(T_I / l) \tan(d / l) = 1 \qquad (34)$$

Fig. 2.14 plots the scaling lengths obtained from our 2nd and 4th order models, as well as the model in Ref. [5], with $T_{ox,eq}$=1,2,5 nm and various dielectric constants. All three models agree very well in the whole range of dielectric constants and T_{ox}, except that the 2nd order model always gives a slightly larger scaling length. The difference between models is small, i.e. about 5% in the scale length. Even in the extreme case of L=3*l*, the difference accounts for only 25mV in $\Delta V_{t,sat}$ and 2% in S while S=104mV. The 4th order model may be essential in future generations where higher S and ΔV_t are tolerated and L is pushed far beyond 4*l*. On the other hand, the 2nd order model captures most of the physics and is still very simple to implement.

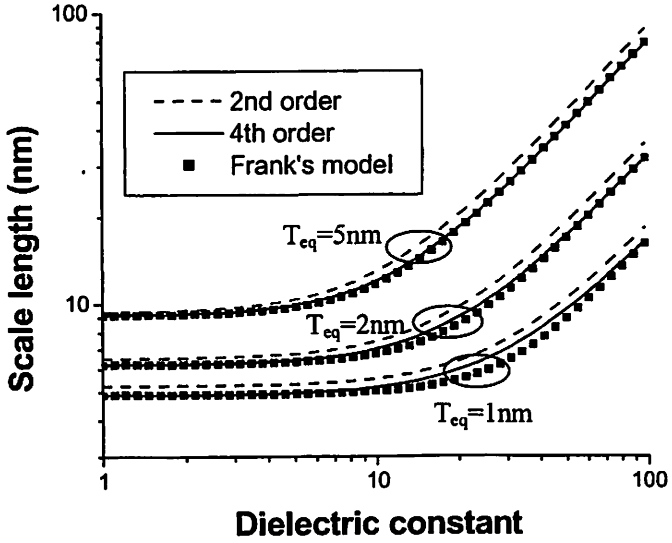


Fig. 2.14. Comparison between the 2nd and 4th order modeling results with Frank's model. The explicit expression here agrees with the numerical solution of the Frank's model. (d=5nm and undoped body)

2.9 Conclusion

A two-dimensional analytical subthreshold model for fully depleted SOI and double-gate MOSFETs has been developed. The subthreshold behaviors (swing, V_t, DIBL) and 2D potential profiles are calculated from the device's physical dimensions and doping profile. The 2D effects in both the channel and high-κ gate dielectric on short channel effects are incorporated into a scale length $l=\sqrt{\frac{\varepsilon_{si}}{\varepsilon_I}T_I d+\frac{T_I^2+d^2-x^2}{2}}$. Pocket doping can be used to improve the V_t roll-off, but not the subthreshold swing or DIBL effect. Body doping engineering becomes ineffective in controlling either SCE or V_t when the L_{min} is scaled below 25nm. If $d >> T_{ox}$, the scaling of the body thickness is more effective than is the scaling of T_{ox} for suppressing SCE. The model also predicts the structural dimensions for very short channel devices (L~10nm) as long as the thermionic current leakage still dominates. Improved prediction capability can be acquired by the use of the more complicated 4th order model. The closed form results can be easily used to guide the design of device dimensions and doping profiles.

2.10 References

[1] T. Toyabe and S. Asai, "Analytical models of threshold voltage and breakdown voltage of short-channel MOSFETs derived from two-dimensional analysis." *IEEE Journal of Solid-State Circuits, Vol. SC-14, No. 2, pp.375-383, April 1979*

[2] C. Caillat, S. Deleonibus, G. Guegan, S. Tedesco, B. Dal'zotto, M. Heitzmann, F. Martin, P. Mur, B. Marchand and F. Balestra, "65 nm physical gate length NMOSFETs with heavy ion implanted pockets and highly reliable 2 nm-thick gate oxide for 1.5 V operation," *Symposium on VLSI Technology, pp. 89 –90, June 1999*

[3] C.H. Choi, S.J. Rhee, T.S. Jeon, N. Lu, J.H. Sime, R. Clark, M. Niwa and D.L. Kwong, "Thermally stable CVD HfO_xN_y advanced gate dielectric with poly-Si gate electrode", *IEEE. Proceedings of the International Electron Device Meeting, pp. 857-860, Dec, 2002*

[4] B. Yu, Y-J. Tung, S. Tang, E. Hui, T-J. King and C. Hu, "Ultra-thin-body silicon on insulator MOSFETs for terabit-scale integration," *Proceedings of the International Semiconductor Device Research Society, pp. 623-628, 1997*

[5] D. Frank, Y. Taur and H. Wong, "Generalized Scale Length for Two-Dimensional Effects in MOSFET's" *IEEE Electron Device Letter, vol. 19, No. 10, pp. 385-387, 1998*

[6] I. Sungjun and K. Banerjee, "Full chip thermal analysis of planar (2-D) and vertically integrated (3-D) high performance ICs." *IEEE. Proceedings of the International Electron Device Meeting, pp. 727-730, Dec. 2000*

[7] International Technology Roadmap for Semicondutors, Semicondutor Industry Association, 2001. http://public.itrs.net/Files/2002Update/2001ITRS/Home.htm.

[8] A. Keshavarzi, S. Narendra, B. Bloechel, S. Borkar and V. De, "Forward body bias for microprocessors in 130nm technology generation and beyond." *IEEE. Symposium on VLSI Circuits. Digest of Technical Papers, pp.312-315. 2002*

[9] Y. Taur and E.J. Nowak, "CMOS devices below 0.1 mu m: how high will performance go?" *IEEE. Proceedings of the International Electron Device Meeting, pp. 215-218, Dec. 1997*

[10] F. Assaderaghi, D. Sinitsky, J. Bokor, P.K. Ko, H. Gaw and C. Hu, "High-field transport of inversion-layer electrons and holes including velocity overshoot" *IEEE.*

Trans. On. Electron devices, Vol. 44, No. 4, pp. 664-671, April, 1997

[11] K. Rim, S. Narasimha, H. Longstreet, A. Mocuta and J. Cai, "Low field mobility characteristics of sub-100nm unstrained and strained Si MOSFET." *IEEE. Proceedings of the International Electron Device Meeting, pp. 43-46, Dec 2002*

[12] T. Tezuka, N. Sugiyama, T. Mizuno and S. Takagi, "Ultrathin body SiGe-on-insulator pMOSFETs with high-mobility SiGe surface channels," *IEEE. Transactions on Electron Devices, Vol. 50, No. 5, pp. 1328 –1333, May 2003*

[13] L.T. Su, J.B. Jacobs, J.E. Chung and D.A. Antoniadis, "Deep-submicrometer channel design in silicon-on-insulator (SOI) MOSFET's," *IEEE. Electron Device Letters, Vol. 15, No. 9, pp.366-369, Sept. 1994.*

[14] N. Lindert, L. Chang, Y-K. Choi, E.H. Anderson, W-C. Lee, T-J. King, J. Bokor and C. Hu, "Sub-60-nm quasi-planar FinFETs fabricated using a simplified process." *IEEE. Electron Device Letters, Vol. 22, No. 10, pp. 487-489, Oct. 2001.*

[15] I.S. Gradshteyn and I.M. Ryzhik, "Table of integrals, series, and products", *p. 40, 1980*

[16] K. Young, "Short-channel effect in fully depleted SOI MOSFET's", *IEEE. Trans. on Electron Devices, Vol. 36, No. 2, pp. 399-402, 1989*

[17] R. Yan, A. Ourmazd and K. Lee, "Scaling the Si MOSFET: from bulk to SOI to bulk", *IEEE. Trans. on Electron Devices, Vol. 39, No. 7, pp. 1704-1710, 1992*

[18] G. Niu, R. Chen and G. Ruan, "Comparisons and extension of recent surface-channel SOI MOSFET's", *IEEE. Trans. on Electron Devices, Vol. 43, No. 11, pp. 2034-2037, 1996*

[19] K. Suzuki, T. Tanaka, Y. Tosaka and H. Horie, "Scaling theory for double-gate

MOSFET's", *IEEE. Trans. on Electron Devices, Vol. 40, No. 12, pp. 2326-2329, 1993*

[20] M. Takamiya, T. Yasuda and T. Hiramoto, "Deep sub-0.1um fully depleted SOI MOSFET's with ultra-thin silicon film and thick buried oxide for low-power applications", *IEEE. Proceedings of International Semiconductor Device Research Symposium, pp. 215-218, 1997*

[21] H. Majima, H. Ishikuro and T. Hiramoto, "Experimental evidence for quantum mechanical narrow channel effect in ultra-narrow MOSFET's", *IEEE. Proceedings of International Electron Devices Meeting, pp. 396-398, Dec. 2000*

[22] W. Liu, X. Jin, Y. King and C. Hu, "An efficient and accurate compact model for thin-oxide-MOSFET intrinsic capacitance considering the finite charge layer thickness" *IEEE. Trans. on Electron Devices, Vol. 46, No. 5, pp. 1070-1072, May 1999*

[23] S. Song, H. 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." *IEEE. Proceedings of International Electron Devices Meeting. pp. 55-58, Dec. 2001*

[24] X. Liu, J. Kang, L. Sun, R. Han and Y. Wang. "Threshold voltage model for MOSFETs with high κ gate dielectrics," *IEEE. Electron Device Letters, Vol. 23, No. 5, pp.270-272, May 2002*

Appendix 2A Simplification of ΔV

Note the fact that the channel doping profile is symmetric around y=L/2, i.e. $\rho(y) = \rho(L-y)$. So D_k's are nonzero only when k is odd.

$$\Delta V = -\sum_{k=1}^{\infty} \frac{D_k l^2}{1+(k\pi\, l/L)^2} \sin\left(k\pi \frac{y}{L}\right) = -\frac{1}{2}\sum_{k=1}^{\infty} \frac{\left(1-(-1)^k\right) D_k l^2}{1+(k\pi\, l/L)^2} \sin\left(k\pi \frac{y}{L}\right)$$

Combining two identities in (7), we can get

$$\sum_k \frac{[1-(-1)^k]\cos k\pi y}{1+(k\pi / a)^2} = \frac{a}{2}\left[\frac{\cosh(a-ay)}{\sinh a} - \frac{\cosh ay}{\sinh a}\right] = \frac{a}{\sinh a}\sinh\frac{a}{2}\sinh\left(\frac{a}{2}-ay\right) = \frac{a}{2}\cdot\frac{\sinh(a/2-ay)}{\cosh a/2}$$

Let $f(t) = \sum_{k=1}^{\infty} A_k \sin(k\pi t)$ and $g(t) = \sum_{k=1}^{\infty} B_k \cos(k\pi t)$. Their convolution can be calculated as: (* represents the convolution of two functions.)

$$h(x) = f(x) * g(x) \equiv 2\int_0^1 f(t) g(x-t)dt = \sum_{j,k=1}^{\infty} A_j B_k \int_0^1 2\sin(j\pi t)\cos[k\pi(x-t)]dt$$

$$= \sum_{j,k=1}^{\infty} A_j B_k \int_0^1 \{\sin[\pi t(j-k)+k\pi x] + \sin[\pi t(j+k)-k\pi x]\}dt = \sum_{j,k=1}^{\infty} A_j B_k \sin(k\pi x)\delta_{jk} = \sum_{k=1}^{\infty} A_k B_k \sin(k\pi x)$$

Therefore, $\Delta V = -\frac{l^2}{2}\sum_{k=1}^{\infty} D_k \frac{1-(-1)^k}{1+(k\pi\, l/L)^2}\sin\left(k\pi\frac{y}{L}\right) = -\frac{l^2}{2}\rho(y) * \frac{L}{2l}\frac{1}{\cosh(L/2l)}\sinh\frac{L/2-|y|}{l}$

Also, due to the symmetry of the doping profile, ΔV is symmetric to y=L/2; therefore only the calculation for y<L/2 is required.

$$\Delta V = -\frac{l^2}{2}\frac{2}{L}\int_0^L \rho(t)\frac{L}{2l}\frac{1}{\cosh(L/2l)}\sinh\frac{L/2-|y-t|}{l}dt$$

$$= \frac{-l}{2\cosh(L/2l)}\left[\int_0^y \rho(t)\sinh\frac{L/2-(y-t)}{l}dt + \left(\int_y^{L/2} + \int_{L/2}^{L-y} + \int_{L-y}^{L}\right)\rho(t)\sinh\frac{L/2-(t-y)}{l}dt\right]$$

$$= \frac{-l}{2\cosh(L/2l)}\left[\int_0^y \rho(t)\left(\sinh\frac{L/2-y+t}{l} + \sinh\frac{y-L/2+t}{l}\right)dt + \int_y^{L/2} \rho(t)\left(\sinh\frac{L/2-t+y}{l} + \sinh\frac{t-L/2+y}{l}\right)dt\right]$$

In the last step, t is transfered to L-t for t>L/2, and $\rho(t) = \rho(L-t)$ is used. Finally,

$$\Delta V = \frac{-l}{\cosh(L/2l)}\left[\cosh\frac{L/2-y}{l}\int_0^y \rho(t)\sinh\frac{t}{l}dt + \sinh\frac{y}{l}\int_y^{L/2} \rho(t)\cosh\frac{L/2-t}{l}dt\right]$$

Appendix 2B Calculation of the barrier and V$_t$ with uniform doping

First, we want to solve for the barrier height in the channel, which is the minimum potential of the profile along the y direction.

$$V(x,y) = V'_{eff} - \frac{1}{2\sinh L/l}\left\{\left[V'_{eff} - V_{ds} - V'_{eff}\exp\left(-\frac{L}{l}\right)\right]\exp\left(\frac{y}{l}\right) + \left[V'_{eff}\exp\left(\frac{L}{l}\right) - (V'_{eff} - V_{ds})\right]\exp\left(-\frac{y}{l}\right)\right\}$$

From $\frac{dV}{dy} = 0$ we get $\exp\left(\frac{y}{l}\right)\left[V'_{eff} - V_{ds} - V'_{eff}\exp\left(-\frac{L}{l}\right)\right] = \exp\left(-\frac{y}{l}\right)\left[V'_{eff}\exp\left(\frac{L}{l}\right) - (V'_{eff} - V_{ds})\right]$

Therefore, $y = \frac{L}{2} + \frac{l}{2}\ln\left\{\left[V'_{eff} - (V'_{eff} - V_{ds})\exp\left(-\frac{L}{l}\right)\right] \Big/ \left[V'_{eff} - V_{ds} - V'_{eff}\exp\left(-\frac{L}{l}\right)\right]\right\}$

$$V_{\min}(x) = V'_{eff} - \frac{1}{\sinh(L/l)}\sqrt{\left[V'_{eff} - V_{ds} - V'_{eff}\exp\left(-\frac{L}{l}\right)\right]\left[V'_{eff}\exp\left(\frac{L}{l}\right) - (V'_{eff} - V_{ds})\right]}$$
$$= V'_{eff} - \frac{1}{\sinh(L/l)}\sqrt{4V'_{eff}(V'_{eff} - V_{ds})\sinh^2\left(\frac{L}{2l}\right) - V_{ds}^2}$$

At the threshold bias, V$_g$=V$_t$, and V$_{min}$(x)=V$_b$. V$_t$ can be solved:

$$4V'_{eff}(V'_{eff} - V_{ds})\sinh^2\frac{L}{2l} - V_{ds}^2 = (V'_{eff} - V_b)^2\sinh^2\frac{L}{l} = 4(V'_{eff} - V_b)^2\sinh^2\frac{L}{2l}\cosh^2\frac{L}{2l}$$
$$V'^{\,2}_{eff}\sinh^4\frac{L}{2l} - V'_{eff}\sinh^2\frac{L}{2l}(2V_b\cosh^2\frac{L}{2l} - V_{ds}) + V_b^2\sinh^2\frac{L}{2l}\cosh^2\frac{L}{2l} + \frac{V_{ds}^2}{4} = 0$$
$$\therefore V'_{eff} = \left[V_b\cosh^2\frac{L}{2l} - \frac{V_{ds}}{2} - \sqrt{V_b(V_b - V_{ds})}\cosh\frac{L}{2l}\right] \Big/ \sinh^2\frac{L}{2l}$$

$$V_t = \phi_{gs} + \rho l^2 + V_b + \left[V_b - \frac{V_{ds}}{2} - \sqrt{V_b(V_b - V_{ds})}\cosh\frac{L}{2l}\right] \Big/ \sinh^2\frac{L}{2l}$$

Chapter 3

Solid Phase Epitaxy for UTB MOSFETs

3.1 Introduction

Fully depleted (FD) structures show promise in suppressing short channel effects (SCE) in sub-50nm transistors, and body thickness is the most critical parameter in device design [1,2]. In Chapter 2, it has been proved that $\mathrm{L}_{\min} \geq 4l = 4\sqrt{\frac{\varepsilon_{si}}{\varepsilon_{ox}} T_{ox} d + \frac{d^2}{2}}$ for an acceptable level of SCE; therefore, $d < L_{\min}/\sqrt{8} \approx L_{\min}/3$. Although, the formation of the ultra-thin channel imposes tremendous process difficulties, several attractive methods have been proposed. In a FinFET, the narrow fin has a limited width of $\mathrm{T}_{si} = 2d < L_{\min}/1.5$, and it is typically created via lithography. In a current CMOS process, the gate patterning is the limiting lithography step. Since the fin must be narrower than the gate, the fabrication process requires a higher lithography capability beyond the scaling limit [3]. In an ultra-thin-body (UTB) MOSFET, the thin film is commonly generated by thinning down a thick SOI (silicon on insulator) film in multiple etches or oxidations [4]. However, the thickness uniformity of the initial SOI wafer is around 5nm, and any non-self limiting process introduces extra thickness variation, so mass production of uniform 10nm thin films using these approaches will be challenging.

LPCVD (low pressure chemical vapor deposition) films are well known to be uniform and controllable in thickness, but typically are amorphous or poly-crystalline. Solid-phase-epitaxy (SPE) is proposed as a practical approach to convert a uniform LPCVD amorphous film into a single crystalline form. Solid-phase-crystallization has been widely used to grow polycrystalline films with large grains in thin-film-transistor (TFT) applications [5]. To get a single crystalline silicon film with a controlled orientation, seeds are required to initiate the crystallization. Since the channel film will lie on top of an oxide, lateral crystallization is required [6].

Fig. 1a shows the basic concept of lateral SPE. First, trenches are created in the initial SOI film, and a thin amorphous silicon film is deposited on both the silicon and the oxide surface. Crystallization in lateral directions from the remaining silicon islands occurs throughout the entire amorphous silicon film. This novel manufacturing process opens a window for improvement of conventional CMOS. It is feasible to use other materials as channel films as long as they can grow epitaxially on silicon surfaces. Strained SiGe film is one promising candidate to improve the performance of the transistors due to its high mobility [7]. Moreover, since the channel film can be created by epitaxy, SPE offers opportunities for multi-layer circuit integration [8]. There are two forms of SPE: non-planarized and planarized. In a planarized scheme, the trenches are filled with oxide, which is then etched back to achieve a planar surface (Fig. 1b). MOSFETs fabricated from non-planarized SPE films have already been published [9], and the ones with the planarized SPE will be presented in this report.

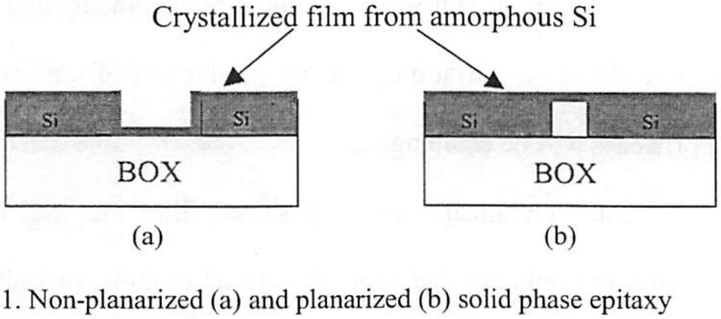


Fig. 1. Non-planarized (a) and planarized (b) solid phase epitaxy

The MOSFET with a planarized SPE film has the following advantages over its counterpart. First, it has a flat topography, leading presumably to better crystalline quality for the final SPE film. Second, it lowers, rather than raises, the source/drain. In this way, the parasitic capacitance between the gate and source/drain can be dramatically reduced [10]. Third, if the trench is made much deeper than S/D junctions, SPE transistors can be fabricated on bulk wafers to avoid the use of SOI wafers, which is a major part of SOI chip cost. Moreover, the SPE process is fully compatible with that of conventional bulk CMOS, so with the addition of a trench mask layer, the SPEFETs can be easily integrated with bulk MOSFETs on the same chip. Fourth, it provides opportunities for novel device structures. A double gate MOSFET becomes possible with a back-gate electrode inserted into the trench [11]. Since the trench can be either larger or smaller than the gate, the influence of the trench size can be also studied.

In this report, UTB MOSFETs are fabricated on planarized lateral SPE channel films. The quality of the resulting SPE film is evaluated via the performance of transistors fabricated on it.

3.2 Fabrication process

3.2.1 Lateral solid-phase-epitaxy

The key issues for the SPE process are surface cleaning, the low temperature deposition of the amorphous film, and low temperature crystallization. Clean surfaces on the seed islands are essential for the successful initiation of solid phase epitaxy. One way of achieving a clean surface is with an HF-last clean. A hydrogen passivation layer exists on the silicon surface after HF cleaning, and it prevents the reaction of the silicon with oxygen or water vapor in the air. Experiments have shown that the hydrogen-terminated silicon surface can be preserved for a few minutes at 400°C [12]. Wafers are dried and put into the furnace right after the HF dip without DI water rinsing. The number of wafers in each batch is limited to two because of the short time available between cleaning and loading into the furnace.

At a high temperature, the amorphous silicon film starts to nucleate at random positions and orientations and develops into a polycrystalline film, so both the deposition and the annealing have to be performed at low temperatures. Because the minimum temperature for SiH_4 deposition is too high (550°C) [13], Si_2H_6 gas is used at 410°C and 300mTorr. Even with these precautions, some quantity of native oxide still forms on the seed surfaces, and a silicon implant is required before crystallization to break it up [14]. The energy (20keV) and dose ($7*10^{15}cm^{-2}$) of the silicon implant are chosen in order to break up the native oxide but not amorphorize the 80nm thick silicon seeds all the way to the BOX, which action is verified by a Monte Carlo simulation (TRIM).

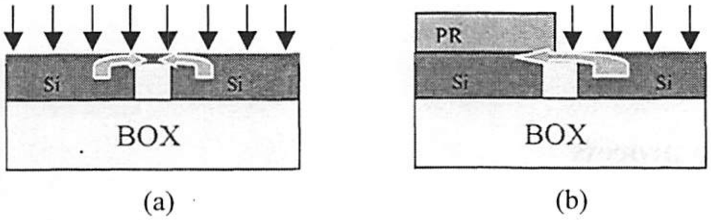


Fig. 2. Crystallization from (a) both side and (b) one side. A twin boundary is expected at the middle of the film in case (a).

In the MOSFET structure, those two seed regions will be the lowered source and the drain. A twin boundary is expected at the location where two crystallization fronts meet. The boundary will introduce trap states, which form a potential barrier and scatter the carriers in the channel. To study the degradation of the drive current resulting from the boundary, the silicon implant is masked to break up the native oxide either on two sides or on only one side. The boundary will be located in the middle of the channel if the film is laterally crystallized from both ends, while it will be out of the channel if the film is crystallized from only one side (Fig. 2).

The crystallization is initiated at 550°C for more than 24 hours and sped up by a subsequent annealing step at 600°C for 12 hours. Finally, a 950°C 30 minutes thermal step removes most of the defects in the final film. The thicker the film, the farther SPE propagates, and the aspect ratio of the crystallization range to the film thickness is about 10~20 [15]. Rutherford Backscattering Spectrometry (RBS) in a channeling geometry shows the influences of the annealing and the Si implant dose on the final film quality (Fig. 3). Low channeling at greater depth is observed after the implantation, because part of the seed has been amorphorized. With a sufficient Si implant dose and long-time SPE anneal at 550°C, a good single crystal film is achieved, indicated by the low count of high energy ions reflected at the film surface.

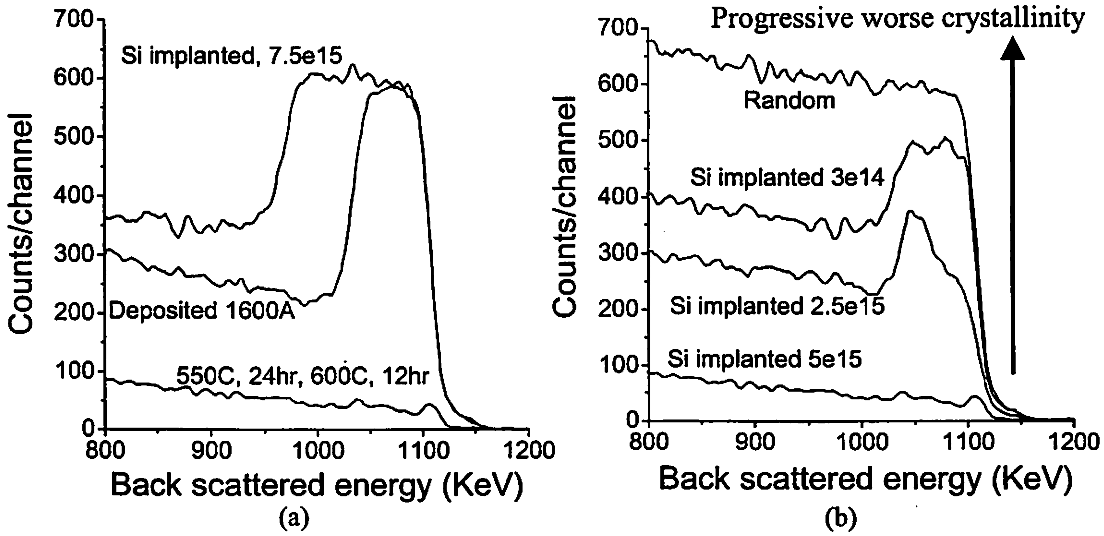
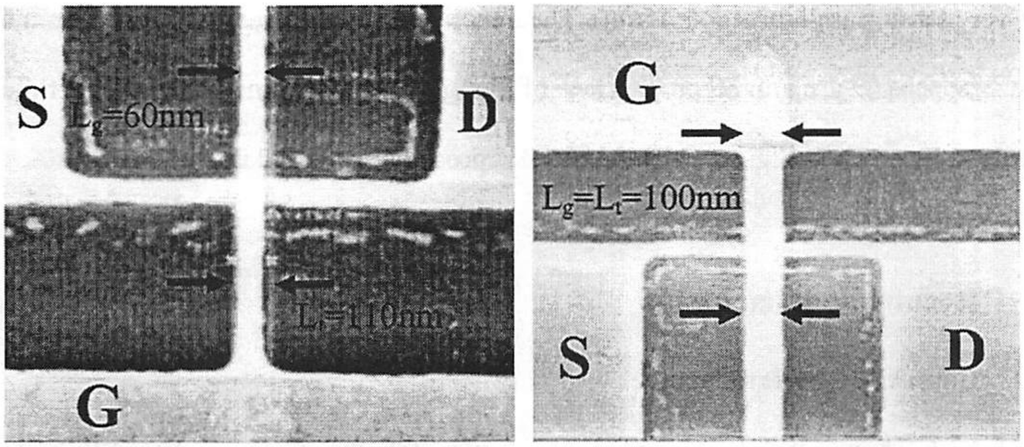


Fig. 3. RBS of the films (a) before and after anneal, and (b) with different Si implant dose after annealing at 550°C for 24hrs and subsequently at 600°C for 12hrs.

3.2.2 SPEFET fabrication

The starting SOITECH wafer had a 100nm (100) silicon film on a 400nm BOX (buried oxide). Since the electron beam (E-beam) lithography facilities at Lawrence Berkeley National Laboratory (LBNL) were used for all small features, such as the trench, mesa and gate, special E-beam alignment marks of 800nm-thick SiGe were created first. Trenches were created in the <110> direction in the SOI film for layout convenience, as <110> orientated trenches should result in better SPE film quality since SPE grows preferentially in the <110> direction [16]. After they were filled with an HTO/LTO (high/low temperature oxide) stack, the oxide was etched-back to create a planar surface using a reactive ion etch followed by a wet etch. A 100Å or 200Å amorphous silicon channel film was deposited on both the trench and silicon seeds, using the same deposition recipe as discussed in Section 3.2.1. Then Si ion implantation was performed to break up the native oxide between the seeds and deposited film. This implant was masked on selected devices for the purpose of positioning the twin boundary. The amorphous film was crystallized in multiple thermal steps, as described in the previous section.

A standard CMOS process was adapted for the following fabrication steps. Simple mesa isolation was used for devices on SOI wafers. The gate stack consisted of a 2.5nm thermal oxide, which was grown at 750°C and annealed at 900°C, and a 130nm P^+ $Si_{0.8}Ge_{0.2}$ gate deposited by LPCVD. After the gate was patterned through E-beam lithography, the source/drain extension was implanted with $1X10^{13}$ cm^{-2} of boron at 5keV for PMOS and $5X10^{13}$ cm^{-2} of arsenic at 10keV for NMOS. Then, a 200Å oxide spacer was formed, followed by heavy S/D implantation. After the source/drain doping was activated with a rapid thermal annealing for 1min at 900°C, the standard back-end process completed the fabrication.



(a) (b)

Fig. 4. SEM images of SPEFETs. (a) The trench is larger than the gate. (b) The trench has the same size as the gate.

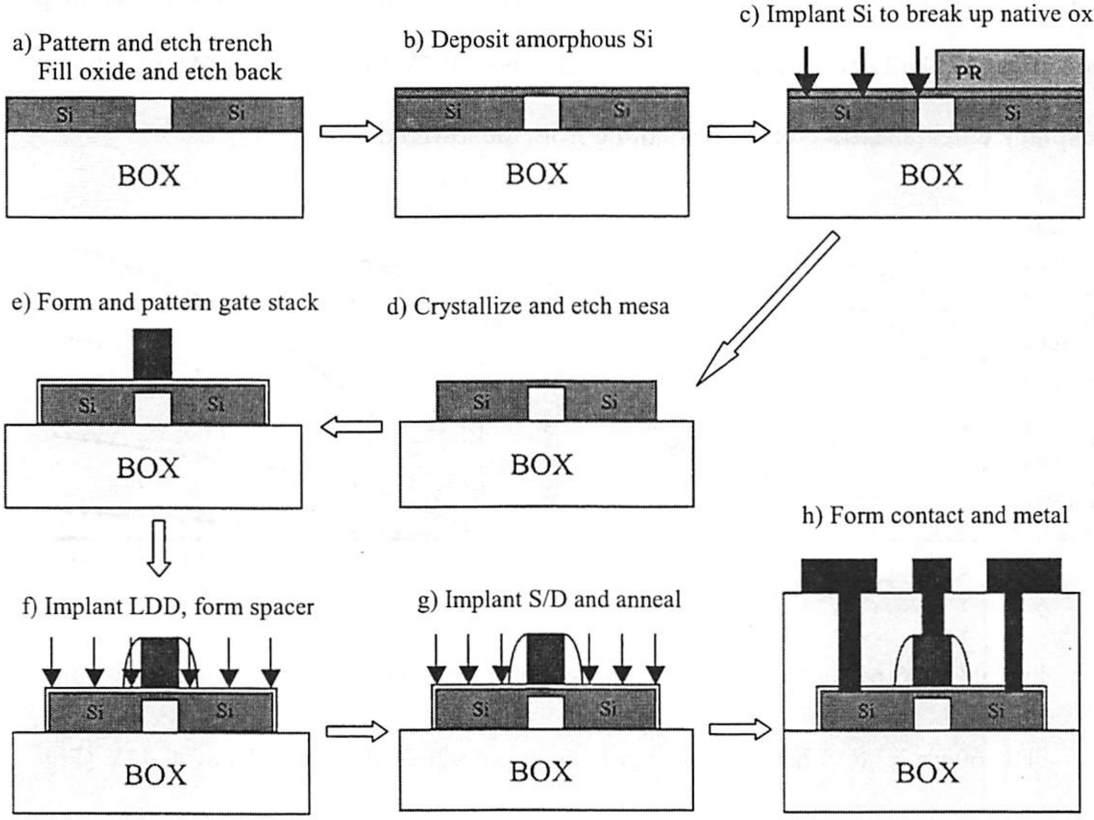

a) Pattern and etch trench
Fill oxide and etch back
Si
Si
BOX
b) Deposit amorphous Si
Si
Si
BOX
c) Implant Si to break up native oxide
PR
Si
Si
BOX
d) Crystallize and etch mesa
Si
Si
BOX
e) Form and pattern gate stack
Si
Si
BOX
f) Implant LDD, form spacer
Si
Si
BOX
g) Implant S/D and anneal
Si
Si
BOX
h) Form contact and metal
Si
Si
BOX

Fig. 5. The process flow of SPEFET fabrication.

Fig. 4 shows two Scanning Electron Microscopy images of the SPEFETs after gate patterning. Smooth small features can be generated through E-Beam lithography, with very small misalignment (<15nm). The trench sizes can be larger or smaller than the gate size, opening up another parameter of design optimization. The process flow schematic is shown in Fig. 5, while the detailed process flow is listed in Appendix 3A.

3.3 Results and discussion

3.3.1 60nm CMOS performance

Fig. 6 shows the I_d-V_g and I_d-V_d curves for the 60nm P-type SPEFET: S=105mV/dec, V_t=-0.13V (defined at I_d=100nA/μm), I_{off}=6nA/μm, and I_{on}=410μA/μm at V_g=1.5V. PMOS shows an excellent drive current, higher than what the roadmap specifies [17]. The DIBL effect is slightly high, about 0.3V for a V_{ds} of 1.5V, which can be explained by the extra electrical coupling from the lowered S/D.

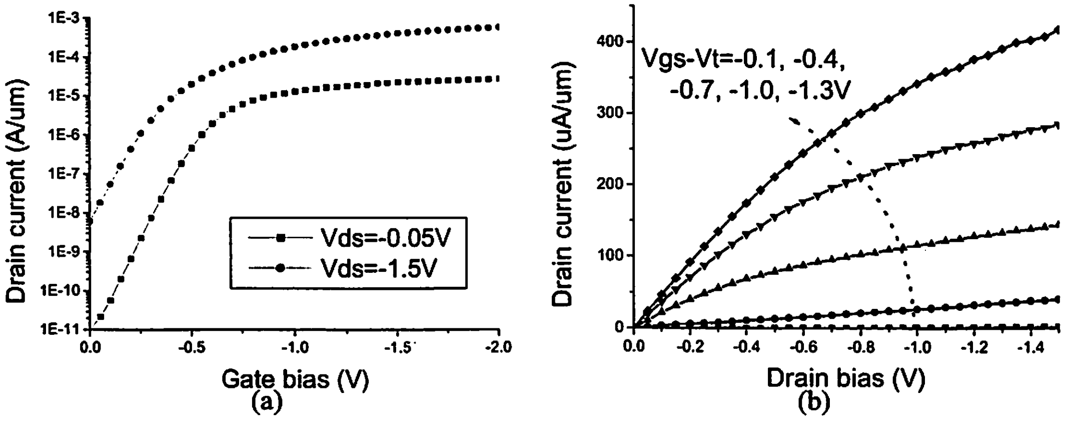


Fig. 6. PMOS performance. W/L=100nm/60nm, T_{ox}=2.5nm, I_{on}=410μA/μm, I_{off}=6nA/μm.

The 60nm NMOS has a subthreshold swing of 92mV/dec and a V_t of -0.47V (Fig. 7). If V_t can be shifted up to 0.2V, then I_{off}=1nA/μm, and I_{on}=480μA/μm for the 1.5V voltage supply. Since the silicon-oxide interface traps cause a V_t shift, the fact that the NMOS V_t deviates from the expected value indicates a high interface state density for NMOS. The NMOS I_{on} is slightly lower than expected, which could also be explained by the scattering of the carriers from the surface states.

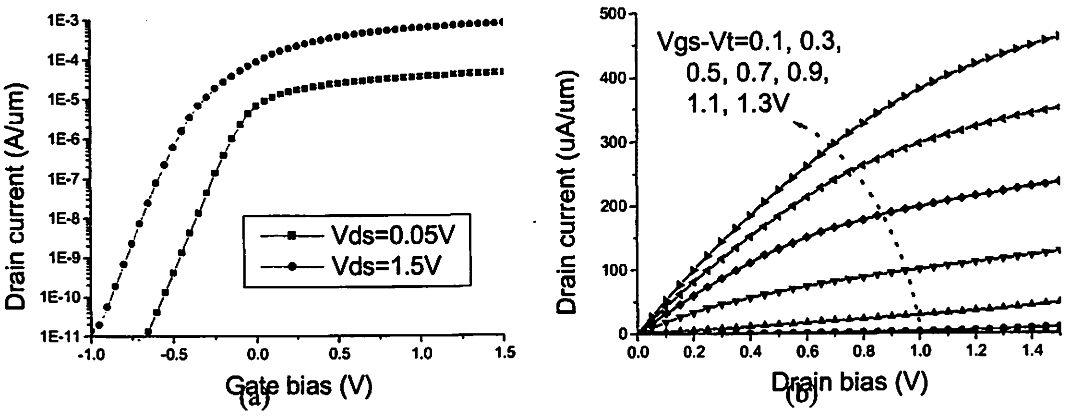


Fig. 7. NMOS performance. W/L=100nm/60nm, T_{ox}=2.5nm, I_{on}=480μA/μm, I_{off}=1nA/μm.

3.3.2 Defects and channel length dependence

Channel length dependence is another important feature of device performance. SPEFETs show very unusual L dependence, as seen in Fig. 8. First, the excellent V_t roll-off of the PMOS demonstrates that the intrinsic structure of SPEFETs can be scaled beyond 60nm. But the traps in the films of long channel devices cause abnormally high NMOS V_t at large channel lengths. Second, the drive-current, which is measured at V_g-V_t=1.3V, decreases much faster than 1/L. The big drop in the drive current between 60nm and 80nm devices suggests that our annealing process can generate a crystalline region with no defects or with a low defect density only at a length of 60nm. When the channel length is beyond 60nm, the defects in the SPE channel film degrade the drive current and increase the threshold voltage. As has been published, dislocation networks dominate within a short SPE range, while microtwins and twins extending through the entire film are generated when the SPE range is long [18].

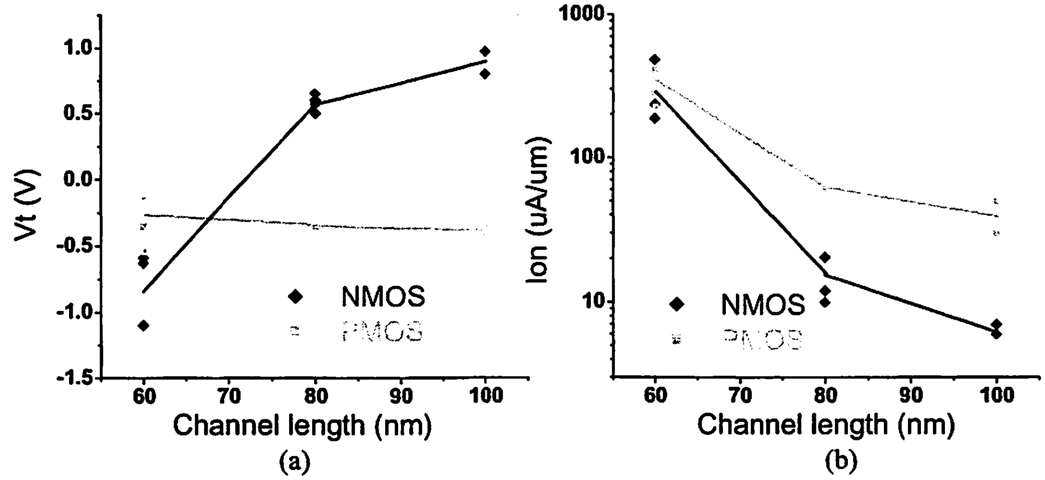


Fig. 8. V_t roll-off and I_{on} dependence on L of (a) NMOS and (b) PMOS. The drive current is measured at V_g-V_t=1.3V.

Fig. 8 also suggests that PMOS has a much better channel length dependence in both I_{on} and V_t. With the use of undoped bodies, PMOS and NMOS have identical device structures, except for the bias condition. Therefore, this asymmetry of polarity strongly suggests non-uniform trap density in the band gap. In an NMOS, the Fermi level (E_f) is near the conduction band (E_c) at the silicon-oxide interface, while it is near the valence band (E_v) in a PMOS (Fig. 9). In the NMOS, the unexpected value of V_t, together with the large performance degradation with L and the low drive current, suggests that the interface trap density is higher near E_c.

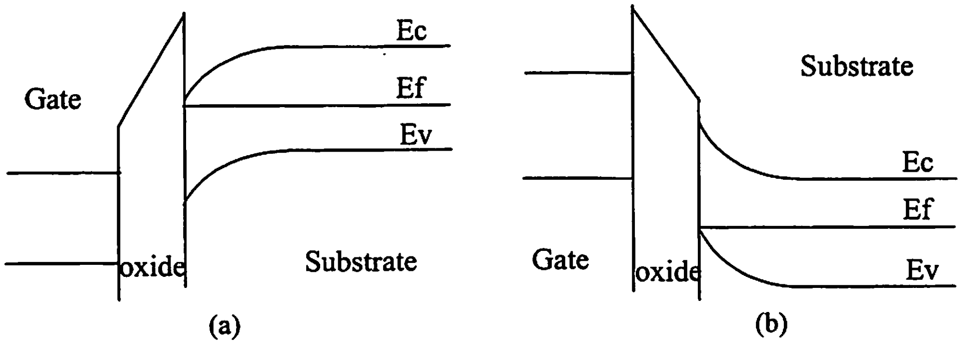


Fig. 9. Gate-oxide-substrate band diagrams for (a) NMOS and (b) PMOS.

3.3.3 SPE twin boundary

As mentioned earlier, a twin boundary is expected where two crystallization fronts merge. The boundary generates an extra barrier in the current path and could be treated simply as an extra resistor. With a masked Si implant, this boundary can be positioned either in the middle or at one end of the channel. In the latter case, the device is asymmetric, and different behaviors are expected when the source and drain are switched. When it is in the source, the boundary reduces the effective V_{gs} and generates the lowest drive current, while a boundary in the drain causes the least degradation of the drive current. The IV curves associated with different locations of the twin boundary are shown in Fig. 10, and they are consistent with the above prediction.

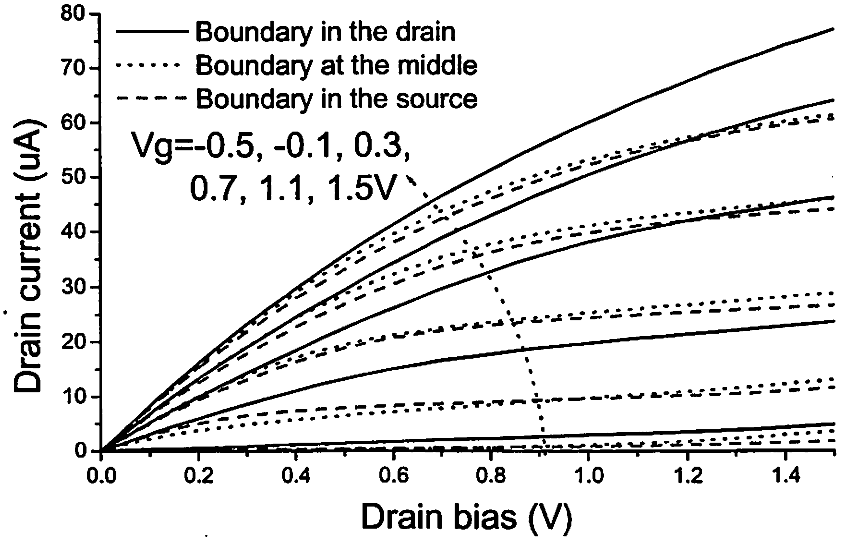


Fig. 10. NMOS with different locations of the twin boundary. W/L=100nm/60nm, V_t=-0.6V

3.3.4 SPE film quality from activation energy

Defects, including twin boundaries, introduce trap states in the band gap. According to the transport model of polycrystalline films, the trap states deplete the nearby semiconductor and form a barrier to the charge carriers in the channel [19]. The activation energy, i.e. the barrier height, depends on the number of traps (N_t) and the mobile charge density (N_D) nearby. Using the simple full-depletion approximation, it can be derived as $E_a \propto \frac{N_t^2}{N_D} \propto \frac{N_t^2}{V_g - V_t}$ in strong inversion.

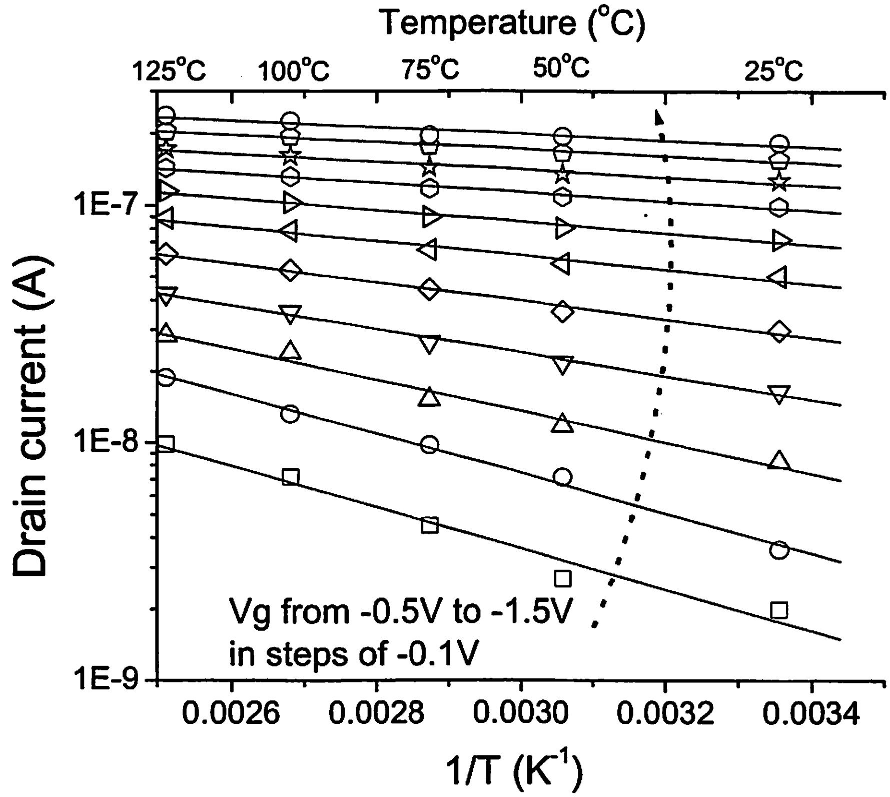


Fig. 11. Extraction of E_a from log(I_d)~1/T plot with various gate biases, V_{ds}=50mV

After IV measurements are performed at different temperatures with a low drain bias (V_{ds}=50mV), the activation energies are extracted from the log(I_d)~1/T curves for each gate bias. Fig. 11 is a sample plot for an 80nm PMOS at a V_g from 0.5 to 1.5V. For each gate bias, the activation energy is just -k_B multiplied by the slope of the linear fitting line, and it is a decreasing function of the gate bias. This approach assumes that the current is limited only by the thermionic emission over the barrier in the channel at a low V_{ds} bias. Actually, the current is also limited by the carrier mobility in the channel, so the measured barrier height includes the degradation of mobility with temperature. Though not completely accurate, E_a still gives qualitatively the barrier height in the channel, and acts as an important index of the channel quality for transport [20].

When the above procedure is repeated for different transistors, the barrier heights associated with different channel lengths and processes can be studied. The E_a curves in Fig. 12 show that E_a decreases with increasing gate bias. It also agrees well with our hypothesis that crystalline quality degrades with increasing SPE range. If the film is crystallized from both ends, the twin boundary in the middle increases the barrier height of the channel more than it does when it is at the drain side. The activation energy of a perfect Si film comes from the fact that V_t drops at a higher temperature. For an FD device with an undoped body, an effective E_a can be derived as $E_a \rightarrow \frac{kTV_{bi}}{V_g - V_t}$ at high gate biases, where V_{bi} is the built-in potential (~0.2V) between the source and the channel. A 60nm SPEFET has a channel film with quality close to that of a perfect Si film.

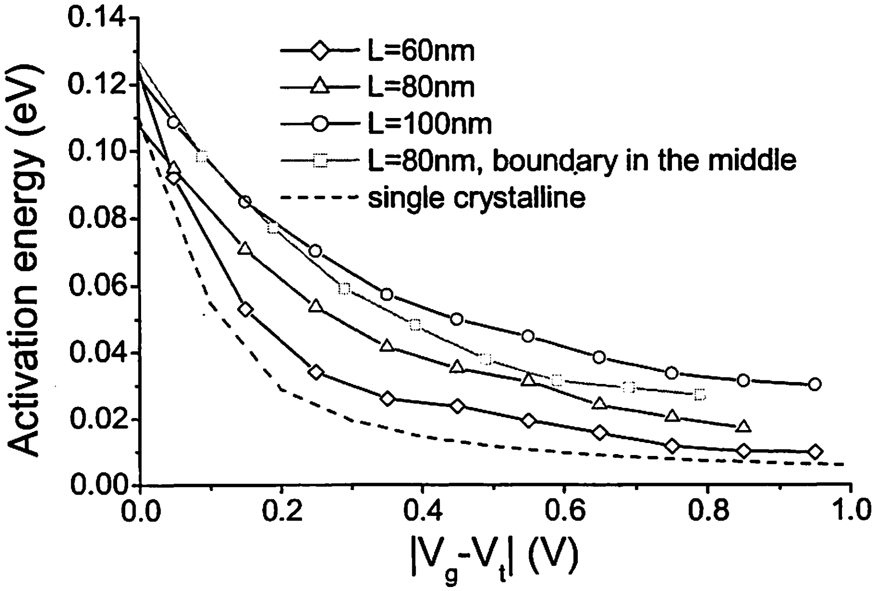


Fig. 12. Activation energy as a function of gate bias for different transistors

3.3.5 Film thickness dependence

As emphasized in Chapter 2, film thickness is the critical parameter for a UTB structure. The thinner the film, the better the suppression of SCE. All the 60nm channel length devices described above function only with 100Å channel films. However, the SPE process requires adequate thickness to yield a high quality crystallized film. The influence of channel film thickness on the device performance has been studied using two different sets of deposited film thicknesses, 100Å and 200Å. The variations of V_t and I_{on}, which is measured at V_g-V_t=0.8V, with the channel lengths for different film thicknesses are shown in Fig. 13. Devices on a 100Å channel film show substantially worse characteristics. Larger degradation of the drive current, almost two decades from channel lengths of 60nm to 100nm, is observed, and there are very few working devices with L greater than 100nm. The V_t roll off is also much larger than that of the 200Å channel film devices. The results indicate that the thinner the film, the shorter the crystallization range, and the more defects left in the final film.

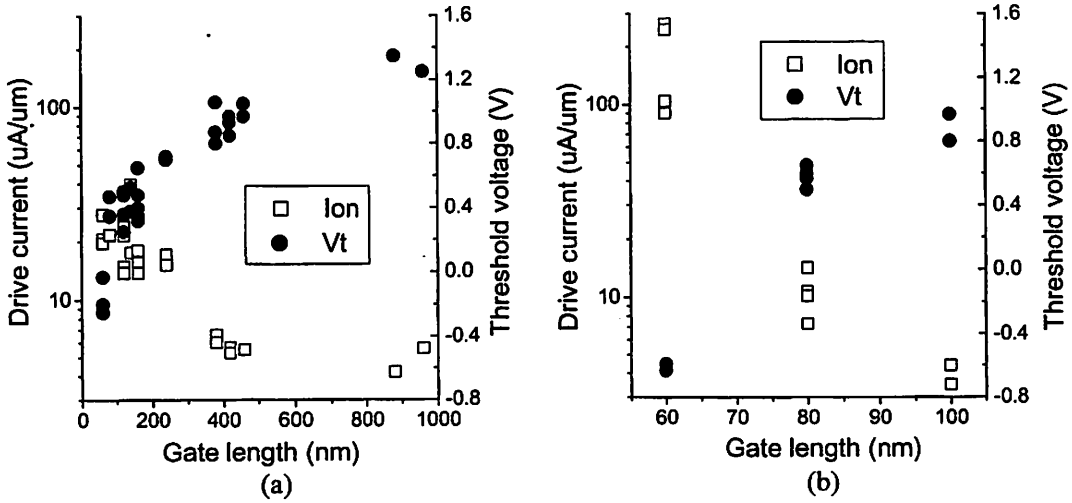
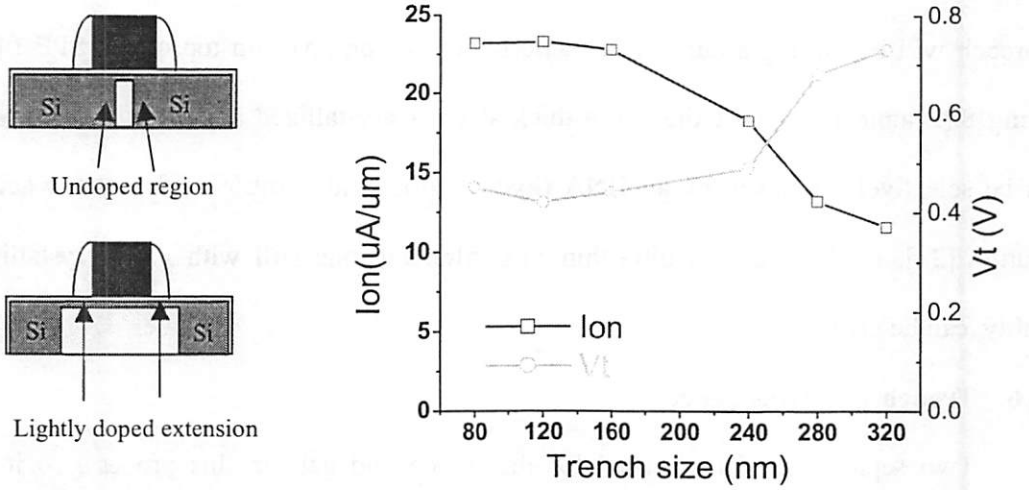


Fig. 13. The effects of channel film thickness on V_t roll-off and drive current measured at V_g-V_t=0.8V. The SPE film thickness is (a) 200Å (b) 100?

Since less than 100Å channel films are required for sub-30nm transistors, better crystallization procedures must be developed for future generations. One promising approach will be putting a dummy N+ amorphous silicon layer on top of the SPE film during SPE annealing. After the entire thick stack is crystallized, the dummy N+ layer can be selectively removed by an HNA (hydrofluoric acid + nitric acid + acetic acid) solution [21]. In this way, an ultra-thin SPE film, but one still with good crystalline quality, can be created.

3.3.6 Trench size dependence

Two separate masks are used for the trench and gate in this process, so it is possible to study the effects of the relative sizes and misalignment between the trench and the gate. In this work, the effect of misalignment has not been studied because it is hard to control the misalignment accurately at the scale of sub-100nm. In Fig. 14, devices with a fixed gate length of 200nm but different trench sizes are probed. If the trench is smaller than the gate, two thick body regions under the gate are not doped in S/D implantation, and act like insulators; therefore, there is not much difference observed in I_{on}, V_t, or I_{off} with various trench sizes. For a trench larger than the gate, the two ultra-thin-body regions outside of the channel are neither heavily doped, because of their thickness, nor driven by the gate. These thin regions acts as two external resistors, and reduce the drive current. On the other hand, a larger trench is beneficial in suppressing SCE, because of the wider separation of the thick source and drain regions. This benefit is manifested in the higher V_t of the devices with larger trench sizes. Another piece of evidence for the better SCE suppression is that the 60nm devices work only with larger trenches (100nm in this study). This study suggests that a trench size slightly larger than the gate results in good SCE without much degradation of the drive current.



(a) (b)

Fig. 14. (a) The device structures with a trench smaller/larger than the gate. (b) For a given channel length, the influence of trench size on device performance

3.4 Conclusion

UTBFETs have been fabricated through the method of solid-phase-epitaxy. The thickness of the deposited film can be controlled precisely, but its crystalline quality after annealing is a crucial issue still to be solved. With the deposition and annealing method presented in this report, 60nm devices have been fabricated on 100Å SPE film with performance comparable to the conventional CMOS. Further improvement is needed to increase the crystallization range for yet thinner film. On the other hand, SPEFET can be relatively easily integrated with bulk CMOS, which is suitable for long channel devices. With separate optimization of some key parameters, such as V_t, good behaviors can be achieved for both long and short channel devices. Therefore, SPEFET is a promising candidate for sub-100nm generations with excellent process controllability.

The SPE process also introduces some unique aspects of device physics. First, the effect of twin boundary locations on the device behavior can be studied. It has been found that pushing the twin boundary into the drain region using a masked silicon implant produces a higher drive current. Second, using two separate masks for trench and gate layers makes possible the investigation of their relative sizes. The results suggest that a trench size slightly larger than the gate gives the best performance.

3.5 References

[1] H-SP. Wong, D.J. Frank, P.M. Solomon, C.H.J. Wann and J.J. Welser. "Nanoscale CMOS." *Proceedings of the IEEE, Vol. 87, No. 4, pp. 537-570, April 1999*

[2] A. Vandooren, D. Jovanovic, S. Egley, M. Sadd, B-Y. Nguyen, B. White, M. Orlowski and J. Mogab, "Scaling assessment of fully-depleted SOI technology at the 30 nm gate length generation," *IEEE International SOI Conference, pp. 25–27, Oct. 2002*

[3] M. Ieong, H-SP. Wong, E. Nowak, J. Kedzierski and E.C. Jones, "High performance double-gate device technology challenges and opportunities," *Proceedings of the Symposium on International Quality Electronic Design, pp. 492-495, March 2002*

[4] Z. Ren, P.M. Solomon, T. Kanarsky, B. Doris, O. Dokumaci, P. Oldiges, R.A. Roy, E.C. Jones, M. Ieong, R.J. Miller, W. Haensch and H-SP. Wong, "Examination of hole mobility in ultra-thin body SOI MOSFETs," *IEEE. Proceedings of International Electron Devices Meeting, pp. 51–54, Dec. 2002*

[5] L. Haji, P. Joubert, J. Stoemenos and N.A. Economou, "Mode of growth and microstructure of polycrystalline silicon obtained by solid-phase crystallization of an amorphous silicon film," *J. Appl. Phys. Vol. 75(8), pp. 3944-3952, April 1994*

[6] M. Miyao, M. Moniwa, K. Kusukawa and W. Sinke, "Low-temperature SOI (Si-on-insulator) formation by lateral solid-phase epitaxy." *Journal of Applied Physics, Vol. 64, No. 6, pp. 3018-3023, Sept. 1988*

[7] Y-C. Yeo, V. Subramanian, J. Kedzierski, P. Xuan, T-J. King, J. Bokor and C. Hu, "Nanoscle ultra-thin-body Silicon-on-Insulator P-MOSFETs with a SiGe/Si heterostructure channel". *IEEE Electron Device Letter, Vol. 21, No. 4, pp. 161-163, 2000*

[8] H. Liu, M. Kumar and J.K.O. Sin, "Device characteristics of the 3-D BiCMOS technology using selective epitaxial growth and lateral solid phase epitaxy," *IEEE. Transactions on Electron Devices, Vol. 49, No. 12, pp. 2359 –2362, Dec. 2002*

[9] V. Subramanian, J. Kedzierski, N. Lindert, H. Tam, Y. Su, J. McHale, K. Cao, T-J. King, J. Bokor and C. Hu, "A bulk-Si-compatible ultra-thin-body SOI technology for sub-100nm MOSFETs". *Device Research Conference, pp. 28-29, 1999*

[10] Y. Nakahara, K. Takeuchi, T. Tatsumi, Y. Ochiai, S. Manako, S. Samukawa and A. Furukawa, "Ultra-shallow in-situ-doped raised source/drain structure for sub-tenth micron CMOS," *Symposium on VLSI Technology, pp. 174 –175, June 1996*

[11] H. Liu, Z. Xiong, J.K.O. Sin, P. Xuan and J. Bokor, "A high performance double-gate SOI MOSFET using lateral solid phase epitaxy," *IEEE International SOI Conference, pp. 28 –29, Oct. 2002*

[12] B.S. Meyerson, F.J. Himpsel and K.J. Uram, "Bistable conditions for low-temperature silicon expitaxy". *Appl. Phys. Letter, Vol. 57(10), pp. 1034-1036, 1990.*

[13] M. Kodama, H. Funabashi and Y. Mitsushima, Y. Taga. "A solid-phase epitaxy with clean surface formation using SiH_4 and evaluation of SOI layer". *Electronics and*

Communications in Japan, Part2, Vol. 79, No. 12, pp. 71-77, 1996

[14] Y-Y Wang, N.W. Cheung, D.K. Sadana, C. Jou and M. Strathman, "The influence of ion implantation on solid phase epitaxy of amorphous silicon deposited by LPCVD". *Advanced Application of Ion Implantation, Vol. 530, pp. 70-74, 1985*

[15] M. Moniwa, K. Kusukawa, E. Murakami, T. Warabisako and M. Miyao. "Influence of Si film thickness on growth enhancement in Si lateral solid phase epitaxy." *Applied Physics Letters, Vol. 52, No. 21, pp. 1788-1790, May 1988*

[16] K. Kusukawa, M. Moniwa, E. Murakami, T. Warabisako and M. Miyao. "Grown-facet-dependent characteristics of silicon-on-insulator by lateral solid phase epitaxy." *Applied Physics Letters, Vol. 52, No. 20, pp. 1681-1683, May 1988*

[17] International Technology Roadmap for Semicondutors, Semicondutor Industry Association, 2001. *http://public.itrs.net/Files/2002Update/2001ITRS/Home.htm*

[18] N. Hirashita, T. Katoh and H. Onoda. "Si-gate CMOS devices on a Si lateral solid-phase epitaxial layer." *IEEE Transactions on Electron Devices, Vol. 36, No. 3, pp. 548-52, March 1989*

[19] J. Levinson, F.R. Shepherd, P.J. Scanlon, W.D. Westwood, G. Este and M. Rider, "Conductivity behavior in polycrystalline semiconductor thin film transistor" *J. Appl. Phys., Vol. 53, pp. 1193-1202, Feb. 1982*

[20] B.A. Khan and R. Pandya, "Activation energy of source-drain current in hydrogenated and unhydrogenated polysilicon thin-film transistors," *IEEE Transactions on Electron Devices, Vol. 37, No. 7, pp. 1727 –1734, July 1990*

[21] P. Kalavade and K.C. Saraswat, "Lateral gate-all-around (GAA) poly-Si transistors," *IEEE. International SOI Conference, pp. 109–110, Oct. 2001*

Appendix 3A: Process flow for SPEFETs

Step	Process	Process specification	Equipment	Comments
0	Starting wafer: 4" SOI wafer, 1000A Si on 4000A buried oxide			
0.1	Scribe	Label the wafers		
1	EBeam alignment marks			
1.1	Cleaning	Piranha (H_2O_2:H_2SO_4=1:5) 120°C, 10min, 25:1 BHF 30s	Sink6	Dewet, up to 16kΩ
1.2	Pad oxide	SDRYOXA 950C, 80min, 20min anneal	Tylan2	SiO_2=35nm Si_{remain}=81nm
1.3	SiGe dep	SiGe.019: Nucleation: T=550°C, P=300mT, SiH_4=200sccm, t=1min. Deposition: T=500°C, P=300mT, SiH_4=186sccm, GeH_4Lo=33sccm, GeH_4Hi=0, t=80min	Tystar19	Ge concentration ~40% 780-800nm SiGe
1.4	LTO cap	VDOLTOC, 450°C, 300mT, SiH_4=25sccm, O_2=75sccm, 10min.	Tystar11	155nm
1.5	Anneal	N2ANNEAL, 1000C, 30min	Tylan7	No thickness change
1.6	Litho	Resist coating: coat=program 1/bake=program 1 Exposure: focus=250, t=0.9s Development: bake=program 1/develop=program 1 Descum: O_2=51sccm, P=50W, t=1min Hard bake: 120°C, 1hr	Svgcoat1 GCAWS Svgdev Technics-c Ovrn	PR=1.2um PEB: 90C, 1min DEV: OPD4226, 1min

1.7	Mark etch	B: p=13mTorr, CF_4=100, P_{top}=200, P_{bot}=40, t=90s
		M: P=15mTorr, Cl_2=50, HBr=150, P_{top}=300, P_{bot}=150, t=55s
		O: P=35mT, HBr=200, O_2=5.0, P_{top}=250, P_{bot}=120, t=25s
			Lam5	ER=20Å/s, SiO_2/Si~1
				ER=100Å/s, SiGe/SiO_2~13
				ER=50Å/s, Si/SiO_2~100.

1.8	Resist strip	O_2 51sccm, 230W, 5min	Technics-c	
1.9	Cleaning	Piranha 120°C, 10min	Sink8	Clean in dirty sink first
2	Trench formation			
2.1	Trench lith	trench.gds. PMMA 120nm, dose=800uC/cm^2	Nanowriter	At LBNL
2.2	Trench etch	B/M/O=20s/20s/20s, see 1.7	Lam5	EBeam resist etched fast
2.3	Resist strip	O_2 51sccm, 300W 6min	Technics-c	
2.4	Cleaning	Piranha 120°C, 10min	Sink8	Clean in dirty sink first
2.4	Cleaning	Piranha 120°C, 10min, 25:1 BHF 60s	Sink6	Dewet
2.6	HTO fill	9HOXN2OD, 800°C, N_2O=100sccm, $Si_2H_2Cl_2$=10sccm. t=6hr	Tylan9.	83-85nm
2.7	LTO fill	VDOLTOC, 450°C, 300mT, SiH_4=25sccm, O_2=75sccm, 20min.	Tylan11	300nm
2.8	Anneal	N2ANNEAL, 1000°C, 30min	Tylan7	Totally ~380nm
2.9	Etch back	Breakthrough. P=13mTorr, CF_4=100, P_{top}=200, P_{bot}=40, 150s,	Lam5	Remain oxide 20-50nm
2.10	Cleaning	Piranha 120°C, 10min	Sink8	
3	SPE film formation			
3.1	Wet etch	Piranha 120°C, 10min, 25:1HF 90s	Sink6	Expose silicon seeds
3.2	a-Si dep	SPESI2.019. Nucleation: T=410°C, P=50mT, SiH_4=12sccm, t=30s, Deposition: T=410°C, P=500mT, Si_2H_6=200sccm, Split1: t=40min. Split2: t=20min	Tystar19	Cool down to 300°C before loading wafers. 21nm and 10min
3.3	Si imp litho	Siim.gds and Siim2.gds. SAL 300nm, dose=80uC/cm^2	Nanowriter	At LBNL
3.4	Si implant	20keV, 7e15. 7 degrees	Innoia	
3.5	Resist strip	O_2 51sccm, 300W 6min	Technics-c	
3.6	Cleaning	Piranha 120°C, 10min	Sink8	
3.7	Cleaning	Piranha 120°C, 10min	Sink6	

3.8	Crystallization	THIN_ANN: 550C, 36hr THIN_ANN: 600C, 12hr N2ANNEAL: 950C, 30min	Tylan7	Nucleation Crystallization speed up Defects removal
4	Mesa and gate definition			
4.1	Mesa litho	SAL 250nm, dose=80uC/cm^2	Nanowriter	At LBNL
4.2	Mesa etch	B/m/o=10s/15s/30s. see 1.7	Lam5	Si mesa 103nm
4.3	Resist strip	O_2 51sccm, 300W 6min	Technics-c	
4.4	Cleaning	Piranha 120°C, 10min	Sink8	
4.5	Cleaning	Piranha 120°C, 10min, 25:1 HF dip 5s	Sink6	remove native oxide.
4.6	Gate oxide	THIN-ANN. 750°C, O_2 15min + N_2 15min + 900°C, N_2 20min	Tylan6	TCA clean, 25±1Å
4.7	Gate dep	SiGeVAR: Nucleation: T=550°C, P=300mT, SiH_4=200sccm, t=1min. Deposition: T=550°C, P=300mT, SiH_4=186sccm, GeH_4Lo=19sccm, GeH_4Hi=0, t=10min	Tystar19	20% Ge concentration 130nm
4.8	Gate imp	BF_2, 40keV, 1e16.	Innovia	
4.9	Anneal	N2ANNEAL 950°C, 30min.	tylan7	
4.10	Gate litho	SAL 150nm, dose=100uC/cm^2.	Nanowriter	At LBNL
4.11	Gate etch	B/m/o=10s/7s/35s. see 1.7	Lam5	20s overetch for 20Å ox
4.12	Resist strip	O_2 51sccm, 300W 6min	Technics-c	
4.13	Cleaning	Piranha 120°C, 10min	Sink8	
5	Source/drain formation			
5.1	LDD	NMOS: As+ 5e13 cm^{-2}, 10keV PMOS: B+ 1e13 cm^{-2}, 5keV	Innovia	
5.2	Cleaning	Piranha 120°C, 10min	Sink8	
5.3	Cleaning	Piranha 120°C, 10min	Sink6	
5.4	LTO spacer	VDOLTOC, 450°C, 300mT, SiH_4=25sccm, O_2=75sccm, 5min	Tylan11	67-71nm
5.5	Spacer etch	breakthrough. P=13mTorr, CF_4=100, P_{top}=200, P_{bot}=40, t=25s.	Lam5	~20nm oxide left
5.6	S/D implant	As+, 3e15 cm^{-2}, 32keV, B+ 3e15 cm^{-2}, 10keV.	Innovia	
5.7	Cleaning	Piranha 120°C, 10min	Sink8	
5.8	Cleaning	Piranha 120°C, 10min	Sink6	
5.9	RTA	900°C, 30s	Heatpulse3	
6	Metal contact			
6.1	Cleaning	Piranha 120°C, 10min	Sink6	
6.2	LTO dep.	VDOLTOC, 450°C, 300mT, SiH_4=25sccm, O_2=75sccm, 25min	Tystar11	350nm
6.3	Cont. litho	Same as 1.6		
6.4	Cont. etch	Breakthrough, P=13mTorr, CF_4=100, P_{top}=200, P_{bot}=40, 170s 5:1 BHF 60s	Lam5 Sink6	Remain oxide 20nm 200nm LTO etched
6.5	Resist strip	O_2 51sccm, 300W 6min	Technics-c	
6.6	Cleaning	Piranha 120°C, 10min	Sink8	
6.7	Cleaning	Piranha 120°C, 10min, 25:1 BHF 80s	Sink6	
6.8	Al dep.	Pure Al: 6mTorr, 4.5kW, 30cm/min, 2passes	CPA	450nm
6.9	Metal litho	Same as 1.6	GCAWS	
6.10	Metal etch	100s in aluminum etchant	Sink8	~50% overetch
6.11	FGA	VSINT400, 400°C, 30min	Tylan13	
7	Calibration			

Chapter 4

Silicide Gates for Workfunction Engineering

4.1 Introduction

As shown in Chapter 2, a UTB or FinFET can achieve excellent SCE suppression with a fully depleted (FD) ultra-thin body, but it also gives rise to many process difficulties and design issues. The most significant challenge is the precise control of the threshold voltage (V_t), which, in a bulk device, is realized by body doping engineering.

As derived in Chapter 2, $V_{t,long} = \phi_{gs} + V_b + qNl^2/\varepsilon_{si}$. In an NMOS with an undoped body and a conventional N+ polysilicon gate, the threshold voltage V_{tn}=-0.2V (and +0.2V for P+ gate PMOS). However, modern circuit technology requires a V_{tn} ≈0.2V [1], so the body doping must shift V_t by 0.4V. On the other hand, the scale length of a device must be less than a quarter of the minimum channel length for decent SCE, so the amount of threshold voltage shift caused by the body doping will be limited. At a gate length of L_{min}=25nm, a 6nm scale length is required, and the channel doping is at least N_{sub}=7×10^{18}cm^{-3} for a 0.4V V_t shift. At such a high channel doping level, the inversion carrier mobility is severely degraded and so is the performance of the transistor [2].

If the threshold voltage is manipulated via the body doping, V_t fluctuation due to the discrete nature of dopant atoms is another concern [3]. As the channel length is scaled, the film thickness also must be scaled for short channel effects (SCE) suppression. The number of doping atoms in such a small channel volume is limited, and its statistical fluctuation can be noticeable. Since V_t relies heavily on the exact doping concentration, any fluctuation of the dopant number causes fluctuation in V_t. For example, at W=L=25nm, and T_{si}= 5nm, even with an N_{sub} as high as $7\times10^{18}cm^{-3}$, the number of the dopant atoms in the small body volume is only 22. The statistical standard deviation of this quantity is $\sqrt{22}=4.7$, which accounts for 21% of the total number. The corresponding V_t fluctuation is 85mV, unacceptably high for most applications.

In short, a new reliable method for V_t control is needed to leave the body undoped or lightly doped. From the V_t formula, it is clear that the only viable method is gate workfunction engineering. In this chapter, silicides are proposed for their continuously adjustable workfunction and relatively easy process integration with CMOS. Meanwhile, body doping can still be used for fine adjustment of the threshold voltage. In a fully depleted body, the polarity of the body doping does not matter, except for the way the threshold voltage is shifted. At the technology node of L_g=25nm, a V_t shift of 0.1V is still possible when the channel doping varies through $1\times10^{18}cm^{-3}$, from n-type to p-type. However, this tuning range also diminishes with further scaling.

4.2 Motivation for silicide gates

4.2.1 Metal gates

Metal gates were widely used in the 1960's and were replaced by polysilicon (poly) gates because poly gates have superior CMOS process compatibility and thermal stability on silicon oxide. Recently, metal gates were reexamined for their merits. First, metal gates can boost the device performance by eliminating the gate depletion layer [4]. When a poly gate is used, a thin layer near the oxide interface is depleted due to its limited doping concentration [5]. This depleted layer increases the effective oxide thickness by about 5Å, which is unacceptable when the gate oxide is scaled below 2nm. Second, metal gates have excellent conductivity. It has been pointed out that the relatively high resistance of the poly gate severely degrades the circuit speed at high frequencies [6]. Moreover, although polysilicon offers superior thermal stability and interface quality on silicon dioxide, it is not advantageous over metal gates on high κ dielectrics, which are implemented to suppress SCE [7]. Over all, metallic gates are believed to be indispensable for future generations.

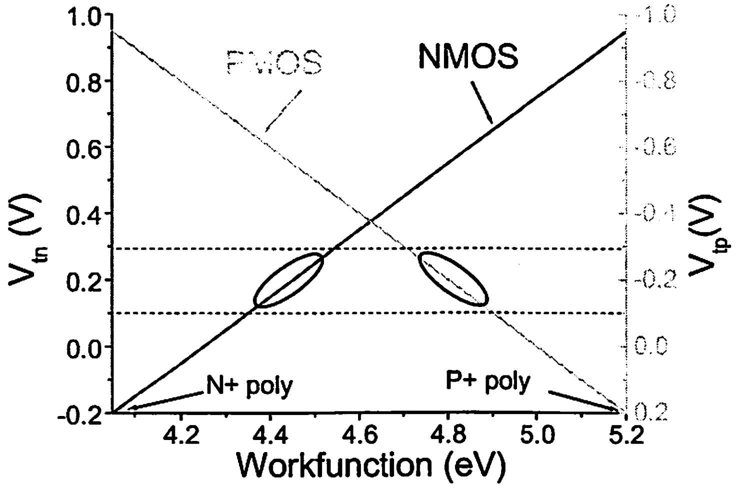


Fig 4.1. Threshold voltages for NMOS/PMOS as functions of gate workfunction on an undoped channel film. $V_{tn} = \phi_g - E_c + V_b$ and $V_{tp} = \phi_g - E_v - V_b$.

In deeply scaled CMOS, gate workfunction engineering is the driving impulse for metal gates. Since the body doping in a fully depleted MOSFET has very limited ability in adjusting V_t (0.1V at 25nm technology), gate materials with correct workfunctions have to be used. Fig. 4.1 shows the requirement for gate workfunctions to achieve the right threshold voltages for both NMOS and PMOS. Two distinct workfunctions, 4.45eV for NMOS and 4.8eV for PMOS, are required. Similar gate workfunction requirements exist even for deeply scaled bulk CMOS [8]. Moreover, modern digital circuit designs typically require two types of transistors on the same chip: high performance FETs on the critical path for the maximum circuit speed, and low power FET on non-critical paths for minimum power consumption [9]. This requires multiple threshold voltages on a single chip and, consequently, multiple workfunctions or materials.

Although offering many advantages, pure metal gates are difficult to integrate into a CMOS process due to the following challenges:

1. The thin gate oxide is exposed and damaged during the gate metal deposition, which is typically performed by plasma sputtering.
2. Due to their extremely small sizes, gates have to be patterned by dry-etching. Dry-etching metals is challenging and may contaminate the wafer or equipment.
3. If more than one gate material is required, multiple depositions and etches are needed. The precious gate oxide is revealed and damaged in each etch. Therefore, it is highly desirable to share one gate material between both NMOS and PMOS, analogous to polysilicon in the current CMOS technology.
4. If the gate is formed before the source/drain, as in the standard process flow, the metal gate has to undergo the high temperature annealing required for dopant activation. The thermal stability of the metal on a thin oxide is crucial.
5. The metal has to be CMOS process compatible, and its potential for diffusion into and contamination of the silicon and oxide require careful attention.

4.2.2 Silicide gates

In this report, silicide gates are proposed to solve most of the above process issues. Silicides are unique in that they are formed via the reaction between silicon and metal. First, silicon can be deposited to protect the underlying gate dielectric in all the following metal depositions or etches even if multiple silicides are required. This eliminates the damage to the oxide and a good interface is possible. Second, the silicide can be formed via a salicide (self-aligned silicide) process, which has become a mature technology in the semiconductor industry [10]. The fine patterning of the gate is carried out by the well-developed dry-etching of poly; therefore, the dry-etching of the metal can be avoided. Third, the silicide can be formed after source/drain doping activation while the poly gate can stand as the self-aligned implant mask. Therefore, the silicide goes through only the backend low temperature steps and the thermal stability requirement can be relaxed to 400°C. Finally, since some silicides have been widely used in the present semiconductor industry as interconnect materials [11,12], their compatibility with the CMOS process has been demonstrated.

In addition, experiments show that most silicides have workfunction in the desired range, which is within ±0.3eV from the mid-gap of the silicon band. It has been shown that the workfunction of some silicides can be manipulated by implanting dopants into the silicon film before the formation of the silicides [13]. The adjustable workfunction of silicides is highly advantageous for both process control and multiple V_t applications.

In this report, it is shown that the workfunction of both NiSi and TiSi can be tuned by ion implantation. The influence of NiSi and TiSi gates on oxide interface quality is investigated.

4.3 NiSi and TiSi workfunction extraction

4.3.1 Fabrication process for MOSCAPs

In order to extract the workfunction of the gate material, MOSCAPs with various oxide thicknesses are required because there is inevitably a certain amount of fixed charges on the interface between the gate dielectric and silicon substrate, which causes a shift in the flat band voltage. To create multiple oxide thicknesses on a single wafer, a thick oxide (80nm) was first grown thermally on p-type blank wafers. One half of the wafer was immersed in HF solution. Then, another timed HF wet-etch was performed after the wafer is rotated by 90 degrees. In this way, four oxide thicknesses were generated on a single wafer (Fig. 4.2).

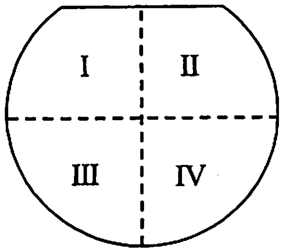


Fig 4.2. A single wafer with four oxide thicknesses.

Experimentally, we found the silicide formed with polysilicon to be unstable on thin oxide. Metal atoms penetrated through the thin oxide and short-circuited the gate to the substrate. However, one of the following two approaches can stop the silicidation process on oxides as thin as 2nm: covering Ni with TiN during the reaction, or forming the silicide using amorphous silicon rather than polysilicon. It is believed that TiN protects devices from environmental contaminants in the silicidation anneal. The grain boundaries in poly cause a non-uniform front as a result of either the reaction or metal diffusion, which results in the thin oxide breakdown. The MOSCAPs silicide gates were fabricated from the reaction of metal with 100nm polysilicon, while TiN covered the devices during silicidation.

Various doses of phosphorus (80keV) or boron (25keV) ions were implanted into the undoped polysilicon film before the formation of the silicides, and the workfunction dependence on the doping level was investigated. It has been shown that the workfunction of NiSi is different when the silicide is formed on N+ or P+ silicon [14].

Although the doping atoms account for only a tiny portion of the gate materials, experiments indicate that the atoms tend to accumulate at the interface to the gate dielectric [15]. Within that thin layer at the interface, the doping concentration is sufficiently high to change the composition of the material and, in turn, its workfunction. The detailed mechanism of the workfunction change is still under investigation.

To guarantee that the entire silicon film is reacted to form silicide all the way down to the oxide interface, sufficient metal must be deposited. The minimum thickness can be calculated from the densities and atomic weights. $\frac{T_m}{T_{si}} = k\frac{m_m}{m_{si}}\frac{\rho_{si}}{\rho_m} = 0.88(\mathrm{TiSi})\,\mathrm{or}\,0.55(\mathrm{NiSi})$ for the target silicide composition of NiSi or TiSi (i.e. k=1). On the other hand, with too much metal deposited, metal-rich silicides (Ni_2Si or Ti_2Si) form and change the property of the resulting silicides. Therefore, the amount of metal must be limited below twice the minimum value. Practically, 100nm of Ti or 80nm of Ni was deposited.

Metals and silicon were patterned into squares of 100μm×100μm in a single-mask lithography. Wet-etch chemicals which would not attack the over-night hard-baked photo resist were developed. Titanium can be easily etched in an RCA SC-1 solution (NH_4OH:H_2O_2:H_2O=1:2:5) with a etch rate of 100Å/min. Nickel can be etched away in a buffered 20:1 HF solution with a similar etch rate. However, the wafer must be rinsed by DI water after every a few seconds of wet etching, or nickel flakes will fall off, resulting in a non-uniform etch.

Low temperature rapid thermal annealing (RTA) was performed to form silicides. Silicide gates attracted great attention in the 80's and were abandoned because they degraded the thin oxide. It was believed that metal atoms reacted with or diffused into the gate dielectric at high temperatures. Therefore, in this work, the silicides were formed at a temperature as low as possible, and in a short time (~2min) with the utilization of an RTA furnace. The capacitors were annealed iteratively at 400°C, 600°C and 800°C for the study of the temperature effect on the silicides. Although the exact composition or phase of the silicides was not analyzed, NiSi and a mixture of Ti-Si were likely formed at 400°C. At high temperatures, NiSi and TiSi may be the dominant components, coexisting with other compositions and phases [16].

4.3.2 Workfunction extraction

Fig 4.3 plots a typical CV measurement between the gate and the substrate. The following approach is taken to extract the workfunction:

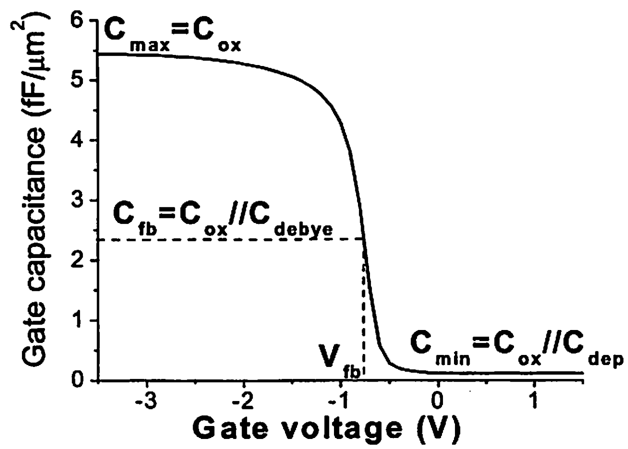


Fig 4.3. A typical CV measurement from a MOSCAP. The oxide thickness, substrate doping, and flat band voltage can be extracted. Here, A//B is fined as AB/(A+B).

a) Extract the electrical oxide thickness in the accumulation region using $\frac{1}{C_{max}}=\frac{T_{ox}}{\varepsilon_{ox}}$. The result is typically 3-5Å thicker than the physical thickness because of the quantum repulsion in the substrate.

b) Extract the depletion depth in the depletion region: $\frac{1}{C_{min}}=\frac{T_{ox}}{\varepsilon_{ox}}+\frac{X_{dep}}{\varepsilon_{si}}$

c) Calculate the substrate doping and Fermi level: $N_a=\frac{2\varepsilon_{si}\Delta\phi}{qX_{dep}^2}$ and $\phi_s=E_i+\frac{kT}{q}\ln\frac{N_a}{n_i}$.

d) Calculate the Debye length, which is the depletion depth at the bias of the flat band voltage. $L_D = \sqrt{\frac{\varepsilon_{si} kT}{q^2 N_a}} = X_{dep}\sqrt{\frac{kT}{2\Delta\phi}}$

e) Calculate the capacitance at the flat band voltage: $\frac{1}{C_{fb}} = \frac{T_{ox}}{\varepsilon_{ox}} + \frac{L_D}{\varepsilon_{si}}$

f) Extract V_{fb} from the CV curve using the value of C_{fb}.

g) Extract the virtual workfunction $\phi'_m = \phi_m - \frac{qN_f}{\varepsilon_{ox}} T_{ox}$ based on $V_{fb} = \phi'_m - \phi_s$.

Fig 4.4a plots the ϕ'_m of TiSi as a function of the gate oxide thickness at different temperatures. The intercept indicates the real workfunction of the gate material, while the slope gives the density of the interface fixed charge.

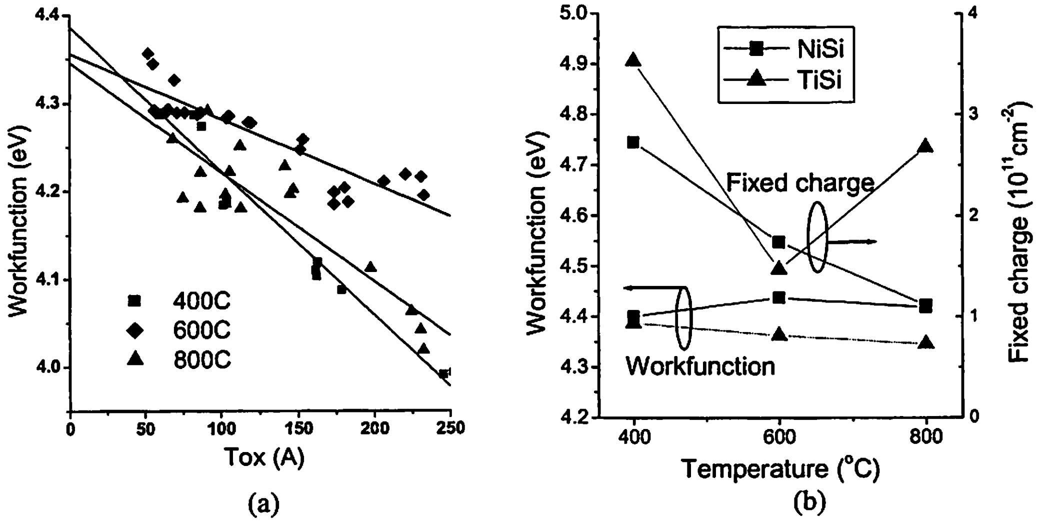


Fig. 4.4. Workfunction extraction for TiSi at different annealing temperatures. (a) ϕ'_m vs. gate oxide thickness. (b) summary of workfunction and fixed charge density.

Fig. 4.4b plots the summarized workfunction results for both NiSi and TiSi, which shows that the workfunctions are insensitive to the anneal temperature between 400°C and 800°C. With a possible degradation in resistivity at 800°C [17], no agglomeration or change of workfunction was observed for the NiSi gate. At 750°C, TiSi changes its phase from C49 to C54, which could have caused the abnormal increase of the fixed charge density [16]. The value of the fixed charge density (N_f=2×10^{11}cm^{-2}) is close to the typical values for metal gate capacitors, but higher than those for poly gate ones, which are ~5×10^{10}cm^{-2}. Nevertheless, when oxide is scaled down to below 2nm, the V_t shift induced by the fixed charge is merely 10mV, which is negligible.

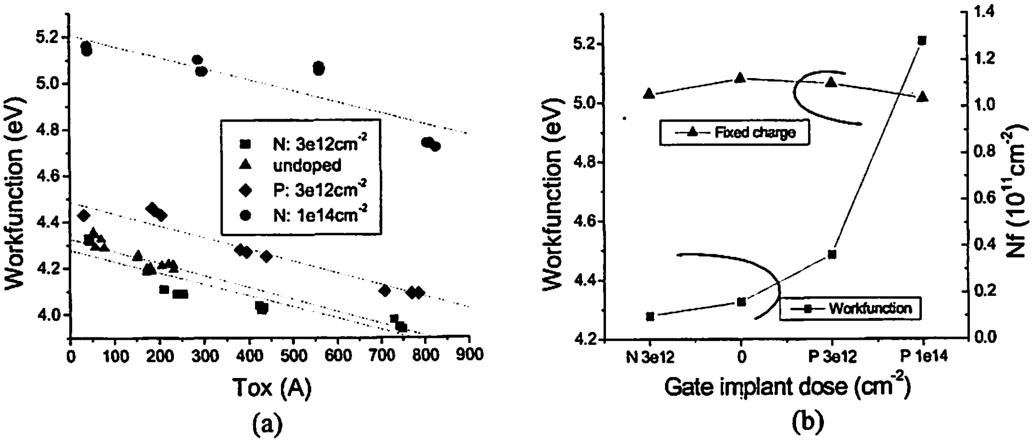


Fig 4.5. Workfunction extraction for TiSi with different gate implant doses. (a) ϕ'_m vs. gate oxide thickness. (b) summary of workfunction and fixed charge density.

The workfunction dependence on the silicon doping concentration was also examined. Wafers were implanted at different doses of either phosphorus or boron. Fig. 4.5 shows the TiSi MOSCAP results with different doping levels into the polysilicon film before silicide formation. Within a dose range below 1×10^{14}cm^{-2}, the workfunction of the silicides can be continuously tuned over a quite large range: from 4.3eV to greater than 5eV. Although the exact mechanism is not well understood, a slightly wider workfunction range than what has been published in [13] is achieved probably because more dopants pile up at the interface due to our higher implantation doses and energies. Therefore multiple V_t values can be easily implemented with the addition of one extra masked gate implant. At higher-level implant doses, highly non-uniform doping profiles are found in the substrate, which make the extraction of the workfunction impossible.

Fig. 4.6 plots the results for NiSi MOSCAPs. From these results, it can be concluded that both NiSi and TiSi workfunctions can be adjusted by the gate doping to meet the threshold voltage requirement for both NMOS and PMOS. NiSi has a higher fixed charge density at the oxide/substrate interface, but one still low enough for circuit applications.

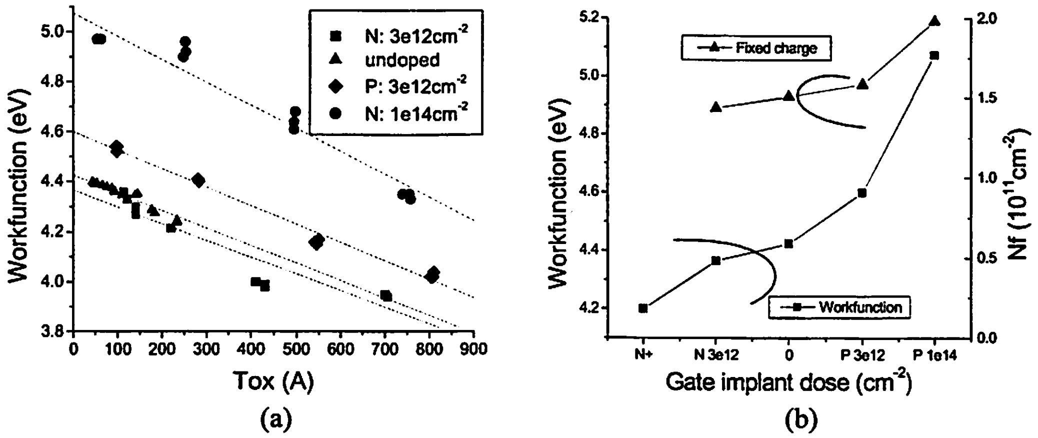


Fig 4.6. Workfunction extraction for TiSi with different gate implant doses. (a) ϕ_m' vs. gate oxide thickness. (b) summary of workfunction and fixed charge density. (The N+ data point is extract from a MOSCAP with a 20Å thin oxide.)

4.4 CMOS with silicide gates

4.4.1 Fabrication process for MOSFETs

While TiSi and NiSi can meet the workfunction requirements for both NMOS and PMOS, their influences on thin oxides had yet to be investigated. The quality of the gate stack was examined via the measurement of gate capacitance, minority mobility and gate current leakage in MOSFETs fabricated through a standard CMOS replacement gate process [18]. After LOCOS isolation, the wafer surface was cured by several sacrificial oxidations. The dummy gate stack consisted of a 13nm thermal oxide grown at 900°C and a 440nm polysilicon film deposited by LPCVD. After the self-aligned source/drain implantation, a 660nm low temperature oxide was deposited and chemical-mechanical polished to exposed the dummy gate. The dummy gate stack was removed by a reactive ion etch and an HF dip; then the real gate oxide (27Å) was thermally grown at 750°C for 15min. Undoped amorphous silicon was deposited and patterned without gate implantation. The MOSFET silicide gates were formed in a salicide anneal at 400°C for 2min in a N_2 ambient. Finally, Ti contacts to the source/drain regions were formed with a lift-off process. The detailed process flow is outlined in Appendix 4A and 4B.

4.4.2 Bulk PMOS results with NiSi gate

PMOS transistors were fabricated with undoped NiSi gates, so they all had a high V_t. Fig. 4.7 shows an example of the I_d-V_g and I_d-V_d curves. Good swing, DIBL and drive current are present: S=70mV/dec and I_{on}=113µA/µm at L=0.75µm. The threshold voltage is not in the desired range (V_t=-1.2V) but agrees with the workfunction extracted previously. A boron implant is needed to shift the V_t to suitable values.

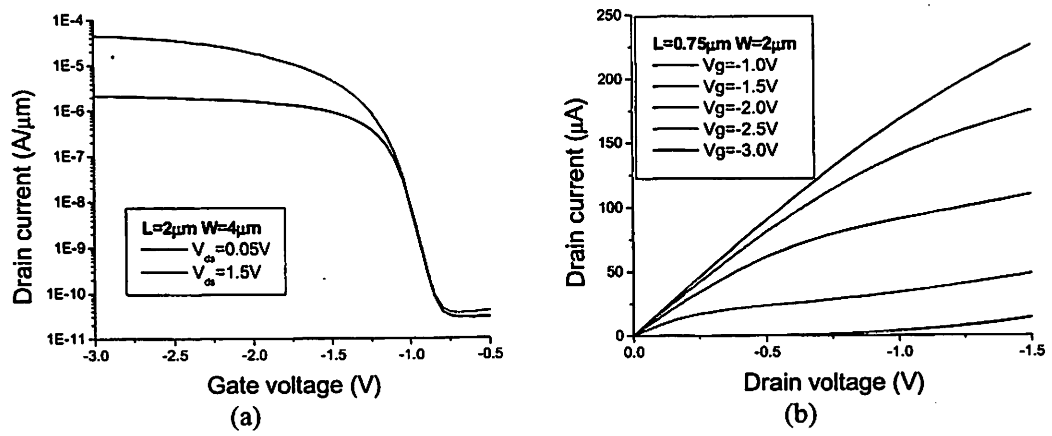


Fig 4.7. Example of PMOS (a) I_d~V_g and (b) I_d~V_d measurement. Good subthreshold swing (S=70mV/dec) and drive current (I_{on}=113µA/um at L=0.75µm) are achieved, indicating excellent interface quality.

The resistance of the transistor at low V_{ds}=50mV is plotted versus the channel length at different gate biases (Fig. 4.8a). The effective channel length and external resistance can be extracted from the interception (ΔL=0.42μm and R_{ext}=3.3kΩ•μm). The results are reasonable since this process is not optimized for short channel devices. Fig. 4.8b compares the V_t roll-off with that of a control device with a polysilicon gate. The shift in the threshold voltage clearly shows that the workfunction of a NiSi gate is ~0.8V lower than that of P+ polysilicon. The V_t roll-off curve is worse for MOSFETs with NiSi gates because they use a gate-last process. The gate oxide is thermally grown after the S/D activation, and this extra thermal step degrades the SCE.

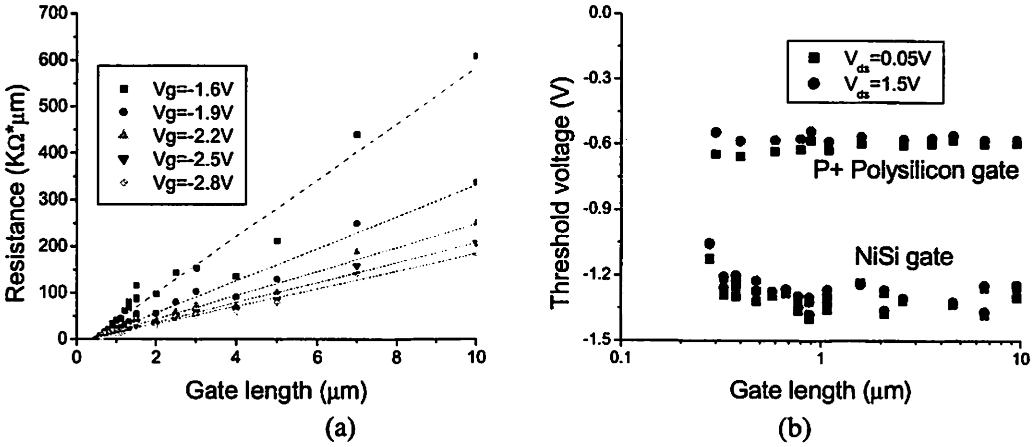


Fig 4.8. (a) Extraction of ΔL=0.42μm and R_{ext}=3.3kΩ•μm. (b) V_t roll-off curve for NiSi gate and P+ polysilicon gate devices. 0.7V V_t difference is observed

A major concern about silicide gates has been the possible degradation of the gate stack due to metal contamination. The gate insulator, together with its interfaces to the substrate and gate, is the key component in a transistor. The density of the trap states at the oxide/substrate interface can be used as the major indicator of the gate stack quality. A high trap density would degrade the swing in the subthreshold region and mobility in the inversion layer by scattering the charge carriers.

Gate capacitance measurement represents a good index of the interface quality since trap states can severely distort the CV curve. Fig. 4.9a shows an excellent match with the quantum simulation results of an ideal MOSCAP structure; meanwhile the oxide thickness, body doping and gate workfunction can be extracted from the fit (T_{ox}=27Å, N_{sub}=3.5×10^{17}cm^{-3}, ϕ_m=4.35eV). Simulation results with a poly depletion effect are also plotted and the improvement with the metallic gate in the inversion region is clearly demonstrated. The absence of the gate depletion effect proves that the gate is metallic at the interface, which means the entire silicon film has been converted into silicide. This increase of the inversion capacitance will significantly improve the drive current and device performance [4]. Fig 4.9b demonstrates that the carrier mobility in the inversion layer agrees with the universal mobility curve [19], as also pointed out in Ref. 20. Since interface traps would scatter carriers and degrade the mobility, the good fitting of the mobility demonstrates a good interface of NiSi on a 27Å oxide.

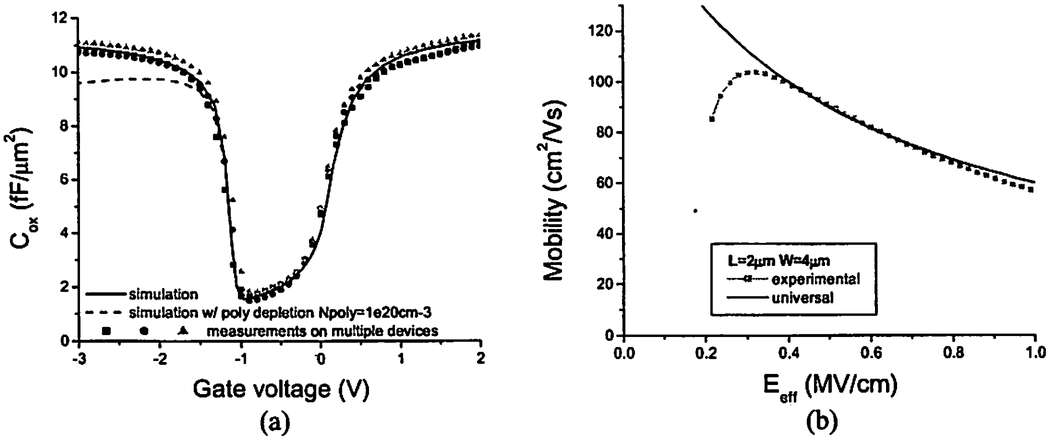


Fig 4.9. (a) The good match between measured and simulated CV results. (T_{ox}=27Å, N_{sub}=3.5e17cm^{-3}, V_{fb}=0.12V, ϕ_m=4.35eV). (b) Carrier mobility in the inversion layer approaches the universal mobility curve in silicon.

Gate current leakage is another important parameter of the oxide integrity because it sets the oxide-scaling limit. With an ideal gate dielectric, quantum tunneling is the only mechanism of current conduction. However, with traps or defects present inside the insulator, the current leakage increases dramatically with defect-aided conduction. Since the gate current is the limiting factor for gate oxide scaling, even a slight degradation of gate leakage increases the minimum oxide thickness. Fig. 4.10 shows that the experimental gate current agrees with published modeling results [21] over the whole bias range from accumulation to inversion. Therefore, an excellent gate oxide with low defects has been achieved with the NiSi gate.

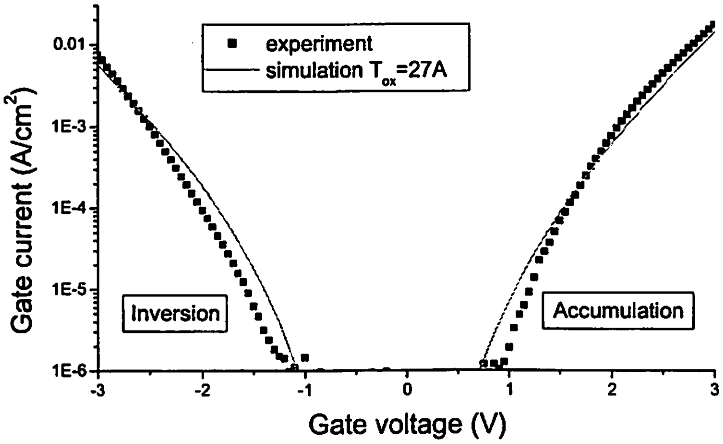


Fig 4.10. Gate current leakage of the NiSi gate device compared with modeling results. The oxide thickness agrees with CV extraction.

4.4.3 SOI PMOS results with NiSi gate

Similar results are obtained from devices fabricated on an SOI wafer. Without a body terminal, the subthreshold behavior for long channel transistors is almost ideal (S=61mV/dec), and no junction leakage is observed (Fig 4.11a). Since the SOI film is thick (30nm) and undoped, the process is not optimized for short channel devices (Fig. 4.11b). The threshold voltage agrees with the formula $V_{tp} = \phi_g - E_v - V_b$, as the model in Chapter 2 predicted.

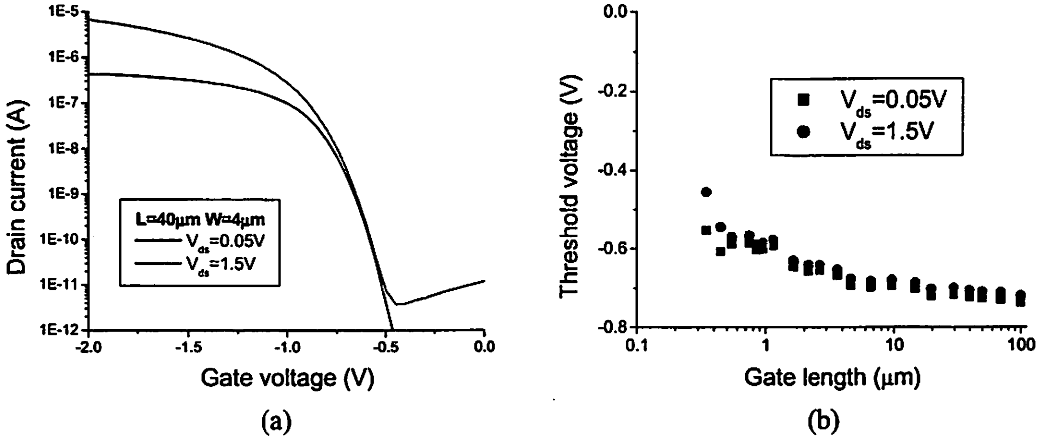


Fig 4.11. Example of (a) I_d~V_g and measurement. Ideal swing is achieved for SOI devices. S=61mV/dec, V_t=-0.7V and ϕ_m=4.33eV. (b) V_t roll-off curve for NiSi gate SOI devices.

Similarly, good matches for CV and mobility can be also obtained on SOI devices (Fig 4.12). Without a neutral body region, only C_{gs} is present in the measurement. Since there is no depletion charge in the undoped silicon film, the vertical electric field is lower in the SOI devices. Although carrier mobilities fall on the same universal curve, they are higher in SOI devices than in bulk MOSFETs at the same gate drive.

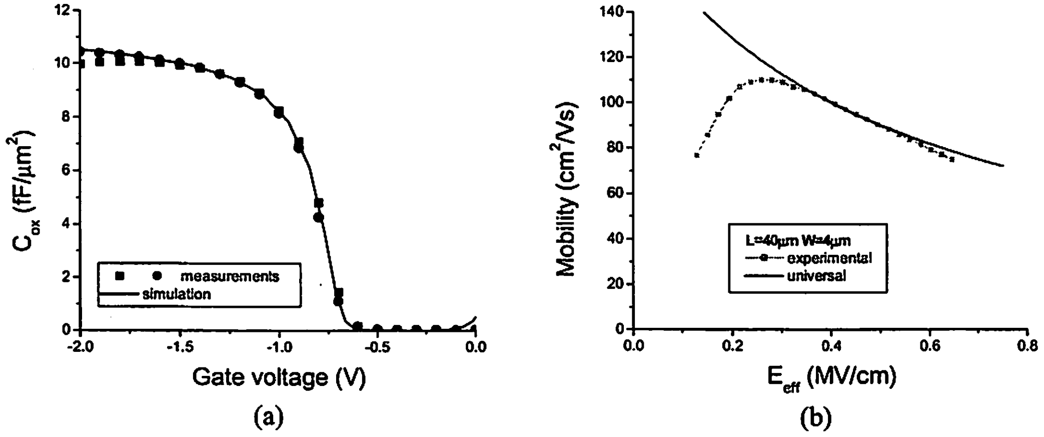


Fig 4.12. (a) Good match between measured and simulated CV results. (T_{ox}=27Å, N_{sub}=1x10^{15}cm^{-3}, V_{fb}=0V, ϕ_m=4.33eV). (b) Carrier mobility in the inversion layer approaches to the universal mobility curve in silicon.

In short, NiSi gate PMOS are fabricated with a good interface quality. The workfunction of the NiSi gate formed from undoped silicon is ~4.4eV, actually suitable for NMOS. A boron implant to the silicon film before it is silicided is required to shift the workfunction to be suitable for PMOS. Good gate capacitance, carrier mobility and gate leakage measurements indicate that no degradation of the thin oxide or its interfaces was caused by the NiSi gate. Therefore, NiSi is applicable as a single metallic gate material to nano-scaled fully depleted CMOS.

4.4.4 Bulk NMOS results with TiSi gate

NMOS transistors were fabricated with TiSi gates formed also from undoped amorphous silicon. Fig. 4.13 shows typical I_d-V_g and I_d-V_d curves. A correct threshold voltage and good drive current are shown (V_t=0.1V, I_{on}=183µA/µm at L=1µm). However, the slightly high subthreshold swing (S=87mV/dec) and the current tail in the low gate bias region indicate the existence of interface traps.

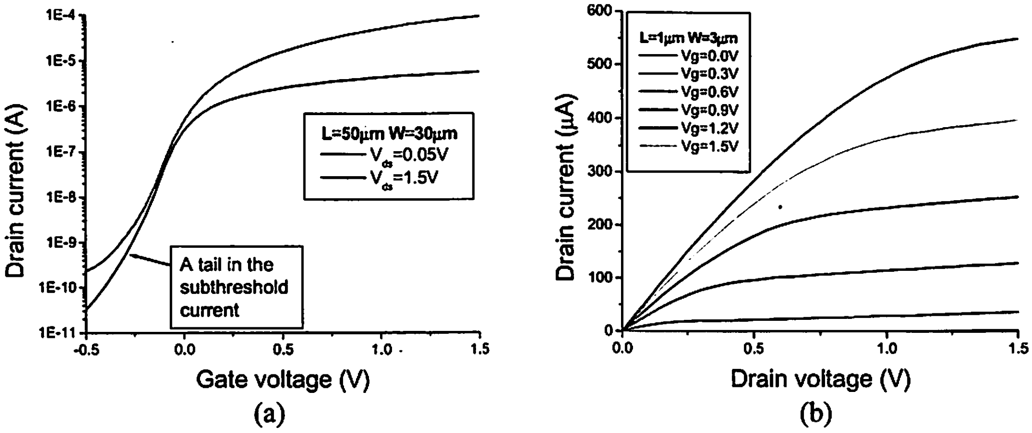


Fig 4.13. Example of (a) I_d~V_g and (b) I_d~V_d measurement. The correct V_t value of 0.1V is obtained with the TiSi gate. S=87mV/dec and the drive current is I_{on}=183µA/µm at L=1µm.

The effective channel length and external resistance can be extracted in the same manner as for PMOS (Fig. 4.14a). Since phosphorus is less diffusive and makes silicon more conductive, a smaller ΔL=0.3μm and R_{ext}=2.6kΩ•μm result. From the V_t roll-off curve (Fig. 4.14b), good SCE are obtained for channel lengths down to L=0.7μm. The threshold has very narrow variation but is slightly lower than the desired value (V_t=0.1V), which can be cured by a very light boron implant into the gate silicon.

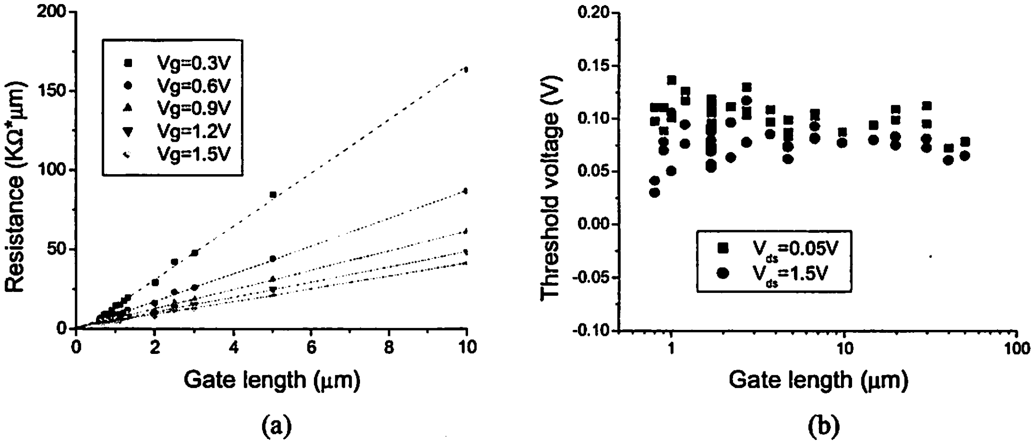
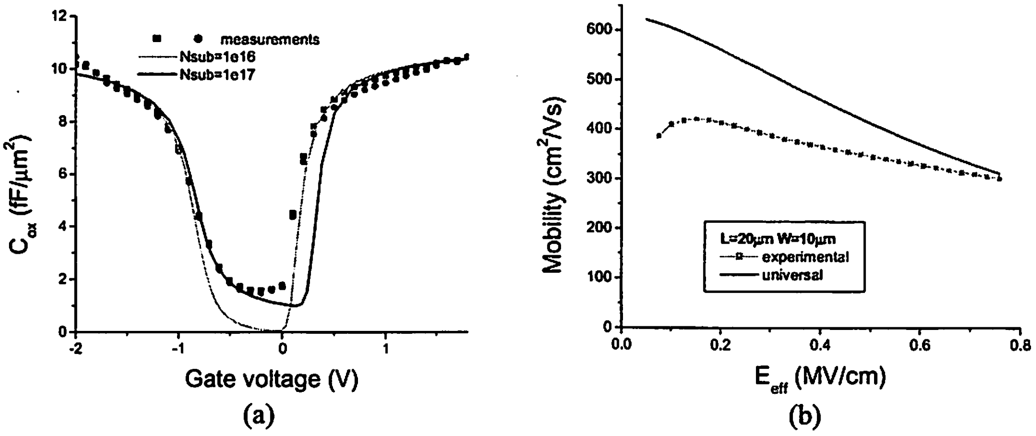


Fig 4.14. (a) Extraction of ΔL=0.3μm and R_{ext}=2.6kΩ•μm. (b) V_t roll-off curve for TiSi gate NMOS.

The quality of the substrate/oxide interface with a TiSi gate was also examined. The measured CV curve has a large distortion, and it is impossible to match simulation results with single body doping level (Fig. 15a). This distortion indicates an elevated trap density, which increases the capacitance in the depletion region and degrades the subthreshold swing as indicated in Fig. 4.13a. The extracted parameters are T_{ox}=29Å, ϕ_m=4.3eV, and N_{sub} between 1×10^{16}-1×10^{17}cm^{-3}. Carrier mobility in the inversion region (Fig. 15b) shows severe degradation at low vertical fields, also indicating elevated interface trap states, which scatter carriers and degrade the mobility.



(a)

(b)

Fig 4.15. (a) Distortion in the CV measured curve. (T_{ox}=29Å, $N_{sub}=1x10^{16}-1x10^{17}cm^{-3}$, V_{fb}=-0.7V, ϕ_m=4.3eV). (b) Severe degradation of inversion carrier mobility at low vertical field.

Finally, the oxide quality with a TiSi gate was checked via the gate current leakage. Severe degradation of the gate current is seen in Fig. 4.16. The gate current is 3 times than that predicted by the model, or equivalently, the tunneling oxide thickness is 1Å less than the value extracted from the gate capacitance measurement. In the intermediate voltage range (~2V), where the devices are actually biased in circuit applications, the degradation is actually even worse (~5X).

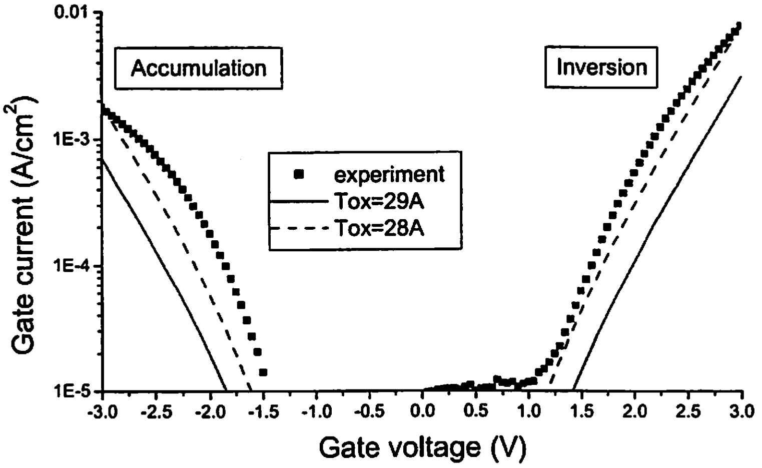


Fig 4.16. 3X degradation is observed in the gate current leakage with a TiSi gate.

It is well known that Ti can reduce SiO_2 to Si at high temperatures. It is believed that Ti atoms either react with or diffuse into the oxide and create traps inside the oxide and at the interface. All experimental results, including the subthreshold swing, CV measurement, carrier mobility and gate leakage, suggest that both the oxide and interface to the substrate are degraded with a TiSi gate. Further investigation and attempts at improvement are needed before TiSi can be implemented in CMOS as a gate material.

4.5 Conclusion

MOS capacitors with both NiSi and TiSi gates were fabricated, from which the workfunctions were extracted experimentally. Both NiSi and TiSi gates eliminate the gate depletion layer for enhanced device performance. Both workfunctions can be continuously manipulated by the implantation of dopants into the silicon film before the silicidation reaction, and both adjustable workfunction ranges cover the requirement of threshold voltages for both NMOS and PMOS with UTB and FinFET structures. The fixed charge density at the oxide/substrate interface is also sufficiently low for circuit applications.

The quality of the silicide gates on thin oxide was investigated for MOSFETs, as well. No degradation compared with polysilicon gates was observed with NiSi gates in PMOS. However, TiSi NMOS showed degraded interface quality. The tail in the subthreshold current, distortion in gate capacitance, degradation of low-field mobility and higher gate leakage all indicate a poor interface with an elevated trap density. Over all, NiSi is a promising gate material for CMOS applications and can be applied to UTB and FinFET to achieve the appropriate threshold voltages.

4.6 References

[1] International Technology Roadmap for Semicondutors, Semicondutor Industry Association, 2001. http://public.itrs.net/Files/2002Update/2001ITRS/Home.htm

[2] S. Takagi, A. Toriumi, M. Iwase and H. Tango, "On the universality of inversion layer mobility in Si MOSFET's: part I – effects of substrate impurity concentration," *IEEE, Trans. on Elec. Devices, Vol 41, No. 12, pp. 2357-2362. Dec. 1994*

[3] I.D. Mayergoyz and P. Andrei, "Statistical analysis of semiconductor devices." *Journal of Applied Physics, Vol. 90, No. 6, pp. 3019-3029. Sept. 2001*

[4] Y. Abe, T. Oishi, K. Shiozawa, Y. Tokuda and S. Satoh. "Simulation study on comparison between metal gate and polysilicon gate for sub-quarter-micron MOSFETs." *IEEE Electron Device Letters, Vol. 20, No. 12, pp. 632-634, Dec. 1999*

[5] B. Yu, D-H. Ju, W-C. Lee, N. Kepler, T-J. King and C. Hu. "Gate engineering for deep-submicron CMOS transistors." *IEEE Transactions on Electron Devices, Vol. 45, No. 6, pp. 1253-1262, June 1998*

[6] T. Hirose, Y. Momiyama, M. Kosugi, H. Kano, Y. Watanabe and T. Sugii, "A 185 GHz f_{max} SOI DTMOS with a new metallic overlay-gate for low-power RF applications." *IEEE. Proceedings of International Electron Devices Meeting, pp. 943-945. Dec. 2001*

[7] I. Polishchuk, P. Ranade, T-J. King and C. Hu. "Dual work function metal gate CMOS transistors by Ni-Ti interdiffusion." *IEEE Electron Device Letters, Vol. 23, No. 4, pp. 200-202, April 2002*

[8] I. De, D. Johri, A. Srivastava and C.M. Osburn, "Impact of gate workfunction on device performance at the 50nm technology node," *Solid-state electronics, Vol. 44,*

pp. 1077-1080, 2000

[9] K. Imai, K. Yamaguchi, T. Kudo, N. Kimizuka, H. Onishi, A. Ono, Y. Nakahara, Y. Goto, K. Noda, S. Masuoka, S. Ito, K. Matsui, K. Ando, E. Hasegawa, T. Ohashi, N. Oda, K. Yokoyama, T. Takewaki, S. Sone and T. Horiuchi. "CMOS device optimization for system-on-a-chip applications." *IEEE. Proceedings of International Electron Devices Meeting, pp. 455-458, Dec. 2000*

[10] S. Inaba, K. Okano, S. Matsuda, M. Fujiwara, A. Hokazono, K. Adachi, K. Ohuchi, H. Suto, H. Fukui, T. Shimizu, S. Mori, H. Oguma, A. Murakoshi, T. Itani, T. Iinuma, T. Kudo, H. Shibata, S. Taniguchi, T. Matsushita, S. Magoshi, Y. Watanabe, M. Takayanagi, A. Azuma, H. Oyamatsu, K. Suguro, Y. Katsumata, Y. Toyoshima and H. Ishiuchi, "High performance 35 nm gate length CMOS with NO oxynitride gate dielectric and Ni SALICIDE." *IEEE. Proceedings of International Electron Devices Meeting, pp. 641-644, Dec. 2001*

[11] M.C. Poon, F. Deng, H. Wong, M. Wong, J.K.O. Sin, S.S. Lan, C.H. Ho and P.G. Han, "Thermal stability of cobalt and nickel silicides in amorphous and crystalline silicon." *IEEE. Proceedings of Hong Kong Electron Devices Meeting, pp. 65-68. 1997*

[12] H.I. Liu, J.A. Burns, C.L. Keast and P.W. Wyatt. "Thin silicide development for fully-depleted SOI CMOS technology." *IEEE. Transactions on Electron Devices, Vol. 45, No. 5, pp. 1099-1104, May 1998*

[13] J. Kedzierski, E. Nowak, T. Kanarsky, Y. Zhang, D. Boyd, R. Carruthers, C. Cabral, R. Amos, C. Lavoie, R. Roy, J. Newbury, E. Sullivan, J. Benedict, P. Saunders, K. Wong, D. Canaperi, M. Krishnan, K-L. Lee, B.A. Rainey, D. Fried, P. Cottrell, H-SP.

Wong, M. Ieong and W. Haensch, "Metal-gate FinFET and fully-depleted SOI devices using total gate silicidation." *IEEE. Proceedings of International Electron Devices Meeting, pp. 247-250, Dec. 2002*

[14] M. Qin, V.M.C. Poon and C.H. Ho, "Investigation of polycrystalline Nickel silicide films as a gate material," *Journal of the Electrochemical Society, 148(5) pp. 271-274, 2001*

[15] 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", *IEEE. Proceedings of International Electron Devices Meeting, pp. 367-370, Dec. 2002*

[16] S. P. Murarka, "Silicides for VLSI applications," Academic Press, New York, 1983

[17] M.C. Poon, F. Deng, M. Chan, W.Y. Chan and S.S. Lau, "Resistivity and thermal stability of nickel mono-silicide." *Applied Surface Science, Vol. 157, No. 1-2, pp. 29-34, March 2000*

[18] A. Yagishita, T. Saito, K. Nakajima, S. Inumiya, Y. Akasaka, Y. Ozawa, K. Hieda, Y. Tsunashima, K. Suguro, T. Arikado and K. Okumura, "High performance damascene metal gate MOSFETs for 0.1 um regime," *IEEE Transactions on Electron Devices, Vol. 47, No. 5, pp. 1028-34, May 2000*

[19] M-S. Liang, J.Y. Choi, P-K. Ko and C. Hu, "Inversion-layer capacitance and mobility of very thin gate-oxide MOSFETs." *IEEE Transactions on Electron Devices, Vol. 33, No.3, pp. 409-413, March 1986*

[20] Z. Krivokapic, W. Maszara, K. Achutan, P. King, J. Gray, M. Sidorow, E. Zhao, J. Zhang, J. Chan, A. Marathe and M.R. Lin, "Nickel silicide metal gate FDSOI

devices with improved gate oxide leakage," *IEEE. Proceedings of International Electron Devices Meeting, pp. 271-274, Dec. 2002*

[21] W-C. Lee and C. Hu, "Modeling gate and substrate currents due to conduction and valance band electron and hole tunneling," *Proceedings of the 2000 symposium on VLSI technology, pp. 98-99, June 2000*

Appendix 4A Process flow for bulk silicide gate MOSFETs

Step	Process	Process specification	Equipment	Comments
0	Starting 4" wafer with resistivity of 5Ω-cm: p type for NMOS, n type for PMOS			
0.1	Scribe	Label the wafers		
1	LOCOS isolation			
1.1	Cleaning	Piranha (H_2O_2:H_2SO_4=1:5) 120°C, 10min, 25:1 BHF 30s	Sink6	Resistance to 16kΩ
1.2	Pad Oxide	SDRYOXA 950°C, 30min, 20min anneal	Tylan2	Oxide=20nm
1.3	Nitride dep	9SNITA, 800°C, 300mTorr, NH_3=75sccm, DSC=25sccm, 40min	Tystar9	Nitride=150-160nm
1.4	LOCOS litho	Resist coating: coat=program 1/bake=program 1 Exposure: focus=255, t=3.9s Development: bake=program 1/develop=program 1 Descum: O_2=51sccm, P=50W, t=1min Hard bake: 120°C, 1hr	Svgcoat1 GCAWS Svgdev Technics-c Ovrn	PR=1.2μm PEB: 90C, 1min DEV: OPD4226, 1min
1.5	Nitride etch	NITSTD1: ME: P=375mTorr, He=50sccm, SF_6=175sccm, RF=150W, t=125s. Overetch: same as ME, t=15%	Lam1	ER=13Å/s
1.6	Resist strip	O_2 ashing, 230W, 5min	Technics-c	
1.7	Cleaning	Piranha 120°C, 10min	Sink8	Resistance to 16kΩ
1.8	Cleaning	Piranha 120°C, 10min	Sink6	Resistance to 16kΩ
1.9	LOCOS	SWETOXB, 1000°C, 80min	Tylan2	Oxide=420nm
1.10	Nitride strip	10:1 BHF 30s, H_3PO_4, 180°C, 3.5hours	Sink7	Dewet, field oxide-15nm
2	Channel preparation			
2.1	Cleaning	Piranha 120°C, 10min, 25:1 BHF, 90s	Sink6	Dewet, field oxide-20nm
2.2	Sac oxide	SWETOXB, 900°C, 15min	Tylan2	Oxide=45nm
2.3	Cleaning	Piranha 120°C, 10min, 25:1 BHF 180s	Sink6	Dewet, field oxide-45nm
2.4	Sac oxide	SGATEOX, 900°C, 30min	Tylan6	Oxide=15nm
2.5	V_t implant	NMOS: B^+, 15keV, 6e12cm^{-2}, PMOS: P^+, 50keV, 2e12cm^{-2}	Core sys.	Foundry. Rp=80nm
2.6	Body litho	See step 1.4		Substrate contact
2.7	Body imp.	NMOS: B^+, 15keV, 5e15cm^{-2}, PMOS: P^+, 50keV, 5e15cm^{-2}	Core sys.	Foundry. Rp=80nm
2.8	Resist strip	O_2 ashing, 230W, 5min	Technics-c	
2.9	Cleaning	Piranha 120°C, 10min	Sink8	Resistance to 16kΩ
2.10	Cleaning	Piranha 120°C, 10min, 25:1 BHF, 90s	Sink6	Dewet, field oxide-15nm
2.11	Sac oxide	THIN_ANN, 800°C, 30min	Tylan6	Oxide=3.2nm
2.12	Measurement	Measure remaining field oxide	Nanoduv	oxide=330nm
3	Dummy gate formation			
3.1	Cleaning	Piranha 120°C, 10min, 25:1 BHF 15s	Sink6	Dewet, field oxide-4nm
3.2	Dummy SiO_2	THIN_ANN, 900°C, 25min	Tylan6	Oxide=13nm
3.3	Dummy gate	10SUPLYA, 615°C, 375mTorr, SiH_4=100sccm, 44min	Tystar10	Poly=440nm
3.4	Gate litho.	See step 1.4		It determines gate length
3.5	Inspection	Measure the resulted gate length from lithography	Leo	Extract gate length

3.6	Gate etch	B: p=13mTorr, CF_4=100, P_{top}=200, P_{bot}=40, t=20s
		M: P=15mTorr, Cl_2=50, HBr=150, P_{top}=300, P_{bot}=150, t=45s
		O: P=35mT, HBr=200, O_2=5.0, Ptop=250, Pbot=120, t=30s
			Lam5	ER=20Å/s, SiO_2/Si~1
				ER=75Å/s, Si/SiO_2~10
				ER=50Å/s, Si/SiO_2~100.

3.7	Resist strip	O_2 ashing, 230W, 5min	Technics-c	
3.8	S/D litho	See step 1.4		Cover the body contact
3.9	S/D implant	NMOS: As^+, 30keV, 4e15cm^{-2}, NMOS: BF_2, 20keV, 4e15cm^{-2}	Core sys.	Foundry. Rp=30nm
3.10	Resist strip	O_2 ashing, 230W, 5min	Technics-c	
4	CMP and dummy gate removal			
4.1	Cleaning	Piranha 120°C, 10min	Sink8	Resistance to 16kΩ
4.2	Cleaning	Piranha 120°C, 10min,	Sink6	Resistance to 16kΩ
4.3	CMP oxide	11SULTOA 450°C,300mTorr, SiH_4=25sccm, O_2=75sccm, 40min	Tylan11	Oxide=660nm
4.4	CMP	Poly.polish, 4×50s, rotate 90° between runs. CMP is non-uniform. Thicker oxide remains at the die corners. Remain poly: 200nm. Remain oxide: 550nm on field, 400nm on active area.	CMP	ER of oxide=30Å/s ER of poly is higher
4.5	Measurement	Measure oxide thickness after each CMP run	Nanoduv	Non-uniform oxide
4.6	CMP clean	a) DI water rinse 1min. b) NH_4OH 1min. c) DI water rinse 1min. d) Piranha 120°C 1min. e) DI water rinse 6min. f) 5:1 BHF 10s. g) DI water rinse 6min. h) SC-1 (NH_4OH:H_2O_2:H_2O=1:1:5) 5min. i) DI water rinse 1min	Sink432	Put wafer in wafer before cleaned. Dry slurry is hard to remove.
4.7	Gate strip	B/m/o=60s/10s/45s. Long breakthrough for non-uniform oxide	Lam5	Stop on dummy oxide
4.8	Cleaning	Piranha 120°C, 10min	Sink8	Resistance to 16kΩ
5	Real gate formation			
5.1	Cleaning	Piranha 120°C, 10min, 25:1 BHF 90s	Sink6	remove dummy oxide
5.2	Gate oxide	THIN-ANN. 750°C, O_2 15min + N_2 5min + 900°C, N_2 5min	Tylan6	TCA clean, 25±1Å
5.3	a-Si gate	SiGeVAR.019, 425°C, 300mTorr, Si_2H_6=100, 30min	Tystar19	Silicon=23-26nm
5.4	Metal sputtering	PMOS: Ni: 15mTorr, 1kW, 60cm/min, 1 pass, NMOS: Ti: 10mTorr, 2kW, 60cm/min, 1 pass	CPA	36nm, 8Ω/ 33nm, 35Ω/
5.5	Gate litho	See step 1.4.		It covers the real gate
5.6	Metal etch	Ni: 5:1 BHF. Fast but non-uniform etch. Rinse frequently Ti: SC-1: NH_4OH:H_2O_2:H_2O=1:2:5. ER=10nm/min	Sink432	PR is baked overnight
5.7	a-Si etch	B/m/o=15s/10s/30s	Lam5	20nm oxide is etched
5.8	Resist strip	O_2 ashing, 230W, 5min	Technics-c	
6	Metal contact			
6.1	Cont. litho.	See step 1.4		S/D/Sub contact holes
6.2	Cont. etch	5:1 BHF, 1min	Sink7	ER=400nm/min
6.3	Ti evap.	E-beam evaporation. Real-time thickness monitored	Ultek	Ti=50nm
6.4	Ti lift-off	Acetone in an ultrasonic bath	Sink432	
6.5	Anneal	N_2: 400°C-600°C, 2min,	Heatpulse1	In accumulative steps
7	Calibration			

Appendix 4B Process flow for SOI silicide gate MOSFETs

Step	Process	Process specification	Equipment	Comments
0	4" SOI wafer, 950Å Si on 4000Å buried oxide			
0.1	Scribe	Label the wafers		
1	Si film thinning			
1.1	Cleaning	Piranha 120°C, 10min, 25:1 BHF 30s	Sink6	Resistance to 16kΩ
1.2	Sac Oxide	SDRYOXA 1000°C, 100min, 20min anneal	Tylan2	SiO_2=70nm, Si_{remain}=64nm
1.3	Cleaning	Piranha 120°C, 10min, 10:1 BHF 300s	Sink6	Dewet
1.4	Sac Oxide	SDRYOXA 950°C, 40min, 20min anneal	Tylan2	SiO_2=27nm, Si_{remain}=52nm
1.5	Cleaning	Piranha 120°C, 10min, 10:1 BHF 120s	Sink6	Dewet
1.6	Sac Oxide	SDRYOXA 1000°C, 45min, 20min anneal	Tylan2	SiO_2=40nm, Si_{remain}=34nm
1.7	Cleaning	Piranha 120°C, 10min, 10:1 BHF 180s	Sink6	Dewet
1.8	Pad Oxide	THIN_ANN 900°C, 20min, 20min anneal	Tylan6	SiO_2=8nm, Si_{remain}=30nm
2	Mesa isolation			
2.1	Mesa litho	See step 1.4 of bulk MOSFET process flow		The same mask as LOCOS
2.2	Si etch	B/m/o=15s/5s/20s	Lam5	
2.3	Resist strip	O_2 ashing, 300W, 7min	Technics-c	
2.4	Measurement	Measure remaining buried oxide	Nanoduv	oxide=390nm
2.5	Cleaning	Piranha 120°C, 10min	Sink8	Resistance to 16kΩ
3	The rest of the process from here is the same as that of bulk MOSFETs			

Chapter 5

FinFET SONOS Flash Memory

5.1 Introduction

5.1.1 Operation of flash memory

Memory devices are an indispensable semiconductor electronic component in digital applications. While volatile memories (SRAM and DRAM) provide fast read/ write operations, they have large cell size or high power consumption. With the boom over the past decade in the market for mobile electronics, such as cellular phones, digital cameras, personal digital assistants, MP3 players, wireless networking and global positioning systems, low power and low cost memory chips have been attracting more and more attention. Although slower than its volatile counterparts, non-volatile memory (NVM) is the most suitable solution for mobile applications because it offers 10 years retention time even without a power supply [1].

Among all NVM structures, flash memory is the mainstream non-volatile memory device in both production and development today. A flash memory cell is simply a MOSFET with an extra poly-silicon [2] or silicon nitride [3] film sandwiched between the tunnel oxide and the inter-poly oxide to form a charge storage layer (Fig. 5.1). In programming, electrons are injected through the tunnel oxide (bottom oxide) into the floating gate by either tunneling (NAND type) [4] or hot carrier injection (NOR type) [5]. The charge tunnels out of the floating gate in an erasing operation. The charge in the floating gate alters the threshold voltage (V_t) of the transistor, through which the stored information can be determined.

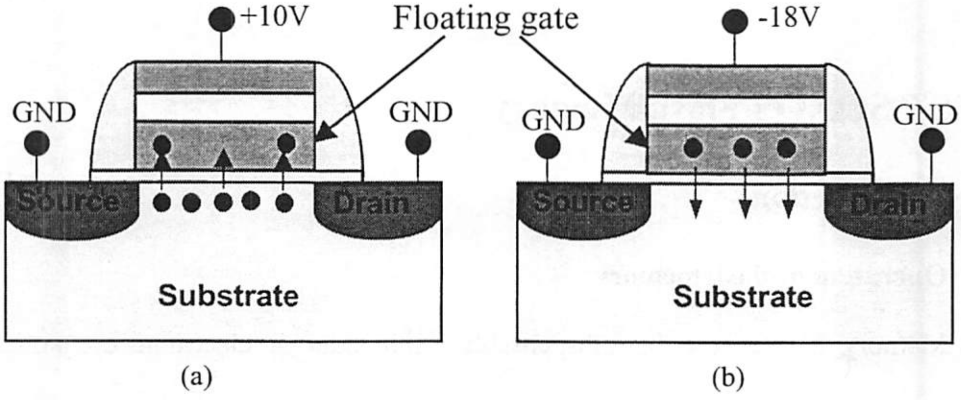


Fig 5.1 The cross section of a NAND flash cell under (a) programming and (b) erasing.

Flash memory dominates the current NVM market because it provides the following advantages. First, flash memory can achieve the highest cell density since a flash memory cell consists solely of a single transistor [6]. Second, flash memory supports multi-bit storage [7], which further increases the memory density and reduces the cost. Mutiple V_t states can be generated through the control of the number of electrons injected into the floating gate. Two-bit per cell (with four V_t states) flash memory has already been commercialized, while four-bit per cell flash memory is in development now [8]. Furthermore, Matrix Semiconductor Inc. has demonstrated multi-layer (or 3D integration) flash memory [9], which offers another possibility for an even higher density and lower cost flash memory solution. Third, flash memory uses a fabrication process compatible with the current CMOS process flow, and it is a perfect solution for embedded memory applications. The integration of flash memory with logic and analog devices for better performance and lower cost has been demonstrated [10].

NAND flash provides higher density and lower power consumption than does

NOR flash, although it requires more complicated periphery circuitry for its proper operation [11]. Currently, 4GB NAND-type flash memory has been demonstrated [12], and NAND flash will be the focus of this report.

5.1.2 Flash memory scaling and SONOS memory

The size and operation voltage of flash memory have been scaled dramatically to achieve high capacity and low power consumption. The reduction of the operation voltage is very closely related to scaling of the cell size because high voltage requires large space for cell isolation [13] and sophisticated periphery circuitry [14]. From experience in logic device scaling, it is well understood that the scaling of gate EOT (equivalent oxide thickness) is crucial in the suppression of short channel effects (SCE). Moreover, a thin gate stack enables the cell to be programmed/erased at a low voltage. However, the scaling of flash memory devices lags far behind that of logic devices due to their unique gate structure and operation mechanism. By 2002, CMOS logic devices had been scaled down to a gate length of 40nm with a gate dielectric EOT of less than 1.5nm and a voltage supply of 1V or even below [15]. Meanwhile, flash memory still has a gate length of above 100nm, a gate stack with an EOT more than 8nm and a voltage supply of above 8V for its normal writing and erasing [3].

In a floating gate flash memory cell, which stores the bit information in a conducting polysilicon floating-gate, the minimum oxide thickness is limited by its reliability requirement. After a flash device is stressed for a million cycles of program and erase (P/E), which is the typical reliability requirement for a NVM device, defects are generated in the insulator. With a single defect path formed inside the tunnel oxide, all charge stored in the conducting floating gate leaks out, and the bit information is lost (Fig.

5.2a). To meet the reliability and retention time requirements, commercial flash memory uses a tunnel oxide of more than 7nm, a significant barrier for memory device scaling. A typical flash memory gate stack consists of a 7nm tunnel oxide, 100nm poly-silicon floating gate and 14nm inter-poly oxide [2]. With the EOT of the whole gate stack up to 21nm, this memory cell would show severe SCE if scaled below 100nm. A large subthreshold swing and drain-induced barrier lowering require a higher V_t window for the distinguishing of programmed and erased states and cause a higher current leakage from unselected cells during a reading operation.

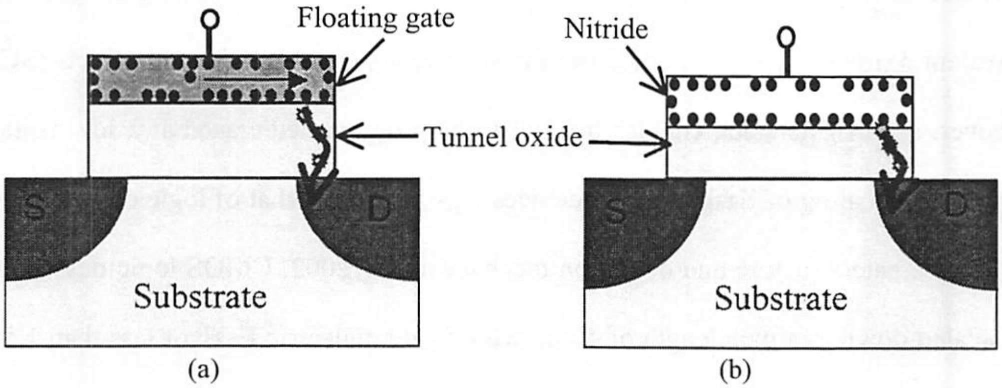


Fig 5.2 A floating gate flash cell (a) is more sensitive to defects than is a SONOS cell (b). Therefore, a thinner tunnel oxide can be used in a SONOS memory.

A SONOS (silicon-oxide-nitride-oxide-silicon) gate stack significantly reduces the minimum thickness of the tunnel oxide by storing charge in trap states inside the sandwiched nitride layer rather than in a conductive floating gate. Since the traps are isolated from each other, even if a defect path forms in the tunnel oxide, most of the charge will remain in the nitride and the information can still be retained (Fig. 5.2b) [16]. In a SONOS device, the minimum tunnel oxide thickness is limited by the requirement of 10 years retention because the stored charge can leak out of the nitride layer through the thin oxide even when the device is at idle. Significant V_t window closure after 10 years retention can be observed when the tunnel oxide is scaled below 2nm [17].

5.1.3 FinFET SONOS memory

Although a SONOS gate stack significantly improves the scalability of flash memory, it still requires the tunnel oxide thickness to be at least 2nm; otherwise, the information can not be retained for 10 years. The inter-poly oxide has to be thicker than the tunnel oxide to prevent the current tunneling from the top gate [18], and the nitride layer has to capture enough trap states for charge storage [19]. Therefore, the minimum EOT of the entire gate stack is about 7nm, which has become the major challenge when flash memory is scaled into sub-100nm regions.

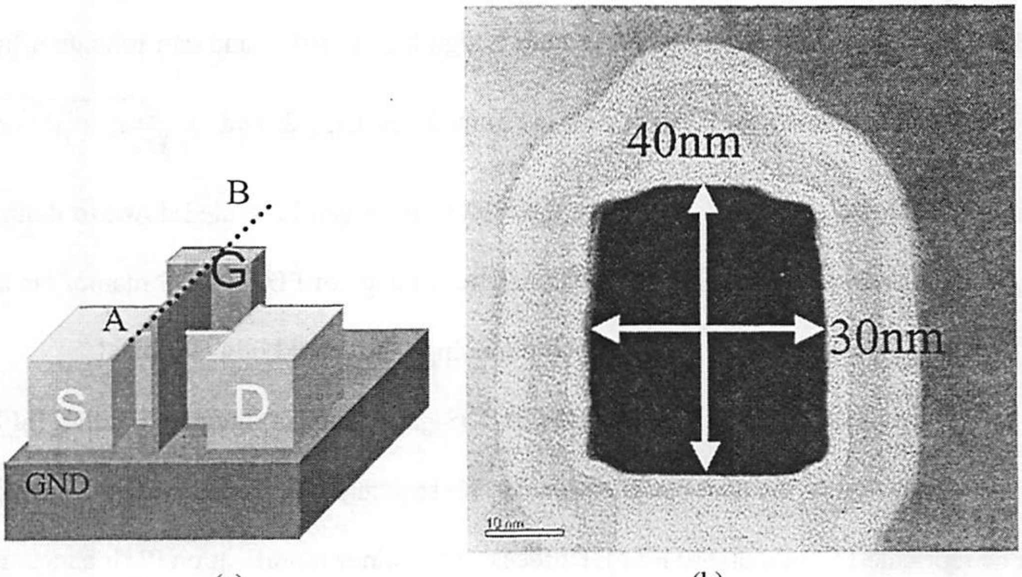
On the other hand, fully depleted (FD) SOI structures have been proposed to suppress SCE for sub-30nm CMOS technologies [20]. Ultra-thin body and double-gate (DG) devices provide an alternative means of device scaling, which is the scaling of the body. Typically, flash memory works under a high bias (~10V), and can tolerate a high swing (S=200mV/dec). In Chapter 2, it is predicted that L_{min}~$2l$ and $l=\sqrt{\frac{\varepsilon_{si}}{\varepsilon_I}T_I d+\frac{1}{2}(T_I^2+d^2)}$. Even with a gate stack EOT of 10nm, the SONOS flash can be scaled down to 40nm if the body is thinned down to d=10nm. One extra advantage of FD SONOS memory is that it eliminates cross talk between devices by removing the shared body terminal.

In this work, for the first time, the SONOS gate stack is integrated with a FinFET device because it is the most manufacturable DG structure. A FinFET SONOS memory can be fabricated from a planar FinFET process with minor modification [21]; hence, it is a good candidate for embedded memory in FinFET integrated circuits. It is found that the FinFET SONOS device has a performance similar to that of a bulk-Si SONOS device

[22], although it does not have a neutral body. Devices fabricated with channels on (100) and (110) silicon surfaces are compared in terms of program/erase speeds, endurance and retention.

5.2 Structure and fabrication of FinFET SONOS memory

Fig. 5.3 shows the structure and cross section TEM (Transmission Electron Microscopy) image of a FinFET SONOS memory cell. The real devices demonstrated in this report have a fin width 10nm narrower than what is shown, i.e. T_{si}=20nm. This device has conducting channels on three surfaces: the sidewalls and a portion of the top surface. Due to the thick dielectric stack, the outer electrode (gate) has much a larger area than does the inner electrode (fin). ISE (Integrated Systems Engineering) capacitance simulation shows an effective channel width of 140nm for each fin.

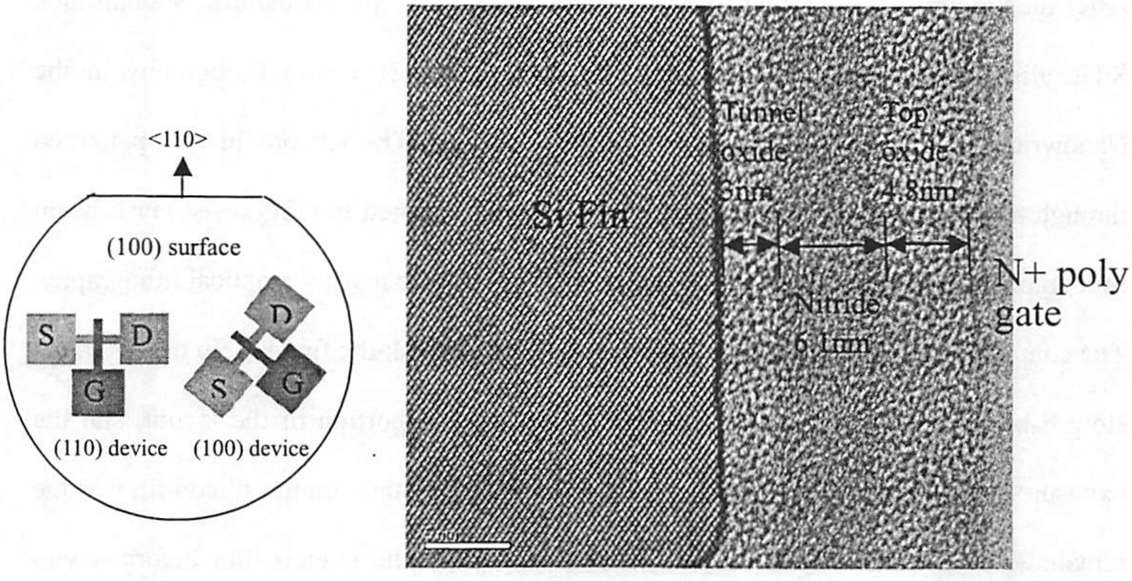


(a) (b)

Fig. 5.3. The FinFET SONOS memory device's (a) structure and (b) cross section along the AB line

The starting SOI wafer had a 100nm silicon film on a 400nm BOX (buried oxide). After the silicon film was thinned down to 40nm with multiple oxidations, 800nm thick SiGe alignment marks were created for electron beam (E-beam) lithography in the Nanowriter in Lawrence Berkeley National Laboratory. The silicon fin was patterned through a double resist exposure: fine features were patterned in HSQ resist via E-beam lithography while big probing pads were patterned in G-line resist via optical lithography. The combination of these two resist layers was used to mask the fin etch. In this way, the slow E-beam lithography was used only for a minimum portion of the layout, and the exposure time could be tremendously reduced. The oxide bump on the silicon fin was the remainder of the oxide hard mask thermally grown on the silicon film before it was patterned (Fig. 5.3b).

The etched sidewall surfaces of the fin, which would become the channels of the FinFET, were cured by a sacrificial oxidation. The crystal orientation of the channel surface on the fin sidewalls was controlled by proper orientation of the fin relative to the major flat, as shown in Fig. 5.4a. The gate stack consisted of a tunnel oxide (3nm), a silicon nitride (6.1nm), an inter-poly oxide (4.8nm) and a N+ in-situ doped polysilicon gate (180nm), as shown in the enlarged TEM image (Fig. 5.4b). The tunnel oxide was thermally grown at 810°C for 24min in diluted oxygen and the others were deposited through low-pressure chemical vapor deposition (LPCVD). Fig 5.4b was obtained from a device with a (110) sidewall surface, and a slightly thinner tunnel oxide was expected on (100) devices [23]. Note that the nitride and inter-poly oxide (HTO) were slightly thinner than expected due to imperfect step coverage on the sidewalls of the LPCVD processes. The TEM result of our experiment shows that the step coverage is 97% and 92% for nitride and HTO, respectively.



(a) (b)

Fig. 5.4. (a) The control of FinFET channel surface orientation. (b) Enlarged TEM image of the gate stack.

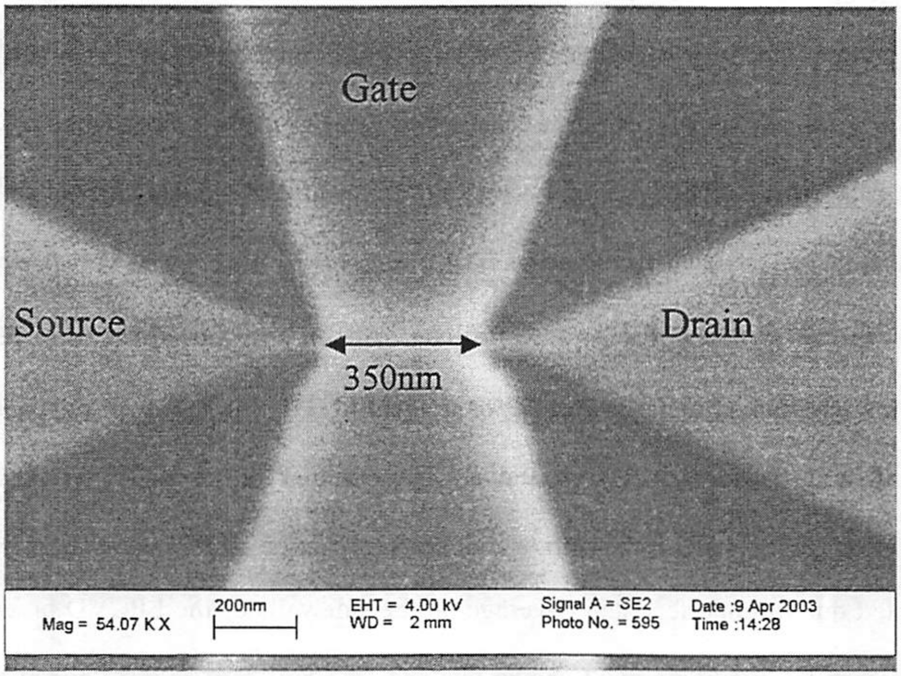


Fig. 5.5 Top view SEM image of a FinFET SONOS memory device. The extracted gate length is 350nm.

The gate was then patterned in the same way as the fin was. After source/drain implantation and activation, contacts were opened for probing. The devices were then annealed in forming gas for 5min at 400°C. The detailed process flow is listed in Appendix 5A. Fig. 5.5 shows a top-view Scanning Electron Microscopy image of the memory cell, from which the gate length was determined to be L_g=350nm.

5.3 Performance characteristics

There are four performance parameters of a flash memory: speed, reliability, retention and scalability. Fast program/erase speeds, but with low voltage operation, are highly preferred. After being stressed for 1 million P/E cycles, which is the typical flash reliability requirement, the cell still shows good characteristics, indicating excellent device endurance. 10 years retention of the stored information is achieved with or without reading disturbance for stressed cells at 85°C. Our FinFET SONOS structure has been demonstrated by simulation to be scalable to L_g=40nm, while still offering good performance.

5.3.1 P/E speeds of FinFET SONOS memory

The FinFET SONOS memory cell can be programmed/erased by a positive/negative gate voltage, with both source and drain grounded. The P/E characteristics of a memory device fabricated on (100) sidewalls are shown in Fig. 5.6. A threshold voltage (V_t) window between 0.9V (erased state) and 2.9V (programmed state) can be achieved with a -11V/10ms erase pulse and a 10V/5ms program pulse, respectively. As expected, higher operation voltages result in faster P/E speeds. The erase characteristics are comparable to those of bulk-Si devices [22], although there is no neutral body as a reservoir of holes. No saturation in the programming is observed in our FinFET SONOS device, so a larger V_t window (>3.5V) can be generated with longer P/E pulses. When the erase voltage is low (magnitude is less than 11V), a negative V_t can be achieved with a long erase pulse. Therefore, it is possible to have holes generated and injected from the channel into the nitride layer, although there is no body contact to the ultra-narrow fin. With a high erase voltage, the V_t finally saturates due to the balance of the electron current tunneling through the inter-poly oxide and the hole current tunneling through the tunnel oxide [24].

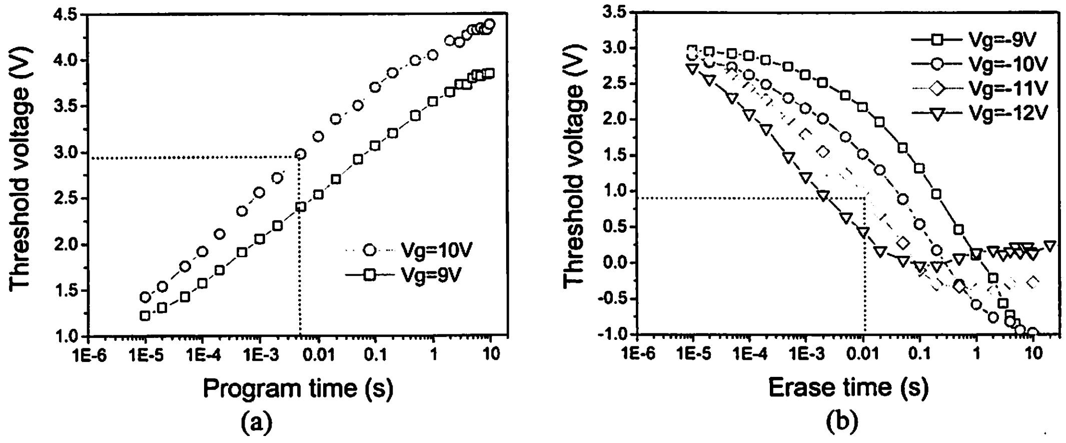


Fig. 5.6. Program (a) and erase (b) characteristics of the FinFET SONOS memory device fabricated on (100) sidewalls

The FinFET SONOS memory device fabricated on (110) sidewalls has also been tested. It has three times slower P/E speeds, as shown in Fig. 5.7: a –11V/35ms erase pulse and 10V/12ms program pulse are required to achieve the same V_t window as for the (100) device. This may be simply due to the thicker tunnel oxide grown on the (110) silicon surface as compared with the (100) silicon surface. From the tunneling current ratio between the (100) and (110) devices, which is approximately the ratio of their P/E speeds, the tunnel oxide on a (110) surface is estimated to be 1Å thicker than that on a

(100) surface [25].

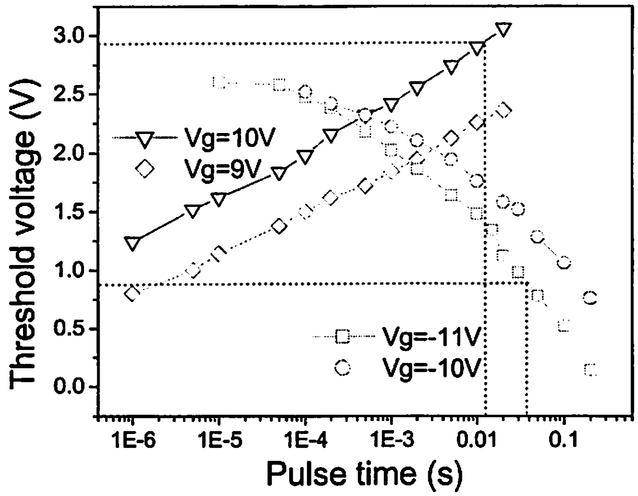


Fig. 5.7. Program and erase characteristics of the FinFET SONOS memory device fabricated on (110) sidewalls

5.3.2 Reliability of FinFET SONOS memory

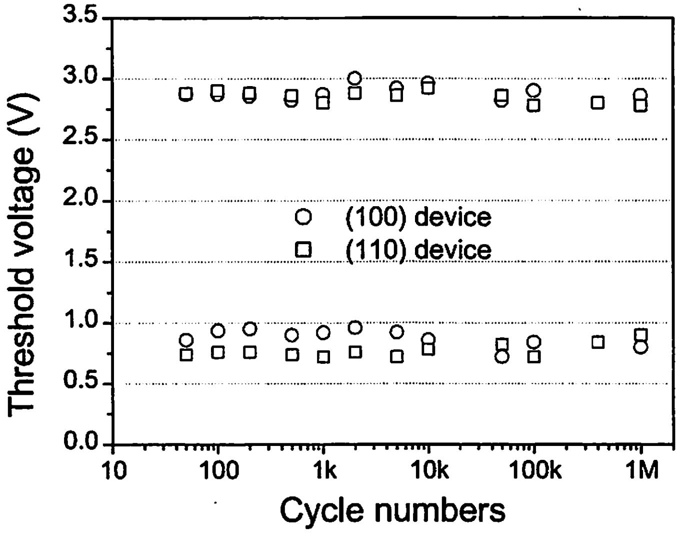


Fig. 5.8. Endurance characteristics of both (100) and (110) devices. No degradation in the V_t window is observed after 1 million P/E cycles. (100) device P/E pulses are 10V/5ms and –11V/10ms. (110) devices P/E pulses are 10V/12ms and –11V/35ms.

The major advantage of a SONOS cell over floating gate flash memory is that it offers high immunity to oxide defects by storing charges in the insulating Si_3N_4 layer. The excellent reliability of the SONOS device is demonstrated in the endurance measurement, as shown in Fig. 5.8. Both (100) and (110) memory devices show excellent endurance up to 1 million P/E cycles without noticeable V_t window degradation. Therefore, our FinFET SONOS structure with a 3nm tunnel oxide is sufficient for 1 million cycles.

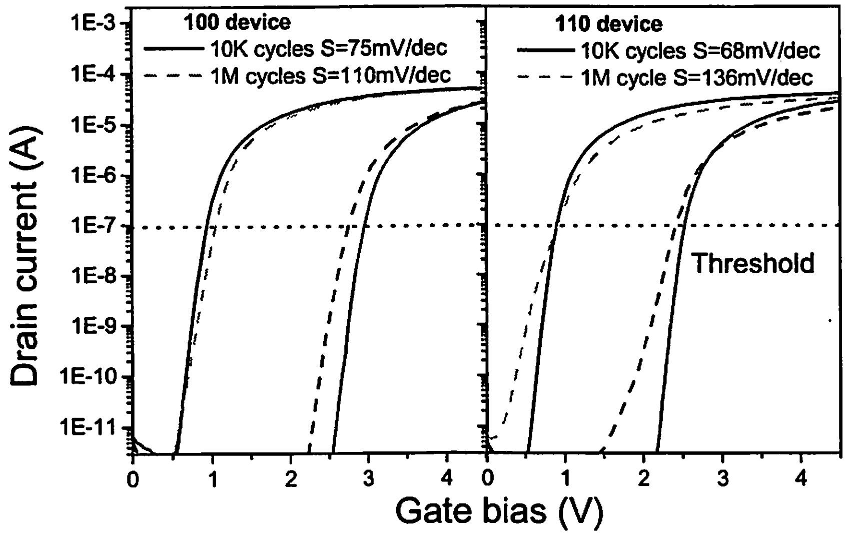


Fig. 5.9. Subthreshold swings degrade after 1 million stress cycles. The initial FinFET SONOS devices show good subthreshold swings. The (100) device shows less degradation after 1 million cycles. (V_{ds}=1V)

Although a SONOS gate stack improves the device's immunity to defects, it does not stop the generation of defects and traps during stress cycles. A significant number of interface traps are generated after the device is stressed beyond 10 thousand P/E cycles [17] and can be detected in the degradation of subthreshold swing. The I_d-V_g characteristics in Fig. 5.9 show the swing difference between devices after 10 thousand and 1 million stress cycles. Although no apparent degradation in the V_t window is observed, significant subthreshold swing degradation occurs after 1 million P/E cycles of stress: from 75mV/dec to 110mV/dec for the (100) memory device, and from 68mV/dec to 136mV/dec for the (110) memory device. While both are degraded after 1 million P/E cycles, it has been concluded that there are fewer interface traps generated in the (100) device. Therefore, the (100) surface has better resistance to stress.

The generation of interface traps also causes degradation of inversion carrier mobility in SONOS memory devices. Fig. 5.10 plots the measured mobility together with the universal mobility curves [26]. First, the universal mobility on the (110) surface is lower due to the higher electron effective mass and surface roughness [27]. Second, the devices after 10K stress cycles, which are almost as good as fresh ones, already show significant mobility degradation, which can be attributed to the surface roughness resulting from the FinFET process [28]. In a FinFET, the channel surfaces are the etched sidewalls created in the fin patterning. Although cured by a sacrificial oxidation, the surfaces are still rough and degrade the carrier mobility. Third, after being stressed for 1 million cycles, (110) devices show substantial mobility degradation, indicating the generation of a large number of interface traps. On the other hand, (100) devices show high resistance to stress on the basis of the somewhat less reduction of mobility.

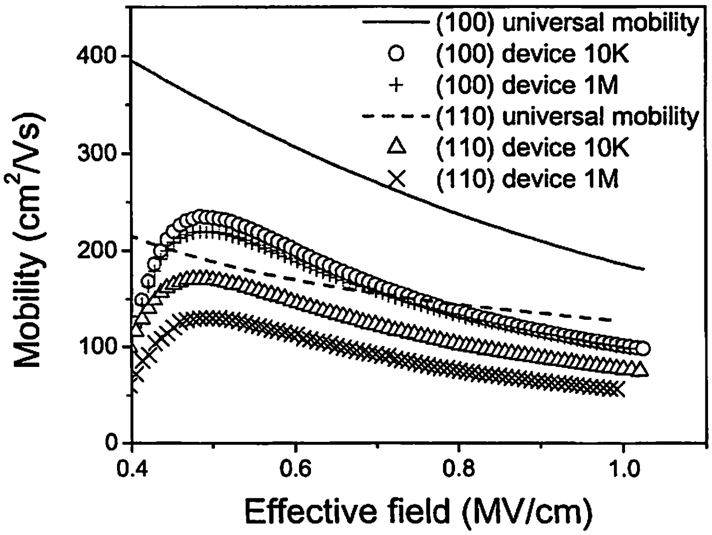


Fig. 5.10. Carrier mobility in the FinFET SONOS devices. The (100) device has high mobility and shows less degradation due to stress.

5.3.3 Retention and reading operation

The retention time has been measured at 85°C, the highest temperature for most FLASH applications, on both (100) and (110) devices after 1 million P/E cycles (Fig. 5.11). The erased state shows virtually no V_t drift with time, while the programmed state shows the leakage of stored charges. The (110) memory device shows better retention time because a thicker tunnel oxide is grown on the (110) silicon sidewalls, but this benefit comes at the price of slower P/E speeds. After 10 years retention, the V_t window in the (100) device is determined to be 1.4V, while it is 1.9V in the (110) device. The large V_t window in our FinFET SONOS devices enables multiple bit storage.

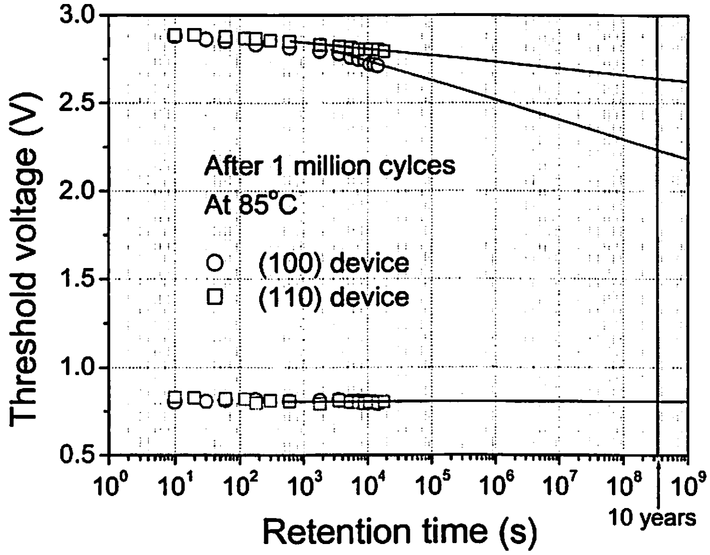


Fig. 5.11. Good retention time is seen on devices after 1 million P/E cycles at 85°C. (V_t window >1.4V)

From the V_t values after 10 years retention, i.e. V_{tL}=0.9V and V_{tH}=2.3V, the gate voltage for a reading operation is set at 1.6V. Fig. 5.12 plots the drain current from a selected cell, in both programmed (V_g=1.6V, V_t=2.15V) and erased (V_g=1.6V, V_t=0.88V) states, together with the current leakage from an unselected cell (V_g=0V, V_t=0.88V) during a reading operation. The current ratio between programmed and erased cells is as high as six orders of magnitude, making the reading of cell information relatively easy. Meanwhile, the leakage of the unselected cell remains very low for drain voltages up to 3.0V. The (110) device has a lower drive current due to its lower carrier mobility, but the current ratio also approaches that of (100) devices. These high current ratios benefit from the excellent subthreshold swing of the DG device and the large V_t window achieved. At high drain voltages (V_{ds}>3V), impact ionization causes abrupt turn-on of the devices. For a good safety margin and low power consumption, the reading voltages are set at V_g=1.6V and V_{ds}=1.2V.

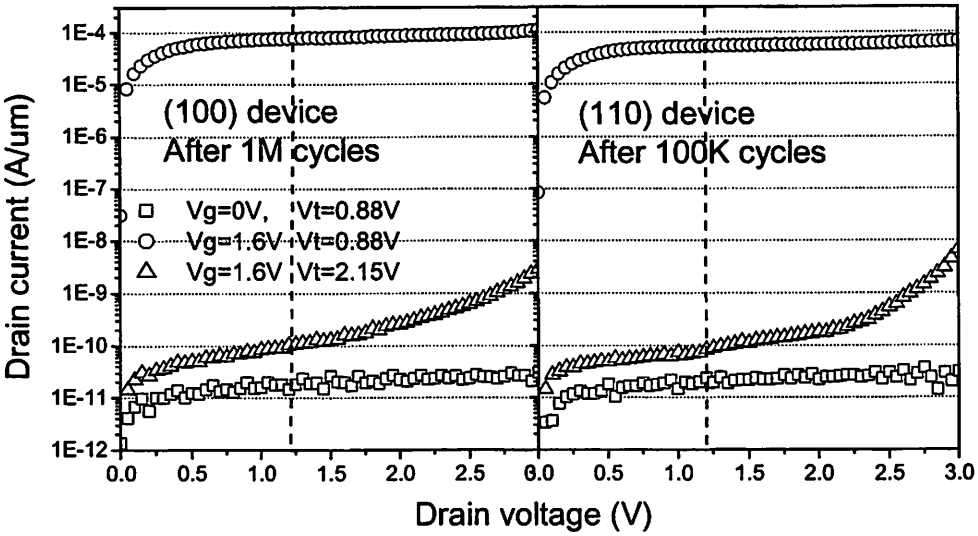


Fig. 5.12. Drain currents of the programmed cell, erased cell and unselected cell.

As opposed to DRAM, the reading operation of a NVM cell has to be non-destructive. Fig. 5.13 demonstrates that the stored information can be retained for 10 years of continuous reading disturbance at 85°C. A V_t window of more than 1.3V is maintained after 10 years of reading for (100) devices. Measured results with V_{ds}=2V are also plotted to demonstrate that impact ionization will not cause the cell to malfunction even at a higher V_{ds}. Again, the (110) device shows better resistance to reading disturbance because of its thicker tunnel oxide. Since the reduction in the V_t window includes both reading disturbance and tunneling leakage during the measurement, the actual influence of the read operation is smaller than what is shown in Fig. 5.13.

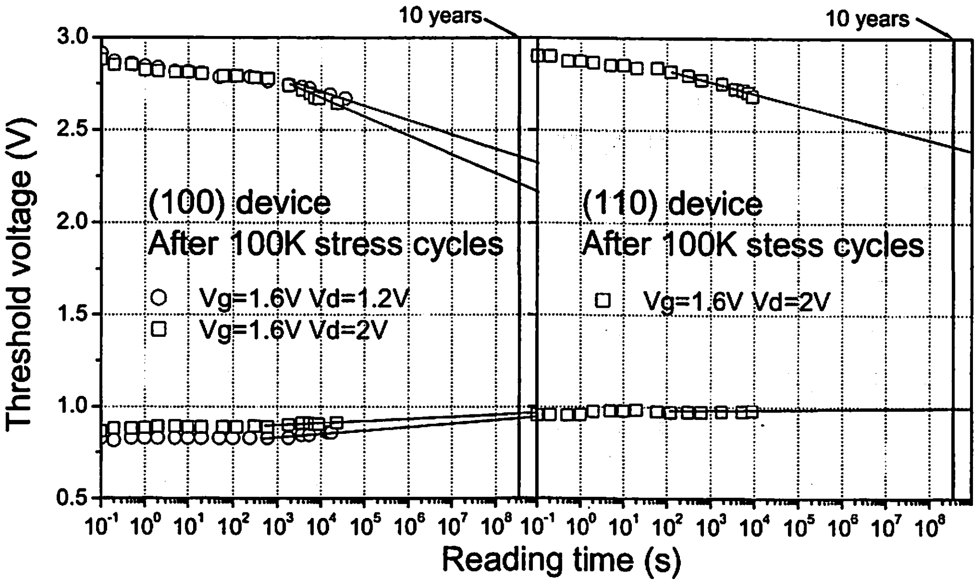


Fig. 5.13. Read disturbance characteristics of devices after 10^5 P/E cycles at 85°C.

5.3.4 Scalability of FinFET SONOS memory

Although devices with only 350nm gate length are demonstrated in the report, our FinFET SONOS memory can be scaled down to L_g=40nm. Fig 5.14a shows the ISE simulation result of a FinFET with a fin width of 20nm. The same V_t window after 10 years retention is assumed, i.e. 1.4V between the programmed and erased states. The scaling of the gate length degrades the subthreshold swing and reduces the current ratio between two states. With our large V_t window, the current ratio is still high enough for state distinction even at L_g=40nm, if the gate bias is set below 1.5V. However, the leakage from unselected cells (V_g=0V, I_d=2nA/μm) may become prohibitive in a large-scale integrated flash chip. When the device is scaled down to L_g=30nm, the high subthreshold swing (380mV/dec) makes the distinction of states extremely difficult. On the other hand, a thinner fin provides better scalability for the FinFET SONOS memory; a 30nm flash cell is possible with the fin scaled to 10nm (Fig. 5.14b).

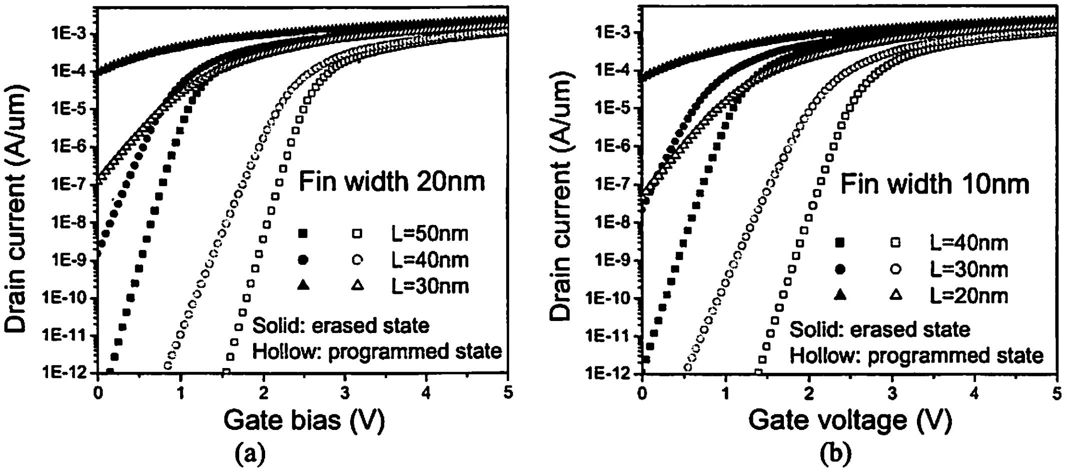


Fig. 5.14. (a) The SONOS FinFET memeory can be scaled to L_g=40nm with a 20nm fin. (b) With a 10nm fin, it can be scaled to L_g=30nm.

5.4 Conclusion

FinFET SONOS flash memory devices on SOI wafers have been demonstrated for the first time. The devices show program/erase speeds comparable to bulk SONOS memory. There is no apparent V_t window degradation up to 1 million P/E cycles, although the generation of interface traps is observed. The stressed memory devices show large V_t windows after 10 years retention at 85°C with or without reading disturbance, and multi-bit storage is possible with such big windows. The ratio of the reading currents between programmed and erased states exceeds 10^6, which enables relatively easy detection of the stored information. The low current leakage of unselected cells makes the devices suitable for low power applications. The FinFET SONOS structure can be successfully scaled to a gate length of 40nm, or even 30nm with a further reduction of the fin width to 10nm. Devices fabricated on (100) and (110) silicon surfaces are compared. Because of its thicker tunnel oxide, the (110) channel memory device has slower P/E speeds but better retention than does the (100) channel memory device. Meanwhile, (100) devices have higher carrier mobility and resistance to stress, so devices fabricated with (100) surfaces are preferred.

5.5 References

[1] K-H. Lee and Y-C. King, "New single-poly EEPROM with cell size down to 8F^2 for high density embedded nonvolatile memory applications," *Symposium on VLSI Technology, pp. 93- 94, June 2003*

[2] D-C. Kim, W-C. Shin, J-D. Lee, J-H. Shin, J-H. Lee, S-H. Hur, I-G. Baik, Y-C. Shin, C-H. Lee, J-S. Yoon, H-G. Lee, K-S. Jo, S-W. Choi, B-K. You, J-H. Choi, D. Park and K. Kim, "A 2Gb NAND flash memory with 0.044 um^2 cell size using 90 nm flash technology," *IEEE. Proceedings of International Electron Devices Meeting, pp. 919-922, Dec. 2002*

[3] Y-K. Lee, S-K. Sung, J-S. Sim, C-J. Lee, T-H. Kim, S-H. Lee, J-D. Lee, B-G. Park, D-H. Lee and Y-W. Kim, "Multi-level vertical channel SONOS nonvolatile memory on SOI," *Symposium on VLSI Technology, pp. 208 -209, June 2002*

[4] J-D. Choi, J-H. Lee, W-H. Lee, K-S. Shin, Y-S. Yim, J-D. Lee, Y-C. Shin, S-N. Chang, K-C. Park, J-W. Park and C-G. Hwang, "A 0.15 μm NAND flash technology with 0.11 μm^2 cell size for 1 Gbit flash memory," *IEEE. Proceedings of*

International Electron Devices Meeting, pp. 767 –770, Dec. 2000

[5] J.H. Kim, I.W. Cho, G.J. Bae, S.S. Kim, K.C. Kim, S.H. Kim, K.W. Kob, N.I. Lee, H-K. Kang, K-P. Suh, S.T. Kang, M.K. Seo, S.H. Lee, M.C. Kim and L.S. Park, "Highly manufacturable SONOS non-volatile memory for the embedded SoC solution," *Symposium on VLSI Technology, pp. 31 -32, June 2003*

[6] R. Bez, E. Camerlenghi, A. Modelli and A. Visconti, "Introduction to Flash memory," *Proceedings of the IEEE, Vol. 91, No. 4, pp. 489-502, April 2000*

[7] L.D. Engh, A.V. Kordesch and C-M. Liu, "A self adaptive programming method with 5 mV accuracy for multi-level storage in FLASH," *Proceedings of the IEEE Custom Integrated Circuits Conference, pp. 115 –118, May 2002*

[8] M. Borgatti, A. Rocchi, M. Bisio, M. Besana, L. Navoni and P.L. Rolandi, "A 64 min single-chip voice recorder/player using embedded 4 bit/cell flash memory," *IEEE. Proceedings of the Custom Integrated Circuits Conference, pp. 219 -222, May 2000*

[9] A.J. Walker, S. Nallamothu, E-H. Chen, M. Mahajani, S.B. Hemer, M. Clark, J.M. Cleeves, S.V. Dunton, V.L. Eckert, J. Gu, S. Hu, J. Knall, M. Konevecki, C. Petti, S. Radigan, U. Raghuram, J. Vienna and A. Michael, "3D TFT-SONOS Memory Cell for Ultra-High Density File Storage Applications," *Symposium on VLSI Technology, pp. 29-30, June 2003*

[10] E. De Fresart, R. De Souza, J. Morrison, P. Parris, J. Heddleson, V. Venkatesan, W. Paulson, D. Collins, G. Nivison, B. Baumert, W. Cowden and D. Blomberg, "Integration of multi-voltage analog and power devices in a 0.25um CMOS + flash memory process," *Symposium on Power Semiconductor Devices and ICs, pp. 305–308, June 2002*

[11] K. Takeuchi, S. Satoh, K. Imamiya, Y. Sugiura, H. Nakamura, T. Himeno, T. Ikehashi, K. Kanda, K. Hosono and K. Sakui, "A source-line programming scheme for low voltage operation NAND flash memories," *Symposium on VLSI Circuits, pp. 37 –38, June 1999*

[12] M. Ichige, Y. Takeuchi, K. Sugimae, A. Sato, M. Matsui, T. Kamigaichi, H. Kutsukake, Y. Ishibashi, M. Saito, S. Mori, H. Meguro, S. Miyazaki, T. Miwa, S. Takahashi, T. Iguchi, N. Kawai, S. Tamon, N. Arai and H. Kamata, "A novel self aligned shallow trench isolation cell for 90nm 4Gbit NAND flash EEPROMs," *Symposium on VLSI Technology, pp. 89 –90, June 2003*

[13] T. Tanzawa, Y. Takano, K. Watanabe and S. Atsumi, "High-voltage transistor scaling circuit techniques for high-density negative-gate channel-erasing NOR flash memories," *IEEE. Journal of Solid-State Circuits, Vol. 37, No. 10, pp. 1318 –1325, Oct. 2002*

[14] T. Tanzawa, T. Tanaka, K. Takeuchi and H. Nakamura, "Circuit techniques for a 1.8V-only NAND flash memory," *IEEE, Journal of Solid-State Circuits, Vol. 37, No. 1, pp. 84 –89, Jan. 2002*

[15] B. Yu, H. Wang, Q. Xiang, J.X. An, J. Jeon and M-R Lin, "Scaling towards 35 nm gate length CMOS," *Symposium on VLSI Technology, pp. 9-10June 2001*

[16] V-Y. Aaron and J-P. Leburton, "Flash memory: towards single-electronics," *IEEE. Potentials, Vol. 21, No. 4, pp. 35 –41, Oct. 2002*

[17] J. Bu and M.H. White, "Retention reliability enhanced SONOS NVSM with scaled programming voltage." *IEEE. Proceedings of Aerospace Conference, Vol. 5, pp. 5.2383-5.2390, 2002*

[18] S-I. Minami and Y. Kamigaki, "A novel MONOS nonvolatile memory device ensuring 10-year data retention after 10^7 erase/write cycles," *IEEE. Transactions on Electron Devices, Vol. 40, No. 11, pp. 2011 –2017, Nov. 1993*

[19] M.L. French, C-Y. Chen, H. Sathianathan and M.H. White, "Design and scaling of a SONOS multi-dielectric device for nonvolatile memory applications," *IEEE. Transactions on Components, Packaging, and Manufacturing Technology, Part A, Vol. 17, No. 3, pp. 390 –397, Sept. 1994*

[20] 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," *IEEE. Proceedings of International Electron Devices Meeting, pp. 251-254. 2002*

[21] N. Lindert, L. Chang, Y-K. Choi, E.H. Anderson, W-C. Lee, T-J. King, J. Bokor and C. Hu, "Sub-60-nm quasi-planar FinFETs fabricated using a simplified process," *IEEE. Electron Device Letters, Vol. 22, No. 10, pp. 487–489, Oct. 2001*

[22] I. Fujiwara, H. Aozasa, A. Nakamura, Y. Hayashi and T. Kobayashi, "MONOS memory cell scalable to 0.1um and beyond, IEEE," *Non-Volatile Semicondudctor Memory Workshop, pp. 117-118, 2000*

[23] H.S. Momose, T. Ohguro, K. Kojima, S. Nakamura and Y. Toyoshima, "1.5-nm gate oxide CMOS on [110] surface-oriented Si substrate," *IEEE Transactions on Electron Devices, Vol. 50, No. 4, pp. 1001 –1008, April 2003*

[24] M.H. White, Y. Yang, A. Purwar and M.L. French, "A low voltage SONOS nonvolatile semiconductor memory technology," *IEEE Transactions on Components, Packaging, and Manufacturing Technology, Part A, Vol. 20, No. 2, pp. 190-195, June*

1997

[25] N. Yang, W.K. Henson and J.J. Wortman, “Analysis of tunneling currents and reliability of NMOSFETs with sub-2nm gate oxides,” *IEEE. Proceedings of International Electron Devices Meeting, pp. 453 -456. Dec. 1999*

[26] M-S. Liang, J.Y. Choi, P-K. Ko and C. Hu. “Inversion-layer capacitance and mobility of very thin gate-oxide MOSFETs.” *IEEE. Transactions on Electron Devices, Vol. 33, No. 3, pp. 409-413, March 1986*

[27] S. Takagi, A. Toriumi, M. Iwase and H. Tango, “On the universality of inversion layer mobility in Si MOSFET's: Part II-effects of surface orientation,” *IEEE. Transactions on Electron Devices, Vol. 41 No. 12, pp. 2363 –2368, Dec. 1994*

[28] Y-K. Choi, L. Chang, P. Ranade, J-S. Lee, D. Ha, S. Balasubramanian, A. Agarwal, M. Ameen, T-J. King and J. Bokor, “FinFET process refinements for improved mobility and gate work function engineering,” *IEEE. Proceedings of International Electron Devices Meeting. pp. 259-262, Dec. 2002*

Appendix 5A: Process flow for FinFET SONOS memory

Step	Process	Process specification	Equipment	Comments
0	4" SOI wafer, 950Å Si on 4000Å buried oxide			
0.1	Scribe	Label the wafers		
1	EBeam alignment marks			
1.1	Cleaning	Piranha (H_2O_2:H_2SO_4=1:5) 120°C, 10min, 25:1 BHF 30s	Sink6	Dewet, up to $16k\Omega$
1.2	Pad Oxide	SDRYOXA 950°C, 80min, 20min N_2 anneal	Tylan2	SiO_2=35nm Si_{remain}=81nm
1.3	SiGe dep	SiGe.019: Nucleation: T=550°C, P=300mT, SiH_4=200sccm, t=1min. Deposition: T=500°C, P=300mT, SiH_4=186sccm, GeH_4Lo=33sccm, GeH_4Hi=0, t=90min	Tystar19	Ge concentration ~40% 790-850nm SiGe
1.4	Cap oxide	11SULTOA 450°C, 300mT, SiH_4=25sccm, O_2=75sccm, 8min	Tystar11	140nm
1.5	Anneal	THINOX, 950°C, 30min	Tylan6	No thickness change
1.6	Alignment mark litho	Resist coating: coat=program 1/bake=program 1 Exposure: focus=250, t=0.9s Development: bake=program 1/develop=program 1 Descum: O_2=51sccm, P=50W, t=1min Hard bake: 120°C, 1hr	Svgcoat1 GCAWS2 Svgdev Technics-c Ovrn	PR=1.2um PEB: 90C, 1min DEV: OPD4226, 1min

1.7	Mark etch	B: p=13mTorr, CF_4=100, P_{top}=200, P_{bot}=40, t=90s
		M: P=15mTorr, Cl_2=50, HBr=150, P_{top}=300, P_{bot}=150, t=55s
		O: P=35mT, HBr=200, O_2=5.0, P_{top}=250, P_{bot}=120, t=25s
			Lam5	ER=20Å/s, SiO_2/Si~1
				ER=100Å/s, SiGe/SiO_2~13
				ER=50Å/s, Si/SiO_2~100.

1.8	Resist strip	O_2 51sccm, 230W, 5min	Technics-c	
1.9	Cleaning	Piranha 120°C, 10min	Sink8	Clean in dirty sink first
2	Mesa formation			
2.1	Cleaning	Piranha 120°C, 10min, 25:1 BHF, 180s	Sink6	Dewet, up to $16k\Omega$
2.2	Oxide mask	SDRYOXA 1000°C, 65min, 20min anneal	Tylan2	SiO_2=65nm Si_{remain}=50nm
2.3	Fin litho	HSQ bi-layer 200nm, dose=1200 (too low, should be ~2000)	Nanowriter	At LBNL
2.4	S/D pad litho	Resist coating: coat=program 2/bake=program 1 Exposure: focus=250, t=1s Development: bake=program 1/develop=program 2 Descum: O_2=51sccm, P=50W, t=1min Hard bake: 120°C, 1hr	Svgcoat1 GCAWS2 Svgdev Technics-c Ovrn	PR=1.2um PEB: 90C, 1min DEV: OPD4226, 1min
2.5	Mesa etch	B/M/O=45s/15s/20s. See 1.7	Lam5	
2.6	Resist strip	100:1 HF 5s O_2 51sccm, 230W, 5min 100:1 HF 20s	Sink7 Technics-c Sink7	Remove the polymer
2.7	Cleaning	Piranha 120°C, 10min	Sink8	
2.8	Cleaning	Piranha 120°C, 10min, 25:1 BHF, 10s	Sink6	Oxide mask recedes
2.9	Sac oxide	THIN_VAR, 830°C, N_2=9, O_2=1, 24min, 900°C 20min in N_2	Tylan6	Oxide=3nm
3	Gate stack definition			
3.1	Cleaning	Piranha 120°C, 10min, 25:1 BHF, 20s	Sink6	Oxide mask recedes
3.2	Tunnel oxide	THIN_VAR, 810°C, N_2=9, O_2=1, 24min, 900°C 20min in N_2	Tylan6	Oxide=3nm

3.3	Inter nitride	9VNITA, 750°C, 300mT, NH_3=24sccm, DCS=25sccm, N_2=100sccm, 5.5min	Tystar9	Nitride=6.3nm
3.4	Top HTO	9VHTOA, 800°C, 300mT, DCS=10sccm, N_2O=100sccm, 13min	Tystar9	HTO=5.2nm
3.5	N+ Gate dep	10SDPLYA 615°C, 375mT, SiH_4=100sccm, PH_3=4sccm, 65min	Tystar10	Poly=178nm
3.6	Gate litho	HSQ bi-layer 200nm, dose=1200 (too low, should be ~2000)	Nanowriter	At LBNL
3.7	Gate pad litho	See step 2.4		
3.8	Gate etch	B/M/O=15s/10s/30s. see 1.7	Lam5	20s overetch for 20A ox
3.9	Nitride etch	P=13mTorr, CF_4=100, P_{top}=200, P_{bot}=40, t=15s	Lam5	ER=20Å/s, SiO_2/Si~1
3.10	Resist strip	100:1 HF 5s O_2 51sccm, 230W, 5min 100:1 HF 5s	Sink7 Technics-c Sink7	Remove the polymer
3.11	Cleaning	Piranha 120°C, 10min	Sink8	Resistance to 16kΩ
4	Source/drain formation			
4.1	Cleaning	Piranha 120°C, 10min, 25:1 BHF, 10s	Sink6	
4.2	Spacer HTO	9VHTOA 800°C, 300mT, N_2O=75sccm, DCS=25sccm, 20min	Tystar9	11nm
4.3	Si_3N_4 dep	9VNITA, 800°C, 300mTorr, NH_3=15sccm, DCS=5sccm, N_2=80sccm, 10min	Tystar9	Nitride=10nm
4.4	Nitride etch	NITSTD1: ME: P=375mTorr, He=50sccm, SF_6=175sccm, RF=150W, t=9s. Overetch: same as ME, t=15%	Lam1	ER=13Å/s
4.5	Imp mask	Resist coating: coat=program 1/bake=program 1 Exposure: t=5s, use half of a wafer as mask Development: bake=program 1/develop=program 1 Descum: O_2=51sccm, P=50W, t=1min Hard bake: 120°C, 1hr	Svgcoat1 Ksaligner Svgdev Technics-c Ovrn	Cover bottom half wafer
4.6	S/D implant	B+, 5e15, 15keV.	Core sys.	Foundry. Rp=30nm
4.7	Resist strip	O_2 51sccm, 230W, 5min	Technics-c	
4.8	Imp mask	See step 4.5		Cover top half wafer
4.9	S/D implant	P+, 5e15, 40keV.	Core sys.	Foundry. Rp=30nm
4.10	Resist strip	O_2 51sccm, 230W, 5min	Technics-c	
4.11	Cleaning	Piranha 120°C, 10min	Sink8	
5	Back end process			
5.1	Cleaning	Piranha 120°C, 10min, 25:1 BHF, 10s	Sink6	
5.2	RTA	N_2, 920°C, 20s	Heatpulse3	
5.3	LTO dep.	VDOLTOC, 15min	Tystar11	260nm
5.4	FGA	N_2/H_2, 400°C, 5min	Heatpulse1	
5.5	Contact litho	Same as 1.6		
5.6	Contact etch	5003 breakthrough, 120s 25:1 BHF 80s	Lam5 Sink6	Remain oxide ~20nm 100nm LTO etched
5.7	Resist strip	O_2 51sccm, 230W, 5min	Technics-c	
6	Calibration			

Chapter 6

Conclusion

6.1 Summary

The exponential growth of the semiconductor market in the past four decades stems from the dramatic miniaturization of silicon-based microelectronic devices. To maintain the low current leakage of a transistor, the vertical dimensions must be scaled together with its channel length [1,2]. However, as devices enter the nano-scaled region, conventional scaling approaches are facing tremendous challenges in both manufacturing process [3] and fundamental physics [4]. New structures, such as ultra-thin-body (UTB) [5] and double-gate (DG) [6], are proposed to extend Moore's law into the future.

Fully depleted (FD) MOSFETs, including UTB and DG devices ensure good gate control of the device by eliminating the current conduction path away from the gate [7]. The effectiveness of the structures can be assessed by the analytic model developed in Chapter 2. Derived from the physical dimensions of a transistor, the scale length $l=\sqrt{\frac{\varepsilon_{si}}{\varepsilon_I}T_I d+\frac{T_I^2+d^2}{2}}$ can be used to guide the design of sub-50nm devices. The 2D effects in both the body and the high κ gate dielectric are included. The influences of body doping and pocket implants on short channel effects (SCE) are also modeled. In addition, the model predicts that a correct threshold voltage can be achieved only by gate workfunction engineering in sub-30nm transistors [8].

One significant fabrication challenge of a UTBFET is the formation of the ultra-thin channel film with good uniformity and crystalline quality. Lateral solid-phase-epitaxy is proposed because a deposited film can be very uniform and its thickness precisely controlled. Several techniques are introduced to improve the quality of the final SPE film, such as implanting silicon atoms and shifting the twin boundary out of the channel. Our experiment shows that the crystalline quality is good for only a short SPE range (≤60nm). Defects are generated and degrade the device performance when the SPE propagates more than 60nm. With its relatively easy integration with bulk CMOS, SPEFET is suitable for sub-50nm device generations [9].

With an N+/P+ poly gate on a FD body, the transistor will have unsuitable threshold voltages (V_t), which prevent its wide use in circuit applications [10]. Nickel silicide is proposed as the CMOS gate material because of its high conductance, elimination of gate depletion, and continuously adjustable workfunction [11]. The workfunction of NiSi can be tuned by dopants implanted into the silicon film before silicidation, and the workfunction range covers the requirement of both NMOS and PMOS. The quality of a thin oxide under a NiSi gate is also investigated, and there is no degradation observed in MOSFET performance compared with devices having a polysilicon gate, which indicates the excellent compatibility of the NiSi gate with the current CMOS process. Therefore, nickel silicide is highly advantageous as a single gate material for a CMOS circuit, even with multiple V_t's [12].

For better scalability, the FD structure is applied to flash memory, which is also facing significant scaling challenges due to its thick gate stack. By providing an alternative way of scaling, i.e. thinning of the body, FinFET SONOS can be successfully scaled to sub-40nm, while excellent device performance can be still achieved. With the absence of a neutral body, the FinFET SONOS memory behaves similarly to a bulk

SONOS cell, except for the improved SCE and scalability. Good program/erase speeds, endurance and retention are demonstrated in FinFET SONOS memory devices. Their large V_t windows enable multi-bit storage, which further increases the storage density. Devices fabricated on (100) and (110) sidewall surfaces are compared, and it is observed that (100) devices show more resistance to electrical stress. The FinFET SONOS device is a promising candidate for large capacity embedded flash memory [13].

6.2 Suggestion for future work

6.2.1 Integration of NiSi gate with FinFET and UTBFET

In Chapter 4, nickel silicide gates are proposed for FD transistors to achieve appropriate threshold voltages, and their correct workfunction and good interface have been demonstrated on bulk MOSCAPs and MOSFETs. However, the integration of this new gate material with UTB and DG devices involves many process challenges. A chemical mechanical polish (CMP) step is required to expose the poly gate while keeping the source/drain protected during silicidation. A manageable CMP process window requires a sufficiently thick poly gate (>400nm). Moreover, a thick oxide hard mask is needed to protect the gate during source/drain implantation because dopants would change the workfunction of the final silicide film. Since electron-beam lithography typically uses a resist only 200nm thick, this thick gate stack in a FinFET is difficult to etch. A CMP step before the hard mask deposition to planarize the surface is highly recommended to relax the over-etch margin. Selective Ge is commonly grown as the raised S/D for a UTBFET. Although Ge will not be removed in the CMP, it will react with Ni and dissolve in piranha if it is exposed in the CMP. Therefore, an etch chemical that can selectively remove unreacted Ni while keeping NiSi and NiGe untouched is highly appreciated for large process windows. If a spacer technology that prevents bridging between the gate and S/D can be developed, the process can be dramatically simplified.

6.2.2 Tunneling FET for sub 60mV/dec subthreshold swing

In a MOSFET, subthreshold swing is the key parameter that determines the minimum V_t and the efficiency of the on/off transition. FD thin body is proposed to improve SCE by reducing the electrostatic coupling from the source and drain. But as long as the drift-diffusion current dominates the current conduction, the optimum limit of the swing is 60mV/dec. Many novel ideas have been proposed to break this limit, and two major approaches are positive feedback, such as impact ionization, and quantum tunneling. While a device taking advantage of impact ionization can achieve S=10mV/dec, it requires a high drain voltage and its application is limited [14]. On the other hand, the tunneling phenomenon sets no fundamental limit on the swing.

Some preliminary results on tunneling FET (TFET) have been obtained. The device structure is shown in Fig. 6.1, and a N-type TFET will be assumed in this discussion. The current conduction mechanism is band-to-band (BTB) tunneling, the same as the origin of gate-induced-drain-leakage (GIDL) [15]. The lightly doped body between the source and drain reduces reversed PN junction leakage. When the vertical band bending in the P+ source region, which is controlled by the gate voltage, is below E_g, only the carrier generation in the depletion region contributes to device leakage. When the band bending exceeds E_g, BTB tunneling occurs. The drive current density can be derived from a simple tunneling model [16] as $I = \mathrm{KE}_s \exp(-B/E_s)$, while

$K \approx \frac{q^2}{4\pi^2\hbar^2}\sqrt{2mqE_g}$ and $B = 4\sqrt{2qmE_g^3}/3\hbar$. The vertical electrical field at the surface E_s can be calculated from the full depletion approximation: $E_s = \sqrt{\left(\frac{qN}{C_{ox}}\right)^2 + \frac{2qN}{\varepsilon_{si}}(V_g - V_{fb})} - \frac{qN}{C_{ox}}$. A high C_{ox} (or thin T_{ox}) means high sensitivity of the E_s to the gate bias.

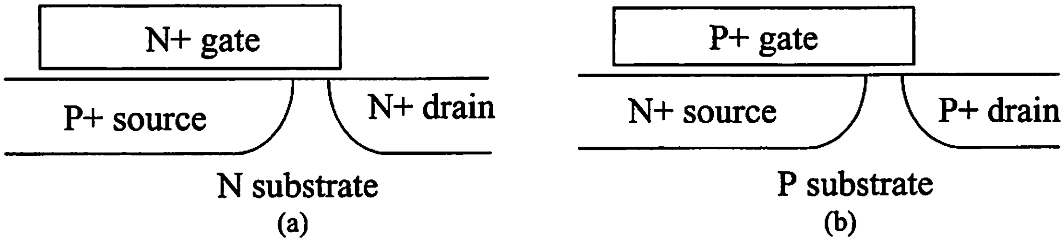
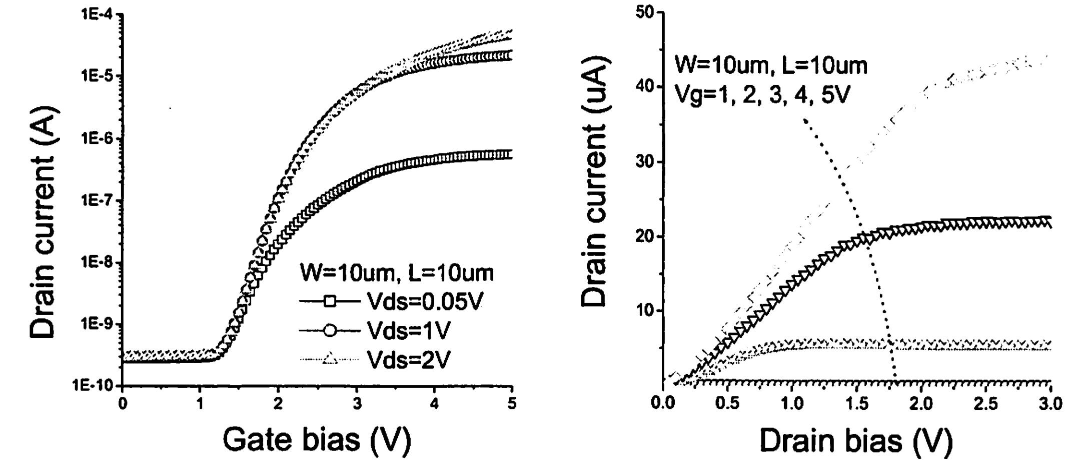


Fig. 6.1. Structures of (a) N-type and (b) P-type tunneling FET.

Fig. 6.2 plots the device performance of such a structure. It shows the behavior of a MOSFET, as expected, but with poor performance. First, the device is slow because the intrinsic delay of this device is estimated to be about 25ns. Since the current is proportional to the area, the intrinsic delay (CV/I) does not scale with device size. A low band-gap material, such as Ge, can dramatically increase the BTB tunneling current, making this structure more attractive. However, with the use of a low E_g material, the junction leakage between the source/drain and body could be a concern. Second, a high V_t is required to bend the band in the heavily doped source region. This problem can be solved by a thinner gate oxide or a lower E_g substrate. Third, the swing (~270mV/dec) does not meet our goal because of the 2D coupling from the drain. With a positive V_{ds}, the total band bending always exceeds E_g. When the gate bias is low, two-step tunneling causes the gradual on/off transition: vertically from the source to the surface, then laterally to the drain. Careful sizing of the dimensions and thorough simulation may reveal the optimal design and break the kT/q limit.

After all, TFETs with a properly designed geometry and a low E_g substrate material, such as Ge, could promise to achieve very low subthreshold swings.



(a)

(b)

Fig. 6.2. Primitive results obtained from a tunneling FET. (a) I_d~V_g characteristics. (b) I_d~V_d characteristics. (T_{ox}=3nm, N_{source}=1e19cm^{-2})

6.3 References

[1] M.T. Bohr, "Nanotechnology goals and challenges for electronic applications," *IEEE, Transaction on Nanotechnology, Vol. 1, No. 1, pp. 56-62, March 2002*

[2] H. Iwai, H.S. Momose and Y. Katsumata, "Si-MOSFET scaling down to deep-sub-0.1-micron range and future of silicon LSI," *Proceedings of International Symposium on VLSI Technology, Systems, and Applications, pp. 262 –267, May 1995*

[3] H-SP. Wong, D.J. Frank, P.M. Solomon, C.H.J. Wann and J.J.Welser, "Nanoscale CMOS." *Proceedings of the IEEE, Vol.87, No.4, pp. 537-570, April 1999*

[4] P.A. Packan: "Device physics: Pushing the limits". *Science. 285(5436), pp. 2079-2081, Sep. 1999*

[5] B. Yu, Y-J. Tung, S. Tang, E. Hui, T-J. King and C. Hu, "Ultra-thin-body silicon on insulator MOSFETs for terabit-scale integration," *Proceedings of the International Semiconductor Device Research Society, pp. 623-628, 1997*

[6] D. Hisamoto, W-C. Lee, J. Kedzierski, H. Takeuchi, K. Asano, C. Kuo, E. Anderson, T-J. King, J. Bokor and C. Hu, "FinFET-a self-aligned double-gate MOSFET scalable to 20 nm," *IEEE. Transactions on Electron Devices, Vol. 47, No. 12, pp. 2320-2325 Dec. 2000*

[7] J. Kedzierski, P. Xuan, V. Subramanian, E. Anderson, J. Bokor, T-J. King and C. Hu, "A 20nm gate-length ultra-thin body P-MOSFET with silicide source/drain". *VLSI Semiconductor Nanoelectronics Workshop, Vol. 28, No. 5/6, pp. 445-452, 2000*

[8] I. De, D. Johri, A. Srivastava and C.M. Osburn, "Impact of gate workfunction on device performance at the 50nm technology node," *Solid-State Electronics, Vol. 44, pp. 1077-1080, 2000*

[9] P. Xuan, J. Kedzierski, V. Subramanian, J. Bokor, T-J. King and C. Hu. "60nm Planarized Ultra-thin Body Solid Phase Epitaxy MOSFETs," *58th Device Research Conference, pp. 67-68, 2000*

[10] L. Chang, S. Tang, T-J. King, J. Bokor and C. Hu, "Gate length scaling and threshold voltage control of double-gate MOSFETs," *Proceedings of the International Electron Device Meeting, pp. 719-722, Dec 10-13 2000*

[11] J. Kedzierski, E. Nowak, T. Kanarsky, Y. Zhang, D. Boyd, R. Carruthers, C. Cabral, R. Amos, C. Lavoie, R. Roy, J. Newbury, E. Sullivan, J. Benedict, P. Saunders, K. Wong, D. Canaperi, M. Krishnan, K-L. Lee, B.A. Rainey, D. Fried, P. Cottrell, H-

SP. Wong, M. Ieong and W. Haensch, "Metal-gate FinFET and fully-depleted SOI devices using total gate silicidation." *IEEE. Proceedings of International Electron Devices Meeting. pp. 247-250. Dec. 2002*

[12] P. Xuan and J. Bokor, "Investigation of NiSi and TiSi as CMOS Gate Materials," *IEEE. Electron Device Letter, Vol. 24, No. 10, pp. 634-636, Oct. 2003*

[13] P. Xuan, M. She, J. Bokor and T-J King, "FinFET SONOS Flash Memory for Embedded Applications," *Proceedings of the International Electron Device Meeting, Dec. 2003, in press*

[14] K. Gopalakrishnan, P.B. Griffin and J.D. Plummer, "I-MOS: a novel semiconductor device with a subthreshold slope lower than kT/q," *IEEE. Proceedings of International Electron Devices Meeting, pp. 289 –292, Dec. 2002*

[15] S.A. Parke, J.E. Moon, H.C. Wann, P.K. Ko and C. Hu, "Design for suppression of gate-induced drain leakage in LDD MOSFETs using a quasi-two-dimensional analytical model," *IEEE. Transactions on Electron Devices, Vol. 39 No. 7, pp. 1694 -1703, July 1992*

[16] E. Takeda, H. Matsuoka, Y. Igura and S. Asai, "A band to band tunneling MOS device (B^2T-MOSFET)-a kind of 'Si quantum device'," *IEEE. Proceedings of International Electron Devices Meeting, pp. 402 –405, Dec. 1988*