TECHNOLOGY AND SCALING OF ULTRATHIN BODY DOUBLE-GATE FETS

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF ELECTRICAL ENGINEERING
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Rohit S. Shenoy
December 2004

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

_____

Krishna C. Saraswat, Principal Advisor

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

_____

Yoshio Nishi

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

_____

James P. McVittie

Approved for the University Committee on Graduate Studies.

This page is intentionally left blank

Abstract

As silicon CMOS technology advances into the sub-50 nm regime, fundamental and manufacturing limits impede the traditional scaling of transistors. Innovations in materials and device structures will be needed for continued transistor miniaturization with commensurate performance improvements. The ultrathin body double-gate (DG) FET is a leading candidate for replacing bulk CMOS transistors in future technology generations. Multiple gates and the ultrathin body enable better electrostatic gate control over the channel, allowing DG FETs to be scaled to smaller dimensions than their conventional bulk counterparts. This research is focused on some of the major issues in the design and fabrication of high performance scaled DG FETs.

The first part of this research deals with the extrinsic factors that limit the overall performance of ultrathin body DG FETs. The impact of parasitic capacitance and resistance is studied and quantified by device simulation. In particular, the importance of optimizing the lateral doping profile in the thin source/drain extension regions to minimize series resistance is discussed in detail.

Next, a novel process is proposed to fabricate a planar DG FET with the following attributes: 1) deposition-controlled uniform ultrathin body, 2) fully self-aligned gates for low parasitic capacitance, and 3) flared-out low resistance source/drain regions. This process thus yields the ideal DG FET that satisfies both intrinsic as well as extrinsic requirements for scalability. The main idea in this process is to use sacrificial SiGe layers as placeholders for the top and bottom gates-to-be. Experimental work has been performed for the process development in order to verify the feasibility of the key steps – 1) epitaxial CVD of high quality trilayer stacks of SiGe/Si/SiGe, 2) enhanced oxidation rate of SiGe with respect to Si, 3) isotropic etching of SiGe with high selectivity to Si, and 4) low resistance in-situ doped source and drain formation with a low thermal budget. As proof of concept, functional double-gate transistors are demonstrated with very good turn-off characteristics (nearly ideal subthreshold swing and very low DIBL).

A new methodology has been developed (in collaboration with co-workers at Stanford University) to optimize the supply voltage, threshold voltage, and effective oxide thickness for minimizing the total power dissipation of digital logic transistors given a target delay and application (switching activity ratio). By employing this concept in conjunction with extensive device simulation and post-processing, a framework has been developed for the comparison of future transistor structures using optimized power-delay curves for benchmarking.

Acknowledgements

I wish to acknowledge the support of many people who have contributed to my Ph.D. research and have made my stay at Stanford a highly enjoyable and useful experience.

First of all, I want to express my sincere gratitude to my advisor, Prof. Krishna Saraswat for his support, guidance, and patience throughout the course of my research. I especially appreciate the latitude and time that he gave me to choose a research topic and explore a large number of areas in novel device physics and technology. It has been a very rewarding experience to work with him.

I would also like to thank my associate advisor Prof. Yoshio Nishi for his insightful suggestions, questions, and comments. I am very grateful to Dr. Jim McVittie for being a very good mentor and friend. He has always been ready to help me, whether it was with equipment issues in the lab, or to share some of his immense depth and breadth of knowledge and experience. I also thank Prof. Paul McIntyre for agreeing to serve as the chair on my Ph. D. oral examination committee.

I have been fortunate to interact on a technical basis with some of the other faculty and research staff at Stanford. I thank Prof. Jim Plummer for his advice during my initial years here. Dr. Baylor Triplett, Dr. Mike Deal, and Dr. Peter Griffin have always been forthcoming with suggestions and questions during group meetings and otherwise. I have learned a lot from them and thank them for that. I sincerely thank Dr. Ann Marshall for introducing me to the fascinating world of transmission electron microscopy (TEM).

I thank Irene Sweeney for her efficient administrative support, and Jason Conroy for making the systems administration of the group computers a painless task.

This Ph. D. research has benefited from the direct help and collaboration of many colleagues. I would like to acknowledge their contributions. Dr. Pranav Kalavade was instrumental in my initial training in the lab and in giving me some of the ideas that went on to form the core of my dissertation research. Some of the work on the hydrogen pre-

bake optimization was shared by Dr. Kailash Gopalakrishnan and Sameer Jain. That really helped in expediting the results and all three of us stood to gain. The work on power-delay optimization was jointly done with Dr. Pawan Kapur and Andy Chao. I wish to acknowledge the help of Dr. Hyoungsub Kim, Raghav Sreenivasan, Nevran Ozguven, and Yaocheng Liu in some of the materials characterization.

A large portion of my Ph. D. was spent in the Stanford Nanofabrication Facility (SNF) cleanroom. I cannot thank the staff enough for training me and keeping most of the equipment running smoothly. In particular, I am especially thankful to Margaret Prisbe, Robin King, Gladys Sarmiento, and Nancy Latta, for their initial training and patience while I learned to operate the cleanroom equipment for the first time; and to Maurice Stevens, Ted Berg, Ray Seymour, Paul Jerabek, Cesar Baxter, and Charley Williams for running and maintaining the tools that were most critical to my experiments. I am also very grateful to the other lab-members for their useful suggestions. In particular, I wish to thank Dr. Aaron Partridge and Dr. Eric Perozziello for their help, on a number of occasions, in rescuing my wafers from a dead machine.

One of the things that make Stanford a great place to be at is the students. I have been privileged to be in the company of some very smart and interesting people. It has been a wonderful experience interacting with my coworkers at CIS, particularly the research groups of Prof. Saraswat and Prof. Plummer. There are far too many people to list here, but I would be amiss if I did not especially acknowledge Dr. Amol Joshi, Dr. Pawan Kapur, Dr. Pranav Kalavade, Dr. Kailash Gopalakrishnan, Dr. Niranjan Talwalkar, Dr. Richard Chang, Dr. Dan Connelly, Rohan Kekatpure, and Sameer Jain. Discussions with them have invariably been very enjoyable and enlightening.

Finally, I would like to thank my family and friends, whose constant support and encouragement were instrumental, in an immeasurable way, in helping me complete this Ph. D.

Financial support for this work, in the form of funding from DARPA and the MARCO MSD Focus Center is gratefully acknowledged.

Contents

CHAPTER 1

INTRODUCTION 1

1.1 Motivation 1

1.2 Challenges in Scaling Conventional MOSFETs 4

1.3 Double-Gate (DG) FET 7

1.4 This Work and Dissertation Organization 9

References 11

CHAPTER 2

DOUBLE-GATE FET OPERATION – INTRINSIC FACTORS 13

2.1 Introduction 13

2.2 Evolution of Novel Device Structures 14

2.2.1 Partially Depleted SOI (PD SOI) 14

2.2.2 Fully Depleted SOI (FD SOI) 15

2.2.3 Double-Gate FET (DG FET) 16

2.2.4 Threshold Voltage Control in FD SOI/DG FET 17

2.2.5 Beyond DG FETs 18

2.3 DG FET Operation and Electrostatics 19

2.3.1 Basic Double-Gate Transistor Operation 20

2.3.2 DG FET Off-state – Gate Electrostatic Control 23

2.3.3 Importance of Body Thickness (t_{si}) 24

2.3.4 Ultrathin Body Scaling Limits 29

2.4 Summary 32

References 33

CHAPTER 3

DOUBLE-GATE FET DEVICE PERFORMANCE – EXTRINSIC FACTORS 37

3.1 Introduction 37

3.2 Parasitic Capacitance 38

3.2.1 Effect of Top-Bottom Gate Misalignment 40

3.2.2 Effect of Bottom Gate Sizing 41

3.2.3 Effect of Gate-Source/Drain Spacer Thickness 42

3.3 Parasitic Resistance 43

3.3.1 Device Design and Simulation Setup 44

3.3.2 Effect of Contact Resistance 46

3.3.3 Ion-Ioff Comparisons and Discussion 48

3.3.4 Optimization of Extension Underlap 53

3.4 Schottky Source/Drain FET 58

3.4.1 Schottky S/D FET Operation 59

3.4.2 Comparison with Doped S/D DG FET 62

3.5 Summary 63

References 65

CHAPTER 4

A NOVEL PROCESS FOR FULLY SELF–ALIGNED PLANAR DOUBLE-GATE FET 69

4.1 Introduction 69

4.2 Double-Gate FET Configurations 70

4.3 Planar DG FET – Prior Art 73

4.3.1 IBM Planar DG FET using Pattern-Constrained Epitaxy 73

4.3.2 MIT Super Self-Aligned Double-Gate (SSDG) FET 75

4.3.3 IBM Pagoda FET 77

4.3.4 STMicro DG FET using Silicon On Nothing (SON) Process 78

4.4 Novel Planar DG FET Process Flow 79

4.4.1 Si/SiGe Trilayer Epitaxy 80
4.4.2 Fin Patterning 81
4.4.3 Sidewall Spacer Formation 82
4.4.4 Source/Drain Deposition 82
4.4.5 Width Patterning 84
4.4.6 Isotropic SiGe Etch 84
4.4.7 Gate Stack Formation 85
4.5 Planar Ultrathin Body DG FET Structure – Salient Features 87
4.6 Process Variations 88
4.6.1 High Mobility Channel DG FET 88
4.6.2 Heterostructure Channel DG FET 89
4.6.3 Source/Drain Material Engineering 90
4.7 Summary 91
References 92

CHAPTER 5

PLANAR DOUBLE-GATE FET PROCESS DEVELOPMENT – EXPERIMENTAL RESULTS 95
5.1 Introduction 95
5.2 Si/SiGe Trilayer Epitaxy 96
5.2.1 Blanket Epitaxy 97
5.2.2 Selective Epitaxy 105
5.3 SiGe Enhanced Oxidation 112
5.4 Isotropic SiGe Selective Etching 114
5.4.1 Ammonium Hydroxide / Hydrogen Peroxide / Water 115
5.4.2 Hydrofluoric Acid / Hydrogen Peroxide / Acetic Acid 116
5.5 In-situ Doped Source/Drain Deposition 120
5.5.1 N-type Source/Drain Process 120
5.5.2 P-type Source/Drain Process 121
5.6 Process Integration 124

5.7 Transistor Results 127
5.8 Summary 132
References 134

CHAPTER 6

A COMPARISON FRAMEWORK FOR FUTURE TRANSISTORS 137
6.1 Introduction 137
6.2 CMOS Performance Metrics – Power and Delay 138
6.2.1 Transistor Delay 139
6.2.2 Power Dissipation 140
6.3 Methodology 142
6.4 Implementation 143
6.5 Results – Total Power Optimization 149
6.6 Results – Optimum Power-Delay Comparisons 152
6.7 Summary 158
References 159

CHAPTER 7

CONCLUSIONS AND RECOMMENDATIONS 161
7.1 Dissertation Summary 161
7.2 Contributions 165
7.3 Recommendations for Future Work 166
References 168

List of Tables

Table 5-1 Trilayer SiGe/Si/SiGe CVD epitaxy recipe details.....104

Table 5-2 Deposition parameters for in-situ Phosphorus doped Si using LPCVD.....121

Table 5-3 Recipe details for in-situ Boron doped Si by RPCVD.....122

Table 5-4 Process sequence for the experimental demonstration of planar DG FET...125

This page is intentionally left blank

List of Figures

Fig. 1.1 Gate length scaling trend over the past 35 years. The transistor gate length has been shrunk from 10 μm in 1970 to about 50 nm today. (From Moore [1.3]).....2

Fig. 1.2 Evolution of microprocessor performance, measured in millions of instructions per second (MIPs). The data points represent different generations of Intel microprocessors. (From Moore [1.3]).....2

Fig. 1.3 Active and standby power density trends plotted from industry data. The extrapolations indicate a cross-over below 20 nm gate length. As devices scale towards that point, it is questionable if the traditional approaches and reasons for scaling will still be valid. (From Nowak [1.4]).....3

Fig. 1.4 Schematic of a conventional bulk MOSFET.....4

Fig. 1.5 Schematic of a symmetric double-gate FET.....8

Fig. 2.1 Schematic of partially depleted SOI (PD SOI) transistor.....15

Fig. 2.2 Schematic of a single-gate fully depleted SOI (FD SOI) transistor.....15

Fig. 2.3 Schematic of double-gate (DG) FET. The top and bottom gates are electrically connected.....17

Fig. 2.4 Schematics of (a) trigate transistor [2.17], and (b) nanowire FinFET [2.19].....19

Fig. 2.5 Medici-simulated conduction band diagrams showing gate-induced lowering of the source barrier at the surface of a DG FET.....21

Fig. 2.6 Lowering of the source barrier across the DG FET body as a function of applied gate voltage. The drain is at 0.9 V.....22

Fig. 2.7 Comparison of gate-modulated lowering of source barrier at the DG FET surface and center of the body. This is extracted from the band diagrams in Fig. 2.6...22

Fig. 2.8 Off-state source barrier lowering as a result of reduced channel length for a DG FET. This leads to short channel effects.....23

Fig. 2.9 Comparison of the off-state source barrier reduction at the surface and body center as the gate length is reduced in a DG FET.....24

Fig. 2.10 Degradation of short channel effects (V_T roll-off, DIBL, and subthreshold swing increase) as the ratio (L/Λ) decreases. (from Frank et. al. [2.25])..... 26

Fig. 2.11 Simulated plots of DIBL as a function of gate length for a DG FET with various oxide thickness teq and body thickness tsi. (From Wong et. al. [2.26])..... 27

Fig. 2.12 Simulated plots of V_T roll-off with reduced gate length comparing single-gate FD SOI and DG FET for various body thickness tsi. (From Wong et. al. [2.26])..... 27

Fig. 2.13 Medici-simulated plots showing the sensitivity of off-state leakage to the body thickness for DG FETs with gate length as a parameter. In both cases, the gate workfunction has been optimized to give the same off-state leakage current at tsi = 7 nm. The gate oxide thickness is 1 nm and the drain bias is 0.9 V..... 28

Fig. 2.14 Plot of calculated threshold voltage (left axis) and threshold sensitivity to body thickness variation as a function of the nominal body thickness t_{si}. The broken lines include the effect of quantum confinement. (From Wong et. al. [2.27])..30

Fig. 2.15 Measured hole mobility as a function of silicon body thickness in a single-gate FD SOI transistor (From Uchida et. al. [2.29])..... 31

Fig. 2.16 Measured electron mobility as a function of silicon body thickness in a single gate FD SOI transistor (From Uchida et. al. [2.29])..... 31

Fig. 3.1 Schematic of the simulated ideal DG FET structure. The gates are perfectly self-aligned and equal sized. In the simulations that follow, the effects of deviations in the positions or size of the bottom gate will be examined. L_{gate}=18nm, T_{ox}=1nm, T_{si}=7nm..... 39

Fig. 3.2 Schematic of simulated inverter chain for mixed mode FO1 delay calculations..... 39

Fig. 3.3 Impact of the bottom gate misalignment on the FO1 inverter delay and the off-state leakage current..... 40

Fig. 3.4 Effect of the bottom gate mis-sizing on the FO1 inverter delay and the off-state leakage current..... 42

Fig. 3.5 Effect of the gate to source/drain sidewall spacer thickness on the FO1 inverter delay..... 43

Fig. 3.6 Schematic cross-section of the simulated double-gate FET..... 44

Fig. 3.7 Lateral doping profile in the source extension region for 3 values of lateral doping gradient (LDG).....46

Fig. 3.8 Effect of increasing either the lateral contact length or the specific contact resistivity on the drive current, with fixed off state leakage.....47

Fig. 3.9 I_{on}-I_{off} curves for a fixed 5 nm extension underlap and varying lateral doping gradients.....49

Fig. 3.10 I_{on}-I_{off} curves for a fixed lateral doping gradient of 0.5 nm/dec. and varying extension underlaps.....50

Fig. 3.11 I_{on}-I_{off} curves for a fixed lateral doping gradient of 3.5 nm/dec. and varying extension underlaps.....51

Fig. 3.12 Degradation of subthreshold swing when the lateral doping gradient is too gradual for a given extension underlap (2 nm).....51

Fig. 3.13 Plots of the doping and the electron concentration from the source to drain near the gate dielectric/silicon interface showing the bottleneck due to insufficient gate-to-source overlap when the lateral doping gradient is too abrupt for a given extension underlap.....53

Fig. 3.14 Optimization of the extension underlap. The leakage current is set to 1μA/μm.....54

Fig. 3.15 Simulated plot of the FO1 inverter delay as a function of the extension underlap as set by the sidewall spacer. The optimum delay is not as good as the delay obtained by decoupling the underlap from the spacer thickness.....55

Fig. 3.16 Effect of the leakage current (I_{off}) constraint on the underlap optimization. The optimal underlap increases for lower required I_{off}.....56

Fig. 3.17 Effect of silicon body thickness (T_{si}) on the underlap optimization. The optimal underlap decreases as T_{si} is reduced.....56

Fig. 3.18 Effect of the gate length on the optimal underlap. For shorter L_{gate}, the optimal shifts to larger values and the optimum also becomes steeper.....57

Fig. 3.19 Effect of the specific contact resistivity on the underlap optimization. The optimal underlap does not change.....58

Fig. 3.20 Schematic of the simulated Schottky source/drain DG FET.....60

Fig. 3.21 Simulated I_d-V_g characteristic for an n-type Schottky S/D DG FET. The metal-semiconductor system is assumed to form a barrier of 0.2 eV to the conduction band.....60

Fig. 3.22 Off-state and on-state band diagrams showing the conduction band edge from the source to drain at a point just beneath the gate dielectric/Si interface.....61

Fig. 3.23 I_{on}-I_{off} comparisons of Schottky S/D DG FETs with different barrier heights with conventional doped S/D DG FETs with different lateral abruptness. In all cases, the extension underlaps have been optimized.....63

Fig. 4.1 Types of double-gate FET configurations. This figure is from a presentation by H.S-P. Wong [4.1].....70

Fig. 4.2 IBM Planar DG FET process using pattern-constrained epitaxy (from [4.15])..74

Fig. 4.3 Key process steps for fabrication of SSDG FETs (from [4.16]) 75

Fig. 4.4 PAGODA DG FET process flow (from [4.17]).....77

Fig. 4.5 ST Microelectronics SON DG FET Process - main steps (from [4.19]).....79

Fig. 4.6 Epitaxial deposition of SiGe/Si/SiGe trilayer stack.....80

Fig. 4.7 Cross section (left) and plan view (right) of trilayer stack after fin patterning...81

Fig. 4.8 Cross section (left) and plan view (right) schematic of trilayer fin after differential oxidation to form sidewall spacers on SiGe.....82

Fig. 4.9 Cross section (left) and plan view (right) schematic of structure after the source/drain formation.....83

Fig. 4.10 Plan view schematic showing fin width patterning. The SiGe is now exposed in the trilayer cross-section.....84

Fig. 4.11 Schematic cross-section (left) and perspective view (right) of the structure after SiGe removal. The Si beam is supported on two sides by the source and drain regions.....85

Fig. 4.12 Cross section (left) and plan view (right) schematic of device structure after gate stack formation and patterning.....86

Fig. 4.13 Schematic cross section of completed planar DG FET structure. The metal lines and vias are not drawn to scale.....86

Fig. 4.14 Novel device structures that can be built by variations around the baseline planar DG FET process.....89

Fig. 5.1 Annotated plots of kinetically-limited critical thickness for strained SiGe growth on (100) Si as a function of growth temperature. This figure is from Houghton [5.4]. CVD data for 625 °C growth shown by the filled squares is perhaps the most representative of the SiGe epitaxy process in our work.....100

Fig. 5.2 x-TEM (off-zone axis) of our initial attempt to grow the trilayer stack. The crystal quality is quite poor as indicated by the large number of defects in the SiGe and the Si layers. There is also evidence of SiGe pile-up near the bottom of the growth interface.....101

Fig. 5.3 Higher resolution lattice image (on (110) zone-axis) showing crystal defects such as misfit dislocations and stacking faults in our initial trilayer stack.....102

Fig. 5.4 X-TEM of another trilayer stack growth that resulted in strain relaxation via non-planar growth (undulations).....103

Fig. 5.5 X-TEM of high quality trilayer stacks grown by the epitaxy process with ramped GeH_4 and lower growth temperature Si cap at the start of the center channel layer. No defects were visible in the sample. The high resolution lattice image shown in (b) indicates excellent crystal quality.....104

Fig. 5.6 Modified starting steps using selective trilayer epitaxy for a bulk-compatible planar DG FET process.....105

Fig. 5.7 x-TEM of selective trilayer epitaxy on a patterned substrate using 1150 °C H_2 bake for 3 min.....107

Fig. 5.8 x-TEM of selective epitaxy of the trilayer with an insufficient H_2 bake. This results in poor quality discontinuous deposition.....107

Fig. 5.9 x-TEM of an oxide-patterned substrate just before loading into the epi reactor for the selective epitaxy.....108

Fig. 5.10 x-TEM of selective Si epitaxy on an oxide patterned substrate after H_2 bake at 1000 °C for 3 min.....109

Fig. 5.11 x-TEM of selective Si epitaxy on an oxide patterned substrate after H_2 bake at 950 °C for 10 min.....109

Fig. 5.12 x-TEM of selective Si epitaxy on an oxide patterned substrate after H_2 bake at 900 °C for 10 min.....110

Fig. 5.13 Plot of the lateral oxide etch rate as a function of H_2 bake temperature. The activation energy of this Arrhenius type of process is 4.78 eV.....110

Fig. 5.14 x-TEM of a sample with an oxide sidewall on Si that has been annealed in H_2 at 925 °C for 2 min. Besides the lateral oxide etching, the Si surface moves into the undercut region due to the enhanced surface mobility of Si atoms under those conditions.....111

Fig. 5.15 x-TEM of a SiGe/Si/SiGe trilayer stack grown by selective epitaxy within a Si active area trench between patterned oxide/nitride isolation regions.....112

Fig. 5.16 Measured oxidation rate enhancement on blanket polycrystalline SiGe films as a function of Ge fraction. The oxidation conditions were 750 °C for 30 min in steam. The oxide thickness on the poly-Si was 17.5 nm.....113

Fig. 5.17 Measured etch selectivity to Si of the SiGe wet etch using $NH_4OH:H_2O_2:H_2O$ (1:1:5) at 75 °C. The samples are polycrystalline films of LPCVD SiGe with varying Ge concentration. The dashed line is a fit to the measured data.....116

Fig. 5.18 Measured etch rate of polycrystalline SiGe films as a function of Ge atomic fraction using $HF:H_2O_2:CH_3COOH$ (1:2:3 by volume).....118

Fig. 5.19 Tilted view SEM images of patterned test structures subjected to isotropic SiGe etching using $HF:H_2O_2:CH_3COOH$. These structures are fins consisting of the epitaxially grown SiGe/Si/SiGe trilayer stack and are surrounded by polysilicon spacers. The spacers protect the sides of the fin, but have been removed from the front, allowing the etchant to access the buried SiGe layers in the trilayer stack. The isotropic SiGe etching can be clearly seen as a time-dependent undercut.....119

Fig. 5.20 x-TEM of an annealed in-situ B-doped RPCVD Si layer on patterned trilayer stack. No H_2 annealing was done before the doped Si deposition. As a result, it crystallizes to form polysilicon everywhere.....123

Fig. 5.21 x-TEM of an annealed in-situ B-doped RPCVD Si layer on the patterned trilayer stack. A 900 °C H_2 bake for 10 min ensured removal of native oxide, leading to single crystal epi formation over the silicon surface and the stack sidewall. Poly-Si is formed over and around the LTO cap.....123

Fig. 5.22 Measured I_d-V_g characteristics of fabricated ultrathin body DG FET. Excellent turn-off characteristics are seen, such as no DIBL and near-ideal 67 mV/dec subthreshold swing.....128

Fig. 5.23 Measured I_d-V_d characteristics of the same device as in Fig 5.22.....129

Fig. 5.24 Schematic of the parasitic bulk transistor created in parallel with the DG FET as a result of the simplified process flow.....130

Fig. 5.25 Effect of the substrate back-bias on the I_d-V_g characteristics of the bulk-only devices. Finite DIBL remains in all cases and the subthreshold swing of 89 mV/dec is far from ideal. The linear threshold voltage increases monotonically with the back-bias.....130

Fig. 5.26 Effect of the substrate back-bias on the I_d-V_g characteristics of composite (DG FET + parasitic bulk FET) devices. Without back-bias, the overall characteristics are dominated by the parasitic bulk FET. Application of back-bias increases the threshold voltage of the bulk FET and the excellent subthreshold characteristics are due to the DG FET. Further application of back-bias does not change the threshold voltage.....131

Fig. 6.1 Schematic illustration of the evolution of static and dynamic power components as a function of V_{dd} for a target delay. The dynamic power grows quadratically, while the static power reduces in an exponential fashion. Due to the opposing trends, the total power has a minimum at some V_{dd}.....142

Fig. 6.2 Schematic of the Medici-simulated double-gate (DG) FET based on the ITRS 2003 45 nm HP/32 nm LSTP node. Baseline DG: L_{gate}=18 nm, T_{si}=7 nm. Back-gate (BG) FET: grounded bottom gate with midgap workfunction.....144

Fig. 6.3 Family of I_d-V_g curves with varying V_{dd} for 2 values of T_{ox}. Negative V_g is used to simulate the effective gate workfunction, which takes on all values within the Si bandgap. The right-side axis shows a 30-40% underestimate of the drive current using drift-diffusion as compared to the energy-balance transport model.....145

Fig. 6.4 Setup for FO1 inverter delay calculation by the numerical integration of Medici-simulated I_d-V_g and C_g-V_g curves.....146

Fig. 6.5 Comparison of inverter FO1 delay calculation methods with mixed mode transient simulations in Medici. Use of the simple CV/I_{on} approximation to calculate FO1 delay causes a large underestimate.....147

Fig. 6.6 Static power (due to source-drain leakage) calculation method. For every delay point, as V_{dd} is swept, implicit V_T adjustment sets the leakage current.....148

Fig. 6.7 Gate tunneling current and oxide voltage drop calculations by post-processing analytical models (solid lines) compared with Medici simulations (symbols).....149

Fig. 6.8 Total power optimization curve for a specific delay and switching activity factor. The dashed curves show the weighted individual components of power – dynamic, source-drain leakage (SDL), and gate leakage power (GLP).....150

Fig. 6.9 Optimized total power as a function of the effective oxide thickness T_{ox} at inversion comparing the SiO_2 with perfect high-k dielectric for the baseline DG FET with moderate (10%) switching activity.....151

Fig. 6.10 Optimized total power as a function of the effective inversion oxide thickness comparing SiO_2 with an ideal high-k dielectric for a device with low (1 %) switching activity.....152

Fig. 6.11 Comparison of devices using optimal power-delay curves. The impact of material change (high-k vs. SiO_2), structure change (DG vs. BG), and parasitic resistance (ideal vs. realistic) is shown. The right-side axis shows the optimal V_{dd} for the DG FET with and without series resistance.....153

Fig. 6.12 Minimum total power-delay curves with corresponding optimal V_{dd} (shown on right-side axis) for devices with different switching activity.....154

Fig. 6.13 Minimum total power-delay curves, along with the corresponding optimal T_{ox} for DG FETs with different Si body thickness (T_{si}).....155

Fig. 6.14 Impact of process-induced gate length variations on the minimum power-delay curves and the corresponding optimum V_{dd}.....156

Fig. 6.15 Optimal gate length that minimizes total power for a given target delay and structure (baseline DG FET with T_{si} = 7 nm). For any delay, the incorporation of process-induced variations increases the optimal L_g and the optimized total power.....157

Chapter 1

Introduction

1.1 Motivation

The silicon metal-oxide-semiconductor field effect transistor (MOSFET) is one of the most important devices in the semiconductor industry. Since its first practical demonstration in 1960 [1.1], the MOSFET has been incorporated in monolithic integrated circuits (ICs) to serve as a basic switching element for digital logic and as an amplifying device for analog applications. While the basic planar structure of the MOSFET has remained mostly unchanged, its size has been shrunk by many orders of magnitude over the past thirty years. The trend showing an exponentially increasing number of transistors on a chip was first predicted in 1965 and has since come to be known as 'Moore's Law' [1.2]. Fig. 1.1 shows a plot of the minimum feature size in the MOSFET, the gate length, over time. The main driving forces behind gate length scaling are the higher switching speed and packing density that result from making the transistors smaller. Device scaling has enabled IC chips to operate faster and with greater functionality each new technology generation. Fig. 1.2 depicts the exponential increase in microprocessor performance, represented by millions of instructions per second (MIPs), over technology generations.

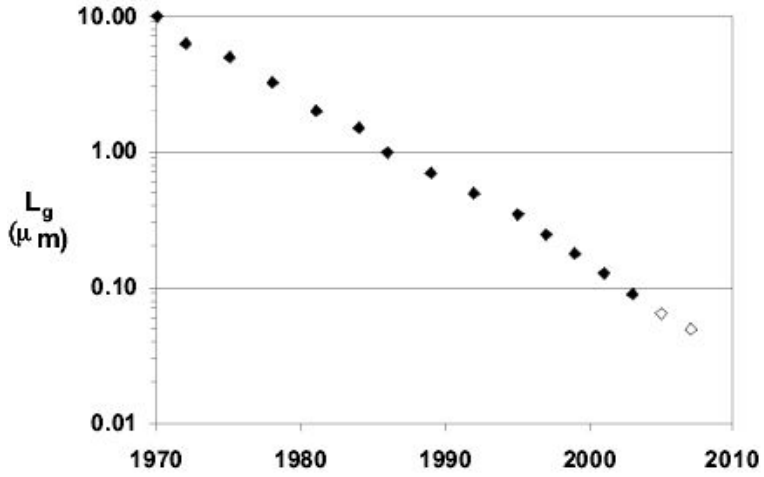


Fig. 1.1 Gate length scaling trend over the past 35 years. The transistor gate length has been shrunk from 10 μm in 1970 to about 50 nm today. (From Moore [1.3])

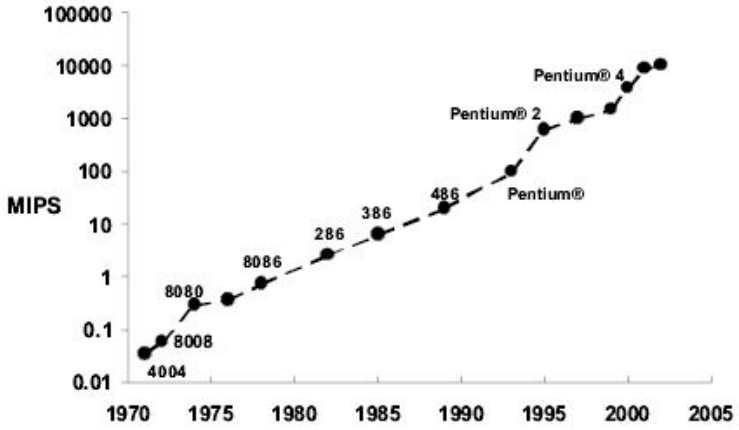


Fig. 1.2 Evolution of microprocessor performance, measured in millions of instructions per second (MIPs). The data points represent different generations of Intel microprocessors. (From Moore [1.3])

The enhanced speed and complexity of IC chips has been accompanied by an increase in power dissipation. Fig. 1.3, from Nowak [1.4], depicts the evolution of power density as the gate length is scaled. The active power arises due to the dissipative switch-

ing of charge between the transistor gates and supply/ground terminals during logic operations. The subthreshold power, also known as static or standby power, is dissipated even in the absence of any switching operation. It arises due to the fact that the MOS transistor is not a perfect switch – there is some leakage current that flows through it in the off-state.

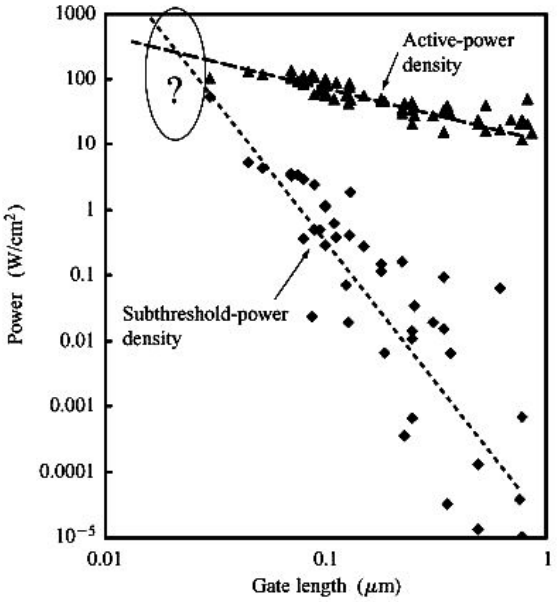


Fig. 1.3 Active and standby power density trends plotted from industry data. The extrapolations indicate a cross-over below 20 nm gate length. As devices scale towards that point, it is questionable if the traditional approaches and reasons for scaling will still be valid. (From Nowak [1.4])

While the active power density has steadily increased with gate length scaling, the static power density has grown at a much faster rate. The latter was a relatively insignificant component of power just a few generations back, but it is now comparable in magnitude to the active power. Management and suppression of static power is one of the major challenges to continued gate length reduction for higher switching speed. Traditional MOSFET scaling has begun to face impediments due to fundamental as well as practical limits. It is now widely accepted [1.5] that novel (i.e. non-classical) transistors

will be needed to prolong device scaling with commensurate improvements in performance. The double-gate (DG) FET is a promising device structure that can potentially replace conventional transistors in future technology generations. This dissertation research is focused on studying, through simulation and experiment, some of the problems and solutions for the technology and scaling of DG FETs.

1.2 Challenges in Scaling Conventional MOSFETs

Fig. 1.4 shows a schematic of a conventional transistor built on bulk-Si substrates (bulk MOSFET). When used in digital logic applications, the gate is used as a control terminal to switch on or off a current conduction channel between the source and drain terminals. The current carriers can be either electrons (in an n-type MOSFET, called NMOS FET) or holes (in a p-type MOSFET, called PMOS FET).

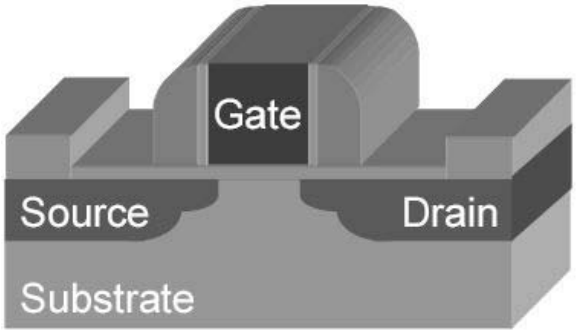


Fig. 1.4 Schematic of a conventional bulk MOSFET.

In the traditional implementation of the bulk MOSFET, the gate electrode is made of heavily (n or p) doped polycrystalline silicon. It is separated from the bulk silicon substrate by a thin insulating dielectric layer of silicon dioxide. The channel region underneath the gate is moderately doped. The source and drain regions, upon which the other electrode contacts are formed, are heavily doped to form n-p (or p-n) junctions to

the oppositely doped substrate. The simplest transistor scaling approach [1.6] involves reducing the vertical and horizontal dimensions, as well as the supply voltage, by the same factor in an attempt to keep the electric fields in the scaled MOSFET the same as before (constant-field scaling). However, actual scaling implementations have been based on slightly modified approaches where the geometry and voltage are reduced by different factors (generalized scaling).

In early generation transistors with large gate lengths (long channel MOSFETs), the vertical electric field in the channel (due to the applied gate to source voltage) is much larger than the lateral channel electric field (due to the applied drain to source voltage). In such a case (gradual channel approximation), the physics of transistor operation can be partitioned into two independent portions, i.e. gate-controlled charge formation in the channel, and drain-controlled charge transport. The threshold voltage V_T, at which the device turns on, is dependent only on the gate voltage and is independent of the drain voltage. Application of the gate voltage lowers a potential barrier near the source and allows carriers (electrons or holes) to flow from source to drain. The subthreshold swing is a measure of how sharply the drain current increases as a function of gate voltage during switching from 0 V to V_T. Measured in mV/dec. (mV of incremental gate voltage required to change the drain current by one decade), under normal transistor operation fundamental thermodynamics constrains the subthreshold swing to be greater than 60 mV/dec. at room temperature. Degraded 2-D electrostatics at short gate lengths worsen (increase) this value – leading to higher off-state leakage current for the same V_T.

In reality, the potential barrier at the source is controlled by the gate as well as the drain through their respective coupling capacitances to that point. The gradual channel approximation is just a simplification of the complicated two-dimensional (2-D) electrostatics in the MOSFET channel. While the simplification holds in long channel devices, as the gate length is reduced, the drain influence becomes stronger. As a result, it becomes harder for the gate to control the source barrier and turn off the channel.

The 2-D effects are manifested in various ways:

i) reduction in threshold voltage with shrinking gate length (V_T roll-off),
ii) V_T reduction with increasing drain voltage (drain induced barrier lowering – DIBL),
iii) degraded subthreshold swing.

Collectively, these phenomena are known as 'short channel effects' (SCE) and they tend to increase the off-state static leakage power. Thus far, device designers have tried to suppress SCE in short gate length devices by a number of methods:

i) reducing the gate oxide thickness to improve the gate control over the channel,
ii) lowering the source/drain junction depth (especially near the gate edge, where the source/drain regions are called 'extensions') to reduce the drain coupling to the source barrier,
iii) increasing the channel doping to terminate the electric field lines which originate from the drain and propagate towards the source. In modern bulk MOSFETs, the channel doping is tailored to have complicated vertical and lateral profiles (super-halo doping) [1.7] so as to minimize the impact of gate length variations on the short channel effects.

Each of these approaches comes at a cost which either degrades transistor performance (speed) or introduces a new static leakage mechanism:

i) As the gate oxide gets very thin, quantum mechanical tunneling allows a gate leakage current to flow. In the direct-tunneling regime, encountered for oxides thinner than about 3 nm, the gate leakage current increases dramatically (~3× for every 1 Å of thickness reduction). The gate leakage can increase standby power as well as compromise proper dynamic logic operation [1.8]. Many people have proposed replacing the silicon dioxide (SiO_2) with higher permittivity (high-k) gate dielectrics [1.9] such as zirconia (ZrO_2) or hafnia (HfO_2). These enable high gate capacitance with physically thick insulators through which tunneling is low. However, the introduction of such new materials without the accompanying degradation of mobility and reliability is very challenging and remains an area of intense ongoing research.

ii) As the source/drain junctions are made shallower, their doping must be increased so as to keep the sheet resistance constant. The solid solubility of dopants puts an upper limit (~ 10^{20} cm^{-3}) on the doping density. Therefore, a further reduction in the junction depth causes an increase in the series resistance encountered in accessing the channel. This degrades the overall transistor performance. Also, from a technological point of view, it becomes difficult to form ultrashallow junctions that remain abrupt after the annealing steps needed to activate the dopants and achieve low resistivity.

iii) As the doping density in the channel is increased for SCE suppression, the carrier mobility is degraded due to increased scattering from the ionized dopant atoms. Besides, the subthreshold swing gets worse due to higher depletion capacitance that 'steals' away part of the gate voltage from the surface potential. For very high channel doping near the source/drain extensions, another component of static leakage, band-to-band tunneling, becomes important. Finally, as the channel volume reduces in extremely scaled transistors, the random placement of discrete dopant atoms cause stochastic inter-device variations [1.10].

As a result of these (and other) problems, it is becoming clear that new materials and/or structures will be needed to supplement or even replace the conventional bulk MOSFET in future technology generations.

1.3 Double-Gate (DG) FET

The double-gate FET is a novel device structure that is a promising candidate for replacing conventional bulk MOSFETs beyond the 45-nm technology node (2007). Fig. 1.5 shows a schematic cross section of this device. The DG FET can be regarded as an evolution of the regular MOSFET structure, with a second gate placed below a thin body in which the channels are formed. In the most effective DG FET implementation, both gates are electrically connected (driven by the same voltage), and the top and bottom gate dielectrics have the same thickness. This is known as a symmetric DG FET. The main

advantage of placing the second gate is the increased electrostatic gate control over the channel. The two gates are more effective at shielding the drain electric field lines so as to prevent them from reaching towards the source and degrading the short channel effects. In addition, by removing part of the bulk substrate and replacing it with a gate, one eliminates sub-surface current leakage paths that are far away from gate control. As a result of the better electrostatics, the DG FET is scalable to shorter gate lengths than bulk FETs. Since the two gates and thin body are sufficient to suppress short channel effects, the body is left undoped. This improves channel carrier transport due to increased mobility resulting from reduced ionized impurity scattering and lower vertical electric field. In addition, the undoped body is more immune to discrete dopant fluctuation effects.

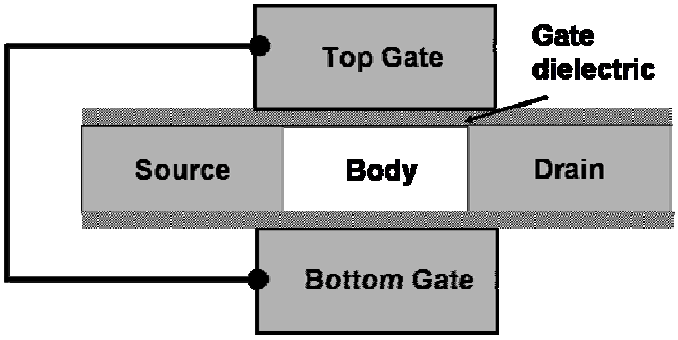


Fig. 1.5 Schematic of a symmetric double-gate FET

For effective suppression of short channel effects, the silicon body between the two gates must be made very thin (typically, it should be less than half of the gate length). Achieving a uniform and controllably ultrathin body is one of the key requirements for a DG FET from an intrinsic scalability viewpoint. In practice the non-classical structure can exacerbate parasitic impedance effects already present in bulk MOSFETs and also introduce new ones. It is therefore very important to understand and minimize these parasitic impedances since such extrinsic effects can degrade the overall DG FET performance even if the intrinsic device is well-designed.

1.4 This Work and Dissertation Organization

The main thrusts of research in this dissertation are simulation-based study of extrinsic impedance and its optimization in DG FETs, and the proposal and experimental fabrication of a novel planar DG FET structure that can, in principle, satisfy most of the intrinsic and extrinsic requirements for an 'ideal' double-gate transistor. Finally, we also develop a methodology to compare various advanced transistor structures and benchmark them using optimized power and delay as metrics.

The next chapter describes the evolution of novel transistors starting from conventional planar bulk MOSFETs to more advanced structures such as fully-depleted SOI single gate FETs and double-gate FETs. The enhanced 2-D electrostatic scalability of the DG FET is explained, motivating the need for an ultrathin body. The limits of scaling the body thickness are also discussed.

Chapter 3 deals with the extrinsic impedances that can limit the overall performance of a DG FET. Using extensive device simulation, we examine the impact of parasitic capacitance resulting from gate misalignment and over-sizing, and the impact of extrinsic series resistance due to contacts and ultrathin body. We point to a way to optimize the lateral doping profile in the ultrathin source/drain extensions. Finally, there is a brief discussion of the applicability and requirements for Schottky source-drain structures in DG FETs.

In chapter 4, we propose a novel process to fabricate the ideal planar DG FET structure with controllably ultrathin and uniform body and potentially low extrinsic impedances. The results of prior attempts in the literature to build such self-aligned planar DG FETs are highlighted. The novel proposed process brings together some of the main ideas from the previous work and eliminates most of their shortcomings. The planar DG FET process in this work also allows for the integration of new materials for the gate electrode, gate dielectric, and channel. Thus, it is a good platform to build interesting device structures by variations on the baseline process.

Chapter 5 deals with the experimental work that was done to fabricate a planar DG FET using the proposed process. The results of the unit process development are discussed, verifying the feasibility of some of the key process steps. Next, the integration scheme to build transistors is described. Finally, electrical results from a proof-of-concept demonstration of functional DG FETs are shown. The fabricated transistors show excellent subthreshold turn-off characteristics, as expected from double-gate transistors of those dimensions.

In chapter 6, we take a step back and look at device characteristics from a system performance point of view. A new methodology is developed to optimize digital logic transistors for minimum total power at a given delay and switching activity using supply voltage, threshold voltage, and oxide thickness as design parameters. The resulting minimum power-delay curves can then be used to compare various transistors and quantify the impact of material and/or structural innovations. While the developed framework is quite general and applicable to a variety of future transistors, we focus specifically on an 18 nm gate length DG FET targeted for 45 nm generation high performance logic.

Finally, in chapter 7, we summarize our work and suggest various directions in which this research can be extended in future.

References

[1.1] D. Kahng and M. M. Atalla, "Silicon-silicon dioxide field induced surface devices," in *IRE-AIEEE Solid-State Device Research Conference* (Carnegie Inst. of Tech., Pittsburgh, PA), 1960.

[1.2] G. E. Moore, "Cramming more components onto integrated circuits," *Electronics,* vol. 38, no. 8, pp. 114-117, April 19th 1965.

[1.3] G. E. Moore, "No exponential is forever: but "forever" can be delayed! [semiconductor industry]," in *Digest of Technical Papers – IEEE International Solid-State Circuits Conference*, 2003, pp. 20-23.

[1.4] E. J. Nowak, "Maintaining the benefits of CMOS scaling when scaling bogs down," *IBM Journal of Research & Development*, Vol. 46, No. 2/3, pp. 169-180, 2002.

[1.5] International Technology Roadmap for Semiconductors, 2003 Edition, SIA, 2003.

[1.6] R. H. Dennard, F. H. Gaensslen, H. Yu, V. L. Rideout, E. Bassons, and A. R. LeBlanc, "Design of ion-implanted MOSFETs with very small physical dimensions," *IEEE Journal of Solid-State Circuits*, vol. 9, no. 5, pp. 256-268, Oct 1974.

[1.7] Y. Taur, C. H. Wann, and D. J. Frank, "25 nm CMOS design considerations," in *IEDM Technical Digest*, 1998, pp. 789-792.

[1.8] P. Wright and K. C. Saraswat, "Thickness limitations of SiO_2 gate dielectrics for MOS ULSI," *IEEE Transactions on Electron Devices*, vol. 37, no. 8, pp. 1884-1892, Aug 1990.

[1.9] G. D. Wilk, R. M. Wallace, and J. M. Anthony, "High-k gate dielectrics: Current status and materials properties considerations," *Journal of Applied Physics*, vol. 89, no. 10, pp. 5243-5275, 2001.

[1.10] A. Asenov, “Random dopant induced threshold voltage lowering and fluctuations in sub-0.1 μm MOSFETs: A 3D “atomistic” simulation study,” *IEEE Transactions on Electron Devices*, vol. 45, No. 12, pp. 2505-2513, Dec 1998.

Chapter 2

Double-Gate FET Operation – Intrinsic Factors

2.1 Introduction

The key advantage of a double-gate FET over a conventional bulk FET is the increased scalability by virtue of better electrostatic gate control over the channel. In addition, the lack of body doping leads to potentially improved channel carrier transport. This is a result of reduced ionized impurity scattering and lower vertical electric fields, both of which lead to higher mobility. However, it is very important to have an ultrathin body for the effective suppression of short channel effects. As the body thickness is reduced, quantum confinement leads to undesirable effects such as higher threshold voltage and lower mobility. This puts an intrinsic lower limit on the body thickness.

In this chapter, the evolution of device structures from conventional planar bulk FETs to ultrathin body DG FETs is outlined. The physics of 2-D electrostatics in DG FETs is then examined in order to motivate the requirement for a thin body. This is followed by a description of the quantum confinement effects that arise in extremely

scaled ultrathin FETs. Finally, mobility issues in ultrathin body transistors are briefly discussed.

2.2 Evolution of Novel Device Structures

As indicated in the previous chapter, it is getting much harder to prolong traditional scaling of the conventional planar bulk MOS transistor with commensurate improvements in performance. It is widely recognized [2.1] that novel device structures will be needed to supplement or even replace bulk transistors.

2.2.1 Partially Depleted SOI (PD SOI)

In the partially depleted silicon on insulator (PD SOI) transistor, shown schematically in Fig. 2.1, a layer of insulating silicon dioxide separates the upper device-containing film from the rest of the bulk Si substrate. The upper Si film is relatively thick (~100 nm or more) and is doped in the same way as a corresponding bulk transistor having similar geometry. In the off-state condition, with zero voltage applied on the gate electrode, the maximum depletion width underneath the gate oxide is less than the Si film thickness. The lower part of the partially depleted layer has a quasineutral region in the body which is typically left uncontacted. The potential of this floating Si region is determined dynamically by capacitive coupling to the various electrodes, and in steady state, by a balance of forward and reverse biased currents to the source and drain junctions. This leads to a variety of floating body effects such as the parasitic bipolar kink effect as well as history-dependent threshold voltage [2.2]. The main advantage of the PD SOI device is somewhat higher speed due to reduced source and drain area junction capacitances. The floating body effects can be mitigated using ion implantation [2.3] or by placing a body contact [2.4]. By clever circuit design, these history-dependent effects can actually be used for increased speed [2.5]. PD SOI technology has been successfully ported from research into high volume manufacturing [2.6]. However, from a static

leakage and electrostatic scalability perspective, the PD SOI transistor looks the same as a bulk FET. It may therefore not be sufficient to prolong device scaling beyond the realm of what can be achieved with well-designed bulk FETs.

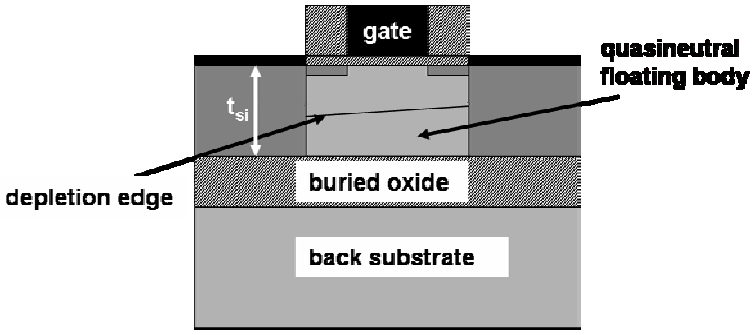


Fig. 2.1 Schematic of partially depleted SOI (PD SOI) transistor.

2.2.2 Fully Depleted SOI (FD SOI)

If the top silicon film thickness of the PD SOI device is reduced, eventually the entire Si body beneath the gate is depleted at zero applied gate bias. This kind of device is called a fully depleted SOI (FD SOI) transistor [2.7]. It has also been referred to as the 'depleted substrate transistor' [2.8] in the published literature. A schematic of an FD SOI FET is shown in Fig. 2.2. The film thickness below which the full-depletion condition occurs depends upon (the square root of) the body doping.

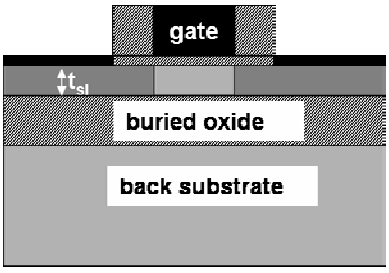


Fig. 2.2 Schematic of a single-gate fully depleted SOI (FD SOI) transistor.

By eliminating the quasineutral floating body, the kink effect and history-dependent behavior are strongly suppressed. More important, the short channel effect

control in the FD SOI device is potentially superior to that in the bulk/PD SOI transistors. As will be shown in section 2.3, smaller vertical depletion widths improve the electro-static gate control in any FET. In bulk/PD SOI FETs, the small vertical depletion width is achieved by using high channel doping. However, the use of very high channel doping has deleterious effects on transistor performance in terms of degraded subthreshold swing, reduced mobility, and enhanced band-to-band tunneling at the drain. In FD SOI transistors, the vertical depletion width can be controlled by the body thickness without requiring any channel doping. This improves carrier transport since the mobility is enhanced, both due to lower ionized impurity scattering, as well as due to lower vertical electric field for the same channel carrier concentration. Furthermore, if the buried oxide is thick compared to the Si body thickness and the gate oxide thickness, the long channel subthreshold swing approaches the ideal value of 60 mV/dec. at room temperature. This is because of the reduced impact of the thick buried oxide in the capacitive potential division between the body and the buried oxide. However, a thick buried oxide can potentially degrade the short channel performance. This is due to the effect of fringing electric field lines originating at the drain, going through the buried oxide, and lowering the barrier for the source to channel current. This is similar to DIBL, but occurs through the buried oxide. One way to reduce this unwanted drain action is to have a thin buried oxide so as to terminate most of the drain electric field lines at the back substrate. This is done at the expense of degraded subthreshold swing due to the capacitive division between the gate, body, and back substrate.

2.2.3 Double-Gate FET (DG FET)

The double-gate FET can be thought of as an enhanced version of an FD SOI transistor with a very thin buried oxide (same thickness as the gate oxide). Only now, the back substrate is heavily doped and electrically connected to the top gate. Since there is no capacitive potential division between the top and bottom gate, i.e. both of them drive the substrate together, the gate to substrate coupling is perfect and the long channel

subthreshold swing is identically 60 mV/dec. In addition, the short channel effect control is very good by virtue of a thin fully depleted body and gate shielding of drain electric field lines from both sides. Due to the action of two gates, the device can now be scaled to shorter gate lengths for the same body (and oxide) thickness. As will be seen in the next chapter, the extra parasitic capacitance due to the bottom gate in the DG FET can be quite high unless the two gates are perfectly self aligned to each other and have the same size. Such an ideal DG FET is schematically shown in Fig. 2.3. There have been a number of experimental implementations of research DG FETs with different orientations - vertical [2.9], fin-type [2.10] and planar [2.11].

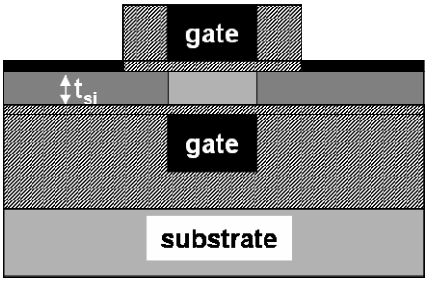


Fig. 2.3 Schematic of double-gate (DG) FET. The top and bottom gates are electrically connected.

2.2.4 Threshold Voltage Control in FD SOI/DG FET

In a conventional bulk FET, the threshold voltage of the transistor is typically tuned by using the channel doping, with higher doping increasing the absolute threshold voltage, V_T of both the n- and p-channel FETs. In a typical dual poly CMOS process, the NMOS devices have N+ polysilicon gate electrodes and medium to high p-type channel doping. The PMOS devices have P+ polysilicon gate electrodes and medium to high n-type channel doping. In the FD SOI or DG FET with undoped body, the ability to use body doping to tune the V_T is diminished. In such cases, using an N+ (P+) gate electrode causes the V_T of the NMOS (PMOS) transistor to be too low (close to 0 V) which leads to channel formation even under zero applied gate bias. On the other hand, using the

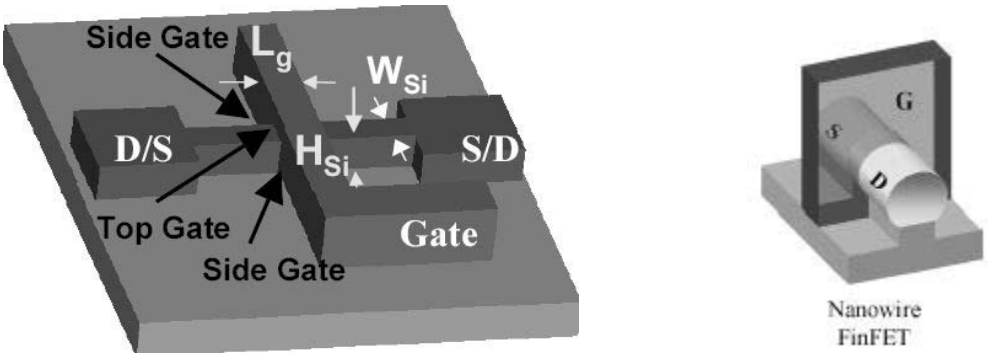
opposite type of gate polarity, i.e. P+ (N+) gate electrode for NMOS (PMOS) FETs leads to unacceptably high (~ 1 V) threshold voltage and hence low drive current. Some researchers [2.12] have proposed using N+ and P+ for the top and bottom gates respectively. This version of the DG FET, called the asymmetric DG FET, has the advantage of using the same kind of gate doping scheme on both NMOS as well as PMOS FETs. However, such devices are not as good [2.13] as the symmetric DG FET due to much higher vertical electric field in the body. Besides, such a scheme allows for only single-threshold voltage transistors, which may not be desirable from a circuit design viewpoint. The best way to tune the V_T in DG FETs is to change the gate workfunction. Theoretical studies have shown [2.14] that for acceptable off-state leakage currents and reasonable drive currents, the optimum gate workfunctions lie around 200 mV above (below) the Si midgap for NMOS (PMOS) DG FETs. In order for FD SOI/DG FETs to be incorporated into mainstream manufacturing, it is imperative to develop technology for metal gate workfunction engineering [2.15] so as to meet these requirements.

2.2.5 Beyond DG FETs

Since two gates provide better electrostatic control than one, it is reasonable to expect that having even more gates can further improve the gate length scalability of FETs. This has been shown rigorously [2.16], with the most electrostatically robust device being a surrounding-gate (or gate-all-around) FET. There have been some experimental demonstrations of tri-gate FETs [2.17, 2.18] and surrounding gate FETs [2.19]. Some of these are shown in Fig. 2.4. While these devices are indeed more scalable than DG FETs, they have a number of problems including gate oxide and transport non-uniformity due to multiple crystal orientations, and layout inefficiency arising from the need to put multiple channels in parallel to get high overall current for driving interconnect capacitance.

The single wall carbon nanotube FET [2.20] is the logical extreme of the surrounding gate FET. It is ideal in many respects: very thin (~ 3 nm) body defined non-

lithograpically (by a material property), perfect self-passivated surface on which gate dielectrics can be deposited, and high mobility leading to near ballistic transport within the tube. However, the problems of defining a priori the chirality (and hence bandgap), location, and orientation of carbon nanotubes remain as yet unsolved.



(a) (b)

Fig. 2.4 Schematics of (a) trigate transistor [2.17], and (b) nanowire FinFET [2.19].

2.3 DG FET Operation and Electrostatics

In this section, we examine the basic operation of the DG FET during the off-state and the transition to the on-state. In addition to utilizing results from some of the literature already existing in this field, device simulations with commercially available software, Medici [2.21], are used for better physical insight. Two dimensional device simulation is a powerful tool widely used both in literature and in this work. This method provides a quick way to enhance the understanding of device physics and to predict trends in futuristic transistor structures without incurring the expense and complication of actually building them. The accuracy of the results from such simulations depends upon the sophistication of the models employed and the degree of calibration. In general, Medici isn't as accurate at predicting drive current as it is in modeling the electrostatic behavior in devices, especially in the sub-100 nm regime.

2.3.1 Basic Double-Gate Transistor Operation

A simple way to think of field-effect transistor operation is to consider the effect of the gate bias on the source-body diode. The MOS FET is a majority carrier device with transport dictated by electrons in the conduction band[1]. In the off-state, the body doping and/or gate workfunction is chosen so as to have a large built-in voltage at the source-body junction. In other words, there is a large barrier for electrons that prevents them from diffusing from the source into the body. As the gate bias increases, the surface potential also increases, which lowers the source barrier. The gate therefore turns on the source-body junction by forward biasing it. The degree of forward bias is highest just beneath the gate oxide and diminishes with increasing depth into the body. Beyond the vertical depletion depth, the gate does not modulate the body potential. As the source barrier is lowered, the electron injection into the body due to thermal emission over the barrier increases exponentially. This is due to the exponentially increasing number of Fermi-Dirac distributed available electrons as the barrier height is lowered. The sub-threshold current thus generated is constrained to increase at a maximum rate set by kT where k is the Boltzmann constant and T is the absolute temperature. At 300 K, the steepest subthreshold swing has to be greater than 60 mV/dec.

Fig. 2.5 shows the simulated conduction band diagrams near the surface (i.e. just below the gate oxide/silicon interface) of a DG FET. These bands are plotted for off-state (zero gate voltage) and on-state (high gate voltage) along a cut-line that runs from source to drain. It should be noted that the positive bias applied to the drain, during both off- and on-states causes the conduction band edge at the drain to be lower than at the source. The collapse of the source barrier as the gate bias is applied can be clearly seen.

---

[1] In all the discussions that follow, we will focus on a NMOS DG FET with heavily doped n+ source/drain and intrinsic (or lightly p-doped) body. The PMOS DG FET is similar if one interchanges the doping polarity and considers hole transport in the valence band.

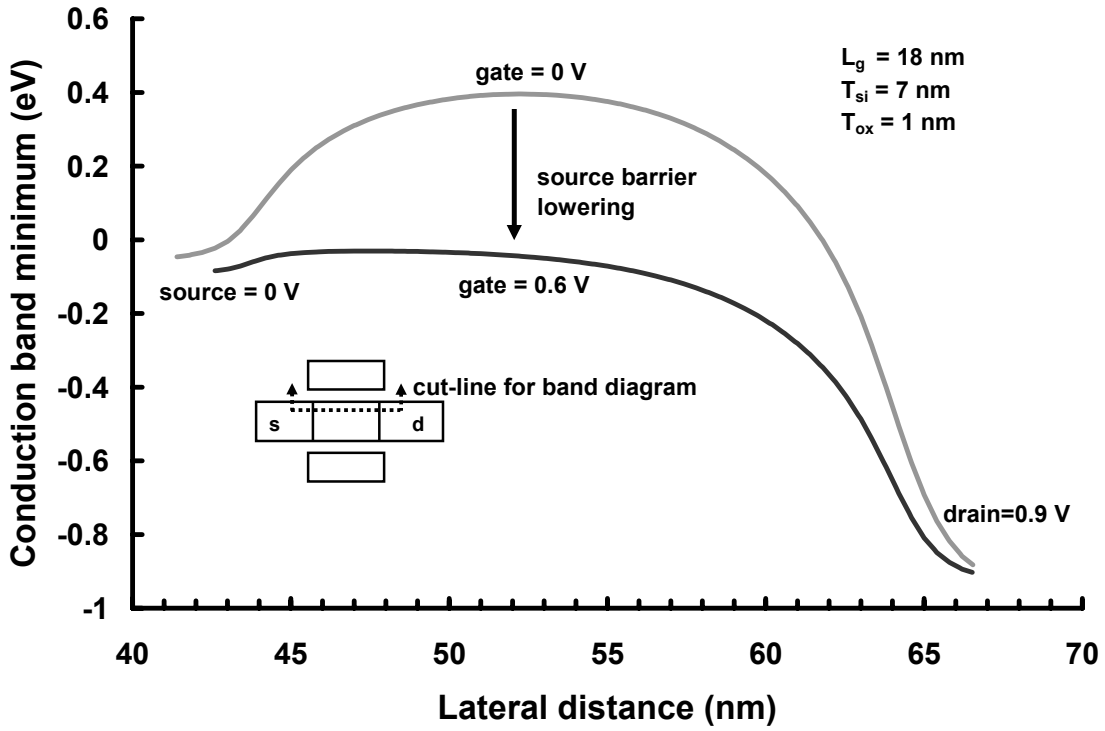


Fig. 2.5 Medici-simulated conduction band diagrams showing gate-induced lowering of the source barrier at the surface of a DG FET.

It is also evident from Fig. 2.5 that the peak of the source barrier occurs at some distance away from the source towards the drain. At this lateral distance, Fig. 2.6 shows the conduction band profiles in the vertical direction, along a cut-line running from the top to the bottom gate. These profiles represent the maximum barrier to electrons as seen from the source (assumed to be at 0 V). As the gate voltage increases, the barrier lowering initially occurs quite uniformly across the body causing electrons to be injected from the source across the entire depth of the device. This phenomenon is called volume inversion [2.22] and it leads to higher transconductance since the carriers traveling at the center of the body have high mobility due to reduced surface roughness scattering and transport in a nearly zero vertical effective electric field. The low vertical electric field arises as a result of the undoped body and due to the symmetry of the DG FET. As the

gate voltage increases, the stronger gate coupling to the surface compared to the center of the body casues the surface barrier to fall slightly faster than the center barrier. At high gate bias, there is a cross-over (shown in Fig. 2.7) after which the barrier at the surface becomes lower than that at the center.

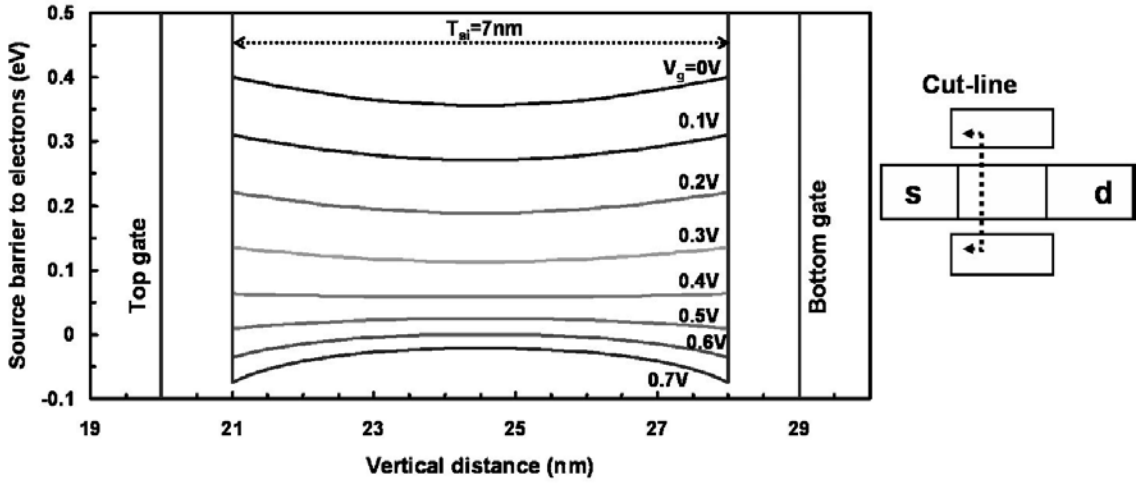


Fig. 2.6 Lowering of the source barrier across the DG FET body as a function of applied gate voltage. The drain is at 0.9 V.

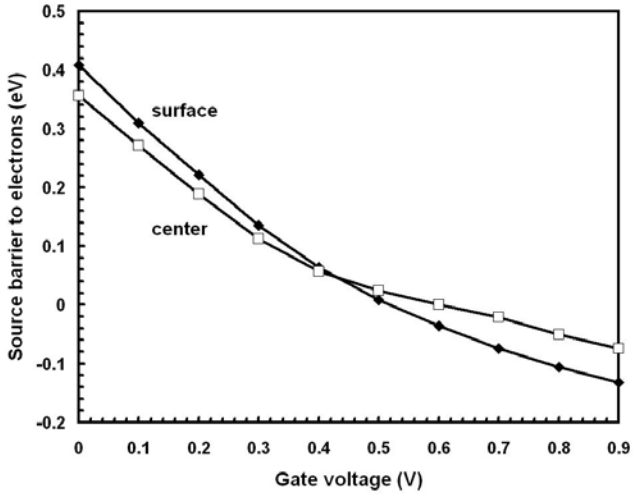


Fig. 2.7 Comparison of gate-modulated lowering of source barrier at the DG FET surface and center of the body. This is extracted from the band diagrams in Fig. 2.6.

Therefore under strong inversion conditions, the carriers predominantly flow near the surface and the benefits of volume inversion are diminished. These simulations neglect inversion layer quantization effects, which actually shift the centroid of inversion charge a few angstroms below the oxide/silicon interface.

2.3.2 DG FET Off-state – Gate Electrostatic Control

In the transistor off-state, the drain voltage is high, but the gate has zero applied voltage. The off-state source-drain leakage current is due to electrons that are able to thermally surmount the source barrier. Therefore, anything that reduces this barrier in the off-state directly causes an increase in the leakage current. Increased drain voltage tends to lower the source barrier (DIBL). Similarly, as the gate length is reduced, keeping everything else the same, the source barrier is reduced. Both these effects are a result of insufficient shielding of drain electric field lines by the gates. Fig. 2.8 shows the simulated conduction band profiles in the vertical direction at the lateral distance corresponding to the location of the source barrier peak.

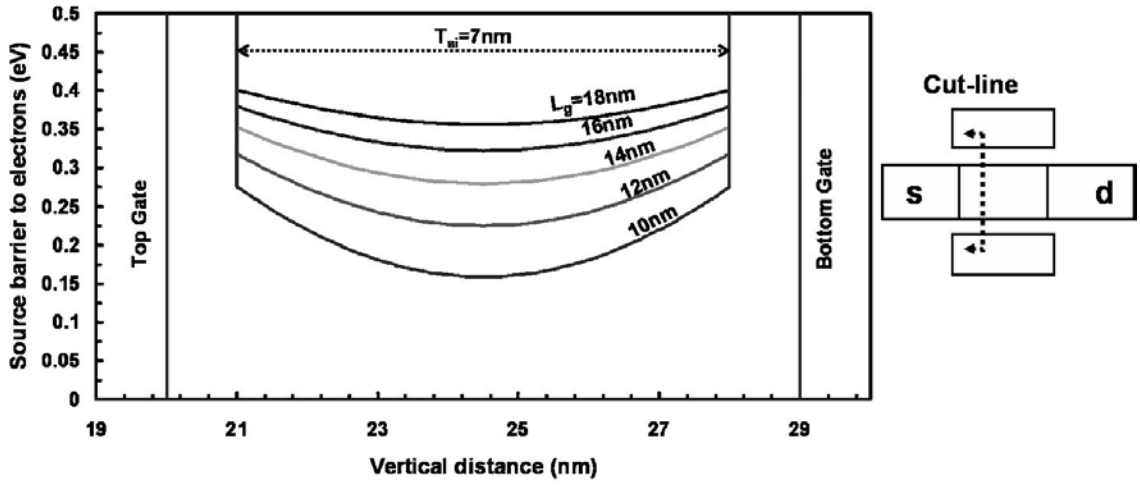


Fig. 2.8 Off-state source barrier lowering as a result of reduced channel length for a DG FET. This leads to short channel effects.

This is along same cut-line as in Fig. 2.6, however in this case, the gate voltage is held fixed at 0 V (representing the off-state condition). The different curves show the effect of reducing the gate length on the source barrier as seen across the body thickness. As the gate length is reduced, the barrier falls faster in the center of the body than at the surface (Fig. 2.9). The weakest point from the perspective of the drain and gate competition for barrier control is thus at the center of the body. This is intuitively expected. Of all the points along the silicon body thickness, the center is furthest away from gate shielding. It is therefore necessary to have an ultrathin body so as to maximize the gate electrostatic control over the entire depth of the body thickness.

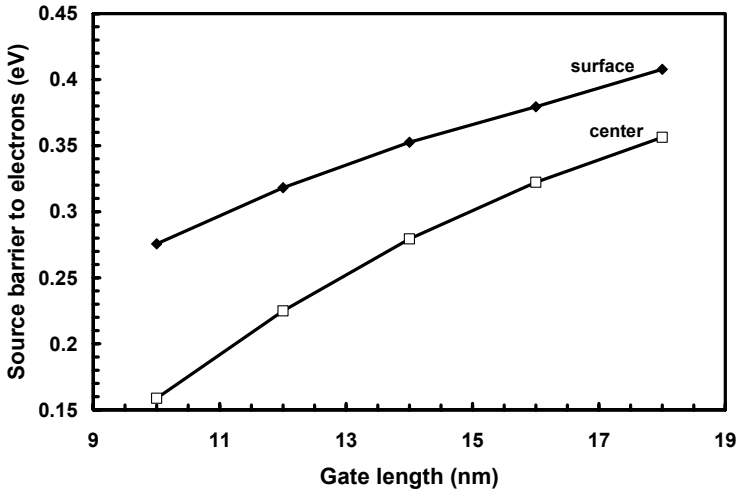


Fig. 2.9 Comparison of the off-state source barrier reduction at the surface and body center as the gate length is reduced in a DG FET.

2.3.3 Importance of Body Thickness (t_{si})

A number of groups [2.16, 2.23, 2.24] have used evanescent-mode analysis to theoretically describe short channel effects due to 2-D electrostatics in bulk, FD SOI, DG FET, and surrounding-gate transistors. In this approach, the channel potential $\psi(x,y)$ is written as the sum of a long channel potential $\psi_L(y)$ and a short channel perturbation term $\psi^*(x,y)$ where the spatial coordinates x and y are taken along the horizontal (from source to drain) and vertical (from top gate to bottom gate) directions respectively.

$$\psi(x,y) = \psi_L(y) + \psi^*(x,y) \tag{2.1}$$

The long channel term $\psi_L(y)$ is assumed to vary only along the vertical direction, in accordance with the gradual channel approximation. It satisfies the Poisson equation and the boundary conditions set by the gate bias. The channel doping, if any, can be accounted for by $\psi_L(y)$. The short channel potential perturbation term $\psi^*(x,y)$ can therefore be written as a solution of the Laplace equation,

$$\nabla^2\psi^*(x,y) = 0 \tag{2.2}$$

For a symmetric DG FET, by writing the perturbation potential $\psi^*(x,y)$ as a Fourier sum of evanescent modes with characteristic decay lengths, and then retaining only the lowest order (most slowly decaying) mode, $\psi^*(x,y)$ can be approximated by the equation,

$$\psi^*(x,y) \approx A_1 \cos(\pi y/\Lambda)\cdot[B_{1+}\exp(\pi x/\Lambda) + B_{1-}\exp(-\pi x/\Lambda)] \tag{2.3}$$

where Λ is called the scale length. It is a measure of how quickly the perturbing short channel potential from the source and drain decays along the channel length direction. By using the appropriate boundary conditions at the top and bottom oxide/silicon interfaces, an implicit equation can be written for Λ,

$$\frac{\varepsilon_{si}}{\varepsilon_{ox}}\tan\left(\frac{\pi t_{ox}}{\Lambda}\right)\cdot\tan\left(\frac{\pi t_{si}}{2\Lambda}\right) = 1 \tag{2.4}$$

where ε_{si} and ε_{ox} are the dielectric permittivities of the channel semiconductor and gate insulator of thickness t_{si} and t_{ox} respectively. For the typical case of $t_{ox} << t_{si}$, eq. (2.4) can be simplified to get a closed form approximate expression for Λ,

$$\Lambda = t_{si} + 2\frac{\varepsilon_{si}}{\varepsilon_{ox}}t_{ox} \tag{2.5}$$

This shows that the vertical confinement of the DG FET, in terms of oxide thickness t_{ox} and silicon body thickness t_{si} directly determines the scale length Λ.

The short channel effects such as threshold voltage roll-off, DIBL, and subthreshold swing degradation depend upon the ratio (L/Λ) of the effective channel length L and the scale length Λ in an exponential manner. As shown in Fig. 2.10, from Frank et. al.

[2.25], for (L/Λ) >>1, the gradual channel approximation is valid, i.e. the FET behaves like an ideal long channel transistor. However, for small (L/Λ), there is substantial degradation of the short channel effects due to the onset of 2-D (drain vs. gate) electrostatics.

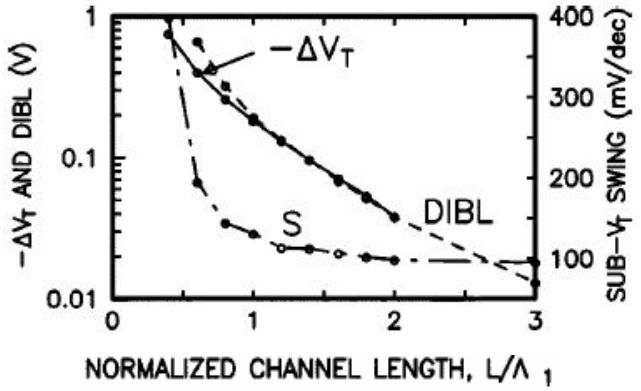


Fig. 2.10 Degradation of short channel effects (V_T roll-off, DIBL, and subthreshold swing increase) as the ratio (L/Λ) decreases. (from Frank et. al. [2.25])

It is clear that reducing the body thickness t_{si} directly reduces the scale length Λ, through eq. (2.5) and hence permits operation at shorter channel lengths. In the DG FET, the body thickness dimension is an additional tuning parameter available to the device designer. It allows for continued device scaling even without necessarily scaling the gate oxide thickness t_{ox}. This is important in light of the fact that tunneling currents limit the physical thickness scaling of silicon dioxide, and high-k gate dielectrics have many (as yet) unsolved problems that preclude their use in manufacturing.

Fig. 2.11 shows device simulations from Wong et. al. [2.26] that plot the extent of DIBL as a function of gate length for a symmetric DG FET with various values of equivalent oxide thickness t_{eq} (same as t_{ox}) and silicon body thickness (t_{si}). As expected from the analytical models, for any given acceptable DIBL, scaling t_{eq} and/or t_{si} down allows for operation at shorter gate length.

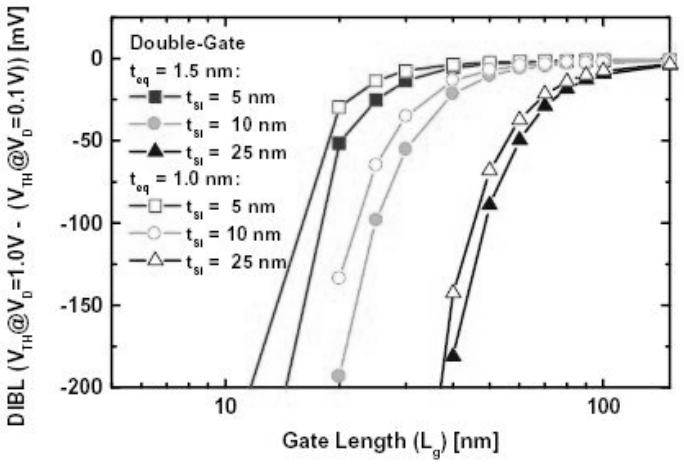


Fig. 2.11 Simulated plots of DIBL as a function of gate length for a DG FET with various oxide thickness t_{eq} and body thickness t_{si}. (From Wong et. al. [2.26])

Fig. 2.12, also from Wong et. al. [2.26] plots the simulated amount of threshold voltage roll-off as a function of gate length for a DG FET and single gate FD SOI transistor with different body thicknesses. It is clear that in both devices, it helps to reduce t_{si}. At any given t_{si}, the improved electrostatic control due to the second gate makes the DG FET more scalable to shorter gate lengths than the single gate FD SOI device.

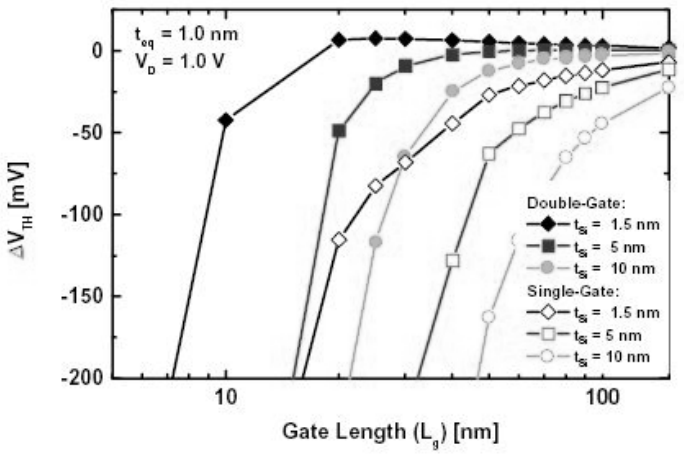


Fig. 2.12 Simulated plots of V_T roll-off with reduced gate length comparing single-gate FD SOI and DG FET for various body thickness t_{si}. (From Wong et. al. [2.26])

Even though the ultrathin body is an enabler for scaling DG FETs, it is necessary to control the body thickness t_{si} with a high degree of precision. This is because the short channel effects, and hence off-state leakage current, are strongly dependent on t_{si}. Any process-induced variations in t_{si} across the chip therefore result in amplified variations in the standby off-state leakage. Fig. 2.13 shows Medici device simulations plotting the off-state leakage current as a function of t_{si}. Even a small increase in t_{si} can cause a huge increase in the off-state leakage current. The sensitivity of the leakage current to t_{si} increases as the gate length is reduced. This is due to operation at a lower (L/Λ) ratio.

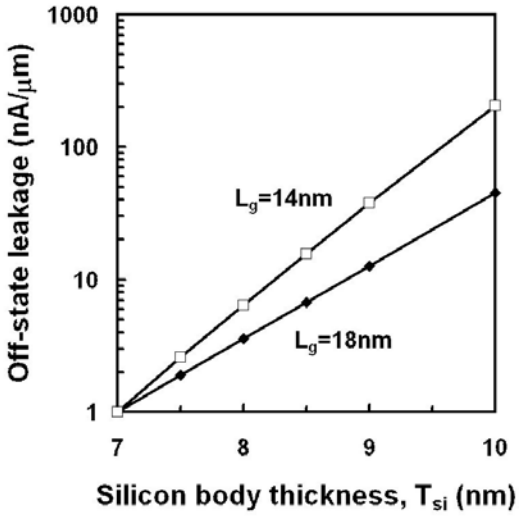


Fig. 2.13 Medici-simulated plots showing the sensitivity of off-state leakage to the body thickness for DG FETs with gate length as a parameter. In both cases, the gate workfunction has been optimized to give the same off-state leakage current at t_{si} = 7 nm. The gate oxide thickness is 1 nm and the drain bias is 0.9 V.

From an intrinsic device performance viewpoint, it is thus clear that one of the key requirements for a DG FET is a uniform ultrathin body. Depending upon the level of tolerance to short channel effects, the ratio of the body thickness to the gate length has to be chosen appropriately. For typical DG FETs, it is necessary to have the body thickness

less than half the gate length. This means that in DG FETs, the body thickness replaces the gate length as the most critical (minimum) dimension in the device.

2.3.4 Ultrathin Body Scaling Limits

Theoretical models of DG FET 2-D electrostatics based on evanescent-mode analysis indicate that it is always better to reduce the body thickness t_{si} for improved gate control and reduced scale length Λ, which in turn enables gate length scaling. However, as t_{si} is reduced below 5 nm, quantum effects due to size confinement in the ultrathin body become non-negligible. These can impact the threshold voltage and mobility in a negative way.

For a very thin body, the quantum confinement increases the ground state energy of the bands. The increase is inversely proportional to t_{si}^2 similar to the case of a simple particle-in-a-box. This directly results in an increase of the threshold voltage V_T. If the body thickness could be controlled with an infinite degree of precision, the increased V_T due to quantum confinement could in principle be compensated for by adjusting the gate workfunction suitably. However, due to the strong inverse square dependence, even small variations in t_{si} can result in large shifts in V_T. According to Wong et. al. [2.27], the variation in threshold voltage, σV_T can be written as,

$$\sigma V_T = -\left(\frac{\hbar^2 \pi^2}{q m^* t_{si}^2}\right) \cdot \left(\frac{\sigma t_{si}}{t_{si}}\right) \tag{2.6}$$

where m^* is the effective mass in the direction of quantization, and the other symbols have their usual meanings.

For a t_{si} of 4 nm with 20 % thickness variation, it leads to a threshold voltage variation of 50 mV, which is unacceptably high. Fig. 2.14, taken from Wong et. al. [2.27] shows the V_T dependence and its sensitivity to t_{si} for ultrathin body transistors. The solid lines are the results of classical calculations while the broken lines include the quantum confinement effects.

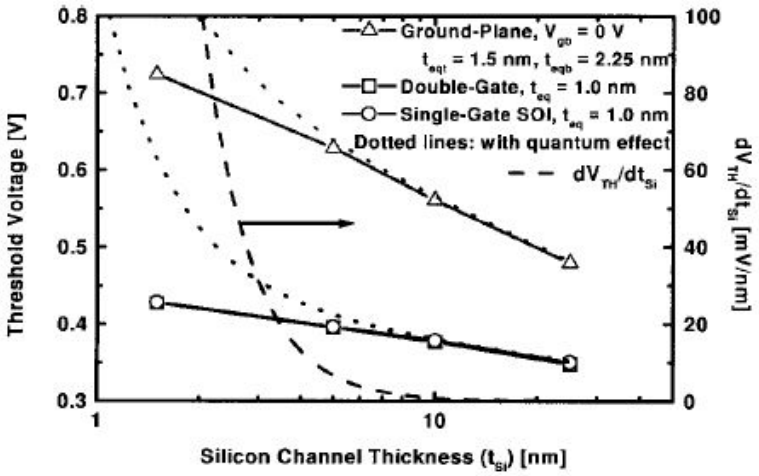


Fig. 2.14 Plot of calculated threshold voltage (left axis) and threshold sensitivity to body thickness variation as a function of the nominal body thickness t_{si}. The broken lines include the effect of quantum confinement. (From Wong et. al. [2.27])

As mentioned earlier, one of the potential advantages of the undoped body FD SOI and DG FET structures is increased carrier mobility. The absence of body doping (and presence of vertical symmetry in the case of the DG FET) causes the effective transverse electric field, E_{eff} to reduce for the same inversion carrier concentration. This enables device operation in a higher region of the universal mobility curve [2.28]. However, as the body thickness is reduced in a FD SOI transistor or DG FET, the mobility drops. Fig. 2.15 from Uchida et. al. [2.29] shows the measured hole mobility as a function of the body thickness t_{si}. The monotonic reduction in hole mobility has been attributed to increased phonon scattering as the t_{si} is reduced.

The electron mobility, on the other hand, shows an interesting non-monotonic behavior. This is shown in the measured data of Uchida et. al. [2.29] (Fig. 2.16), where the electron mobility is seen to first decrease as t_{si} is reduced down to 4.5 nm, increase slightly in the 4.5 to 3.5 nm range, and then sharply decrease for t_{si} lower than 3.5 nm. This is in accordance with theoretical predictions by Takagi et. al. [2.30], where the

electron mobility enhancement in that t_{si} range is explained as being due to sub-band level modulation by quantum confinement effects in the ultrathin silicon film.

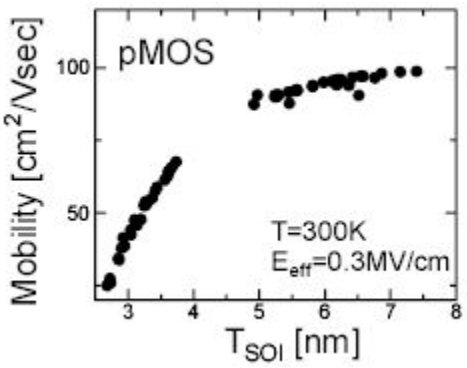


Fig. 2.15 Measured hole mobility as a function of silicon body thickness in a single-gate FD SOI transistor (From Uchida et. al. [2.29])

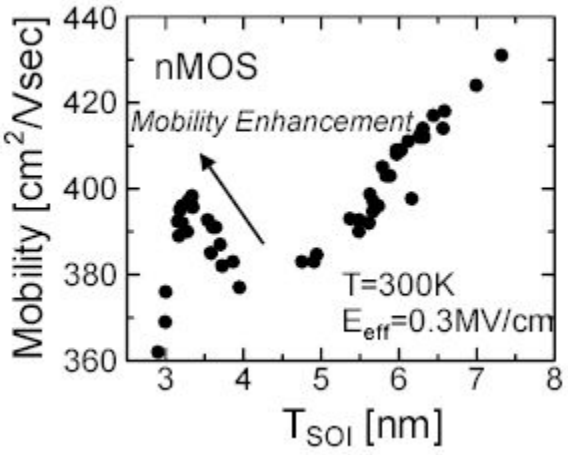


Fig. 2.16 Measured electron mobility as a function of silicon body thickness in a single gate FD SOI transistor (From Uchida et. al. [2.29])

In addition to the phonon scattering, Uchida et. al. [2.29] also measured the mobility reduction in very thin Si films due to t_{si} variation-induced scattering. In ultrathin films, the thickness variations result in potential fluctuations due to spatially-varying ground

state energy in the quantum-confined body. These potential fluctuations lead to a mobility component that is proportional to t_{si}^{6}. Unless the thickness is controlled to a precision corresponding to a few atomic layers, the thickness fluctuation can degrade mobility in Si films with t_{si} less than 5 nm.

2.4 Summary

The increasing difficulty in scaling gate lengths for future generations of conventional bulk FETs has led to the evolution of various advanced transistor structures. Single-gate FD SOI and double-gate FETs offer a path to extend transistor scaling. In these devices, an undoped ultrathin body and multiple gates are key to achieving better electrostatic gate control over the channel for the suppression of short channel effects. The role of body thickness (t_{si}) scaling is explained by theoretical models using evanescent-mode analysis and 2-D device simulations. From an intrinsic device operation viewpoint, one of the main requirements in DG FETs is a uniform and ultrathin body. The lower limit on body thickness scaling is dictated by the onset of quantum confinement effects which cause the threshold voltage and mobility to become very sensitive to t_{si} variations.

References

[2.1] International Technology Roadmap for Semiconductors, 2003 edition, SIA.

[2.2] M. M. Pelella et. al., "Advantages and challenges of high performance CMOS on SOI," in *2001 IEEE International SOI Conference Proceedings*, 2001, pp. 1-4.

[2.3] H.-F. Wei, J. E. Chung, N. M. Kalkhoran, and F. Namavar, "Suppression of parasitic bipolar effects and off-state leakage in fully-depleted SOI n-MOSFET's using Ge-implantation," *IEEE Transactions on Electron Devices*, Vol. 42, No. 12, pp. 2096-2103, Dec 1995.

[2.4] J. Sleight and K. Mistry, "A compact Schottky body contact technology for SOI transistors," in *IEDM Technical Digest*, 1997, pp. 419-422.

[2.5] G. Shahidi, "SOI technology for the GHz era," *IBM Journal of Research and Development*, Vol. 46, Nos. 2/3, pp. 121-132, 2002.

[2.6] M. Horstmann et. al., "Advanced transistor structures for high performance microprocessors," in *2004 International Conference on Integrated Circuit Design and Technology*, 2004, pp. 65-71.

[2.7] D. Hisamoto et. al., "A compact FD-SOI MOSFETs fabrication process featuring Si_xGe_{1-x} gate and damascene-dummy SAC," in *Digest of Technical Papers - Symposium on VLSI Technology*, 2000, pp. 208-209.

[2.8] R. Chau et. al., "A 50nm depleted-substrate CMOS transistor (DST)," in *IEDM Technical Digest*, 2001, pp. 621-624.

[2.9] C. Auth, "Physics and technology of vertical surrounding gate MOSFETs," Ph. D. Thesis, Stanford University, 1998.

[2.10] D. Hisamoto et. al., "Folded-channel MOSFET for deep-sub-tenth micron era," in *IEDM Technical Digest*, 1998, pp. 1032-1034.

[2.11] H.-S. P. Wong, K. K. Chan, and Y. Taur, "Self-aligned (top and bottom) double-gate MOSFET with a 25 nm thick silicon channel," in *IEDM Technical Digest*, 1997, pp. 427-430.

[2.12] K. Kim and J. G. Fossum, "Optimal double-gate MOSFETs: Symmetrical or asymmetrical gates?" in *1999 IEEE International SOI Conference Proceedings*, 1999, pp. 98-99.

[2.13] H.-S. P. Wong, "Double-gate FET – device design and performance analysis," in *Short Course, 2000 IEEE International SOI Conference*, 2000.

[2.14] L. Chang, S. Tang, T.-J. King, J. Bokor, and C. Hu, "Gate length scaling and threshold-voltage control of double-gate MOSFETs," in *IEDM Technical Digest*, 2000, pp. 719-722.

[2.15] Q. Lu, R. Lin, P. Ranade, T.-J. King, and C. Hu, "Metal gate workfunction adjustment for future CMOS technology," in *Digest of Technical Papers – Symposium on VLSI Technology*, 2001, pp. 45-46.

[2.16] S.-H. Oh, D. Monroe, and J. M. Hergenrother, "Analytic description of short-channel effects in fully-depleted double-gate and cylindrical, surrounding-gate MOSFETs," *IEEE Electron Device Letters*, Vol. 21, No. 9, pp. 445-447, Sep 2000.

[2.17] B. Doyle et. al., "Tri-gate fully-depleted CMOS transistors: fabrication, design and layout," in *Digest of Technical Papers – Symposium on VLSI Technology*, 2003, pp. 133-134.

[2.18] F.-L. Yang et. al., "25 nm CMOS omega FETs," in *IEDM Technical Digest*, 2002, pp. 255-258.

[2.19] F.-L. Yang et. al., "5nm-gate nanowire FinFET," in *Digest of Technical Papers – Symposium on VLSI Technology*, 2004, pp. 196-197.

[2.20] A. Javey, Q. Wang, W. Kim, and H. Dai, "Advancements in complementary carbon nanotube field-effect transistors," in *IEDM Technical Digest*, 2003, pp. 741-744.

[2.21] Synopsys Corporation, Mountain View, CA, *Medici version 2003.12 User Guide*, 2003.

[2.22] J. P. Colinge, "Silicon-on-insulator technology: materials to VLSI," 2nd edition, Kluwer Academic Publishers, Boston, 1997.

[2.23] D. Monroe and J. M. Hergenrother, "Evanescent-mode analysis of short-channel effects in fully-depleted SOI and related MOSFETs," in *1998 IEEE International SOI Conference Proceedings*, 1998, pp. 157-158.

[2.24] D. J. Frank, Y. Taur, and H.-S. P. Wong, "Generalized scale length for two-dimensional effects in MOSFETs," *IEEE Electron Device Letters*, vol. 19, no. 10, pp. 385-387, Oct 1998.

[2.25] D. J. Frank, R. H. Dennard, E. Nowak, P. M. Solomon, Y. Taur, and H.-S. P. Wong, "Device scaling limits of Si MOSFETs and their application dependencies," *Proceedings of the IEEE*, vol. 89, no. 3, pp. 259-288, Mar 2001.

[2.26] H.-S. P. Wong, D. J. Frank, and P. M. Solomon, "Device design considerations for double-gate, ground-plane, and single-gated ultra-thin SOI MOSFET's at the 25 nm channel length generation," in *IEDM Technical Digest*, 1998, pp. 407-410.

[2.27] H.-S. P. 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, Apr 1999.

[2.28] S.-I. Takagi, A. Toriumi, M. Iwase, and H. Tango, "On the universality of inversion layer mobility in Si MOSFET's: Part 1-Effects of substrate impurity concentration," *IEEE Transactions on Electron Devices*, vol. 41, no. 12, pp. 2357-2362, Dec 1994.

[2.29] K. Uchida, H. Watanabe, A. Kinoshita, J. Koga, T. Numata, and S.-I. Takagi, "Experimental study on carrier transport mechanism in ultrathin-body SOI n- and p-MOSFETs with SOI thickness less than 5 nm," in *IEDM Technical Digest*, 2002, pp. 47-50.

[2.30] S. Takagi, J. Koga, and A. Toriumi, "Subband structure engineering for performance enhancement of Si MOSFETs," in *IEDM Technical Digest*, 1997, pp. 219-222.

Chapter 3

Double-Gate FET Device Performance – Extrinsic Factors

3.1 Introduction

The intrinsic performance of a transistor can be improved by engineering its structure and/or materials for better electrostatics and carrier transport. However, it is possible, and usually the case, that the overall device performance is limited by extrinsic factors. Parasitic capacitance and resistance are the two main quantities that cause a real transistor to switch more slowly and dissipate higher power than expected from the purely intrinsic device. It is therefore very important to understand the ways in which these extrinsic effects degrade performance. The insights thus gained can be used during device design so as to minimize the impact of parasitic elements.

In this chapter, the main sources of parasitic capacitance and resistance in double-gate FETs are examined through device simulations. First, parasitic gate capacitance and its effects on inverter delay are studied. Next, there is an extensive discussion of series resistance due to non-ideal contacts and finite doping in the ultrathin source/drain extension regions used to access the channels. The chapter concludes with a section on

Schottky source/drain double-gate FETs, in which electrical contact is made to the channel via metal-semiconductor Schottky junctions instead of using conventional p-n junctions.

3.2 Parasitic Capacitance

As discussed in the previous chapter, digital circuit switching speed is dictated by how fast various node capacitances can be charged and discharged by the transistors. The node capacitances originate from transistors as well as from the interconnects between them. Suitably sized buffer chains of logic gates are used to drive high capacitance wires. This causes most logic gates to be device-loaded, i.e. the load at their output is mainly due to transistor capacitance. There are three major components of this capacitance: 1) gate to channel, 2) drain to substrate, and 3) gate to drain/source (overlap and fringe components). Of these, the first is intrinsic to any FET since it directly contributes to channel charge and thus drive current. The second, while relatively significant in bulk MOSFETs, is not important in most double-gate FETs due to the presence of a thick buried oxide layer between the drain and the substrate wafer. The last component is the major source of parasitic capacitance in DG FETs. In planar-type DG FETs, where the two gates are placed above and below a horizontally-oriented thin silicon layer, any misalignment between top and bottom gates can lead to increased gate to source/drain capacitance. Also, in some implementations of planar DG FETs, the bottom gate may be sized somewhat differently than the top gate. This too leads to increased parasitic gate overlap capacitance. Finally, in DG FETs with source/drain regions that are flared out to reduce series resistance, if the insulating spacer separating the gate electrode from the source/drain is too thin, that can lead to excessive parallel-plate type gate to source/drain capacitance along the gate sidewall. These cases will be examined one by one using device simulations. The simulated structure is based on an ideal 18 nm gate length DG FET shown in Fig. 3.1. For the moment, abrupt junctions are assumed along with ideal

(zero resistance) contacts to the source and drain. A gate workfunction of 4.5 eV is used for the NMOS DG FET. That results in an off-state leakage of about 100 nA/μm. A gate workfunction of 4.96 eV for the PMOS DG FET results in somewhat similar off-state leakage. The supply voltage is set to 0.9 V.

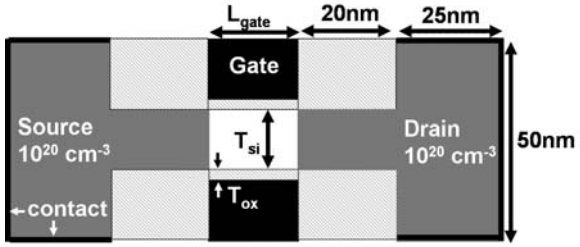


Fig. 3.1 Schematic of the simulated ideal DG FET structure. The gates are perfectly self-aligned and equal sized. In the simulations that follow, the effects of deviations in the positions or size of the bottom gate will be examined. L_{gate}=18nm, T_{ox}=1nm, T_{si}=7nm.

Deviations in the positions or sizes of the gates typically result in changes in drive current and off-state leakage as compared to the ideal baseline device. In order to capture the effects of parasitic capacitance on switching speed, mixed-mode (device + circuit) simulations are performed in Medici using a chain of 3 equal sized inverters shown in Fig. 3.2.

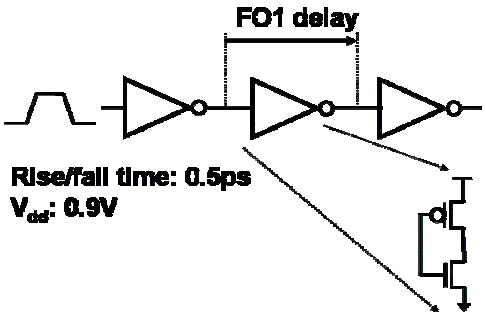


Fig. 3.2 Schematic of simulated inverter chain for mixed mode FO1 delay calculations.

The average of rising and falling delay from input to output of the second stage is reported as the fanout-of-1 (FO1) inverter delay. This value is a measure of the transistor switching speed in a real circuit.

3.2.1 Effect of Top-Bottom Gate Misalignment

One of the major challenges in building planar DG FETs is finding a way to align the gates to one another. Usually, the top gate acts as a mask for the ion implantation for source/drain doping. Thus the source and drain are self-aligned to the top gate. If the bottom gate is misaligned with respect to the top gate, it will overlap either the source or the drain by some amount. Fig. 3.3 shows a plot of FO1 inverter delay and NMOS off state leakage as a function of the bottom gate misalignment (taken to be positive for misalignment towards the drain.)

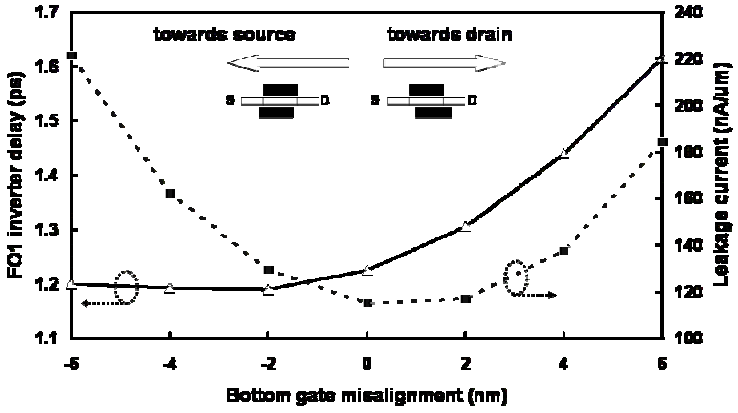


Fig. 3.3 Impact of the bottom gate misalignment on the FO1 inverter delay and the off-state leakage current.

For misalignment towards the source, the FO1 delay is mostly unchanged, but the off-state leakage increases. For misalignment towards the drain, both FO1 delay as well as off-state leakage increase. The drain-side misalignment impacts the delay in two ways: first, there is an offset between the bottom gate and the source extension doping. This

leads to high source resistance and effectively decreases the gate overdrive of the bottom channel, reducing the drive current. In addition, the drain to bottom gate capacitance is increased. In an inverter, this has a much more significant impact on delay than source to gate capacitance (due to the Miller effect.) For the case of source-side misalignment, the drive current isn't affected to first order since it is mostly limited by the effective injection velocity at the source. A small offset between the gate and drain doping does not affect the carriers near the drain which are anyway traveling at saturation velocity in the high-field region there. In fact, the FO1 delay slightly reduces as a result of decreased Miller capacitance between the bottom gate and drain. In both cases, the increased off-state leakage current is due to poorer shielding of the drain electric field lines by the bottom gate. This makes it easier for the drain bias to lower the barrier near the source (DIBL), causing the leakage current over that barrier to increase.

3.2.2 Effect of Bottom Gate Sizing

In some planar DG FET implementations, the bottom gate is not the same size as the top gate. This may be the case even though the centers of the two gates are the same – making them, in a sense, 'self-aligned' to each other. For instance, the IBM 'Pagoda FET' [3.1] has a bottom gate that may be smaller than the top gate. On the other hand, the planar DG FET fabricated by the Silicon-on-Nothing (SON) process [3.2] has a bottom gate that is much larger than the top gate. Fig. 3.4 shows a plot of FO1 inverter delay and off-state leakage as a function of the bottom gate size mismatch. The sign of the mismatch is taken to be positive if the bottom gate is larger than the top gate. In the case of such oversized bottom gates, the delay increases while the off state leakage remains unchanged (slightly decreases). For the undersized bottom gate, the delay is somewhat less. However, the off-state leakage increases almost by a factor of 5, negating the apparent advantage of such a configuration. Once again, the reason for these effects is easy to understand. The oversized bottom gate provides better shielding of drain field-lines resulting in slightly better electrostatic gate control of the channel. Thus the off-state

leakage current is low. However, the extra gate overlap over source and drain causes increased parasitic capacitance, and hence higher delay. The undersized bottom gate has much lower bottom gate capacitance, both intrinsic as well as gate to drain. This offsets the reduced bottom channel drive current due to the offset between source doping and bottom gate and results in a reduced inverter delay. However, this is achieved at the cost of much worse electrostatic shielding from the smaller bottom gate. This degrades the subthreshold swing and DIBL, leading to higher off-state leakage current.

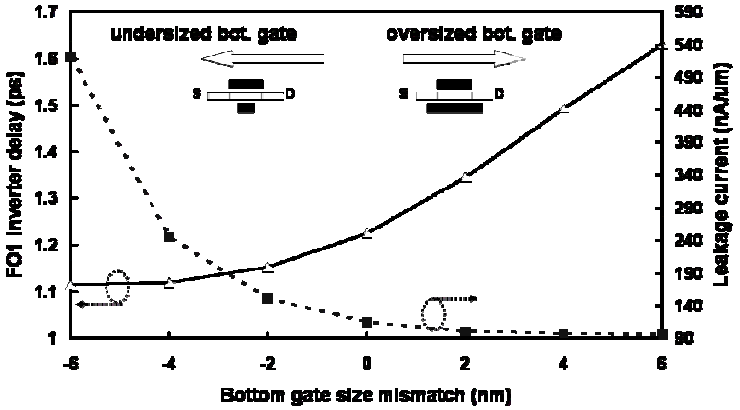


Fig. 3.4 Effect of the bottom gate mis-sizing on the FO1 inverter delay and the off-state leakage current.

3.2.3 Effect of Gate-Source/Drain Spacer Thickness

It is difficult to form a silicide and make a low resistance ohmic contact to the ultrathin body in single or double-gate FETs. Therefore it is desirable to have raised or flared-out source/drain regions that are typically grown by a selective epitaxy step prior to contact formation. During this step and before salicide formation, it is necessary to have an insulating sidewall spacer that prevents gate to source/drain short circuits due to silicon or silicide bridging. If this sidewall spacer is too thin, it leads to increased gate to source/drain overlap capacitance due to the parallel plate capacitor formed along the transistor width and height of the gate electrode (or raised source/drain, whichever is smaller.) Fig. 3.5 shows a plot of FO1 inverter delay as a function of sidewall spacer

thickness. In this case, the aspect ratio of the source/drain overlapping gate electrode is roughly 1:1. It is seen that for spacer thickness less than 10 nm, the delay begins to increase, going up steeply below 6 nm. This points to the need for having sidewall spacers that are much thicker than the effective gate oxide.

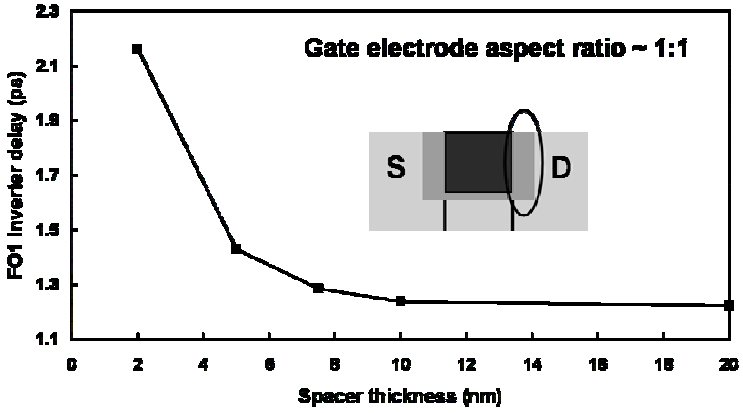


Fig. 3.5 Effect of the gate to source/drain sidewall spacer thickness on the FO1 inverter delay.

Some implementations of planar DG FETs [3.3, 3.4] do not meet this requirement since the sidewall spacer is about the same thickness as the gate oxide. A reduction in the sidewall height to gate length aspect ratio could mitigate this problem somewhat. However, those solutions have the associated drawbacks of increased contact resistance or increased gate sheet resistance. The off-state leakage, not shown here, does not change. This is to be expected since there is no change in gate control (relative to the drain) over the source barrier.

3.3 Parasitic Resistance

We have seen that, in DG FETs, the ultrathin body, whose thickness is typically 1/3 to 1/2 of the gate length, is key to suppressing short channel effects such as V_T roll-off, drain-induced barrier lowering (DIBL), and degraded subthreshold swing. However, it

also introduces an extrinsic parasitic resistance R_s in series with the channel and the source/drain electrodes. The effective gate overdrive is reduced by an amount $I_d \cdot R_s$, where I_d is the drain-source current when the transistor is turned on and in saturation. As a result, the transconductance and performance, as measured by drive current I_{on} and intrinsic switching delay (CV/I_{on}), is degraded even though the intrinsic device has nearly ballistic carrier transport [3.5]. This problem is even more severe in a DG FET since the presence of two channels implies that twice the current flows through the series resistance, leading to higher potential drop across the extrinsic resistance. Previous work on analysis [3.6, 3.7] of extrinsic resistance has been focused on the conventional planar bulk MOSFET structure. In [3.7], the authors suggest the existence of optimal lateral abruptness of the source/drain extension doping profile. It is not clear whether the same results and conclusions would hold for advanced ultrathin body DG FET structures where the process constraints and tunable parameters are quite different. In this section, we use 2-D device simulations to study and optimize the extrinsic resistance in an N-channel DG FET designed in a scaled technology.

3.3.1 Device Design and Simulation Setup

Fig. 3.6 shows a schematic cross-section of the device simulated in Medici [3.8].

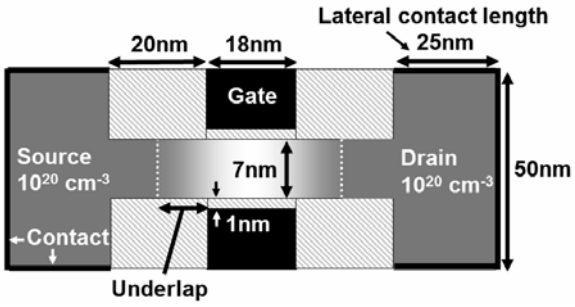


Fig. 3.6 Schematic cross-section of the simulated double-gate FET

Values from the 45 nm node high performance logic device of the ITRS 2001 [3.9] are used to guide device design. Most of the device dimensions and doping values are

chosen to be as close as possible to a novel planar DG FET structure that will be described in subsequent chapters. An 18 nm gate length is used. Neither gate tunneling nor inversion layer quantization are modeled here. Therefore, the gate dielectric thickness T_{ox} value of 1 nm should be interpreted as the equivalent electrical T_{ox} extracted in inversion. In order to achieve this at acceptable gate current levels, a high-K dielectric material might be needed [3.10]. The silicon channel thickness T_{si} is set to 7 nm. In the section on optimization, this is varied from 9 nm down to 5 nm. Values below 5 nm are not simulated since in those cases neglecting quantum effects would introduce significant errors. Symmetric workfunction metal gates are used to set the threshold voltage. The gate workfunction Φ_m is assumed to be continuously variable within the silicon bandgap. This assumption may be justified based on experimental results on tunable gate workfunction [3.11, 3.12] and fully-silicided gates [3.13].

A flared out source/drain structure with constant 10^{20} cm^{-3} N-type doping is used along with a surrounding contact geometry. The specific contact resistivity ρ_c is chosen to be 5 Ω-μm^2 and a distributed contact resistance model is employed. At a certain distance, defined as the 'underlap', away from the gate edge, the constant doping rolls off laterally into the channel with a Gaussian profile whose abruptness is characterized by the lateral doping gradient LDG. The doping profile in this ultrathin extension region is the key component that is optimized in this study. The channel is nearly undoped, with constant 10^{14} cm^{-3} P-type doping. Fig. 3.7 illustrates the lateral extension doping profile for three values of LDG. Low values of LDG indicate more abrupt lateral doping profiles. Though only the part near the source-side is shown, the doping is assumed to be laterally symmetric, and the drain-side lateral doping profile is similar.

Carrier transport is modeled using a drift-diffusion model (LUCMOB) based on work by Darwish [3.14]. In general drift-diffusion models are known to underestimate the drive current in MOSFETs at deep submicron geometries [3.15]. Therefore the absolute values of drive current in these simulations may not be accurate. However, the trends with varying extrinsic resistance are captured well. One expects that inclusion of more

accurate transport models (such as energy-balance or Monte Carlo) would make the impact of extrinsic resistance even more severe since the best case drive current (low parasitic resistance) will increase while the parasitic resistance-limited drive current will not change much across different transport models.

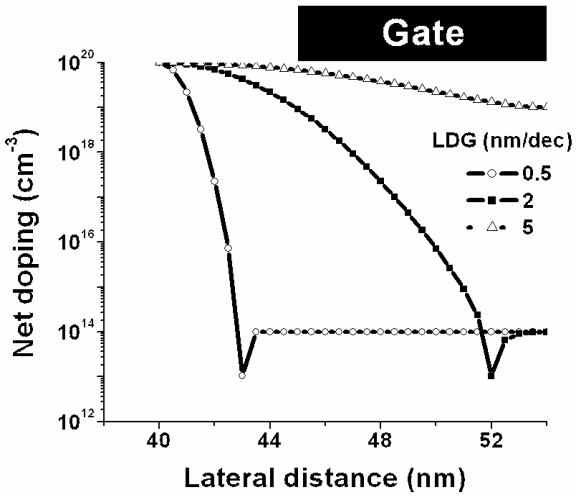


Fig. 3.7 Lateral doping profile in the source extension region for 3 values of lateral doping gradient (LDG).

3.3.2 Effect of Contact Resistance

In this section, we examine the impact of contact resistance on the drive current of the DG FET. There are two ways to reduce the total contact resistance. One way is to lower the barrier from the contact material to the doped silicon. Increasing the doping concentration at the contact-silicon interface thins the barrier and makes the tunneling process easier, which can be interpreted as a lowering of the effective barrier height. Alternatively, the actual barrier height to carriers can be changed by choosing a different metal (or silicide) as the contact material. Either of these approaches essentially results in a lowering of the specific contact resistivity ρ_c. The other way to reduce the total contact resistance is to increase the area of the contact. This method is effective as long as the

contact length is approximately on the scale of (less than) the contact transfer length. The contact transfer length l_t is defined by, $l_t = (\rho_c/R_s)^{1/2}$, where R_s is the sheet resistance of the doped layer through which current flows before reaching the contact. Physically, this is a scale length (in a distributed 1-D model) which corresponds to the effective length of the contact available for current flow. Increasing the contact length far beyond this amount yields diminishing returns due to current crowding effects. In conventional bulk Si MOSFETs, silicide consumption increases the sheet resistance of the junction, causing the contact transfer length to reduce (for the same specific contact resistivity.) There have been efforts to develop elevated source/drain processes [3.16] to alleviate this issue. In the simulated DG FET structure, the flared-out source/drain architecture extends the elevated source/drain concept and results in relatively large contact transfer length. This enables the increase of effective contact area. Fig. 3.8 shows results of simulations that confirm the above statements.

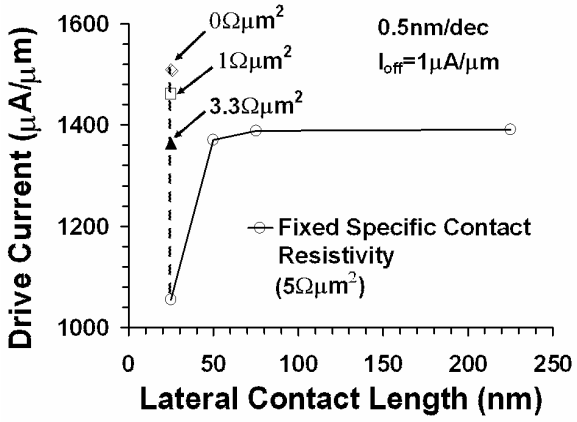


Fig. 3.8 Effect of increasing either the lateral contact length or the specific contact resistivity on the drive current, with fixed off state leakage.

Here, the points with the solid line indicate values of drive current as the extent of the lateral contact is increased from 25 nm to 225 nm. The height of the vertical contact remains fixed at 50 nm. The off-state leakage current I_{off} is kept constant by tuning the

gate workfunction. The drive current I_{on} increases with lateral contact length until around 100 nm, where it saturates. For each of these cases, the value of ρ_c is kept fixed at 5 Ω-μm^2. The points with the broken line show the increase in drive current as the value of ρ_c is lowered down to 0 Ω-μm^2 while keeping the lateral contact extent fixed at 25 nm.

It is seen that simply increasing the contact size can be quite effective in increasing the drive current to a value close to the case where there is no contact resistance ($\rho_c = 0$). In practice, design rules and pitch requirements might constrain the maximum extent of the contact. Therefore, it still makes sense to pursue approaches to lower the ρ_c even though that is a much harder technological problem. This becomes absolutely critical in the Schottky-barrier version of the ultrathin body FET or DG FET where the "contact" size is constrained to be as small as the silicon body thickness T_{si} [3.17, 3.18]. The Schottky source/drain DG FET will be examined in more detail in the last section of this chapter.

In the following sections, unless otherwise specified, contact resistance is kept fixed at a realistic value, assuming ρ_c = 5 Ω-μm^2 and lateral contact length = 25 nm, while we focus on the impact of the lateral doping gradient in the extension region.

3.3.3 I_{on}-I_{off} Comparisons and Discussion

In this section, I_{on}-I_{off} plots are used for benchmarking devices with different extension lateral doping profiles. For a given structure, the gate workfunction is swept, changing the V_T, to generate a plot of I_{off} as a function of I_{on}. Such plots are created for different devices and the I_{on} is compared at some fixed I_{off}. Therefore, the curves further out to the right represent 'better' devices. This approach is similar to that used by Kwong et. al. [3.7]. However, in that work, the authors use a modified I_{on}-I_{off} curve with the sub nominal gate length used for I_{off} and the super nominal gate length used for I_{on}. That approach essentially aims to include process variations instead of comparing devices at a fixed gate length. At the present time, since there is no consensus on which process will eventually be used to fabricate the DG FET in manufacturing, we do not know reasonable

values of gate length variation to use in the simulations. In addition, there will be other factors such as T_{si} variation which may be equally important. For simplicity, we choose to omit device variations from the present study. It should be noted however that the aforementioned dual gate length benchmarking approach rewards devices which have superior electrostatics and hence greater immunity to short-channel effects.

Fig. 3.9 shows a typical set of I_{on}-I_{off} curves as the doping abruptness is varied for a fixed underlap of 5 nm. Qualitatively, one can notice two main features. First, as the LDG is reduced from 5 nm/dec. to 3.5 nm/dec., the devices get better. Below 2 nm/dec. however, the curves bend steeply upwards, indicating that the I_{on} does not change much with reducing V_T even though the leakage current increases exponentially. This is a signature of a current-limiting series resistance.

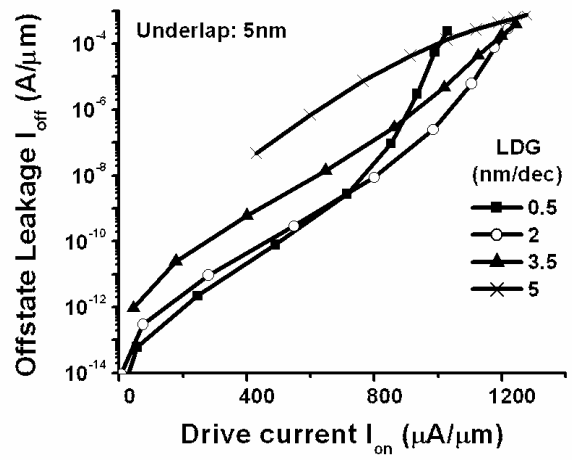


Fig. 3.9 I_{on}-I_{off} curves for a fixed 5 nm extension underlap and varying lateral doping gradients.

Such a set of I_{on}-I_{off} curves may be useful in inverse modeling of experimental transistors where the lateral doping profile needs to be extracted. However from a device design perspective, it makes more sense to examine I_{on}-I_{off} curves where devices with the same LDG and different underlaps are compared. This is because typically there isn't much control over the minimum value of LDG – it is set by the thermal budget and the

diffusivity of the dopant species. On the other hand, the underlap can be changed by a number of techniques.

Fig. 3.10 shows I_{on}-I_{off} plots for a case where the LDG is set to 0.5 nm/dec. (very abrupt) and the underlap is varied from 0 nm to 12.5 nm. Such small values of LDG are indeed very aggressive and may need non-standard processes such as solid phase epitaxy [3.19] or laser thermal annealing [3.20] in order to be achievable.

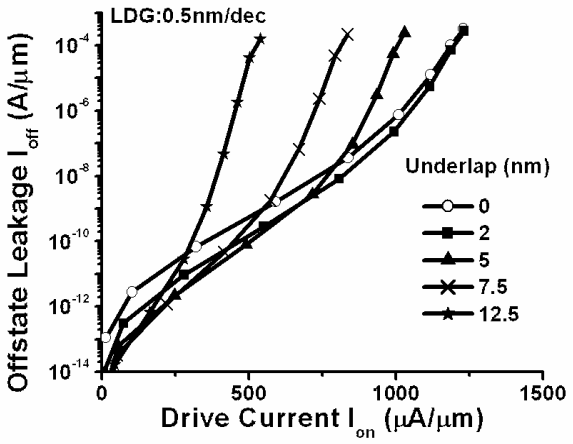


Fig. 3.10 I_{on}-I_{off} curves for a fixed lateral doping gradient of 0.5 nm/dec. and varying extension underlaps.

Fig. 3.11 shows a similar plot where the LDG is set to a much more realistic value of 3.5 nm/dec. These figures show very different pictures in terms of which devices are the best. The device with no underlap, which is nearly the best device for the LDG = 0.5 nm/dec. case, cannot be turned off when the LDG is increased to 3.5 nm/dec.

From these sets of I_{on}-I_{off} curves, one learns that for a given underlap, a lateral doping profile that is too abrupt or too gradual can reduce the drive current. We now look at these cases separately.

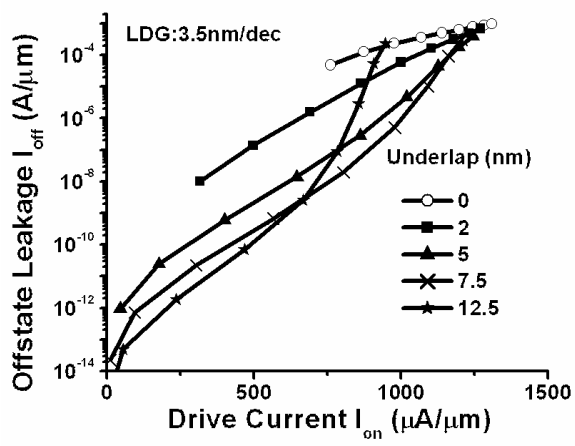


Fig. 3.11 I_{on}-I_{off} curves for a fixed lateral doping gradient of 3.5 nm/dec. and varying extension underlaps.

Case 1. Doping profile too gradual (large LDG)

At large values of LDG (very gradual lateral doping profile,) a significant amount of dopant spills over into the channel. This degrades the short channel performance of the device since the effective channel length is now reduced. Fig. 3.12 shows the subthreshold drain current plots of devices with 2 different values of LDG .

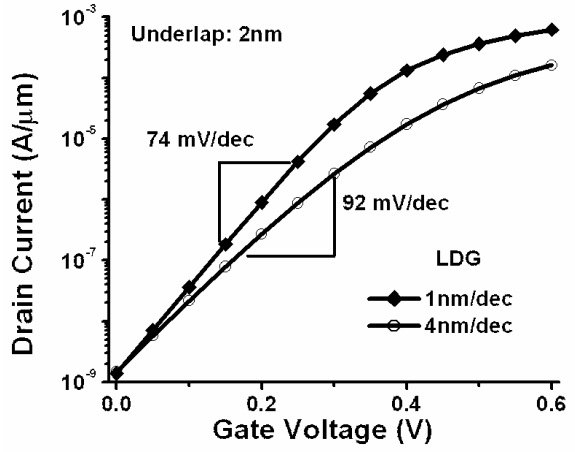


Fig. 3.12 Degradation of subthreshold swing when the lateral doping gradient is too gradual for a given extension underlap (2 nm).

The underlap is fixed to 2 nm and the gate workfunctions of both transistors are adjusted so as to hold the off-state drain leakage current value constant at 1 nA/μm for each device.

It is easily apparent that the subthreshold swing is degraded in the device with higher LDG as compared to the device with more abrupt lateral doping. This leads to lower I_{on} since the I_{off} is fixed. In addition to worse short channel effects, dopant spillover into the channel could degrade the carrier mobility due to ionized impurity scattering. Furthermore, there could also be increased inter-device variations due to discrete dopant effects [3.21]. The main point here is that the introduction of dopants into the channel region negates some of the advantages of the nearly-undoped body and should be avoided.

Case 2. Doping profile too abrupt (small LDG)

On the other hand, if the LDG is too small, the doping profile is too abrupt for the given underlap. This leads to an inadequate supply of carriers from the lightly accumulated extension region into the channel. At higher values of V_T (or low gate overdrive), the inversion charge is not very high and this effect is not significant. As the gate workfunction is decreased, the required inversion charge increases and the thin extension region with insufficient carriers becomes a bottleneck. The drive current does not increase much with higher gate overdrive, causing the I_{on}-I_{off} curve to sharply curve upward. Fig. 3.13 shows plots of doping and electron concentration from the source to drain along a cross section near the gate dielectric/silicon interface.

The carrier starvation near the source can be clearly seen in the case of the device with too abrupt doping profile (LDG = 0.5 nm/dec.) while it is absent in the 2 nm/dec. case. In order to avoid the series resistance from this effect, one should ensure that there is sufficient doping near the gate edge.

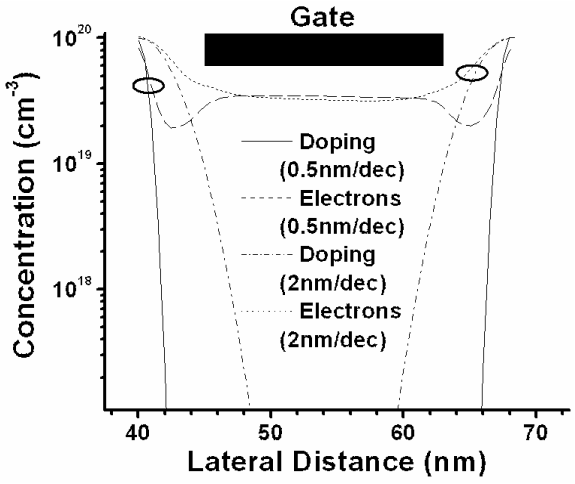


Fig. 3.13 Plots of the doping and the electron concentration from the source to drain near the gate dielectric/silicon interface showing the bottleneck due to insufficient gate-to-source overlap when the lateral doping gradient is too abrupt for a given extension underlap.

3.3.4 Optimization of Extension Underlap

From the I_{on}-I_{off} plots, especially Fig. 3.10 and Fig. 3.11, it follows that the drive current can be maximized by optimizing the extension underlap. Such an optimization would depend upon the lateral doping gradient and the maximum leakage current I_{off} that can be tolerated. Fig. 3.14 shows one such set of optimizations for the case where the I_{off} is set to 1μA/μm. The optimization essentially trades off short channel effects (sub-optimal underlap) for increased series resistance (underlap greater than optimal value). In both cases, drive current is degraded through the reduction of effective gate overdrive. As the lateral doping abruptness decreases, the optimal underlap shifts to higher values in an attempt to keep the dopants out of the channel. By proper optimization, the devices non-abrupt lateral doping profiles can be tuned to give nearly the same (within 10 %) drive current as the ideal devices with very abrupt junctions.

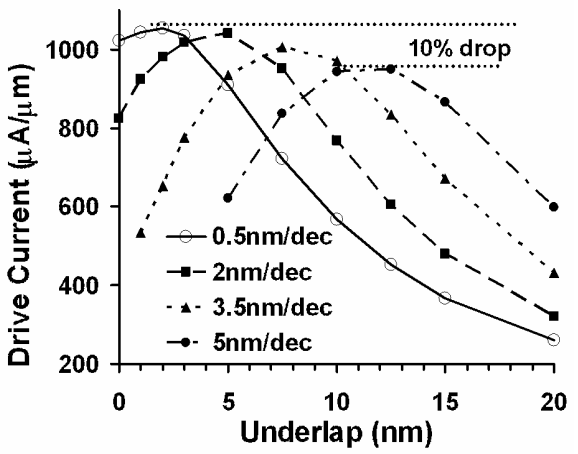


Fig. 3.14 Optimization of the extension underlap. The leakage current is set to 1μA/μm.

One important point that needs to be made here is that optimization of drive current I_{on} implicitly optimizes the intrinsic device delay given by CV/I_{on}. The capacitance term in the delay expression includes the parasitic capacitance as well as the usual gate to channel capacitance. If the underlap is set by changing the gate-source/drain spacer thickness, the parasitic gate to source/drain capacitance will change to first order since the parallel plate thickness is changed. That will impact the optimum delay and cause it to occur at a different underlap from that which maximizes I_{on}. Based on process concepts such as in [3.22] and [3.23] where the source and drain are grown epitaxially or deposited independent of the channel, it is possible to think of some ways to decouple the underlap from the sidewall spacer thickness. We assume such a case and keep the spacer fixed at 20 nm thickness as the underlap is changed. Fig. 3.15 shows that, if the underlap is set instead by the gate to source/drain sidewall spacer thickness, the optimal underlap which minimizes delay is larger than that which maximizes drive current. As a result of the sub-optimal drive current, the minimum delay that is now achieved is higher than the case in which the underlap and spacer thickness are decoupled.

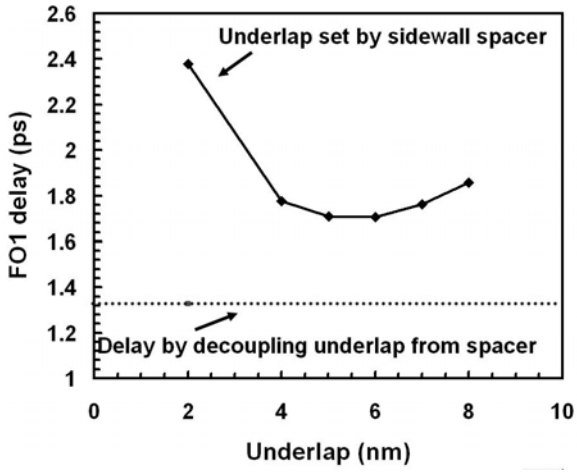


Fig. 3.15 Simulated plot of the FO1 inverter delay as a function of the extension underlap as set by the sidewall spacer. The optimum delay is not as good as the delay obtained by decoupling the underlap from the spacer thickness.

We now look at the effect of different parameters on the optimal underlap. In all cases, unless otherwise specified, the LDG is set to 2.5 nm/dec and the leakage current is 1 μA/μm.

Effect of I_{off}

Fig. 3.16 shows the effect of the choice of leakage current I_{off} on the underlap optimization. In the 1 nA/μm leakage device, the short channel effect needs to be controlled more effectively than in the 1 μA/μm leakage device. Moreover, the low leakage device also has a lower drive current. This reduces the impact of the $I_d \cdot R_s$ series resistance drop. Coupled together, these facts imply that the short channel effect is weighted more strongly than the series resistance in the underlap optimization for the low leakage transistor. The low leakage device requires steeper subthreshold swing and lower DIBL than the high leakage DG FET. Consequently, the optimal underlap amount is increased in that device as compared to the 1 μA/μm transistor.

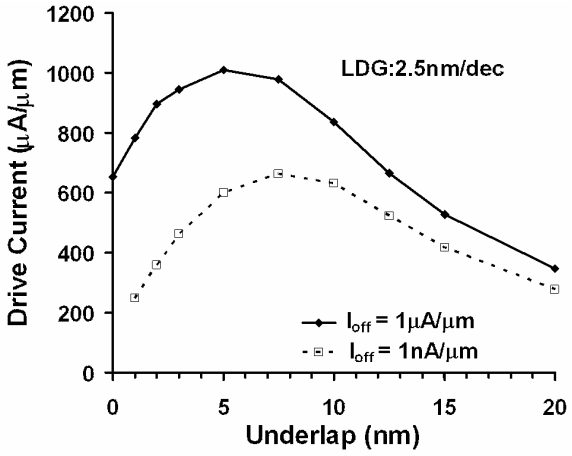


Fig. 3.16 Effect of the leakage current (I_{off}) constraint on the underlap optimization. The optimal underlap increases for lower required I_{off}.

Effect of T_{si}

Fig. 3.17 shows the effect of the silicon body thickness T_{si} on the optimal underlap. It is seen that lowering the T_{si} causes the optimal underlap amount to decrease.

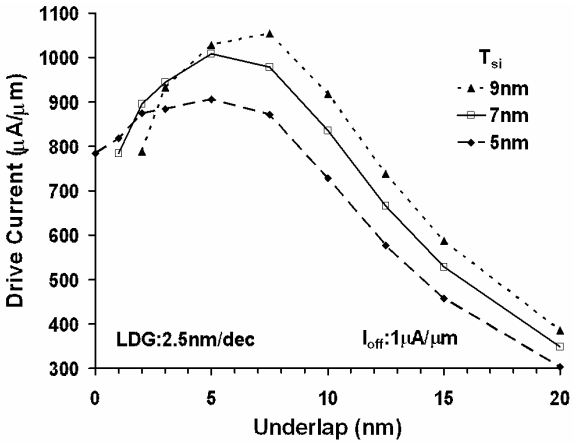


Fig. 3.17 Effect of silicon body thickness (T_{si}) on underlap optimization. The optimal underlap decreases as T_{si} is reduced.

Once again, this can be explained in terms of the short channel effect – series resistance trade-off. Thinner body devices are more immune to short channel effects due to

their superior electrostatic channel control. Therefore, they can tolerate a greater amount of dopant spilling into the channel. Also, the thinner body leads to a higher extension resistance. Both these factors point toward pushing the dopant roll-off point closer to the gate edge – a lower optimal underlap.

Effect of L_{gate}

Fig. 3.18 shows the effect of the gate length L_{gate} on the underlap optimization. As the gate length is reduced, the optimal underlap increases. This essentially counters the electrostatic degradation by increasing effective channel length (L_{eff}). For the given body thickness and effective gate oxide thickness, the 22 nm L_{gate} device is already quite robust against short channel effects. Therefore in this case, not only is the optimal underlap very low, but the optimum is also quite flat.

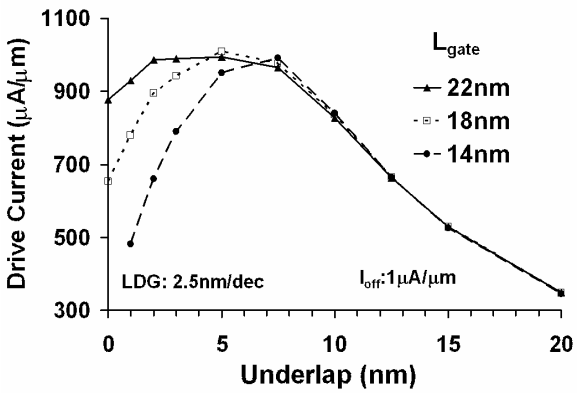


Fig. 3.18 Effect of the gate length on the optimal underlap. For shorter L_{gate}, the optimal shifts to larger values and the optimum also becomes steeper.

Effect of Contact Resistance

Here, we examine the effect of the contact resistance on the underlap optimization. Fig. 3.19 shows that the optimal underlap does not change as the specific contact resistivity is reduced. However, the value of optimized drive current increases. The optimum becomes sharper in the case with lower specific contact resistivity since the drive current

is now entirely limited by the extension series resistance and is thus more sensitive to underlap. Unlike lateral doping gradient and underlap, the contact resistance does not impact short channel effects. Therefore the extension doping profile optimization is independent of the contact resistance. However, it must be emphasized that it is always beneficial to reduce the contact resistance since the overall drive current, and hence performance, is increased.

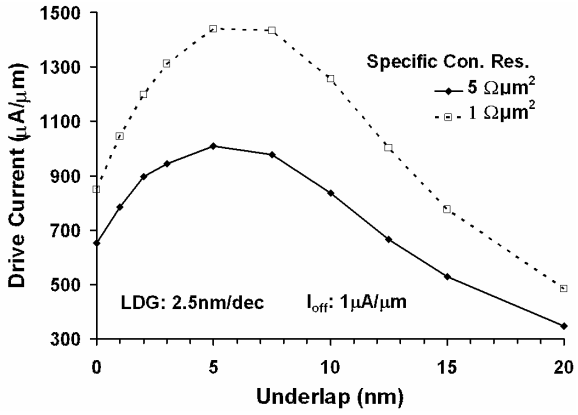


Fig. 3.19 Effect of the specific contact resistivity on the underlap optimization. The optimal underlap does not change.

3.4 Schottky Source/Drain FET

A Schottky source/drain FET is obtained by replacing the doped semiconductor source and drain regions by metals or metal-silicides. Such structures have been investigated since the late 1960's [3.24], with more recent resurgent interest [3.25, 3.26]. These devices have many compelling advantages such as immunity to latch-up and parasitic bipolar action, low extrinsic resistance, and relative simplicity of processing. In ultrathin undoped body DG FETs, we have seen that metal gates will most likely be needed in order to tune the transistor threshold voltage. If, in addition, the source and drain are also formed using metals, that would eliminate the need for any doping and the associated

problems of controlling dopant placement and diffusion during subsequent thermal cycles. These factors make it interesting to investigate the potential of Schottky source/drain in DG FETs. In this section, Schottky S/D DG FETs are investigated theoretically by device simulation. The operating principles are discussed and comparisons are made with conventional doped S/D devices.

3.4.1 Schottky S/D FET Operation

Fig. 3.20 shows a schematic of the Medici-simulated Schottky S/D DG FET. The main source of series resistance in such a device is the non-zero barrier from the metal to the undoped Si channel region. When a metal is brought in contact with a semiconductor, there is typically a potential barrier formed at the interface due to the mismatch in the metal and semiconductor workfunctions. In the simplest picture, when the Fermi levels line up at equilibrium, there is a discontinuity in the bands that is equal to the difference in the workfunctions. However, in real metal-semiconductor junctions, the Fermi level in the semiconductor gets pinned within the bandgap due to a high density of interface states. The origin of these states can be either extrinsic or intrinsic [3.27]. In most metal-silicon or silicide-silicon systems, the Fermi-level pinning occurs deep inside the Si bandgap, giving rise to high barriers to both the conduction band as well as the valence band. In some cases, the barrier height can be made low. For example, $ErSi_2$ and PtSi [3.26] have barriers of 0.28 eV and 0.24 eV to the Si conduction and valence band respectively. These make them potential candidates for NMOS and PMOS transistors respectively.

Fig. 3.21 shows the I_d-V_g characteristics of an n-type Schottky S/D DG FET with a barrier of 0.2 eV to the conduction band. Since the resistance of the metallic regions is much less than that of heavily doped Si, it is neglected in the simulation. The transport across the metal-Si junction is modeled using a combination of thermionic and field emission based on the work of Ieong et. al. [3.28].

Fig. 3.22 shows the conduction band edge profile along the lateral direction, from source to drain, at a very short distance beneath the gate dielectric/Si interface. In the off-state, there is a 0.2 eV barrier at the source due to the metal-Si junction. In addition, there is a small (in this case) barrier due to the action of the gate on the Si. This second barrier can be modulated by the gate workfunction, with higher values of workfunction yielding higher barriers.

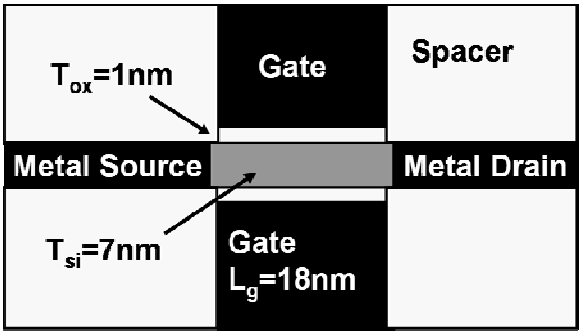


Fig. 3.20 Schematic of the simulated Schottky source/drain DG FET.

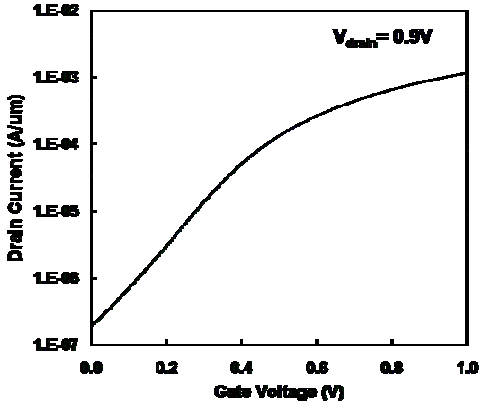


Fig. 3.21 Simulated I_d-V_g characteristic for an n-type Schottky S/D DG FET. The metal-semiconductor system is assumed to form a barrier of 0.2 eV to the conduction band.

In the on-state, the gate-induced barrier in Si disappears, while the metal-Si barrier remains. However, due to the high vertical electric field near the source, even this barrier becomes very thin. This allows current to flow through the barrier by tunneling (field emission) in addition to the conventional thermionic emission above it. Thus, the sub-threshold I_d-V_g characteristic of the Schottky S/D DG FET has 2 regions – for low gate voltages, the conduction is mostly thermionic emission over the combined metal-Si and gate-induced Si barrier.

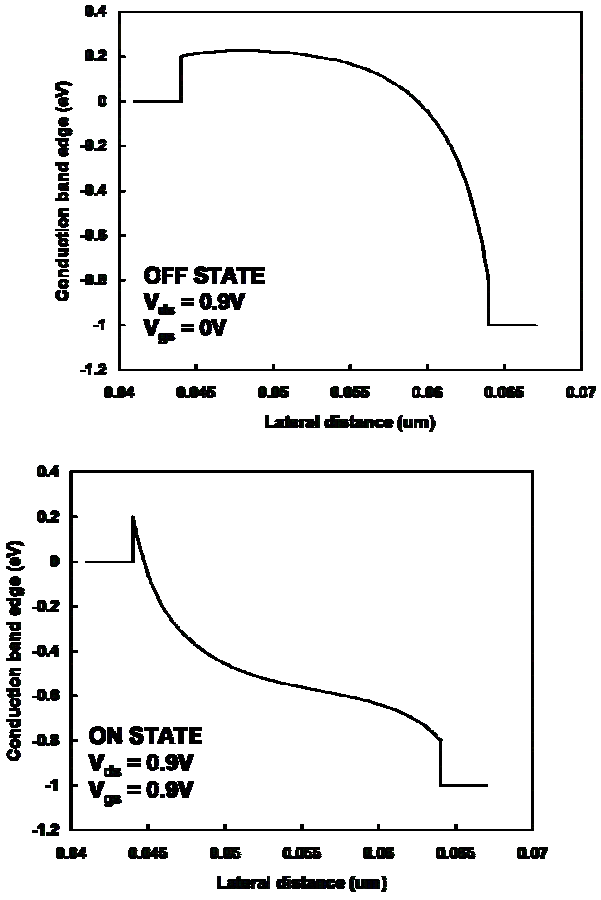


Fig. 3.22 Off-state and on-state band diagrams showing the conduction band edge from the source to drain at a point just beneath the gate dielectric/Si interface.

As the gate voltage increases, if there is perfect capacitive coupling from the gate to the semiconductor surface, this barrier falls at the same rate as in a conventional doped S/D DG FET – yielding a KT/q-limited subthreshold slope that has a theoretical minimum of 60 mV/decade at 300 K. Once the gate-induced barrier has vanished, the effect of the gate is to modulate the thickness of the metal-Si barrier. An important point that should be noted is that the gate is effective in changing the effective barrier height only in the gate-depleted portion of the substrate. Therefore, using a low metal-Si barrier height junction in a bulk MOSFET would cause a large amount of thermionic emission leakage along the bottom area of the junction. On the other hand, in ultrathin body fully depleted transistors (single or double-gate), the gate has control over the entire metal-Si interface and therefore the leakage can be suppressed even for a zero metal-Si barrier height contact.

3.4.2 Comparison with Doped S/D DG FET

Having seen the operating principles of Schottky S/D DG FETs, it is now instructive to compare them with conventional doped S/D FETs. As before, we use simulated I_{on}-I_{off} curves for the comparison where the gate workfunction is swept to generate the curves. In case of the doped S/D devices, the extension underlap has been optimized as in the previous section for each lateral doping gradient. In the Schottky S/D devices too, the underlap has been optimized for each barrier height. Here, the underlap is defined as the offset between the gate edge and the metal-semiconductor junction. Fig. 3.23 shows the graphs of the different I_{on}-I_{off} curves.

Only the zero-barrier Schottky S/D DG FET outperforms all of the doped S/D DG FETs. Even if one accounts for simulation inaccuracies, it is clear that the aforementioned $ErSi_2$ and PtSi systems will not be sufficient to compete with conventional doped S/D DG FETs that have relatively non-abrupt (5 nm/decade) lateral doping profiles. Therefore it is very important to realize alternative ways of obtaining zero or near-zero barrier height metal-semiconductor junctions.

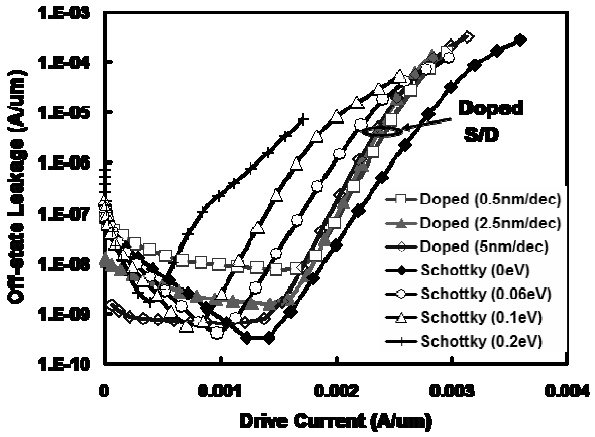


Fig. 3.23 I_{on}-I_{off} comparisons of Schottky S/D DG FETs with different barrier heights with conventional doped S/D DG FETs with different lateral abruptness. In all cases, the extension underlaps have been optimized.

Connelly and co-workers [3.29] have shown that inserting a very thin dielectric barrier layer at the metal-semiconductor interface can reduce the density of intrinsic gap states and effectively de-pin the Fermi level. This enables them to get barriers as low as 45 meV to the Si conduction band. Finally, it should be noted that even non-zero metal-semiconductors barriers may be an attractive option for building Schottky S/D DG FETs on high-mobility Germanium channels where the control and incorporation of dopants (especially n-type) is not as good as it is in Si [3.30].

3.5 Summary

Parasitic capacitance and resistance can limit the overall device performance in ultrathin body DG FETs. For minimum parasitic capacitance, the ideal DG FET should have the top and bottom gates perfectly self-aligned to one another. They should have the same size and be separated from the source/drain regions by sidewall spacers that are much thicker than the gate oxide. In order to minimize the impact of parasitic resistance,

very low specific contact resistivity and extremely abrupt lateral doping profile are ideally desired. In the absence of the former, the flared out source/drain structure allows increased effective contact area and reduces the total contact resistance. In the presence of finite lateral doping abruptness, there is an optimal value of extension underlap at which the drive current is maximized. This optimization involves balancing the impact of short channel effects and series resistance and hence depends on the allowed leakage current and the electrostatic integrity of the device structure. Schottky source/drain structures may offer a path to achieving low access resistance to the channel. However, the metal-semiconductor barrier height needs to be nearly zero eV for such a device to outperform DG FETs with optimized doped source/drain.

References

[3.1] K.W. Guarini et. al., "Triple-self-aligned, planar double-gate MOSFETs: Devices and circuits," in *IEDM Technical Digest*, 2001, pp. 425-428.

[3.2] S. Harrison et. al., "Highly performant double gate MOSFET realized with SON process," in *IEDM Technical Digest*, 2003, pp. 449-452.

[3.3] H.-S. P. Wong, K. K. Chan, and Y. Taur, "Self-aligned (top and bottom) double-gate MOSFET with a 25 nm thick silicon channel," in *IEDM Technical Digest*, 1997, pp. 427-430.

[3.4] S.-Y. Lee et. al., "A novel sub-50 nm multi-bridge-channel MOSFET (MBCFET) with extremely high performance," in *Digest of Technical Papers – Symposium on VLSI Technology*, 2004, pp. 200-201.

[3.5] Z. Ren, R. Venugopal, S. Datta, and M. Lundstrom, "Examination of design and manufacturing issues in a 10 nm double gate MOSFET using nonequilibrium Green's function simulation," in *IEDM Technical Digest*, 2001, pp. 107-110.

[3.6] S.-D. Kim, C.-M. Park, and J. C. S. Woo, "Advanced model and analysis for series resistance in sub-100nm CMOS including poly depletion and overlap doping gradient effect," in *IEDM Technical Digest*, 2000, pp. 723-726.

[3.7] M. Y. Kwong, R. Kasnavi, P. Griffin, J. D. Plummer, and R. W. Dutton, "Impact of lateral source/drain abruptness on device performance," *IEEE Transactions on Electron Devices*, vol. 49, no. 11, pp. 1882-1890, Nov 2002.

[3.8] Synopsys Corporation, Mountain View, CA, MEDICI^{TM} version 2002.2 User Guide, 2002.

[3.9] International Technology Roadmap for Semiconductors, 2001 edition, SIA.

[3.10] G. D. Wilk, R. M. Wallace, and J. M. Anthony, "High-k gate dielectrics: Current status and materials properties considerations," *Journal of Applied Physics*, Vol. 89, no. 10, pp. 5243-5275, May 2001.

[3.11] T.-J. King, "Applications of polycrystalline silicon-germanium thin films in metal-oxide-semiconductor technologies," Ph. D. Thesis, Stanford University, 1994.

[3.12] P. Ranade, Y.-K. Choi, D. Ha, A. Agarwal, M. Ameen, and T.-J. King, "Tunable work function molybdenum gate technology for FDSOI-CMOS," in *IEDM Technical Digest*, 2002, pp. 363-366.

[3.13] J. Kedzierski et. al., "Metal-gate FinFET and fully-depleted SOI devices using total gate silicidation," in *IEDM Technical Digest*, 2002, pp. 247-250.

[3.14] M. N. Darwish, J. L. Lentz, M. R. Pinto, P. M. Zeitzoff, T. J. Krutsick, and H. H. Vuong, "An improved electron and hole mobility model for general purpose device simulation," *IEEE Transactions on Electron Devices*, vol. 44, no. 9, pp. 1529-1538, Sep 1997.

[3.15] J. D. Bude, "MOSFET modeling into the ballistic regime," in *2000 International SISPAD Conference*, 2000, pp. 23-26.

[3.16] J. R. Pfiester, R. D. Sivan, H. Ming Liaw, C. A. Seelbach, and C. D. Gunderson, "A self-aligned elevated source/drain MOSFET", *IEEE Electron Device Letters*, vol. 11, no. 9, pp. 365-367, Sep 1990.

[3.17] J. Kedzierski, P. Xuan, E. H. Anderson, J. Bokor, T.-J. King, C. Hu, "Complementary silicide source/drain thin-body MOSFETs for the 20nm gate length regime," in *IEDM Technical Digest*, 2000, pp. 57-60.

[3.18] D. Connelly, C. Faulkner, and D. E. Grupp, "Performance advantage of Schottky source/drain in ultrathin-body silicon-on-insulator and dual-gate CMOS," *IEEE Transactions on Electron Devices*, vol. 50, no. 5, pp. 1340-1345, May 2003.

[3.19] S. Gannavaram, N. Pesovic, and M. C. Öztürk, "Low temperature (≤800°C) recessed junction selective silicon-germanium source/drain technology for sub-70 nm CMOS," in *IEDM Technical Digest*, 2000, pp. 437-440.

[3.20] B. Yu, Y. Wang, H. Wang, Q. Xiang, C. Riccobene, S. Talwar, and M.-R. Lin, "70 nm MOSFET with ultra-shallow, abrupt, and super-doped S/D extension implemented by laser thermal process (LTP)," in *IEDM Technical Digest*, 1999, pp. 509-512.

[3.21] A. Asenov, "Random dopant induced threshold voltage lowering and fluctuations in sub-0.1 μm MOSFETs: A 3D 'atomistic' simulation study," *IEEE Transactions on Electron Devices*, vol. 45, no. 12, pp. 2505-2513, Dec 1998.

[3.22] J.-H. Lee, G. Taraschi, A. Wei, T. A. Langdo, E. A. Fitzgerald, and D. A. Antoniadis, "Super self-aligned double-gate (SSDG) MOSFETs utilizing oxidation rate difference and selective epitaxy," in *IEDM Technical Digest*, 1999, pp. 71-74.

[3.23] T. Yoshitomi, M. Saito, T. Ohguro, M. Ono, H. S. Momose, and H. Iwai, "Silicided silicon-sidewall source and drain (S^4D) structure for high-performance 75-nm gate length pMOSFETs," in *Digest of Technical Papers – Symposium on VLSI Technology*, 1995, pp. 11-12.

[3.24] M. P. Lepselter and S. M. Sze, "SB-IGFET: An insulated-gate field-effect transistor using Schottky barrier contacts as source and drain," *Proceedings of the IEEE*, vol. 56, no. 8, p. 1088, 1968.

[3.25] J. P. Snyder, C. R. Helms, and Y. Nishi, "Experimental investigation of a PtSi source and drain field emission transistor," *Applied Physics Letters*, vol. 67, no. 10, pp. 1420-1422, 1995.

[3.26] J. Kedzierski, P. Xuan, E. H. Anderson, J. Bokor, T.-J. King, and C. Hu, "Complementary silicide source/drain thin-body MOSFETs for the 20 nm gate length regime," in *IEDM Technical Digest*, 2000, pp. 57-60.

[3.27] J. Tersoff, “Schottky barrier heights and the continuum of gap states,” *Physical Review Letters*, vol. 52, no. 6, pp. 465-468, 1984.

[3.28] M. K. Ieong, P. M. Solomon, S. E. Laux, H-S. P. Wong, and D. Chidambarrao, “Comparison of raised and Schottky source/drain MOSFETs using a novel tunneling contact model,” in *IEDM Technical Digest*, 1998, pp. 733-736.

[3.29] D. Connelly, C. Faulkner, D. E. Grupp, and J. S. Harris, “A new route to zero-barrier metal source/drain MOSFETs”, *IEEE Transactions on Nanotechnology*, vol. 3, no.1, pp. 98-104, Mar 2004.

[3.30] C. O. Chui, K. Gopalakrishnan, P. B. Griffin, J. D. Plummer, and K. C. Saraswat, "Activation and diffusion studies of ion-implanted p and n dopants in germanium," *Applied Physics Letters*, vol. 83, no. 16, pp. 3275-3277, 2003.

Chapter 4

A Novel Process for Fully Self–Aligned Planar Double-Gate FET

4.1 Introduction

In the preceding chapters, we have seen the requirements for double-gate FET structures. These arise from the need to optimize the intrinsic as well as extrinsic performance of the transistor. From the intrinsic device point of view, one of the key necessities is an ultrathin silicon body whose thickness needs to be less than half the gate length. While it is important to able to define this dimension precisely, there are also stringent constraints on achieving body thickness uniformity across the device as well as across the wafer and from one lot to another. In order to minimize the impact of extrinsic parasitic elements, the two gates need to be perfectly self-aligned to each other with low capacitance to the source and drain. In addition, the source and drain extension regions need to be optimized in order to maximize the drive current, and hence performance, subject to constraints on short channel effects. In this chapter, a process flow is proposed to build a planar double-gate FET which satisfies all of the above requirements. In

addition, the process is sufficiently flexible so as to allow the fabrication of novel device structures based on minor variations around the basic flow.

The present chapter is organized as follows. First of all, the motivation for planar versions of the DG FET is discussed. Next, there is a survey of prior art in this field. Previous attempts to implement planar DG FETs are briefly described along with their advantages and shortcomings. The next section then describes in more detail the main steps in the novel planar DG FET process flow. Finally, the extensibility of the basic process is discussed, which can result in the realization of various interesting device structures.

4.2 Double-Gate FET Configurations

As the name implies, a double-gate FET has two gates, one on either side of a thin silicon body. Depending on the orientation of the gates and the direction of current flow with respect to the wafer surface, there are three kinds of configurations that may be used to implement DG FETs. Fig. 4.1 depicts these three types of DG FET structures.

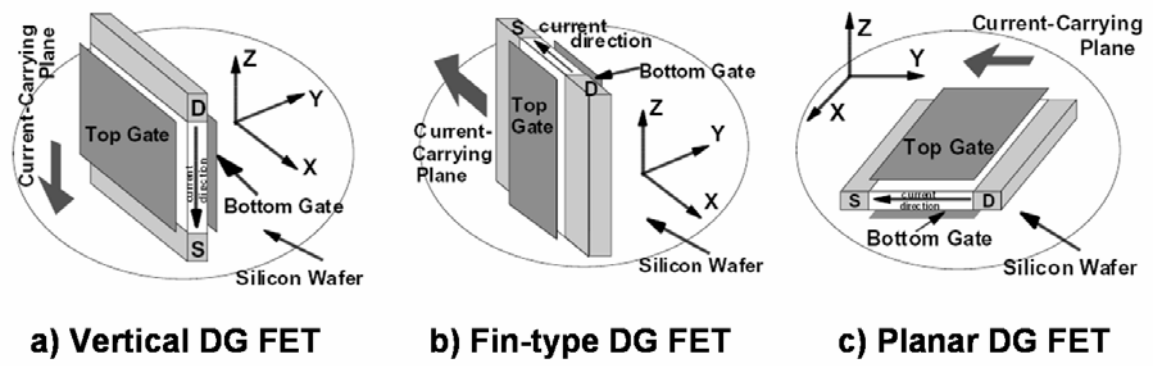


Fig. 4.1 Types of double-gate FET configurations. This figure is from a presentation by H.S-P. Wong [4.1].

In the first structure (Fig. 4.1 (a)), the current flows along the sides of a vertical silicon wall. Such a device is also referred to as a 'pillar FET' if the width dimension (along the x-axis) is comparable to the wall thickness (along the y-axis). These vertical transistors have been extensively studied [4.2, 4.3] and have a reduced footprint compared to conventional planar transistors with the same current-carrying capability. This property makes them very attractive for memory applications [4.2, 4.4], where the reduced cell-size enables higher bit density. In addition, since the channel length is defined by the height of the pillar, it is possible to set this dimension based on a deposited film thickness [4.5]. This allows sub-lithographic channel length definition with very good uniformity. In addition, if the pillar is formed by epitaxy, the channel materials and/or doping can be engineered during the growth process in order to get devices with improved properties [4.6] or new functionality [4.7]. In spite of all these advantages, vertical FETs have some shortcomings. Some of these are described next.

Vertical DG FETs which are built on (100)-oriented Si substrates have the channels typically formed along the (110) sidewalls. It is known [4.8] that acoustic phonon-limited electron mobility in the (110) plane is degraded as compared to that in the (100) plane. Moreover, due to the higher areal density of Si atoms in the (110) plane, thermal oxides grown on those sidewalls have a higher density of interface fixed charge. Finally, the sidewall roughness and damage resulting from the plasma etching used to form the vertical walls can also degrade the carrier mobility [4.9]. As a result of these effects, the channel carrier transport in vertical FETs is not as efficient as it is in conventional bulk FETs. This results in degraded current drive, and hence reduced speed.

For highly scalable DG FETs with fully depleted bodies, we have seen that the silicon thickness needs to be less than the channel length. In vertical DG FETs, even though the device length can be set sub-lithographically, the body thickness definition still needs a lithography step. While there have been attempts to resolve this problem by using deposited channel films [4.10], those have resulted in single-gate ultrathin body transistors whose electrostatic robustness is not as good as DG FETs.

From an extrinsic resistance viewpoint, the relative inaccessibility of the bottom electrode can make it difficult to achieve low series resistance to the channel. In addition, unless the vertical FET process can be integrated with the conventional planar process, all transistors on the chip are constrained to have the same gate length. This may be a problem in circuit designs which require transistors with multiple gate lengths.

Fig. 4.1 (b) shows a device called the FinFET. In this implementation, the current flows in the horizontal direction (along the x-axis) similar to a conventional bulk FET. However, the current carrying plane is perpendicular to the wafer surface, similar to the vertical FET. The main advantage of such a structure is that it is highly manufacturable. All electrodes (source, drain, gates) are accessible from the top and the process flow is similar to that for planar bulk FETs. The relative ease of fabrication has made the FinFET the most studied version of the double-gate FET with extensive research done in universities [4.11] and industry [4.12]. However, this structure has some problems as described below.

As in the case of the vertical FET, carrier transport along etched (110) sidewalls can result in degraded electron current [4.13]. In addition, the definition of ultrathin Si fins with a high degree of thickness uniformity is challenging. The use of spacer-lithography [4.14] addresses this issue to some extent. In FinFETs, the device width is defined by the fin height (it is equal to the sum of the fin thickness and twice the fin height); multiple fins must be used in parallel to get effectively wider devices. The achievable widths are thus quantized in integer multiples of a fixed fin height. This may be problematic for circuit designs in which finer granularity in width definition is needed.

Fig. 4.1 (c), on the right side, shows a planar DG FET. In this structure, the current flows in the horizontal direction (along the x-axis), and in the same plane as the wafer surface. The carrier transport is thus in the well understood Si (100) plane. The body thickness is perpendicular to the substrate (along the z-axis) and can therefore be defined sub-lithographically and uniformly by using a deposited film thickness. Assuming that the top and bottom gates wrap around the Si body, all electrodes are accessible

from the top, similar to bulk FETs. In fact, in plan view, this structure appears identical to the conventional single-gate bulk FET. This enables easier portability of circuit design layouts. The major challenge in planar DG FETs has been bottom gate placement. As we have seen in the previous chapter, the bottom gate needs to be of the same size and perfectly self-aligned to the top gate. Any misalignment or mis-sizing of the bottom gate results in increased delay and/or higher off-state leakage.

In the following sections, previous attempts at fabricating planar DG FETs are described followed by the proposal of a novel process flow to build a fully self-aligned planar DG FET.

4.3 Planar DG FET – Prior Art

Due to the advantages of the planar version of the DG FET over other configurations, there have been a number of attempts to fabricate it. This section contains a survey of some of the planar DG FET implementations that have been reported in the literature.

4.3.1 IBM Planar DG FET using Pattern-Constrained Epitaxy

In [4.15], Wong et. al. of IBM have described a planar DG FET process flow that uses pattern-constrained Si epitaxy to form the transistor body in between replacement gates. The main steps in the fabrication sequence are shown in Fig. 4.2. The basic idea involves patterning a dielectric stack to the dimensions that will eventually correspond to the device length and width. A tunnel is then opened up in between the dielectric layers by isotropically etching away the core material. The remaining layers above and below the tunnel serve as temporary sacrificial gates. Selective epitaxy is then used to grow silicon out of a seed hole, through the tunnel and onto the larger area on the other side of the tunnel as shown in Fig. 4.2 (a). Chemical mechanical polishing (CMP) is then used to remove the excess epi-grown silicon, resulting in the structure shown in Fig. 4.2 (b). The seed hole and the other recessed area across the tunnel form the source and drain while

the Si that fills the tunnel is the transistor body. Following a source/drain ion-implant masked by the replacement gate, the dielectric layers are removed, leaving behind a thin Si bridge suspended between the thicker source and drain regions (Fig. 4.2 (c)). Finally a gate oxide is grown and followed by a conformal polysilicon gate deposition that surrounds the Si bridge (Fig. 4.2 (d)).

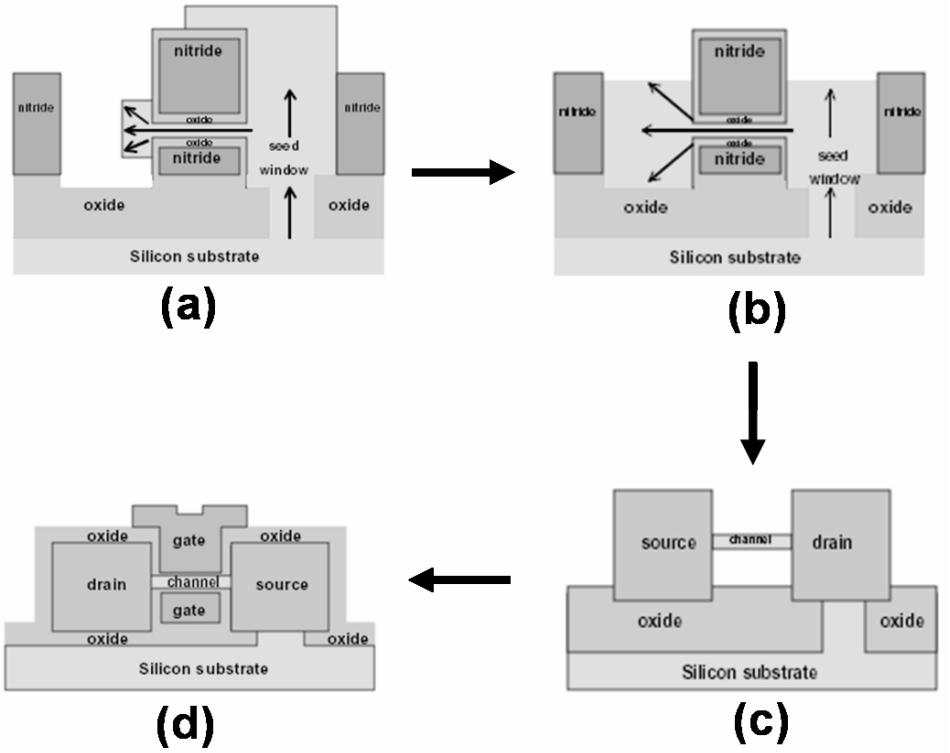


Fig. 4.2 IBM Planar DG FET process using pattern-constrained epitaxy (from [4.15])

The pattern-constrained epitaxy process thus results in a planar DG FET with deposition-controlled body thickness, self-aligned top and bottom gates, and flared-out source/drain regions for low series resistance. Using this process, Wong et. al. [4.15] have demonstrated N-channel DG FETs with gate lengths down to 0.66 μm and a body thickness of 25 nm.

One of the potential shortcomings of this process scheme is that the sidewall spacer separating the gate and source/drain fanout regions is grown at the same time as the gate oxide. Such a thin sidewall spacer can increase the gate to source/drain capacitance, compromising the device switching speed.

4.3.2 MIT Super Self-Aligned Double-Gate (SSDG) FET

In [4.16], Lee et. al. of MIT have proposed a planar DG FET process (Fig. 4.3) that utilizes a combination of wafer bonding, differential oxidation rates, and selective epitaxy.

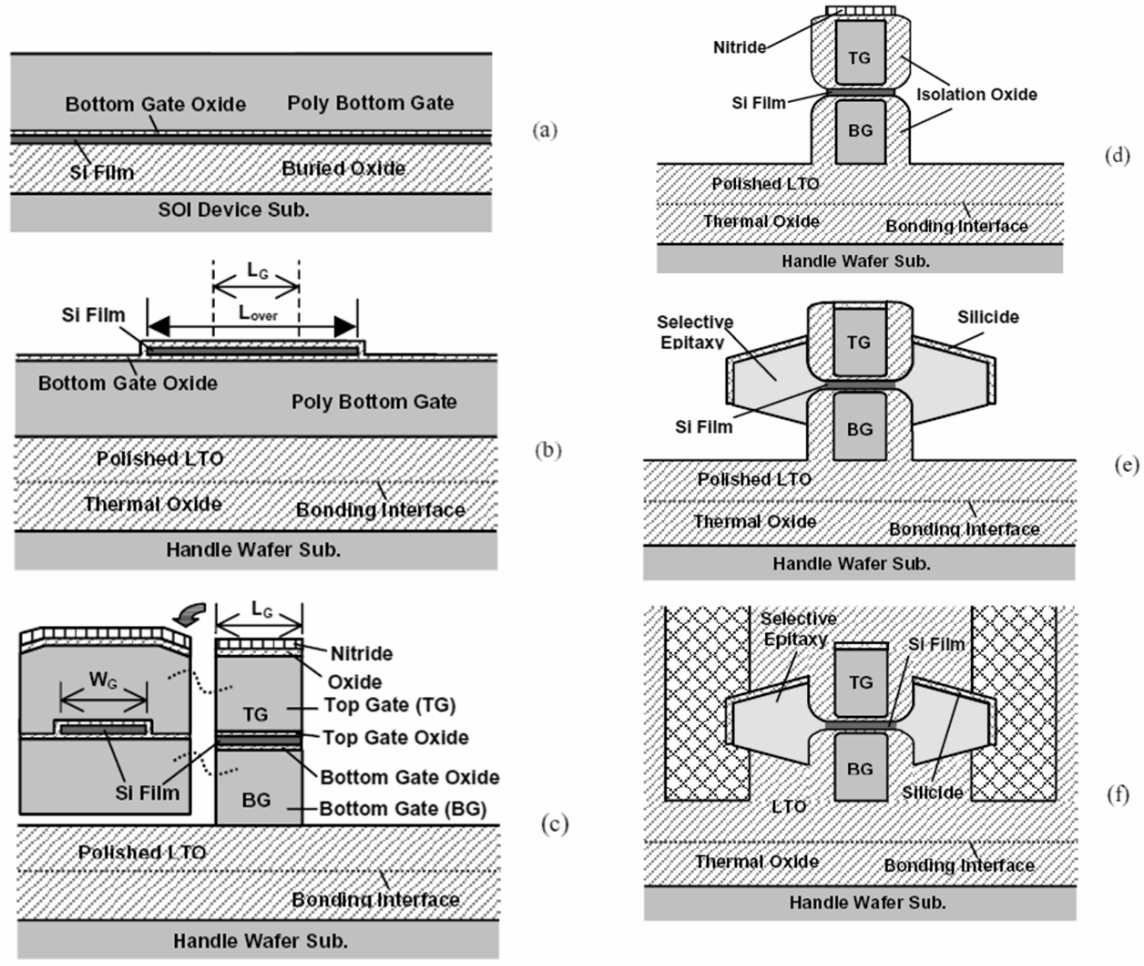


Fig. 4.3 Key process steps for fabrication of SSDG FETs (from [4.16])

An SOI wafer with a thin Si layer is used as the starting substrate. Thermal oxidation followed by polysilicon deposition and doping is used to form the to-be bottom gate stack (Fig. 4.3 (a)). This is then covered with LTO, polished, and bonded to a handle Si wafer covered with thick thermal oxide. The SOI device wafer is then etched back to remove the Si substrate and the buried oxide. An oversized (greater than the final device gate length) active layer patterning step is then used to etch the thin Si film and stop on the bottom gate oxide (Fig. 4.3 (b)). The top gate oxide is grown followed by top polysilicon gate deposition, doping, and a nitride hard mask. Using a single mask for gate length definition, both gates are etched together (Fig. 4.3 (c)) for self-alignment. Next, differential oxidation rate on the doped polysilicon is used to form thicker oxide spacers on the gate sidewalls (Fig. 4.3 (d)). After removing the thinner oxide formed on the channel sidewall, Si is selectively grown by epitaxy to form fanned out source/drain regions (Fig. 4.3 (e)). This is followed by the source/drain implant, annealing, and bottom gate contact patterning. Silicide formation and back-end processing complete the process to yield the structure shown in Fig. 4.3 (f).

The advantages of this process are fully self-aligned top and bottom gates, fanout source/drain for low series resistance, and potentially low parasitic gate capacitance due to the differential oxidation spacer process. In addition, since the top and bottom gate stack are formed independently, such a structure could be tuned for operation in the variable threshold back-gate mode. One of the main drawbacks of the SSDG FET is that the critical body thickness dimension depends upon the starting SOI layer thickness. This may not be easy to control in terms of achieving target thickness and across-wafer uniformity. Also, the use of bonding and etch back makes this process rather complicated. The authors have reported experimental verification of the critical enabling process steps, but no electrical data on transistor characteristics have been shown.

4.3.3 IBM Pagoda FET

In [4.17], Guarini et. al. of IBM have described a way to fabricate planar DG FET structures that uses wafer bonding and chemical mechanical polishing (CMP). The main process steps are shown in Fig. 4.4. The process starts by forming the eventual back-gate stack (thermal oxide covered with doped polysilicon) on a thinned SOI wafer. This is bonded to a handle wafer, flipped over and etched back. The back-gate stack is patterned and planarized by CMP (Fig. 4.4 (a)). The front gate oxide is then grown and followed by the front gate polysilicon deposition. The front gate stack is patterned, and planarized by a second CMP step (Fig. 4.4 (b)). The device gate length critical dimension is then defined and the front gate stack is etched, stopping on the front gate oxide. After forming oxide sidewall spacers on the front gate edge, amorphous Si is deposited and recrystallized by solid phase epitaxy off the thin silicon channel. The source and drain regions are implanted with dopants and etched to form doped sidewall spacers which are then silicided. The raised source/drain thus formed have low contact resistance to the channel (Fig. 4.4 (c)). The silicide on the sidewall spacers is now used as a mask while the back-gate is isotropically etched and undercut. This ensures that the top and bottom gates are self-aligned to the source/drain. After passivating the bottom gate with a nitride liner, the gate contacts are etched and silicided.

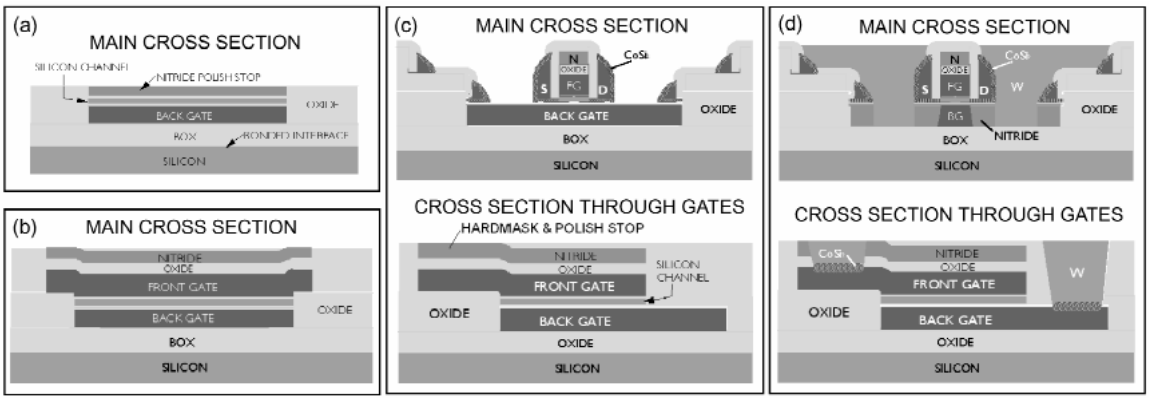


Fig. 4.4 PAGODA DG FET process flow (from [4.17])

Finally, tungsten plugs are formed to make contact to the source, drain and the two gates (Fig. 4.4 (d).)

The main advantages of this process are the low resistance source/drains and self-aligned gates which can be independently accessed. This device can thus be effectively used as a variable threshold back-gate (BG) FET. Guarini et. al. [4.17] have shown transistors operated in both DG as well as BG modes. Though the gates are self-aligned, the bottom gate size depends upon a timed etch and is therefore not exactly the same as the top gate. As we have seen in the previous chapter, this size mismatch can adversely affect the delay or the off-state leakage. Another potential concern with this rather complicated process is that the critical body thickness (T_{si}) dimension depends upon the starting SOI film thickness. This may lead to performance-limiting T_{si} variations across the wafer.

4.3.4 STMicro DG FET using Silicon On Nothing (SON) Process

The 'Silicon on Nothing' (SON) process was originally proposed [4.18] to fabricate SOI-like single-gate fully depleted FETs on bulk-Si wafers. This process (Fig. 4.5) has been extended by Harrison et. al. of ST Microelectronics [4.19] to fabricate planar DG FETs.

The process begins by epitaxially growing SiGe selectively over the patterned active areas (AA) of a bulk Si wafer. This is followed by non-selective epitaxial growth of a thin Si body layer (Fig. 4.5 (a)). A mask is used to pattern the single crystal Si over the SiGe and the polysilicon formed over the oxide isolation areas (Fig. 4.5 (b)). The SiGe underneath the Si channel film is then removed using an isotropic etch selective to Si. This leaves behind a Si bridge with an empty space below it - hence 'silicon on nothing' (Fig. 4.5 (c)). Gate oxidation followed by in-situ doped polysilicon deposition results in a gate electrode that fills the empty space and wraps around the Si bridge. Gate lithography

and etching defines the device gate length. The source and drain are then implanted self-aligned to this patterned polysilicon (Fig. 4.5 (d).)

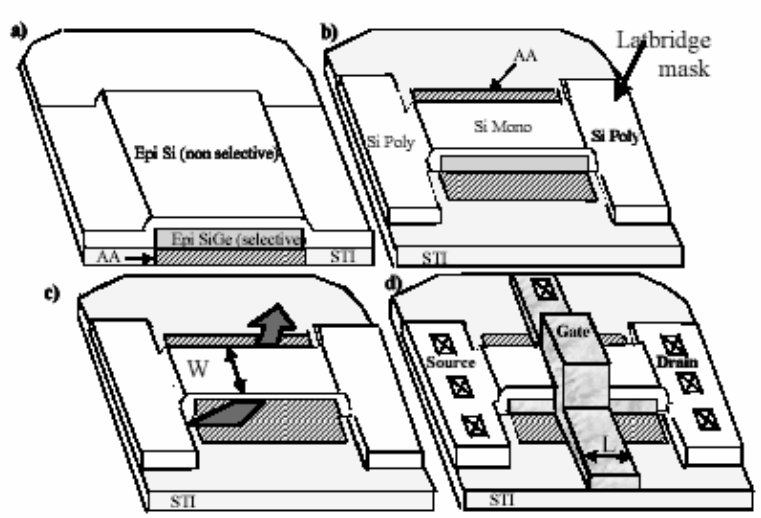


Fig. 4.5 ST Microelectronics SON DG FET Process - main steps (from [4.19])

The key benefit of this structure is that the critical body thickness is set by an epitaxial deposition, thus enabling, in principle, very good T_{si} dimension control and uniformity. However, there is a large overlap between the bottom gate and source/drain. The resulting parasitic capacitance increases the effective device delay.

4.4 Novel Planar DG FET Process Flow

We now present a novel process flow to fabricate the ideal planar DG FET. A sacrificial SiGe layer is used to enable epitaxial deposition of the Si DG FET body above it. This SiGe layer is selectively removed from beneath the Si to create a space for the bottom gate. In addition, enhanced oxidation rate of SiGe compared to Si is used to form sidewall spacers that separate the gate and source/drain regions. The source/drain structure is flared out for low extrinsic resistance. By thus combining the key concepts behind

the SON DG FET and the MIT SSDG FET, we can realize a planar ultrathin body DG FET with low parasitic impedance. The main process steps are described next.

4.4.1 Si/SiGe Trilayer Epitaxy

The first step of this process involves epitaxial chemical vapor deposition (CVD) of a trilayer stack composed of an ultrathin Si layer sandwiched between two pseudomorphic SiGe layers. These SiGe layers serve as placeholders for the gate stack that will be formed later. It is important that they have very good crystalline quality. The lower SiGe layer acts as a template for epitaxy of the ultrathin Si layer above it. This Si layer will eventually be the body of the DG FET and thus needs to be of very high quality with no crystal defects in the bulk or at the interface with SiGe. The CVD process has the advantage, in principle, of allowing extremely good control over film thickness and uniformity. As we have seen earlier, this is perhaps the most critical requirement for a fully depleted or ultrathin body DG FET.

Fig. 4.6 shows a cross section of the trilayer stack. The starting substrate used is a silicon on insulator (SOI) wafer with a thin top Si layer. Ideally, the SOI film should be heavily doped with Phosphorus. The importance of this requirement will be seen later during the spacer formation step. The epitaxial trilayer is capped by an insulating film, such as low temperature deposited silicon dioxide (LTO), which serves as a hard mask as well as an etch stop for source/drain spacer formation during subsequent steps.

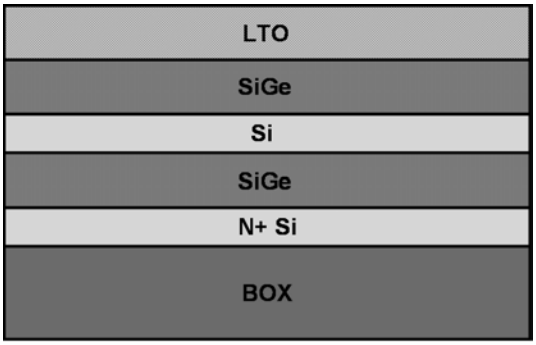


Fig. 4.6 Epitaxial deposition of SiGe/Si/SiGe trilayer stack

The thickness and the Ge atomic fraction in the SiGe layers are chosen so as to maintain their in-plane biaxial compressive strain. It is useful to have a high Ge atomic fraction from the perspective of maximizing etch selectivity to Si during the removal of the SiGe films. For a given Ge atomic fraction, if the SiGe thickness exceeds a critical value, the strain is relaxed via the formation of crystal defects such as misfit and threading dislocations. These could compromise the quality of the Si body layer. On the other hand, if the SiGe is too thin, the sheet resistance of gate electrode increases. The choice of Ge fraction and SiGe thickness is thus one of the design trade-offs in this process. A value of 20 nm for $Si_{0.75}Ge_{0.25}$ can result in pseudomorphic strained films, and at the same time, allows for high etch selectivity to Si.

4.4.2 Fin Patterning

The trilayer stack is patterned into fins using lithography and anisotropic plasma etching. This is a critical lithography step since fin width defines the eventual channel length of the DG FET. For the same reason, the etch needs to be very anisotropic, i.e. with vertical sidewalls. Any slope in the sidewall will eventually result in the top and bottom gates having slightly different sizes and consequently degrade performance. The buried oxide of the SOI substrate acts as an etch stop for the fin etch. Fig. 4.7 shows a cross section and plan view of the structure at the end of this step.

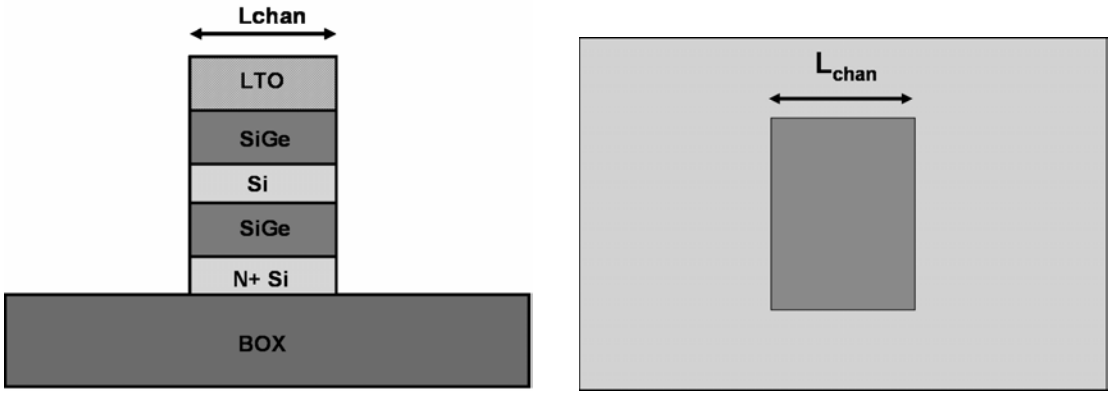


Fig. 4.7 Cross section (left) and plan view (right) of trilayer stack after fin patterning

4.4.3 Sidewall Spacer Formation

After the fins have been patterned, a short oxidation step is carried out. Under certain conditions, the oxidation rate of SiGe is higher than that of Si. The oxidation rate enhancement increases with the atomic concentration of Ge in the SiGe films. A combination of this differential oxidation and HF etching of the thinner oxide formed on Si results in the formation of oxide spacers along the SiGe sidewalls in the trilayer stack. It is well known that Phosphorus doping enhances oxidation rate on Si. Therefore an oxide spacer is also formed around N+ Si layer of the starting SOI substrate. This spacer is important since it insulates the source/drain from the starting Si substrate and prevents the formation of a parasitic single gate FD SOI transistor beneath the bottom gate. If this option is not feasible, the alternative options are a) to design the thickness of the starting SOI layer so that it is fully consumed after the gate oxidation, or b) to implant the starting SOI film so as to increase the threshold voltage of the parasitic FD SOI transistor and prevent it from turning on. The resulting structure is depicted in Fig. 4.8. In plan view, the oxide spacers appear as a thin halo surrounding the fin.

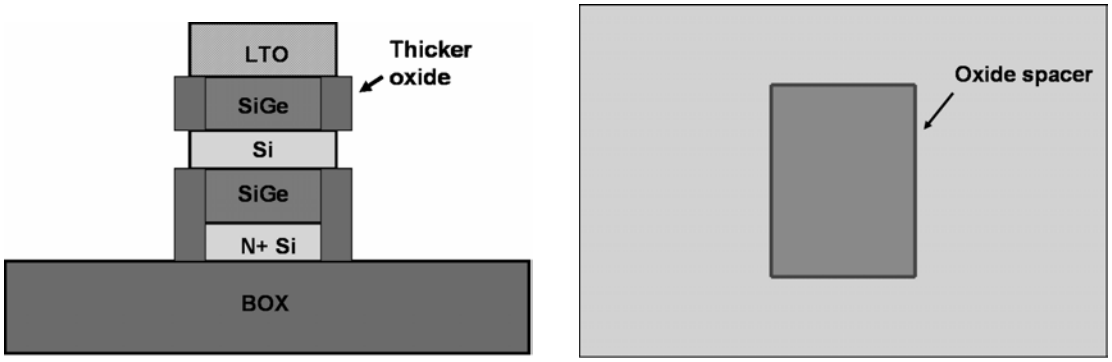


Fig. 4.8 Cross section (left) and plan view (right) schematic of trilayer fin after differential oxidation to form sidewall spacers on SiGe.

4.4.4 Source/Drain Deposition

The source and drain are deposited around the fin using CVD. For this step, one can use selective epitaxial growth of Si off the oxide-free sidewall of the Si body. This

approach is similar to that used for source/drain formation in the MIT SSDG FET. Alternatively, a blanket CVD step could be used (to deposit the source/drain silicon everywhere) followed by an anisotropic spacer etch. In either case, this results in Si source/drain regions that are flared-out from the fin sidewalls. Such structures are useful since they provide a large area over which contacts can be made – reducing the effective contact resistance to the channel. Fig. 4.9 shows the cross section and plan view schematic of the structure after the source/drain formation.

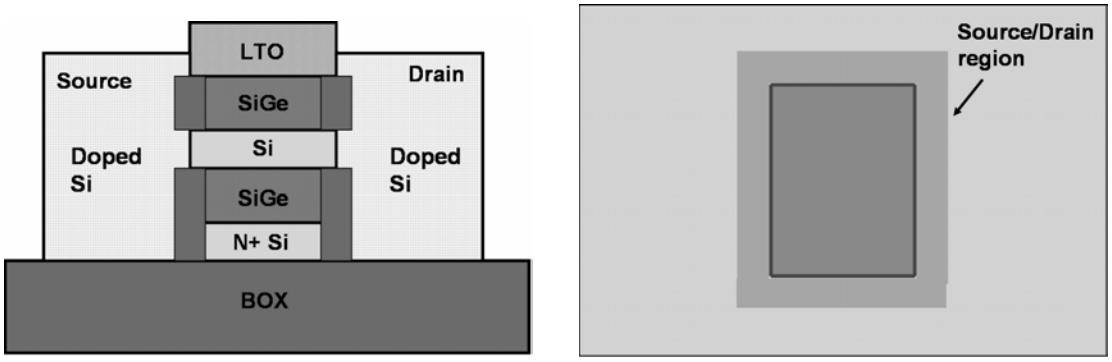


Fig. 4.9 Cross section (left) and plan view (right) schematic of structure after the source/drain formation.

If the source/drain regions are intrinsic as deposited, conventional masked implants can then be used for doping them N and P-type for CMOS integration. However, at the expense of increased process complexity, the source/drain can be doped in-situ during the deposition. This reduces the thermal budget of the process while maintaining low sheet resistance due to a high degree of dopant activation. Also, the use of in-situ doping allows for the tuning of extension underlap by choosing the point at which the dopant source is turned on during the deposition. This enables the proper optimization of extension underlap without changing the gate to source/drain spacer thickness (and hence capacitance.) As we have seen in the previous chapter, the extension underlap must be decoupled from the spacer thickness in order to maximize drive current and at the same time, minimize switching delay.

4.4.5 Width Patterning

At the end of the source/drain formation step, there is a doped Si spacer that surrounds the fin and completely encapsulates the SiGe. The next step uses lithography followed by anisotropic etching to cut off the spacer along the fin edges that are perpendicular to the channel length direction. This step therefore defines the device width as shown in Fig. 4.10. The width patterning step also exposes the SiGe layers in the fin cross section, making them available for selective removal in the next step.

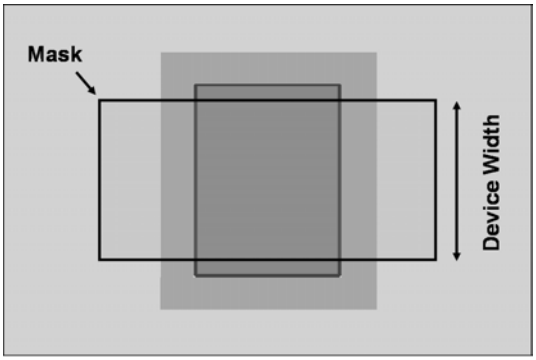


Fig. 4.10 Plan view schematic showing fin width patterning. The SiGe is now exposed in the trilayer cross-section.

4.4.6 Isotropic SiGe Etch

As a consequence of the previous width patterning step, there is now direct access to the SiGe layers in the fin cross section on the two faces in the width direction. An isotropic etch is used to selectively remove the SiGe layers from above and below the Si body layer. This leaves behind a Si beam supported along the length direction by the source and drain regions. Fig. 4.11 shows the cross-section and perspective view of the structure at this point in the process. The maximum width of the Si beam is dictated by the aspect ratio of the tunnels left behind and the selectivity of the SiGe etch to Si. Larger effective device widths can be obtained by forming multiple bridges in parallel, supported by wide source/drain regions. On the other hand, if the width of the Si beam is made very small (comparable to T_{si}), the resulting device is a gate-all-around MOSFET.

Such a structure offers maximum gate control and therefore even higher scalability than the DG FET.

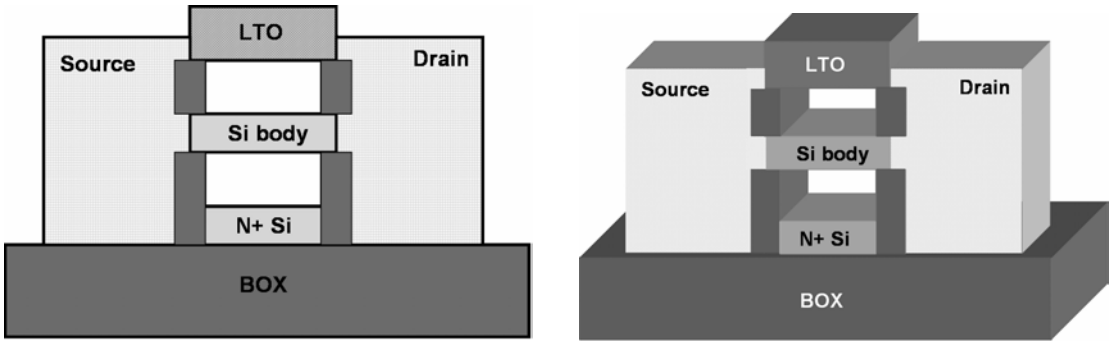


Fig. 4.11 Schematic cross-section (left) and perspective view (right) of the structure after SiGe removal. The Si beam is supported on two sides by the source and drain regions.

4.4.7 Gate Stack Formation

The gate dielectric is formed either by thermal oxidation or by a conformal deposition. Atomic layer deposition (ALD) can be used to deposit a high-k dielectric film. Since ALD is a highly surface reaction-limited process, the conformality of the deposited film is extremely good – leading to uniform dielectric thickness on all surfaces and edges of the Si beam. The gate electrode is deposited using a conformal process in order to fill the high aspect ratio tunnels formed above and below the Si body. Conventional CVD of polysilicon or poly-SiGe from SiH_4 or GeH_4 precursors results in a very conformal film due to the low sticking coefficient of SiH_4. Since there is no easy way to ion-implant the bottom gate, in-situ doping must be used. Alternatively, ALD could be used to deposit a metal gate. The gate material wraps around the Si body layer, resulting in the top and bottom gates being electrically connected. This is acceptable for operation of the transistor in the symmetric DG mode where both gates are driven together. It is possible to electrically isolate the two gates for operation in the variable-threshold back-gate mode. However, this requires additional etching steps and greatly increases the process com-

plexity. Gate electrode patterning is performed next, etching the gate material to leave behind pads on which contacts can be made. The resulting structure is shown in Fig. 4.12. Silicide is formed on the flared-out source/drain regions for low contact resistance. Back-end processing similar to that in conventional bulk FETs completes the process. The schematic of the final structure is shown in Fig. 4.13.

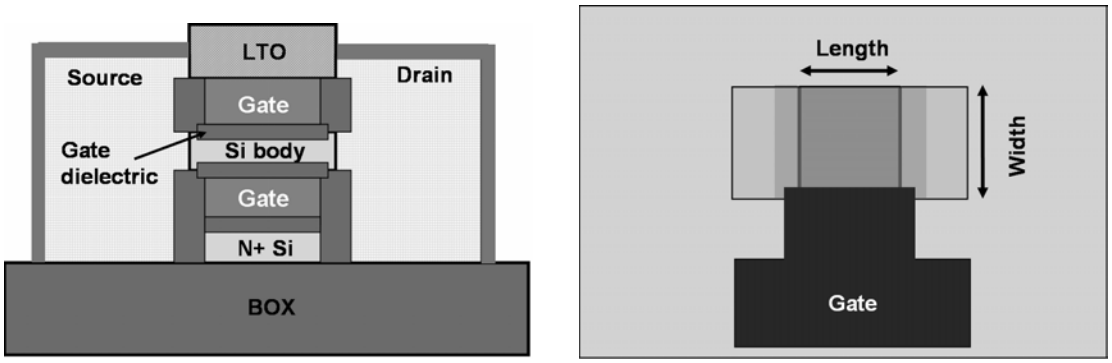


Fig. 4.12 Cross section (left) and plan view (right) schematic of device structure after gate stack formation and patterning.

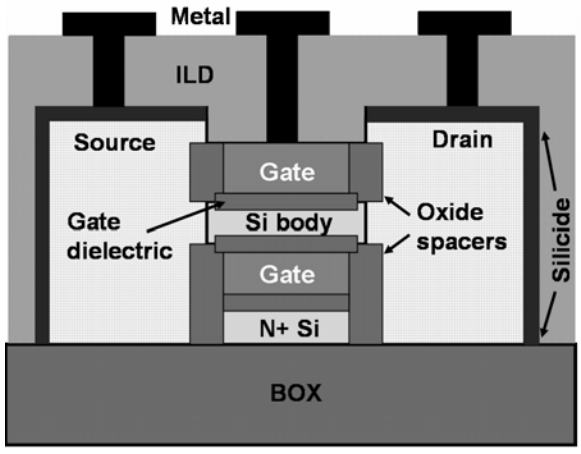


Fig. 4.13 Schematic cross section of completed planar DG FET structure. The metal lines and vias are not drawn to scale.

4.5 Planar Ultrathin Body DG FET Structure – Salient Features

The process flow just outlined results in an 'ideal' planar double-gate FET structure with the following salient features:

- **CVD-defined silicon body**

  This enables sub-lithographic control of the critical body thickness (T_{si}) dimension. This method of forming the channel exploits the high degree of across-wafer film thickness uniformity that is possible with CVD. This should result in better circuit performance due to lesser variation across devices.

- **Fully-self aligned gates**

  The top and bottom gates are self-aligned to each other, equal-sized, and separated from the source and drain by oxide spacers that are much thicker than the gate dielectric. Thus, the parasitic capacitance is minimized in this structure, leading to lower delay.

- **Flared-out source/drain**

  The source and drain regions can be engineered for minimizing extrinsic resistance. The flared-out structure allows for larger contact area and hence reduced contact resistance to the channel. In addition, in-situ doping enables a low thermal budget process for minimizing dopant diffusion and keeping the junctions abrupt. A unique feature of this process is the ability to tune the extension doping underlap independent of the gate-source/drain capacitance. The latter is determined by the differential oxidation of the SiGe sidewall, while the underlap can be set by the turn-on of the dopant source during source/drain film deposition.

- **Gate-last process**

  The use of sacrificial SiGe makes this a replacement gate process. This enables the integration of high-k dielectrics and metal gates which do not experience any high temperature steps in the subsequent processing.

In addition to meeting some of the key requirements for double-gate FET structures that were pointed out in the earlier chapters, the present structure is also an excellent vehicle for research. With this device, fundamental studies can be carried out on carrier transport, dopant diffusion, and quantum confinement in ultrathin body transistors.

4.6 Process Variations

The planar DG FET process described in the earlier sections is quite modular. The channel and source/drain formation steps are relatively independent of one another. This feature can be used to fabricate a number of interesting device structures by variations on the basic theme. Fig. 4.14 shows a schematic of some of these novel devices.

4.6.1 High Mobility Channel DG FET

In place of SOI substrates, if the same process is carried out on SiGe on insulator (SGOI) wafers with the SiGe film relaxed, the middle Si layer in the trilayer stack will conform to the larger in-plane lattice constant of the relaxed SiGe. That will induce in-plane biaxial tensile strain in the DG FET channel. Such strain has been known to increase the low field electron and hole mobility [4.20, 4.21], and hence drive current in bulk Si FETs. At this point, it is not clear if the strain in the Si beam will be maintained after selectively removing the SiGe layers which induced the strain to begin with. However, based on results where others have obtained SiGe-free strained Si on insulator [4.22], it may be possible that the source and drain sidewall regions preserve the strain in the Si body layer.

Instead of using an etch to remove SiGe selectively to Si, one can use an isotropic Si etch that is selective to SiGe to fabricate a planar DG FET with the same structure, but with SiGe channel and source/drain regions. TMAH and NH_4OH are examples of such etchants that attack Si but not SiGe. In the modified process, the trilayer stack sequence is

reversed, with a SiGe film sandwiched between two Si layers. The SiGe film thus grown is under in-plane biaxial compressive strain which improves hole mobility [4.23].

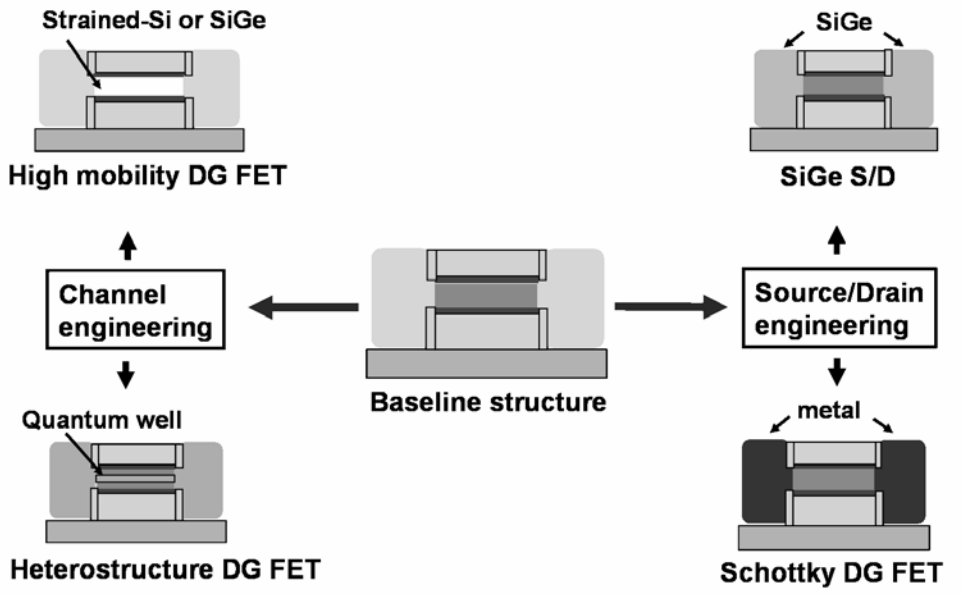


Fig. 4.14 Novel device structures that can be built by variations around the baseline planar DG FET process.

Such a structure may thus yield a higher performance PFET. However, since the enhanced oxidation rate of SiGe over Si works in the 'wrong' way now, the SiGe DG FET formed in this manner will not have the advantage of thick sidewall spacers between the gate and source/drain. In addition, gate dielectrics formed directly on SiGe channels have been reported to have poorer interface quality [4.24] than those on Si channels. In spite of these drawbacks, it may still be interesting to build such transistors for experimental studies of carrier transport in ultrathin SiGe channels.

4.6.2 Heterostructure Channel DG FET

Since the channel Si is formed by an epitaxy step, it is easy to incorporate Si-based heterostructures into the growth sequence. Thus, instead of forming a Si body DG FET,

this process can be used to fabricate heterostructure DG FETs. Depending upon the band offsets in the heterostructure, carriers can be constrained to flow in a quantum well in the center. For example, a Si/strained-SiGe/Si heterostructure has a quantum well formed in the valence band – this leads to center-channel operation for PFETs. On the other hand, a SiGe/strained-Si/SiGe heterostructure forms a quantum well in the conduction band where electron confinement allows center-channel mode operation for NFETs. Simulation studies [4.25] indicate potentially better performance in these kinds of heterostructure DG FET configurations. This is due to the enhanced carrier mobility due to reduced surface scattering and transport in a low transverse electric field region.

In the fabrication of heterostructure DG FETs using this process, the exponential etch rate dependence on Ge atomic fraction can be used to selectively etch the sacrificial top and bottom replacement gate layers while preserving the heterostructure layers in the center.

4.6.3 Source/Drain Material Engineering

In this process, the source and drain regions are formed by a deposition step. This fact can be used to engineer the materials used for the source and drain for better device performance.

In place of Si, SiGe can be used as the material for the source and drain regions. Using SiGe, a higher concentration of active Boron doping can be achieved, thus leading to lower series resistance. In addition, the use of SiGe at the source reduces the impact of the parasitic bipolar action in N-type DG FETs. The valence band offset between the SiGe source and Si body lowers the injection efficiency of the forward biased source-body n-p junction and consequently decreases the bipolar gain. If SiGe is to be used as the source/drain material, it should be completely encapsulated by a thin Si layer in order to protect it during the sacrificial SiGe gate removal step.

It is possible to incorporate metallic source and drain regions in order to operate the device as a Schottky barrier DG FET. One way to achieve this is to use total silicidation

[4.26] of the source and drain silicon regions after the gate stack processing. Alternatively, in place of Si, metals can be deposited after the SiGe sidewall spacer formation step early on in the process. Once again, these metals should either be able to withstand the SiGe isotropic etching or they should be protected from the etchant by using a thin Si layer.

4.7 Summary

In this chapter, the motivation for the planar double-gate FET configuration is discussed. Following that, there is a survey of prior attempts to fabricate planar DG FET[2] structures. The key advantages and shortcomings of those structures are highlighted. Next, a novel process is proposed that combines the best features of some of the earlier planar DG FET implementations. This process results in an ideal planar DG FET structure with uniform ultrathin body and low parasitic impedance. Finally, extensions to this process are described. The modular nature of the process allows interesting device structures to be realized by variations on the basic flow.

[2] Towards the final stages of this Ph.D. research, we found that Samsung [4.27] independently showed a planar multi-bridge DG FET using a process very similar to that proposed in this work. That structure however has some disadvantages compared to our device. Most important, since the sidewall spacer between the gate and source/drain is grown during the gate oxidation step, it is not much thicker than the gate oxide – leading to large parasitic capacitance. That notwithstanding, this demonstration increases our confidence about the industrial manufacturability of planar DG FETs using the process proposed in this dissertation.

References

[4.1] H.-S. P. Wong, "Double-Gate FET – Device design and performance analysis," in *Short Course, 2000 IEEE International SOI Conference*, 2000.

[4.2] H. Pein and J. D. Plummer, "A 3-D Flash EPROM cell and memory array," *IEEE Electron Device Letters*, vol. 14, no. 8, pp. 415-417, Aug 1993.

[4.3] C. P. Auth and J. D. Plummer, "Vertical, fully-depleted, surrounding gate MOSFETs on sub-0.1μm thick silicon pillars," in *Device Research Conference Digest*, 1996, pp. 108-109.

[4.4] H.-J. Cho, F. Nemati, P. B. Griffin, and J. D. Plummer, "A novel pillar DRAM cell for 4 Gbit and beyond," in *Digest of Technical Papers – Symposium on VLSI Technology*, 1998, pp. 38-39.

[4.5] J. M. Hergenrother et. al., "The vertical replacement-gate (VRG) MOSFET: A 50-nm vertical MOSFET with lithography-independent gate length," in *IEDM Technical Digest*, 1999, pp. 75-78.

[4.6] C. K. Date and J. D. Plummer, "SiGe heterojunctions in epitaxial vertical surrounding-gate MOSFETs," in *Digest of Technical Papers – Symposium on VLSI Technology*, 2000, pp. 36-37.

[4.7] F. Nemati and J. D. Plummer, "A novel high density, low voltage SRAM cell with a vertical NDR device," in *Digest of Technical Papers – Symposium on VLSI Technology*, 1998, pp. 66-67.

[4.8] 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.

[4.9] C. J. Petti, J. P. McVittie, and J. D. Plummer, "Characterization of surface mobility on the sidewalls of dry-etched trenches," in *IEDM Technical Digest*, 1988, pp. 104-107.

[4.10] P. Kalavade et. al., "The ultrathin-body vertical replacement-gate MOSFET: A highly-scalable, fully-depleted MOSFET with a deposition-defined ultrathin (< 15 nm) silicon body," in *Proceedings of the IEEE 2002 Silicon Nanoelectronics Workshop*, 2002, pp. 3-4.

[4.11] D. Hisamoto et. al., "A folded-channel MOSFET for deep-sub-tenth micron era," in *IEDM Technical Digest*, 1998, pp. 1032-1034.

[4.12] J. Kedzierski et. al., "High-performance symmetric-gate and CMOS compatible V_t asymmetric-gate FinFET devices," in *IEDM Technical Digest*, 2001, pp. 437-440.

[4.13] Y.-K. Choi et. al., "FinFET process refinements for improved mobility and gate work function engineering," in *IEDM Technical Digest*, 2002, pp. 259-262.

[4.14] Y.-K. Choi et. al., "Sub-20 nm CMOS FinFET technologies," in *IEDM Technical Digest*, 2001, pp. 421-424.

[4.15] H.-S. P. Wong, K. K. Chan, and Y. Taur, "Self-aligned (top and bottom) double-gate MOSFET with a 25 nm thick silicon channel," in *IEDM Technical Dig*est, 1997, pp. 427-430.

[4.16] J.-H. Lee, G. Taraschi, A. Wei, T. A. Langdo, E. A. Fitzgerald, and D. A. Antoniadis, "Super self-aligned double-gate (SSDG) MOSFETs utilizing oxidation rate difference and selective epitaxy," in *IEDM Technical Digest*, 1999, pp. 71-74.

[4.17] K. W. Guarini et. al., "Triple-self-aligned, planar double-gate MOSFETs: Devices and circuits," in *IEDM Technical Digest*, 2001, pp. 425-428.

[4.18] M. Jurczak et. al., "SON (silicon on nothing) – a new device architecture for the ULSI era," in *Digest of Technical Papers – Symposium on VLSI Technology*, 1999, pp. 29-30.

[4.19] S. Harrison et. al., "Highly performant double gate MOSFET realized with SON process," in *IEDM Technical Digest*, 2003, pp. 449-452.

[4.20] J. J. Welser, J. L. Hoyt, and J. F. Gibbons, "Strain dependence of the performance enhancement in strained-Si n-MOSFETs," in *IEDM Technical Digest*, 1994, pp. 373-376.

[4.21] K. Rim, J. Welser, J. L. Hoyt, and J. F. Gibbons, "Enhanced hole mobilities in surface-channel strained-Si p-MOSFETs," in *IEDM Technical Digest*, 1995, pp.517-520.

[4.22] T. A. Langdo et. al., "Preparation of novel SiGe-free strained Si on insulator substrates," in *2002 IEEE International SOI Conference Proceedings*, 2002, pp. 211-212.

[4.23] S. K. Chun and K. L. Wang, "Effective mass and mobility of holes in strained $Si_{1-x}Ge_x$ layers on (001) $Si_{1-y}Ge_y$ substrate," *IEEE Transactions on Electron Devices*, vol. 39, no. 9, pp. 2153-2164, Sep 1992.

[4.24] 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.

[4.25] T. Krishnamohan, C. Jungemann, and K. C. Saraswat, "A novel, very high performance, sub-20nm depletion-mode double-gate (DMDG) $Si/Si_xGe_{1-x}/Si$ channel PMOSFET," in *IEDM Technical Digest*, 2003, pp. 687-690.

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

[4.27] S.-Y. Lee et. al., "A novel sub-50 nm multi-bridge-channel MOSFET (MBCFET) with extremely high performance," in *Digest of Technical Papers - Symposium on VLSI Technology*, 2004, pp. 200-201.

Chapter 5

Planar Double-Gate FET Process Development – Experimental Results

5.1 Introduction

In the previous chapter, we proposed a novel process to fabricate the ideal planar double-gate FET with an ultrathin body and low extrinsic impedance. The key enabling steps in this process are Si/SiGe trilayer epitaxy, enhanced oxidation rate of SiGe over Si, isotropic etching of SiGe selective to Si, and formation of in-situ doped source/drain regions. The process development for these critical steps was carried out at Stanford Nanofabrication Facility (SNF). That was followed by a fabrication run to make proof-of-concept transistors in order to demonstrate the feasibility of the process. The double-gate FETs that resulted exhibit excellent electrostatics as evidenced by near-ideal subthreshold swing and no measurable DIBL. The unit process development and integration to fabricate these transistors constitute the experimental portion of this dissertation research.

The chapter is organized as follows. The process development results for the critical unit steps are described in detail in separate sections. This is followed by a summary of the process integration scheme to build DG FETs. Finally, the measured transistor data

are presented along with analysis to distinguish the double-gate FETs from parasitic bulk transistors that were also formed in the process.

5.2 Si/SiGe Trilayer Epitaxy

Epitaxial growth of high quality SiGe/Si/SiGe trilayers is the first step in the novel planar DG FET process. These films were grown by reduced pressure chemical vapor deposition (RP-CVD) in an ASM Epsilon II single wafer (4 inch) reactor. This tool has a quartz process chamber with tungsten halogen lamps in linear arrays above and below the wafer as well as spot lamps below the wafer. During deposition, the wafer sits on a graphite susceptor which rotates in order to improve uniformity. The reactor has a nominal base pressure in the 6 mtorr range and can typically run processes from atmospheric pressure down to approximately 1 torr. The wafer transfer between the load locks and the process chamber is carried out automatically using a Bernoulli-effect wand. The exchange chamber and the load locks are constantly purged with nitrogen (N_2) to reduce oxygen contamination. A number of gases are plumbed into the reactor - hydrogen (H_2) and N_2 carriers, silane (SiH_4) and dichlorosilane (SiH_2Cl_2) silicon sources, germane (GeH_4) germanium source, 1% phosphine (PH_3) in H_2 and 100 ppm arsine (AsH_3) in H_2 N-type dopant sources, 1% and 100 ppm diborane (B_2H_6) in H_2 P-type dopant sources, and gaseous hydrochloric acid (HCl) for chamber cleaning, Si wafer etching, and selective Si epitaxy. There are four thermocouples - one center thermocouple embedded in the susceptor and one each in the rear, front, and side zones. Temperature control is achieved by using automatic feedback from the thermocouples to set the lamp power in order to achieve the desired temperature setpoint in all zones.

Prior to loading into the epi reactor, the wafers were cleaning using a 4:1 mixture of sulfuric acid (H_2SO_4) and hydrogen peroxide (H_2O_2) to strip organics, a 5:1:1 mixture of DI water, HCl, and H_2O_2 to remove trace metallic contaminants, and hydrofluoric acid (HF) diluted 50:1 in water. The last step in the clean was always a 30-60 seconds dilute

HF dip until the wafer surface turned hydrophobic. This was done to minimize the amount of chemical and native oxide on the Si surface. The wafers were then immediately loaded into the N_2-purged loadlocks. Before every deposition, the process chamber was cleaned at a high temperature using HCl. This was done to remove any previously deposited films from the susceptor and chamber walls so as to maintain the same starting conditions for each wafer.

Once the wafer was transferred into the process chamber, it was heated up to a high temperature in H_2 in order to desorb any remaining native oxide. This H_2 bake step was always performed at reduced pressure (50 torr). The bake temperature was set to 1150 °C for blanket epitaxy and lower temperatures for selective epitaxy on pre-patterned substrates.

5.2.1 Blanket Epitaxy

5.2.1.1 Recipe Development

Recipe development for the trilayer epitaxy was first carried out using blanket deposition on bare Si substrates. These recipes were based on modifications to existing SiGe processes using SiH_2Cl_2 and GeH_4 that had been developed earlier [5.1] in the same reactor. The process pressure was set to 15 torr. The temperature (625-700 °C) and reactive gas flows (GeH_4: 0.5-4.5 sccm, SiH_2Cl_2: 50-75 sccm) were varied in order to obtain target film thickness and Ge atomic fraction. The carrier gas used was H_2. No dopant gases were used during the trilayer stack growth.

Under all these conditions, the deposition of both Si as well as SiGe is surface-reaction limited. This ensures that the epitaxial film growth is, in principle, highly uniform assuming that the temperature can be controlled uniformly across the wafer surface. In this regime, the deposition rate of Si depends linearly upon the partial pressure

of SiH_2Cl_2. However, since the deposition is a thermally activated process, the deposition rate varies exponentially with temperature.

SiGe deposition is a complicated process. In general, with everything else kept constant, increasing the GeH_4 flow, and hence the GeH_4 mole fraction, increases the deposition rate and the Ge atomic fraction of the resulting SiGe film. The SiGe depositions were done at 625 °C. Since the Si deposition rate at that temperature is very low, the Si layers were grown at 700 °C. In addition, the SiH_2Cl_2 flow was increased during the Si growth.

As mentioned in the previous chapter, the trilayer design ideally requires, for a given SiGe layer thickness, as high a Ge atomic fraction as possible while maintaining pseudomorphic growth. The lower limit of the SiGe thickness is set by the conformality of the deposited gate electrode and its acceptable sheet resistance. Higher Ge atomic fraction increases the selectivity of the isotropic SiGe etch as well as the oxidation rate enhancement. The SiGe layers in this work were targeted to be 20 nm thick with 25 % Ge.

5.2.1.2 Trilayer Stack Characterization Method

Cross-sectional transmission electron microscopy (x-TEM) was extensively used to characterize film thickness, composition, and crystal quality. The x-TEM samples were prepared by cleaving the wafer at a suitable location and gluing two small pieces to one another face to face. This sandwich was then thinned down and polished from both sides to end up with a semi-transparent sliver about 10 um thick. Low angle ion milling was then used to further thin the glued interface region down to the region of electron transparency (< 100 nm). A Philips CM-20 TEM with 200 kV electrons was then used to image the sample at high magnifications of up to 600,000X. By tilting the sample with the interface plane exactly perpendicular to the (110) zone axis, high resolution lattice fringes could be obtained. Other tilt angles were used to image crystalline defects in the films using diffraction contrast. Spatially localized film composition was determined

using energy dispersive spectroscopy (EDS). In this method, the focused high energy electron beam impinging on the substrate excites inner shell electrons from the sample atoms to higher levels. These then relax back to the inner shell, emitting x-rays in the process. The x-rays so generated are characteristic of the material. They are detected and analyzed to obtain an intensity spectrum, quantitative analysis of which reveals the composition information. The Ge atomic fraction obtained by EDS was cross-checked and found to be in reasonable agreement (within a couple of percent) with that extracted using other techniques such as x-ray photoelectron spectroscopy (XPS), x-ray diffraction (XRD), and Rutherford backscattering (RBS).

5.2.1.3 Pseudomorphic SiGe Growth on Si – Critical Thickness

The lattice constant of Ge is 4% larger than that of Si [5.2]. In the case of relaxed SiGe alloys, the lattice constant is a linear interpolation between the Si and Ge lattice constants. When SiGe films are epitaxially grown on bulk Si substrates, they are initially under in-plane biaxial compressive strain since the crystal conforms to the smaller lattice spacing of the Si atoms below. This is called pseudomorphic growth. As the SiGe film thickness increases, the strain is relieved by forming misfit dislocations. These lead to defects such as threading dislocations and stacking faults which can enter the middle Si channel layer and compromise its quality. It is therefore very important to ensure that the SiGe layers are fully strained and free of defects.

The critical thickness at which strain relaxation occurs depends inversely on the Ge atomic fraction. The Matthews-Blakeslee (M-B) [5.3] criterion for critical thickness is based on balancing the force due to the misfit strain and the tension in the dislocation. Above the critical thickness, it is energetically more favorable to form misfit dislocations than to maintain film strain. However, there is an activation energy for dislocation nucleation that must be surmounted before the dislocations can form. Thus, by performing the epitaxy at lower temperatures and by limiting the thermal budget during subsequent processing, the SiGe layer can be pseudomorphically grown beyond the

equilibrium critical thickness predicted by the M-B criterion. Such a layer is metastably strained and will eventually relax given a high enough thermal budget. Fig. 5.1 from Houghton [5.4] shows theoretical plots with some experimental data points for the kinetically-limited critical thickness for SiGe strained layer growth on Si at various temperatures.

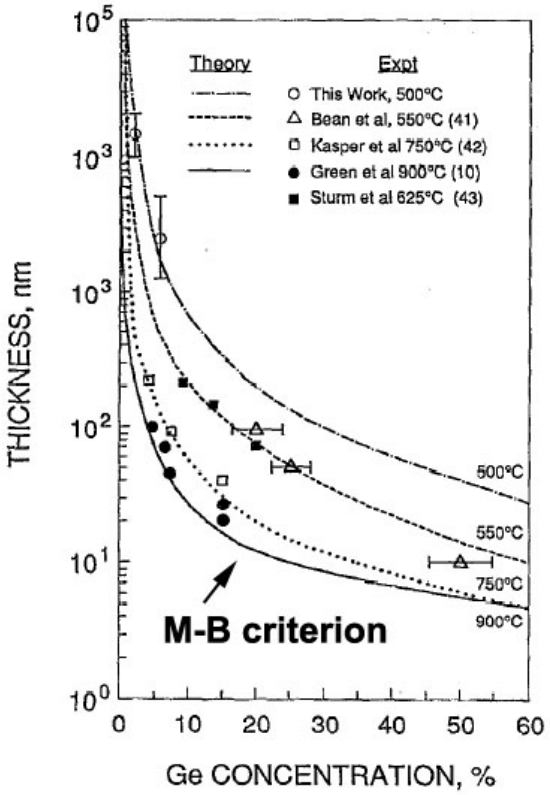


Fig. 5.1 Annotated plots of kinetically-limited critical thickness for strained SiGe growth on (100) Si as a function of growth temperature. This figure is from Houghton [5.4]. CVD data for 625 °C growth shown by the filled squares is perhaps the most representative of the SiGe epitaxy process in our work.

The lowermost solid curve is the equilibrium critical thickness from the M-B criterion. Each of the dashed curves above this represent the maximum thickness beyond which strain relaxation will occur at that growth temperature. The metastable regime lies

in between the critical thickness curves derived from kinetically-limited and equilibrium M-B criterion considerations. From the plots in Fig. 5.1, it is clear that the target SiGe layers in our work (20 nm, 25 % Ge) lie in the metastable regime.

5.2.1.4 Trilayer Epitaxy - Results

Our initial attempts at growing the trilayer stack used 700 °C for the Si layer and 625 °C for the SiGe layers. Before the lower SiGe film growth, a thin 10 nm seed layer of Si was grown, also at 700 °C, to bury any defects that might have resulted at the end of the H_2 pre-bake step. Both the Si layers were grown using SiH_2Cl_2. The GeH_4 to SiH_2Cl_2 ratio for the SiGe growth was set to 4.5 sccm: 50 sccm. These conditions resulted in poor quality film growth with a lot of defects visible in the x-TEM. Fig. 5.2 and Fig. 5.3 depict low and higher magnification images of those films.

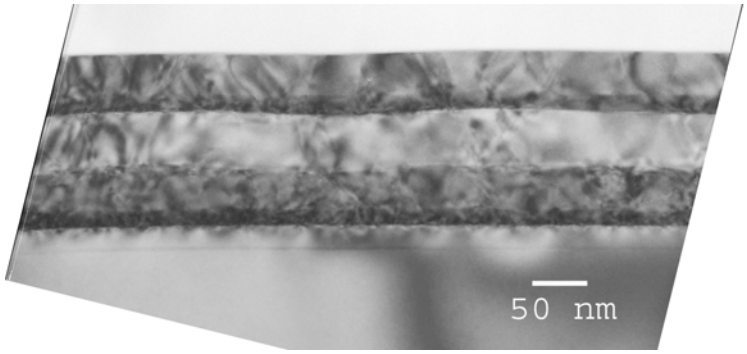


Fig. 5.2 x-TEM (off-zone axis) of our initial attempt to grow the trilayer stack. The crystal quality is quite poor as indicated by the large number of defects in the SiGe and the Si layers. There is also evidence of SiGe pile-up near the bottom of the growth interface.

In Fig. 5.2, taken off the zone axis, there is also a Ge pile-up seen as a darker layer at the bottom of the SiGe layers. This corresponded to an atomic concentration of about 40-45% Ge. The rest of the SiGe layer had close to the 25 % target Ge fraction. The above-target SiGe film thickness and Ge fraction resulted in films which were above the

kinetically-limited critical thickness and hence had dislocations as deposited. Fig. 5.3, taken by tilting the sample onto the (110) zone axis, shows a higher resolution lattice image in which the dislocations and other defects such as stacking faults are seen more clearly.

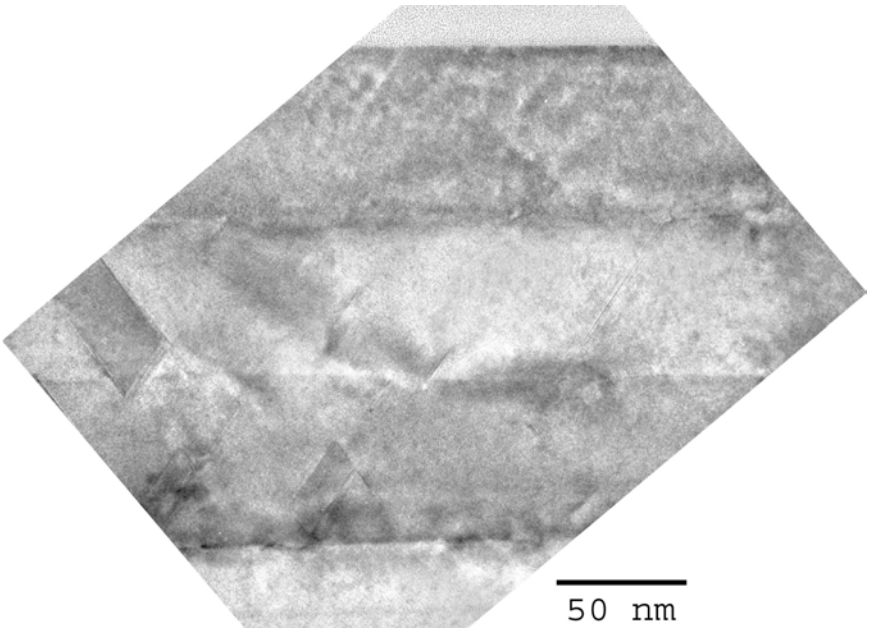


Fig. 5.3 Higher resolution lattice image (on (110) zone-axis) showing crystal defects such as misfit dislocations and stacking faults in our initial trilayer stack.

In some other of our early trilayer films, there was also evidence of strain relaxation by non-planar film growth. Fig. 5.4 shows an x-TEM of such a stack which had quasi-periodic undulations. The reason for the high concentration SiGe piled up layer was found to be improper pressure matching in the reactor between the GeH_4 vent and the deposition chamber. This led to a burst of GeH_4 at the beginning of the SiGe deposition, causing the pile-up. As a work-around to this problem, the recipe was modified to linearly ramp up the GeH_4 flow from 0.5 sccm to 4.5 sccm during the initial stage of the SiGe layer growth. The optimum ramp time was empirically determined using x-TEM characterization. In addition, the growth temperature for the middle Si channel layer was dropped down to 625 °C for the first 5 nm above the lower SiGe film. The rest of the Si

channel layer was then grown at 650 °C. It is known [5.4] that the metastability of a strained SiGe layer can be enhanced by adding an unstrained Si cap layer above it. Therefore, a thin Si cap was deposited at the growth temperature (625 °C) of the lower SiGe before completing the rest of the channel Si layer at a higher temperature and growth rate.

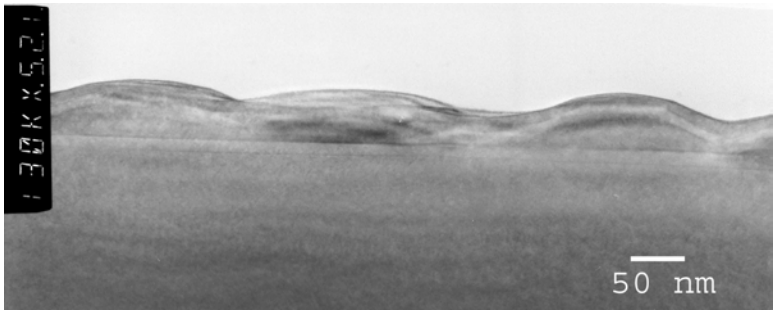


Fig. 5.4 X-TEM of another trilayer stack growth that resulted in strain relaxation via non-planar growth (undulations).

These modifications finally resulted in high quality trilayer stacks with layer thickness and composition close to the target values. Fig. 5.5 shows x-TEMs of these films. No defects were found in these samples. The Ge piled-up layer could not be completely eliminated. However, its thickness (< 5 nm) and Ge atomic fraction (30 %) are well within the critical thickness. Therefore the strain is maintained in the SiGe layers and the entire stack has perfect crystallinity as seen in the high resolution lattice image in Fig. 5.5 (b). The final recipe details are listed in Table 5-1

The epi process was complicated by reactor conditions drifting over time. The age of the SNF epi tool, coupled with several instances of machine down-time and quartz-ware change, caused significant changes in the effective temperature at the wafer surface. This led to changes, mainly in target deposition rates (upto 100 % variation over a 2 year time-frame) The recipe had therefore to be recalibrated every few months, significantly adding to the overall process development time.

Table 5-1 Trilayer SiGe/Si/SiGe CVD epitaxy recipe details

Layer/Parameter	*Temperature*	*SiH_2Cl_2 flow*	*GeH_4 flow*	*Time*
Si seed	700 °C	75 sccm	0 sccm	72 sec
Bottom SiGe ramp	625 °C	50 sccm	0.5-4.5 sccm	7 sec
Bottom SiGe	625 °C	50 sccm	4.5 sccm	22 sec
Si cap	625 °C	75 sccm	0 sccm	5 min
Si channel	650 °C	75 sccm	0 sccm	20 min
Top SiGe ramp	625 °C	50 sccm	0.5-4.5 sccm	7 sec
Top SiGe	625 °C	50 sccm	4.5 sccm	22 sec

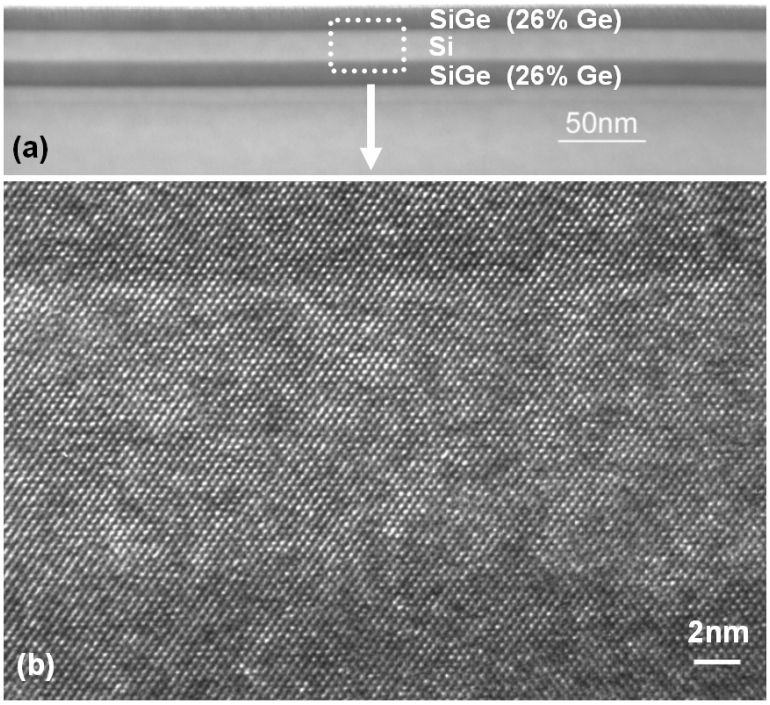


Fig. 5.5 X-TEM of high quality trilayer stacks grown by the epitaxy process with ramped GeH_4 and lower growth temperature Si cap at the start of the center channel layer. No defects were visible in the sample. The high resolution lattice image shown in (b) indicates excellent crystal quality.

5.2.2 Selective Epitaxy

The results shown in the previous section were for trilayer epitaxy by blanket CVD over the entire wafer. It is also useful to be able to deposit the trilayer stack selectively in patterned Si areas without any deposition over the isolation regions. If this can be done, it becomes possible to make the planar DG FET process compatible with conventional bulk-Si starting substrates instead of requiring thinned SOI wafers. This can be accomplished by a minor variation on the basic process described in the previous chapter. Fig. 5.6 shows a schematic of the changed portion of the flow.

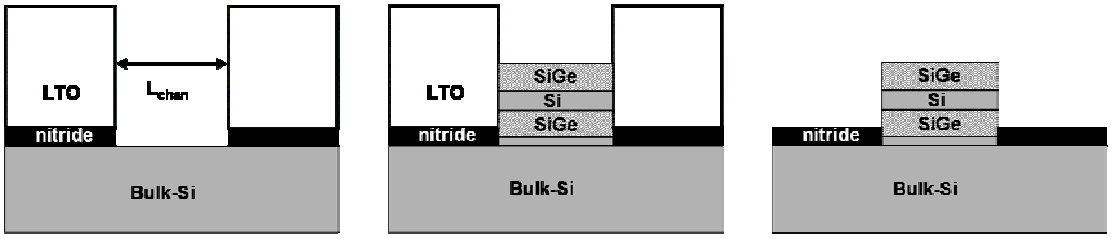


Fig. 5.6 Modified starting steps using selective trilayer epitaxy for a bulk-compatible planar DG FET process.

The process begins on a bulk Si wafer by depositing a stack of silicon nitride capped with low temperature deposited oxide (LTO). This stack is patterned to open active areas where the Si is exposed. This patterning defines the eventual channel length of the DG FET (Fig. 5.6 (a)). Next, selective epitaxy is used to grow the trilayer SiGe/Si/SiGe stack only inside this active area trench (Fig. 5.6 (b)). The LTO is then removed using a wet etch in HF. This leaves behind a fin of the trilayer stack surrounded by nitride (Fig. 5.6 (c)). The rest of the process is carried out in the same way as in the original flow. The insulating nitride layer will now isolate the source and drain regions from the substrate. The selective epitaxy thus enables a quasi-SOI isolation scheme while allowing conventional bulk Si wafers to be used as starting substrates.

In order to test this process variant, patterned substrates were prepared with active area openings in a nitride (20 nm) / LTO (100 nm) stack. The width of the open active area regions ranged from several microns down to 1 μm. The nitride and LTO films were both deposited by low pressure CVD at 785 °C and 400 °C respectively. The openings were made using anisotropic plasma etching of LTO and nitride selective to Si in a CHF_3/O_2 chemistry. After stripping the photoresist and residual polymer that is a byproduct of the plasma etching, the wafers were then subjected to the same wet cleaning prior to epitaxy as the blanket Si wafers. It is known [5.5] that for thin Si films epitaxially grown using SiH_2Cl_2 at 700 °C, the deposition is selective, i.e. the epitaxy occurs only over Si and not over oxide. The Cl atoms that are present in the precursor gas etch the Si nuclei deposited by CVD. The competition between this etching and the deposition results in much longer incubation (nucleation) time for film deposition over oxide than over Si. Hence, for thin layers, the Si films can be selectively grown only over exposed Si regions. In a similar way, SiGe epitaxy from SiH_2Cl_2 and GeH_4 results in selective deposition over Si and not over oxide due to the etching property of Cl as well as GeH_4.

The trilayer epitaxy recipe was carried out on the patterned substrates using the same in-situ 1150 °C H_2 pre-bake and deposition parameters that gave high quality epi layers on blanket substrates. Optical microscope inspection revealed that the size of the openings had increased significantly (on the order of a micron) after the epitaxy process.

Fig. 5.7 shows an x-TEM of the resulting trilayer. It is clear that there is a lot of lateral etching of the oxide/nitride regions. The lateral etching was attributed to the high temperature 1150 °C H_2 bake. Reducing the bake temperature to 900 °C solved the lateral oxide etching problem, but resulted in poor quality of the selective epitaxy layers.

Fig. 5.8 shows an x-TEM of the films grown with an insufficient H_2 bake. In this case, the native oxide is not completely removed from the bare Si surface. This inhibits the epitaxy in certain places and the result is a poor quality discontinuous film.

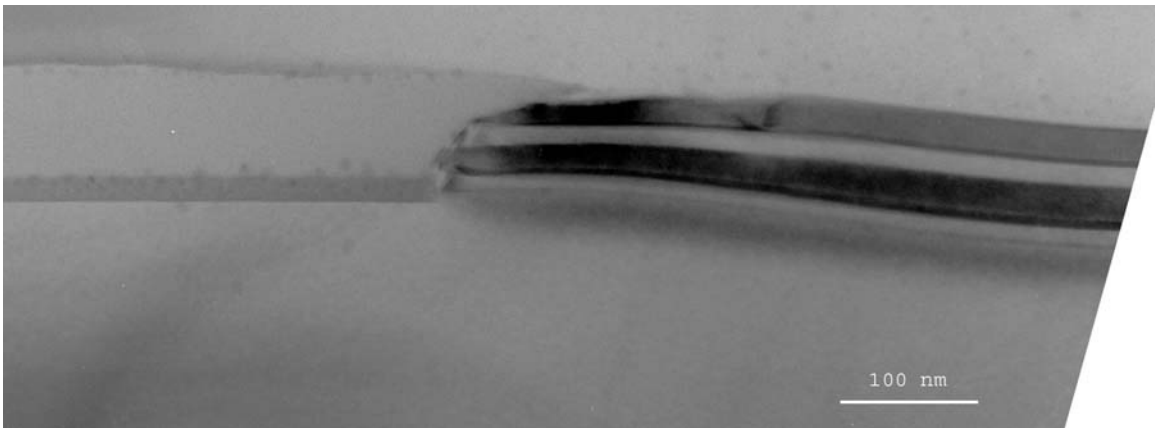


Fig. 5.7 x-TEM of selective trilayer epitaxy on a patterned substrate using 1150 °C H_2 bake for 3 min.

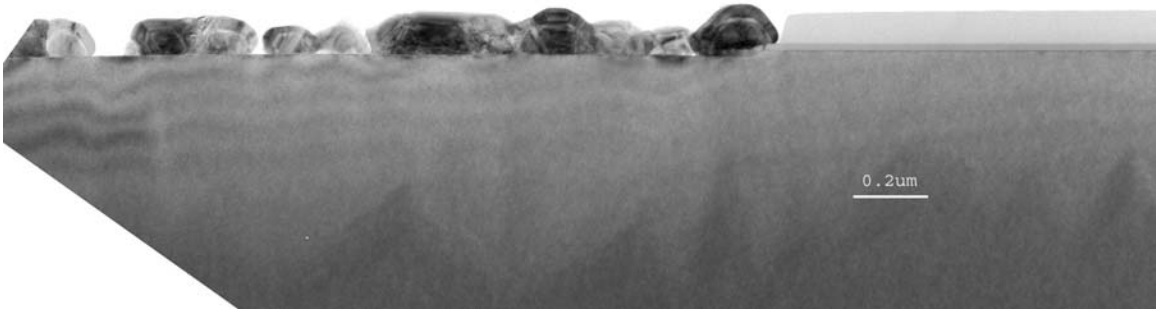


Fig. 5.8 x-TEM of selective epitaxy of the trilayer with an insufficient H_2 bake. This results in poor quality discontinuous deposition.

From these results, it is clear that the H_2 bake needs to be optimized. It should be sufficient to remove the native oxide completely from the Si surface. At the same time, it should not be so strong as to cause too much lateral etching of the isolation oxide. The H_2 prebake optimization was studied experimentally[3] using substrates with patterned thermal oxide.

[3] The H_2 prebake optimization study was done in collaboration with Sameer Jain and Kailash Gopalakrishnan of Stanford University.

Fig. 5.9 shows an x-TEM of the patterned region after the pre-diffusion wafer cleaning and before loading into the epi reactor. The oxide sidewalls are quite vertical, but the Si surface is slightly recessed below the oxide/Si interface. This is a result of the over-etch step during the oxide plasma etching and also due to the post-etch polymer removal process.

Fig. 5.9 x-TEM of an oxide-patterned substrate just before loading into the epi reactor for the selective epitaxy.

Fig. 5.10, Fig. 5.11, and Fig. 5.12 show x-TEMs of samples where the same thickness of Si was grown by selective epitaxy after H_2 bake steps of 1000 °C (3 min), 950 °C (10 min), and 900 °C (10 min) respectively. It is observed that the extent of the lateral oxide etching varies inversely with the bake temperature. The mechanism for this lateral etching can be explained [5.5] as follows.

When the Si wafer is baked at high temperature in H_2, the native oxide is reduced to form volatile silicon monoxide and water vapor through the reaction,

$$SiO_2\ (s) + H_2\ (g) \leftrightarrow SiO\ (g) + H_2O\ (g)$$

where the ‘s’ and ‘g’ in parentheses stand for solid and gas phase respectively. If the reaction chamber has a low partial pressure of oxygen and water vapor contaminants, the reaction proceeds in the forward direction, and the native oxide is desorbed. Since the reaction is thermally activated, the SiO_2 removal rate exponentially increases with the

bake temperature. However, in addition to this reaction, the native oxide also reacts with the Si via the reaction,

$$Si\ (s) + SiO_2\ (s) \leftrightarrow 2\ SiO\ (g)$$

This reaction is self-limiting. Once the native oxide gets consumed, the reaction stops. However, in the patterned substrates, where there is an oxide/silicon interface, this reaction proceeds, consuming both Si as well as SiO_2. It is easy to see how such a prolonged reaction can lead to lateral etching of the oxide sidewalls. Fig. 5.13 shows a plot of the measured lateral etch rate as a function of the bake temperature. The activation energy of the lateral etching process is found to be 4.78 eV.

Fig. 5.10 x-TEM of selective Si epitaxy on an oxide patterned substrate after H_2 bake at 1000 °C for 3 min.

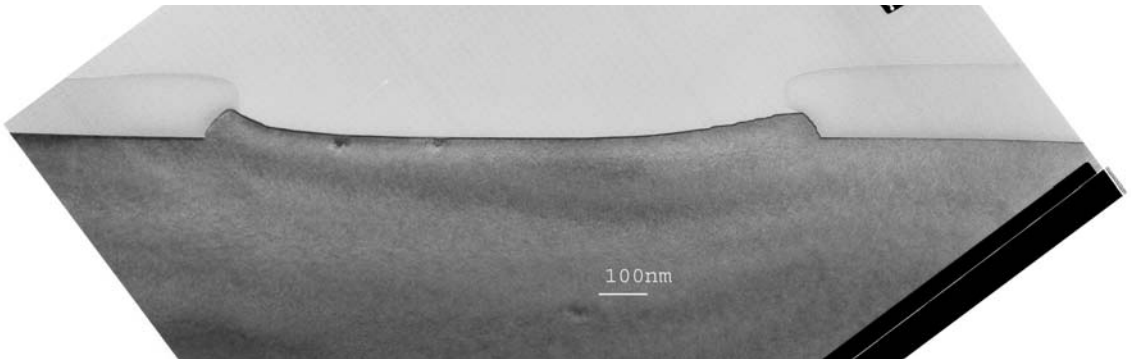


Fig. 5.11 x-TEM of selective Si epitaxy on an oxide patterned substrate after H_2 bake at 950 °C for 10 min.

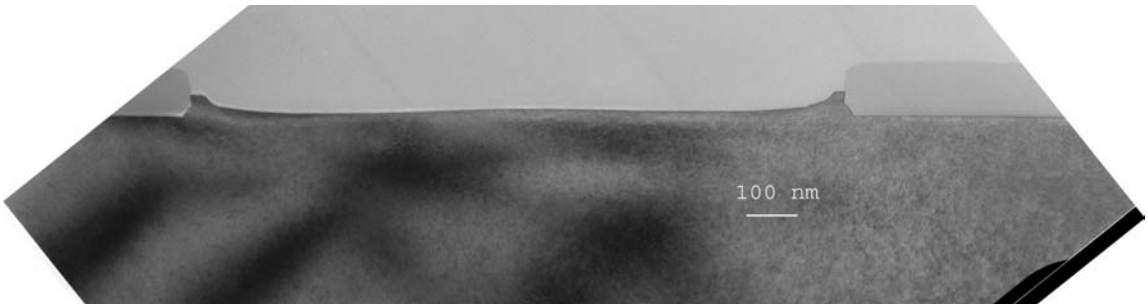


Fig. 5.12 x-TEM of selective Si epitaxy on an oxide patterned substrate after H_2 bake at 900 °C for 10 min.

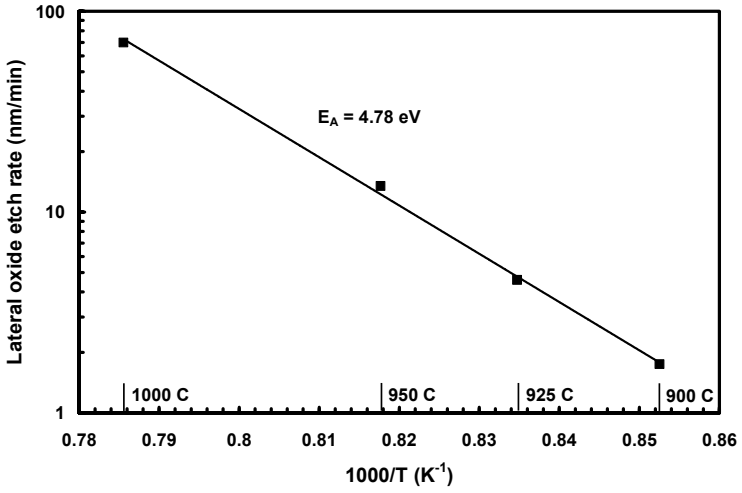


Fig. 5.13 Plot of the lateral oxide etch rate as a function of H_2 bake temperature. The activation energy of this Arrhenius type of process is 4.78 eV.

From the x-TEM images, it is also evident that besides the lateral oxide etching, the Si surface is not flat in the open active area. It seems to curve upwards near the sidewalls. In order to verify if this was due to the selective epitaxy process or due to the H_2 prebake, some samples were subjected to just the bake conditions and then unloaded. Fig. 5.14 shows an x-TEM of one such sample which underwent a H_2 bake at 950 °C for 2 min. It is clear that the Si surface curvature is due to the H_2 bake process. High temperature annealing in H_2 is known to enhance the surface mobility of Si atoms [5.6]. In our

sample, Si atoms from the center of the active area trench move laterally and try to fill the void left behind due to the consumption of Si and oxide during the lateral etching process. This leads to the surface curvature, which is then preserved during the selective epitaxy process that is conformal to Si.

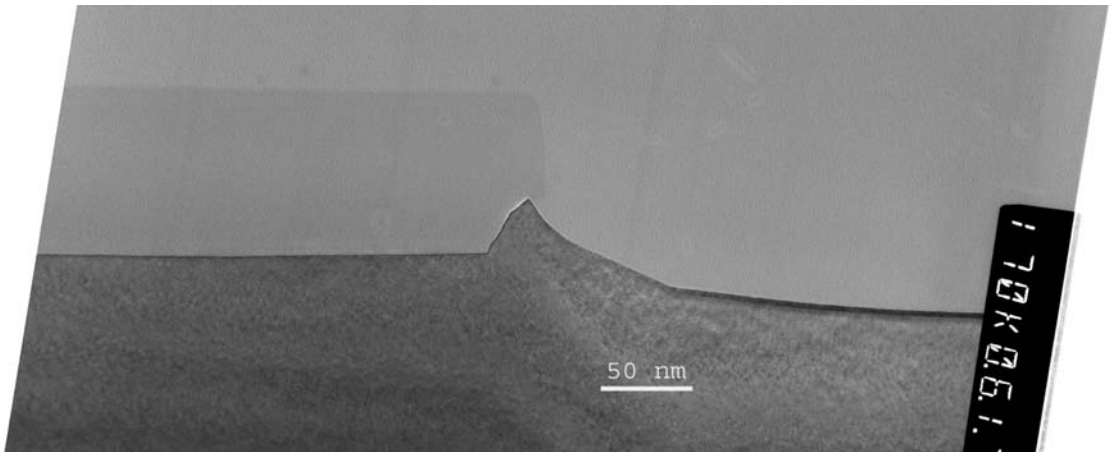


Fig. 5.14 x-TEM of a sample with an oxide sidewall on Si that has been annealed in H_2 at 925 °C for 2 min. Besides the lateral oxide etching, the Si surface moves into the undercut region due to the enhanced surface mobility of Si atoms under those conditions.

The minimum H_2 bake temperature/time required for the complete removal of the native oxide is a function of the initial thickness of native oxide present on the samples as well as the residual oxygen and water vapor contamination in the process chamber. In our case, since the wafers came out of the wet clean with an HF-last step, there should be very little or no native oxide remaining on the hydrogen-terminated surface. However, the oxygen background levels in the reactor still necessitate a high temperature H_2 bake. In our case, we found the minimum bake condition for good epitaxial film deposition was 900 °C for 10 min. Under these conditions, the lateral oxide etching is less than 20 nm. By using this minimum bake step, SiGe/Si/SiGe trilayer films were selectively deposited in the Si active areas of the oxide/nitride patterned substrates. This finally resulted in high quality selective epitaxy as shown in Fig. 5.15.

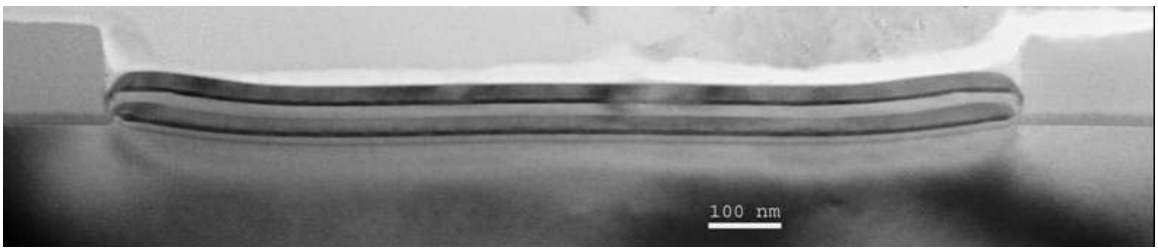


Fig. 5.15 x-TEM of a SiGe/Si/SiGe trilayer stack grown by selective epitaxy within a Si active area trench between patterned oxide/nitride isolation regions.

5.3 SiGe Enhanced Oxidation

The oxidation of SiGe is a complicated process and depends upon the temperature, oxidizing ambient and the Ge atomic concentration. Researchers have investigated SiGe oxidation under dry [5.7, 5.8] and wet conditions [5.9, 5.10]. For dry oxidation, the oxide growth rates are similar for SiGe and Si. However, if wet oxidation is done, the SiGe oxidation rate is higher than that of Si. The oxidation rate enhancement is the highest in the linear regime which is surface reaction limited. In the parabolic regime, where the oxidation rate is limited by the diffusion of steam through the oxide, both SiGe and Si have similar oxidation rates. At low temperatures (< 700 °C) or for high Ge content (> 50%) layers, Ge is incorporated into the growing oxide. At higher temperatures (> 800 °C) and lower Ge content, the oxide grown is pure SiO_2 and Ge piles up below it ('snow plough effect').

The mechanism for the linear-regime oxidation rate enhancement of SiGe is due to the catalytic action of Ge which promotes the decomposition of steam or Si-OH bonds to form GeO. The GeO so formed, being less stable than SiO_2, is reduced by Si at the growth interface to form SiO_2, with the rejection of Ge atoms. If the supply of Si atoms is reduced, either due to low temperature or high Ge atomic concentration, the Ge gets

oxidized and a mixed SiGe oxide is formed. In this case, the oxidation rate is even higher, since both Ge and Si are being oxidized.

In order to experimentally verify the oxidation rate enhancement of SiGe, blanket films of amorphous SiGe with different Ge concentration were deposited on thermally oxidized Si substrates using LPCVD. SiH_4 and GeH_4 were used to deposit SiGe at temperatures below 500 °C. These films were then crystallized by annealing in N_2 at 800 °C. The film thickness was measured using spectro-reflectrometry and the Ge concentration was determined using XPS. All the samples were oxidized together in an atmospheric pressure furnace at 750 °C for 30 min in pyrogenic steam. The oxides so grown were then removed in dilute hydrofluoric acid and the remaining Si/SiGe thickness was measured. By assuming the thickness of the grown oxide to be 2.2 times the Si/SiGe consumption, the oxide thickness was extracted. The measured oxidation rate enhancement ratio (oxide thickness over SiGe divided by oxide thickness over Si) is plotted in Fig. 5.16 as a function of the starting Ge atomic fraction in the SiGe.

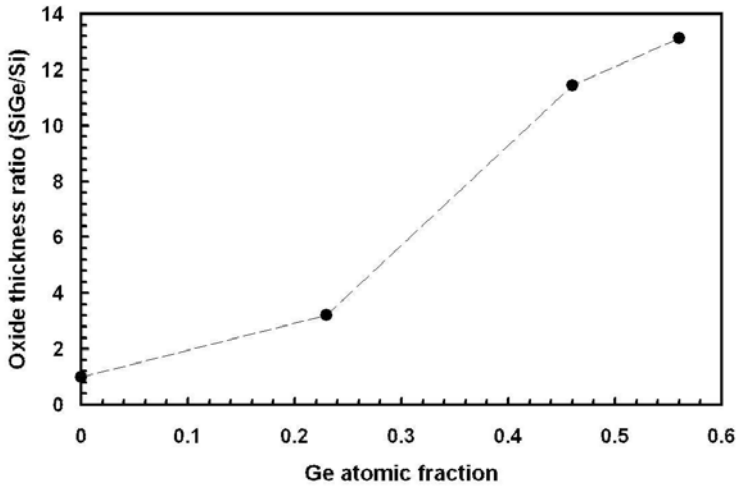


Fig. 5.16 Measured oxidation rate enhancement on blanket polycrystalline SiGe films as a function of Ge fraction. The oxidation conditions were 750 °C for 30 min in steam. The oxide thickness on the poly-Si was 17.5 nm.

From this graph, we see that the oxidation rate at the target 0.25 Ge fraction is about 3-4 times higher than that of Si under the conditions used. Also, the etch rate in HF of the oxide on the $Si_{0.75}Ge_{0.25}$ layer was very similar to that of the oxide on Si, indicating that it was most likely pure SiO_2 as would be expected from previous data in the literature.

5.4 Isotropic SiGe Selective Etching

Selective removal of the sacrificial SiGe layers above and below the ultrathin Si body is one of the critical steps of our proposed planar DG FET process. The SiGe etch needs to be isotropic since it has to reach the SiGe that is buried under other layers and not accessible from the top. It is important to maximize the selectivity of this etch to Si since that constrains the maximum width of the basic device.[4] For example, with a Si body thickness of 10 nm and a SiGe to Si etch selectivity of 10:1, the maximum undercut of SiGe is 50 nm before the Si at the end gets etched. Since the SiGe etch occurs from two sides, and there are two channels in the DG FET, the maximum width in this situation will be 200 nm. In practice, either higher selectivity or narrower devices will be desirable in order to minimize etching of the ultrathin Si body.

The literature contains reports of isotropic SiGe etch recipes that are selective to Si. These approaches include wet etching [5.11, 5.12, 5.13] as well as chemical dry etching in a downstream plasma [5.14]. In order to develop a suitable SiGe etch process for our purpose, we carried out wet etching experiments using two kinds of etch recipes. These are discussed next.

[4] Even so, effectively wide DG FETs can be fabricated, perhaps at the cost of pitch, by placing multiple basic FETs in parallel across the source and drain regions.

5.4.1 Ammonium Hydroxide / Hydrogen Peroxide / Water

Johnson et. al. [5.12] reported selective chemical etching of SiGe with respect to Si using a mixture of ammonium hydroxide (NH_4OH), hydrogen peroxide (H_2O_2), and water. This solution, also known as 'SC-1', is one of the components of the standard RCA clean [5.15]. It is typically used to remove particle contamination from the wafer surface. The mechanism of this etch involves oxidizing the surface using the H_2O_2 and removing the surface germanium oxide using NH_4OH. Since Ge oxidizes faster than Si, the net etch rate of SiGe due to this repeated oxidation-etch sequence is greater than that of Si. In [5.12], the authors reported SiGe etch rates as a function of Ge concentration, temperature, and the composition of the etching solution. They also found that the etching leaves behind a smooth Si surface after the SiGe removal.

Our etching experiment using this recipe was done using a 1:1:5 dilution of NH_4OH:H_2O_2:H_2O at 75 °C. Polycrystalline Si and SiGe samples of varying Ge concentration were generated by blanket LPCVD over thermally oxidized substrates. These were then broken into smaller pieces for the etching experiments. Before etching, the samples were dipped in 50:1 HF to remove native oxide. The initial film thickness was then measured using spectro-reflectometry. The samples were then placed in the etching solution for a certain time interval. After that, they were once again dipped in 50:1 HF to remove any chemical oxide. The post-etch thickness was measured in the same way as before. By varying the etching time, a linear etch rate was extracted for each sample.

The measured etch selectivity to Si is plotted in Fig. 5.17. The etch rate of SiGe depends exponentially on the Ge atomic concentration. The high-Ge content etch rates are similar to those reported in [5.12]. However, for pure Si and the low-Ge content layers, our measured etch rates are about twice that of [5.12]. The reasons for this discrepancy are not very clear, but it could be related to the changing concentration of H_2O_2 in our heated solution as a function of time. Also, Johnson et. al. [5.12] do not mention if they removed the native oxide prior to the wet etching. We observed that the presence of the native oxide on Si suppresses the etch rate in the NH_4OH/H_2O_2 mixture but does not

affect the SiGe etch rate. This could be another reason why their Si etch rate, and hence etch selectivity, is different from what we measured.

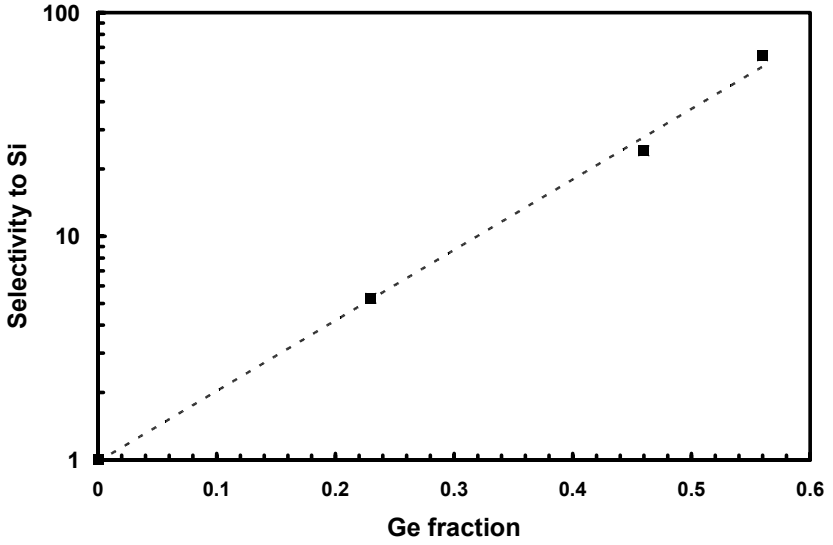


Fig. 5.17 Measured etch selectivity to Si of the SiGe wet etch using $NH_4OH:H_2O_2:H_2O$ (1:1:5) at 75 °C. The samples are polycrystalline films of LPCVD SiGe with varying Ge concentration. The dashed line is a fit to the measured data.

From our data, it is clear that for SiGe etch selectivity higher than 10:1, we need to use SiGe films containing more than 30 % Ge. With an aim to get higher etch selectivity using the 25 % Ge layers for which we had already developed the trilayer epitaxy process, we investigated another wet etch recipe.

5.4.2 Hydrofluoric Acid / Hydrogen Peroxide / Acetic Acid

Carns et. al. [5.11] published data for the chemical etching of SiGe with high selectivity to Si. They used a room temperature mixture of HF and H_2O_2 diluted with acetic acid (CH_3COOH). The basic etch mechanism is similar to the case of NH_4OH/H_2O_2. The H_2O_2 oxidizes the surface while the HF etches that surface oxide. Since the surface oxidation is promoted in the presence of Ge atoms, SiGe etches faster than Si. Using

CH_3COOH instead of water as the diluent greatly enhances the SiGe etch rate, and hence selectivity. In addition, it helps leave behind a smooth Si surface. The mixture needs to be stabilized for a certain period of time before the higher etch rate due to CH_3COOH dilution can be observed.

The exact mechanism for SiGe etching using this mixture is quite complicated. One possible explanation, due to [5.11], involves a reaction between H_2O_2 and CH_3COOH to form a peroxy acid.

$$2\ CH_3COOH + 2\ H_2O_2 \rightarrow 2\ CH_3COOOH + 2\ H_2O$$

This peroxy acid then ionizes Ge atoms via the reaction,

$$Ge + 2\ CH_3COOOH + 2\ e^- \rightarrow Ge^{+2} + 2\ CH_3COO^- + 2\ OH^-$$

Since the first reaction is slow, the solution needs to be stabilized for the etch rate to reach steady state. Also, the need for electrons in the second reaction causes n-type samples to etch faster than p-type.

In our etching experiments with this solution, we used 49 % HF, 30 % H_2O_2, and glacial acetic acid in the ratio 1:2:3 by volume. This solution was allowed to stabilize for 3 hours after mixing. The stability was confirmed by checking that the etch rates were not time dependent after waiting for this period. As in the case of the previous wet etching experiments, polycrystalline Si and SiGe samples with different Ge atomic fraction were prepared by blanket LPCVD over thermally oxidized Si wafers. These were then broken into smaller pieces for the etching experiments. The actual etch was preceded and followed by dips in dilute HF to remove the native and chemical oxide. All the etching was done at room temperature. The thickness of the surface oxide-free samples was measured using spectro-reflectometry before and after etching. The measured etch rate of SiGe is plotted as a function of Ge atomic fraction in Fig. 5.18.

Similar to the NH_4OH/H_2O_2 case, the SiGe etch rate, and hence selectivity to Si, increases exponentially with Ge atomic fraction. At the target 25 % Ge content, the etch selectivity is greater than 60:1 – much higher than what was obtained with the NH_4OH/H_2O_2 solution.

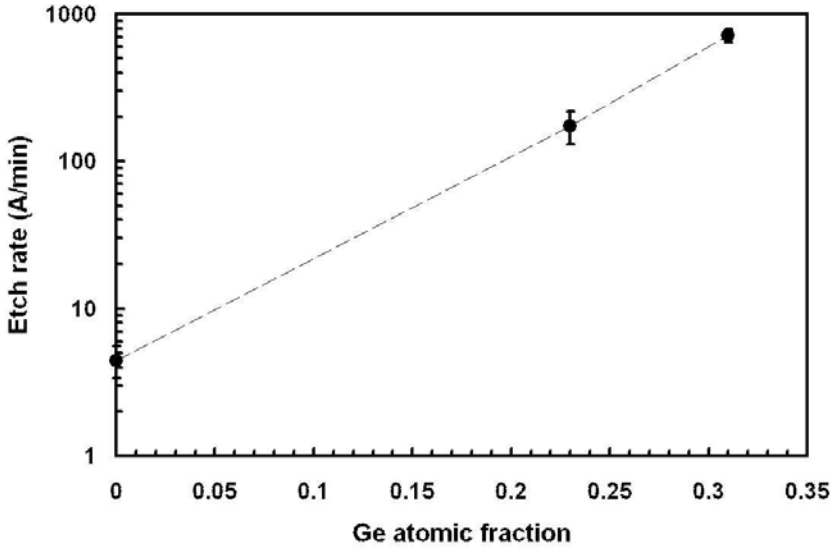


Fig. 5.18 Measured etch rate of polycrystalline SiGe films as a function of Ge atomic fraction using $HF:H_2O_2:CH_3COOH$ (1:2:3 by volume).

This etching method was also carried out on patterned test structures similar to the actual structures in the planar DG FET process flow. Trilayer epitaxial SiGe/Si/SiGe stacks were formed on Si. These were patterned into fins several microns long and 1.5 μm wide. The fins were surrounded by polysilicon spacers created by conformal LPCVD of polysilicon at 620 °C, followed by an anisotropic Si etch. After another photolithography step, an anisotropic etch was used to cut the fins perpendicular to their length and expose the cross section. These samples were then etched in the $HF:H_2O_2:CH_3COOH$ mixtures for varying lengths of time. Tilted view scanning electron microscopy (SEM) was used to observe the samples after etching. Fig. 5.19 shows SEM images of one such sample after etching for 10 and 20 minutes respectively.

The extent of SiGe isotropic etching beneath the thin epi Si layer can be seen as an undercut in the images. Two things can be learned from these experiments. First, the lateral etch rate of the nominally $Si_{0.75}Ge_{0.25}$ layers, about 30 nm/min, is close to that obtained on the blanket films (see Fig. 5.18). Also, the etching proceeds linearly with

time – the 20 min etched sample shows twice the undercut as the one etched for 10 min. This indicates that, at least up to the extent etched in this experiment, there is no evidence of self-limited etching due to possible capillary action.

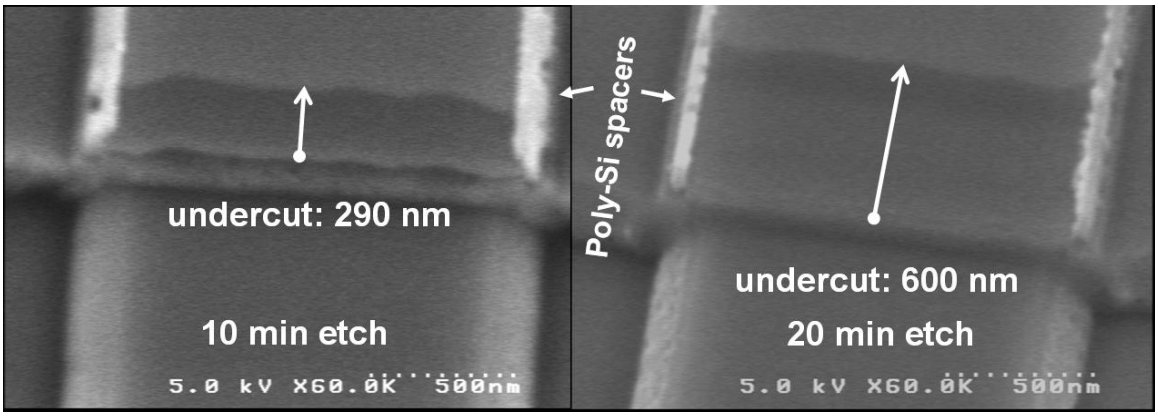


Fig. 5.19 Tilted view SEM images of patterned test structures subjected to isotropic SiGe etching using $HF:H_2O_2:CH_3COOH$. These structures are fins consisting of the epitaxially grown SiGe/Si/SiGe trilayer stack and are surrounded by polysilicon spacers. The spacers protect the sides of the fin, but have been removed from the front, allowing the etchant to access the buried SiGe layers in the trilayer stack. The isotropic SiGe etching can be clearly seen as a time-dependent undercut.

While the higher selectivity to Si makes this a better option than the SC-1 etch, there are two drawbacks to using this recipe. The presence of HF causes etching of oxide. This means that the spacers grown by enhanced SiGe sidewall oxidation will not survive this etch. This is not a catastrophic failure mode, since during gate oxidation, there will be an oxide grown on the doped Si source/drain regions that isolates them from the gate electrode. But this may not be much thicker than the gate oxide. The high-speed performance of such a device would suffer, but the dc behavior will not be affected. Secondly, while the selectivity to undoped Si is very high, heavily doped n-type Si etches at a faster rate, due to the ready availability of electrons. The selectivity of SiGe etching to N^+-Si isn't as good. As a result, special steps would be required to protect the source/drain

regions during the SiGe etch if they are in-situ doped n-type. This problem does not apply to heavily p-doped Si.

5.5 In-situ Doped Source/Drain Deposition

As indicated in the previous chapter, the source and drain in the planar DG FET flow can be formed either by blanket CVD followed by a spacer etch, or by selective epitaxy off the channel sidewall. Ideally, these should be doped in-situ. The main advantages of the in-situ doped source/drain are (a) low thermal budget process which minimizes dopant diffusion, and at the same time gives high degree of dopant activation, and (b) the ability to tune extension doping underlap independent of the gate to source/drain sidewall spacer thickness.

N and P-type in-situ doped source/drain processes were developed experimentally using LPCVD and RPCVD respectively.

5.5.1 N-type Source/Drain Process

Polysilicon source and drain deposition using in-situ phosphorus doping was done using LPCVD from SiH_4 and PH_3 in a hot wall Tystar furnace. The recipe details are listed in Table 5-2. 150 nm thick films were deposited on thermally oxidized Si substrates. The doped Si layer was amorphous as deposited and had very high resistivity as indicated by four point probe measurements. These wafers were then furnace annealed at 650 °C in an inert N_2 ambient for 30 min. After annealing, four point probe measurements showed a large reduction in sheet resistivity. Hall-effect measurements were made on the annealed samples to extract a high active doping concentration of 2.5×10^{20} cm^{-3}.

Table 5-2 Deposition parameters for in-situ Phosphorus doped Si using LPCVD.

Recipe Parameter	*Value*
Temperature	580 °C
Pressure	400 mtorr
SiH_4 flow	100 sccm
PH_3 flow	1.5 sccm
Deposition time	45 min

5.5.2 P-type Source/Drain Process

Since there was no P-type dopant source available in the LPCVD tube, the in-situ P-type doped source/drain process was developed in the ASM epi reactor described earlier in section 5.2. The process was carried out at 580 °C under reduced pressure (25 torr). SiH_4 was used to deposit 150 nm of Si non-selectively with B_2H_6 as the dopant source. The recipe details are given in Table 5-3. After deposition, the films were annealed in an inert N_2 ambient either in a furnace or in-situ at 650 °C for 30 min.

The samples used for the Boron-doped Si deposition were wafers with patterned fins of the trilayer epi SiGe/Si/SiGe stack capped with 120 nm of LTO. Before loading into the epi reactor, the wafers were subjected to a 50:1 HF-dip until they became hydrophobic. Once they were transferred to the process chamber, some wafers directly had the doped-Si deposition, while others were first baked at 900 °C for 10 min in H_2 to remove any remaining native oxide.

Table 5-3 Recipe details for in-situ Boron doped Si by RPCVD.

Recipe Parameter	*Value*
Temperature	580 °C
Pressure	25 torr
SiH_4 flow	50 sccm
B_2H_6 flow (1% in H_2) (source)	25 sccm
B_2H_6 flow (1% in H_2) (injector)	20 sccm
Diluent flow	0
Carrier gas and flow	N_2, 10 slm
Deposition Time	7 min, 10 seconds

Fig. 5.20 shows an x-TEM of an annealed sample that had the doped Si deposition without the H_2 bake step. The presence of native oxide inhibits epitaxy and the layer becomes polycrystalline upon annealing. Fig. 5.21 shows an x-TEM of an annealed sample with doped Si deposition after the H_2 pre-bake. In this case, we see that the doped layer is single crystal in region where it is in contact with the substrate and sidewall Si, whereas it is polycrystalline over and around the oxide cap. The sloped trilayer sidewall is due to the Si surface mobility during the H_2 bake step. Hall effect measurements on the single-crystal portion of the B-doped Si indicated a high active B concentration of 3 x 10^{20} cm^{-3}.

One of the potential concerns with the high temperature H_2 bake step is the possibility of Si/Ge interdiffusion and/or strain relaxation of the metastable SiGe layers with accompanying dislocation generation. X-TEM imaging was used to examine the H_2-baked layers for crystalline defects and interdiffusion. No defects were seen in the films. Also, there was no noticeable evidence of Si/Ge interdiffusion. This is consistent with predictions from theoretical calculations [5.16] of interdiffusion under these annealing

conditions. In addition, XRD measurements were made to check for strain relaxation. These indicated a very low degree (~ 0.5 %) of strain relaxation, confirming that the SiGe layers did indeed maintain their strain even after 900 °C annealing in H_2.

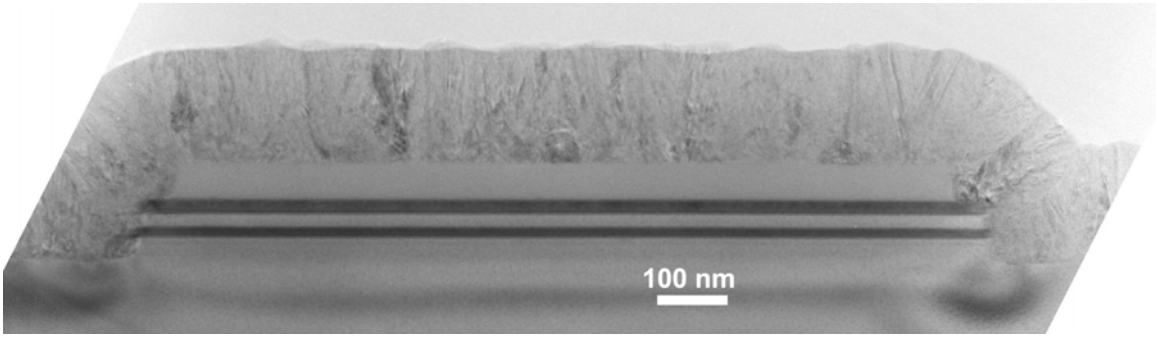


Fig. 5.20 x-TEM of an annealed in-situ B-doped RPCVD Si layer on patterned trilayer stack. No H_2 annealing was done before the doped Si deposition. As a result, it crystallizes to form polysilicon everywhere.

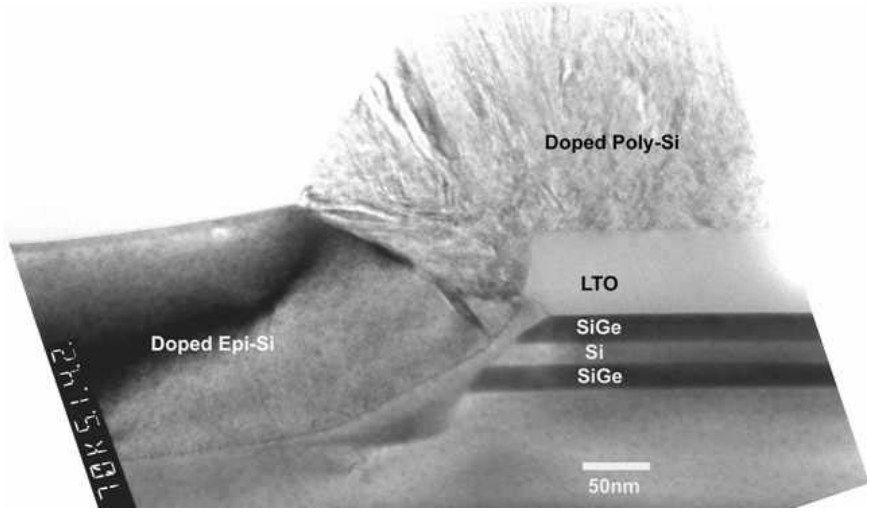


Fig. 5.21 x-TEM of an annealed in-situ B-doped RPCVD Si layer on the patterned trilayer stack. A 900 °C H_2 bake for 10 min ensured removal of native oxide, leading to single crystal epi formation over the silicon surface and the stack sidewall. Poly-Si is formed over and around the LTO cap.

5.6 Process Integration

Having verified the key unit process steps individually, the next step was to put them all together to demonstrate functional DG FETs as proof of concept. The process flow was a somewhat simplified version of the one described in the previous chapter. Table 5-4 lists the sequence of fabrication steps. The steps listed in bold font were described in detail in the previous section. Some of the main highlights of the process integration are described below.

- Process simplifications included a) use of bulk instead of thinned SOI substrates to keep costs down for the prototype transistor demonstration, (This had some implications on the measured data as we will describe in the next section.) b) Omitting the SiGe enhanced oxidation step to form oxide spacers from gate to source/drain. Since we decided to use the $HF/H_2O_2/CH_3COOH$ recipe for high selectivity SiGe isotropic etching, the loss of selectivity to oxide meant that the differential oxide spacer would have been etched during this step. If the spacers have to be retained, the lower selectivity NH_4OH/H_2O_2 recipe could be used. The absence of these spacers does not affect the dc DG FET performance. Finally, c) Only p-channel devices were made. This was due to the process complications that would have arisen in order to protect the in-situ doped n-type source/drain which do not withstand the SiGe wet etchant as well as the heavily doped p-type source/drains.

- A mixed optical/e-beam lithography scheme was used. Optical lithography (minimum resolution of 1 μm in our lab) was used for zero level alignment target definition and for the final metal pad patterning. (The use of e-beam for this high area coverage level would have been very time-consuming.) E-beam lithography (minimum resolution of 200 nm) was used for the other levels. The high resolution was required for the width patterning step, where SiGe etch selectivity considerations limited the maximum patterned width to 300 nm. Also, the good overlay alignment capability (<

50 nm) of the e-beam system was required for the source/drain patterning step, where a misalignment between the source/drain finger and spacer could have resulted in open circuits.

- Relatively conservative value of gate oxide (6 nm) was used in order to avoid further complications due to tunneling through thin oxide. Also, N-type doped polysilicon was used as a gate electrode since the LPCVD process used to deposit the layer was known to have excellent conformality. The Tystar deposition tool did not have the capability to deposit in-situ P-type doped Si. The use of low workfunction N-poly gate electrodes makes the threshold voltage of the DG FET very high (> 1 V).

Table 5-4 Process sequence for the experimental demonstration of planar DG FET.

Process Step	*Recipe/Process details*	*Notes*
1. Wafer scribe	10 ohm-cm (100) N-type Si.	
2. Trilayer stack deposition	Epi reactor, SiGe /Si/SiGe, each 20 nm. SiGe films have 25% Ge.	Blanket RPCVD with H_2 pre-bake of 1150 °C, 3 min.
3. LTO cap deposition	Tylan BPSG furnace, Recipe: LTO400. Target 150 nm undoped LTO.	LPCVD deposition of SiO_2 from SiH_4/O_2 at 400 °C.
4. Litho Mask 1 (Zero level)	Ultratech stepper, 1 μm SPR 3612 resist.	Optical lithography defines alignment targets for subsequent e-beam lithography.
5. Alignment target etch 1 (oxide etch)	AMT 8100 etcher. Standard oxide etch using CHF_3/O_2 and stop on Si.	
6. Alignment target etch 2 (Si etch)	Lam TCP 9400 etcher, Standard Si etch with $HBr/O_2/Cl_2$. Timed etch for 1 μm deep trenches.	Resist left on after step 5.

Process Step	*Recipe/Process details*	*Notes*
7. Litho Mask 2 (Fin patterning)	Hitachi H700-F e-beam writer, 0.4 μm Ma-N 2403 resist.	E-beam lithography to define fins in the trilayer stack. Device channel length is set here.
8. Fin etch	AMT 8100 etcher. Anisotropic timed etch using NF_3.	Similar etch rates of resist, oxide, and Si.
9. Source/Drain layer deposition	Epi reactor, blanket Si with in-situ B doping.	RPCVD with 900 °C, 10 min H_2 pre-bake. Deposition at 580 °C, 25 torr from SiH_4 and B_2H_6
10. Source/Drain activation	Tylan atmospheric furnace, Recipe: 650AN for 30 min.	Crystallization/dopant activation at 650 °C in N_2.
11. Litho Mask 3 (Source/Drain patterning)	Hitachi H700-F e-beam writer, 0.4 μm Ma-N 2403 resist.	E-beam lithography to define source/drain regions. 50 nm overlap with fin.
12. Source/Drain etch	Lam TCP 9400 etcher, Si etch with HBr/O_2. Stop using visual endpoint plot.	High etch selectivity to oxide. Leaves behind S/D spacer and pads.
13. Litho Mask 4 (Width cut)	Hitachi H700-F e-beam writer, 0.3 μm ZEP 520-22 resist.	E-beam lithography to cut fins perpendicular to channel length direction. Defines device width.
14. Fin width patterning	AMT 8100 etcher. Anisotropic timed etch using NF_3.	Cuts off fin and S/D spacers at the edges to expose trilayer cross section.
15. Isotropic SiGe etch	$HF:H_2O_2:CH_3COOH$ (1:2:3) for 8 min.	High selectivity to Si, target undercut 0.24 μm from each end.
16. Gate oxidation	Tylan atmospheric furnace, Recipe: DRY800 for 1 hr, target 6 nm oxide.	Dry oxidation.
17. Gate electrode deposition	Tystar doped poly furnace, Target 150 nm of in-situ Phosphorus doped amorphous-Si.	LPCVD at 580 °C and 400 mtorr using SiH_4 and PH_3. No clean between previous step and this one.

Process Step	*Recipe/Process details*	*Notes*
18. Gate electrode activation	Tylan atmospheric furnace, Recipe: 650AN for 30 min.	Inert anneal at 650 °C to activate Phosphorus in gate.
19. Litho Mask 5 (Gate patterning)	Hitachi H700-F e-beam writer, 0.4 μm Ma-N 2403 resist.	E-beam lithography to define gate poly pads.
20. Poly patterning	Lam TCP 9400 etcher, Si etch with HBr/O_2. Stop using visual endpoint plot.	High etch selectivity to oxide.
21. ILD0 deposition	Tylan BPSG furnace, Recipe: LTO400. Target 150 nm undoped LTO.	LPCVD deposition of SiO_2 Passivation oxide before metal.
22. Litho Mask 6 (Contact hole definition)	Hitachi H700-F e-beam writer, 0.3 μm ZEP 520-22 resist.	
23. Contact etch	AMT 8100 etcher. Standard oxide etch using CHF_3/O_2 and stop on Si.	Open contact holes. 50 % overetch to make sure oxide is cleared.
24. Metal deposition	Gryphon sputter system. 0.5 μm Al (with 1% Si) deposition.	In-situ Ar sputter before metal deposition.
25. Litho Mask 7 (Metal patterning)	Ultratech stepper, 1 μm SPR 3612 resist.	Metal lines and probe pads defined.
26. Metal etch	Applied P5000 Etcher. Standard Al anisotropic plasma etch using Cl chemistry. Stop on oxide using automatic endpoint.	
27. Forming gas anneal	Tylan FGA furnace. Recipe: FGA400 for 45 min.	Anneal at 400 °C in N_2/H_2 to reduce oxide interface trapped charge.

5.7 Transistor Results

The experimental DG FET process run resulted in some functional transistors. Fig. 5.22 shows the gate characteristics of one of the working devices. The I_d-V_g curve shows excellent turn-off characteristics – no measurable drain induced barrier lowering (DIBL),

and near-ideal subthreshold swing of 67 mV/decade. These characteristics are as expected from ultrathin body DG FETs of these dimensions due to their superior electrostatic robustness. The slight deviation from ideal subthreshold swing of 60 mV/decade can be explained by the presence of gate oxide interface states. These states can provide some of the charge that compensates the gate electrode charge. This can be thought of as a capacitance in series with the gate to body capacitance. Due to the capacitive division, the subthreshold swing increases from its ideal value in the absence of interface states. Comparison to Medici device simulation showed that a density D_{it} of about 5 x 10^{11} cm^{-2} explains the slight degradation of subthreshold slope to 67 mV/dec. This is a high interface trap density, but it is not very surprising given that there was no special surface treatment (sacrificial oxidation etc.) in between the SiGe isotropic etch and the gate oxidation step. The Si surface damage from the SiGe etching could explain the high density of interface states.

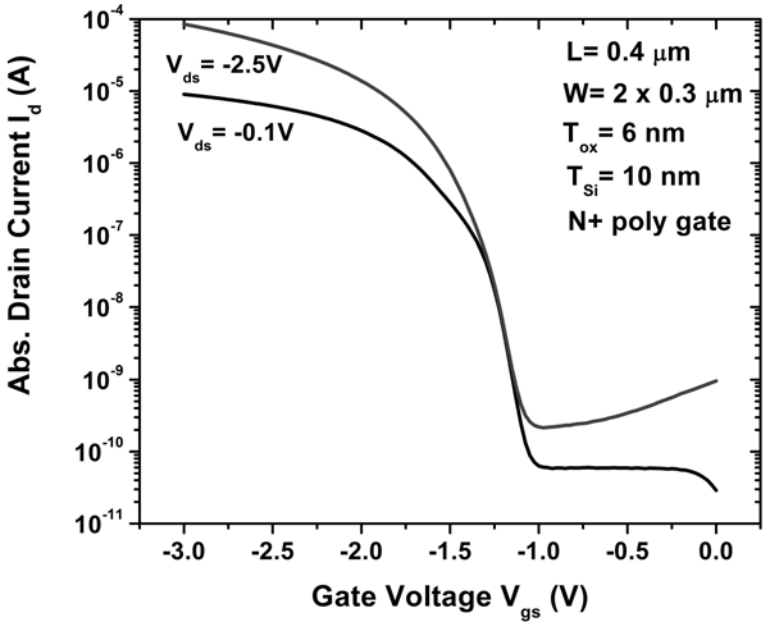


Fig. 5.22 Measured I_d-V_g characteristics of fabricated ultrathin body DG FET. Excellent turn-off characteristics are seen, such as no DIBL and near-ideal 67 mV/dec subthreshold swing.

The drain characteristics shown in Fig. 5.23 look well behaved, with no indication of excessive series resistance in the linear regime or kink-effect in the saturation regime.

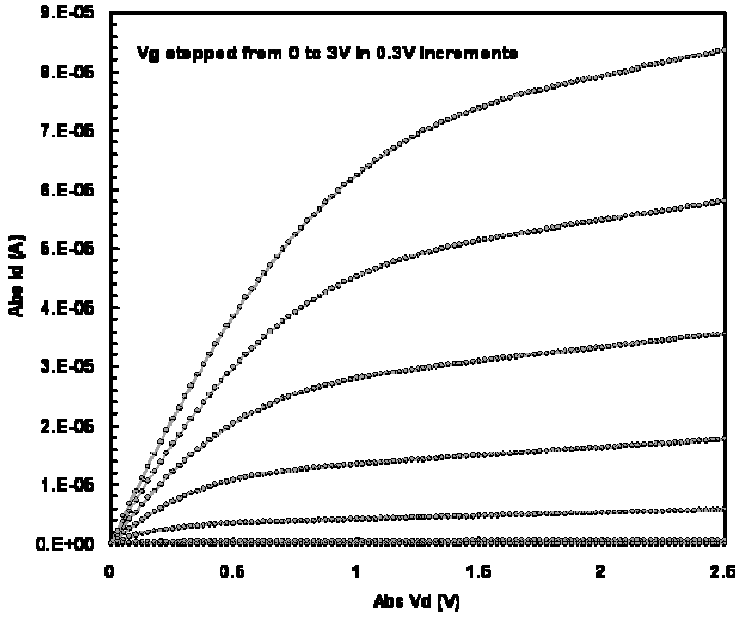


Fig. 5.23 Measured I_d-V_d characteristics of the same device as in Fig 5.22.

Substrate back-bias was applied by using the back surface of the wafer. Fig. 5.25 shows the effect of back-bias on the I_d-V_g characteristics of devices which did not have the DG FET. These characteristics are solely due to the parasitic transistor (Fig. 5.24) in the bulk. For the case with no applied back-bias (Fig. 5.25 (a)), the linear threshold voltage (V_T) is 1.075 V and there is a large amount of DIBL observed. This is most likely sub-surface DIBL due to the short channel length (0.4 μm) and the absence of any channel implant in the low doped bulk substrate. Upon the application of back-bias (Fig. 5.25 (b)), the sub-surface DIBL is reduced and what remains is the conventional surface DIBL. The V_T increases as expected due to the body effect. Further application of back-bias (Fig. 5.25 (c)) does not change the DIBL value too much. Once again, the V_T increases. In all cases, the best subthreshold swing is 89 mV/dec.

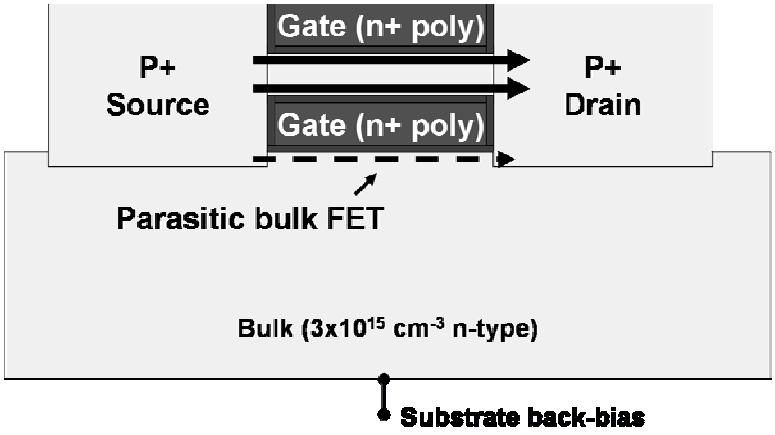


Fig. 5.24 Schematic of the parasitic bulk transistor created in parallel with the DG FET as a result of the simplified process flow.

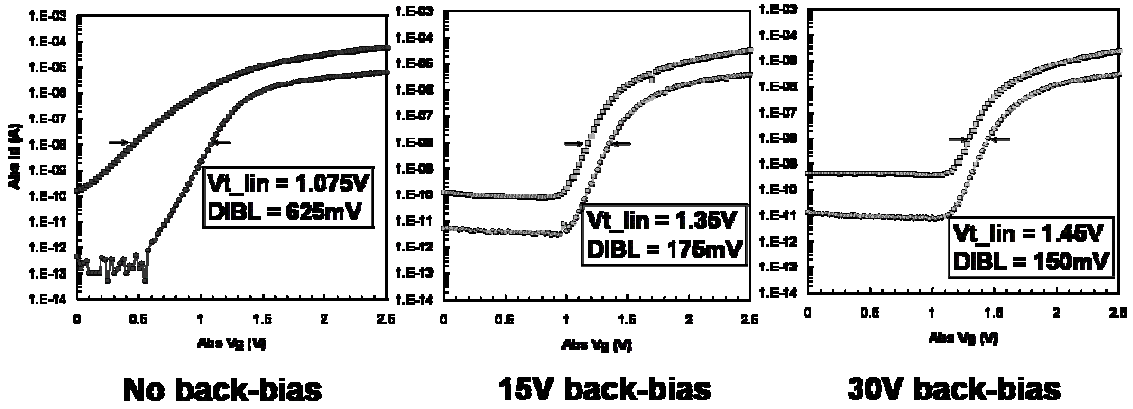


Fig. 5.25 Effect of the substrate back-bias on the I_d-V_g characteristics of the bulk-only devices. Finite DIBL remains in all cases and the subthreshold swing of 89 mV/dec is far from ideal. The linear threshold voltage increases monotonically with the back-bias.

The same analysis was done on the composite (DG FET in parallel with the parasitic bulk FET) devices. This is shown in Fig. 5.26.

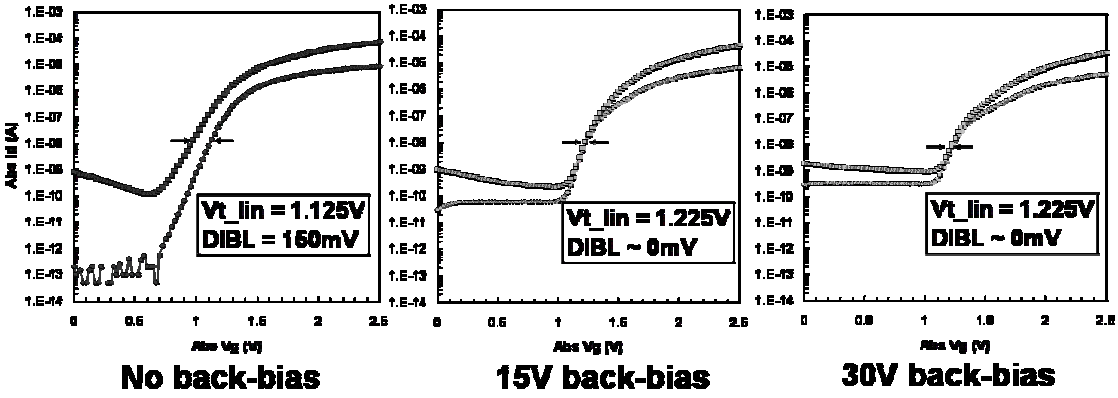


Fig. 5.26 Effect of the substrate back-bias on the I_d-V_g characteristics of composite (DG FET + parasitic bulk FET) devices. Without back-bias, the overall characteristics are dominated by the parasitic bulk FET. Application of back-bias increases the threshold voltage of the bulk FET and the excellent subthreshold characteristics are due to the DG FET. Further application of back-bias does not change the threshold voltage.

If no back-bias is applied (Fig. 5.26 (a)), the threshold voltage of the DG FET is comparable to that of the bulk FET. Coupled with the larger area of the latter device, it causes the I_d-V_g characteristics to look poor (high subthreshold swing, DIBL). When back-bias is applied (Fig. 5.26 (b)), the V_T of the bulk FET is pushed out, and the subthreshold characteristics are now solely due to the DG FET. Therefore, the DIBL drops dramatically and the subthreshold swing reduces. Further application of back-bias (Fig. 5.26 (c)) does not change the overall V_T since that is now dictated the DG FET which is not affected by back-bias. The minimum off-state current does increase with back-bias. That is because of the increasing source/drain to substrate junction leakage current with increasing back-bias.

Even though the back-bias technique increases the V_T of parasitic bulk FET so as to clean up the overall subthreshold characteristics, it does eventually turn on and affect the overall measured drive current. Due to the low substrate doping, the coupling from the back-contact isn't very efficient and large voltages are needed to increase the V_T. The back-bias could not be increased to more than 30 V since that caused source junction

breakdown. Since the area of the bulk parasitic FET is not known accurately, it is very difficult to separate the contributions to drive current of the DG FET and the bulk FET. Therefore, unfortunately no quantitative estimation of DG FET mobility could be made with confidence.

The complicated nature of the process, and the fact that it was being done for the first time, caused the yield to be very poor. Very few devices showed the DG FET characteristics described above. Of these, many were destroyed during measurement due to source/drain and/or gate oxide breakdown during the application of back-bias. The major factor contributing to the poor initial yield was most likely overetching of the thin Si membrane and its mechanical stability during the SiGe wet etching and cleaning/drying steps thereafter. This can be resolved by further optimization of the wet etch for higher selectivity and the use of critical point drying. Alternatively, a downstream plasma etch could be used to remove the SiGe isotropically as done in the Silicon-on-Nothing process [5.14]. The breakdown problem is much easier to solve – increasing the substrate doping will allow much higher V_T increase with a smaller applied back-bias. The substrate doping cannot be too high, or the source/drain junctions will leak heavily and break down due to band-to-band tunneling.

5.8 Summary

Process development was carried out to experimentally test the novel planar DG FET flow proposed in the previous chapter. The key enabling process steps were first individually studied in detail. Recipes were developed to grow high quality SiGe/Si/SiGe trilayer stacks by blanket RPCVD. By optimizing the H_2 pre-bake, these stacks were also deposited selectively in trenches patterned into isolation oxide layers. The enhanced oxidation rate of SiGe compared to Si was verified. By using steam oxidation at 750 °C, we observed about 3-4 times thicker oxides on 25% Ge-containing SiGe. Different approaches for isotropically etching SiGe with high selectivity to Si were studied and the

etch rates quantified. Recipes were developed to deposit in-situ doped Si, both n and p-type for the source/drain regions. In both cases, a high degree of dopant activation was measured with low thermal budget. Finally, process integration was carried out to fabricate transistors. Using a slightly simplified version of the proposed flow, functional planar ultrathin body DG FETs were obtained. The effect of a parasitic bulk FET was analyzed and the proof of concept DG FETs were shown to have excellent turn-off characteristics.

References

[5.1] C Date, "Vertical surrounding-gate MOSFETs incorporating silicon germanium heterojunctions," Ph. D. Thesis, Stanford University, 2000.

[5.2] S. M. Sze, "Physics of Semiconductor Devices," 2nd edition, John Wiley & Sons, 1981.

[5.3] J. W. Matthews and A. E. Blakeslee, "Defects in epitaxial multilayers," *Journal of Crystal Growth*, vol. 27, pp. 118-125, 1974.

[5.4] D. C. Houghton, "Strain relaxation kinetics in $Si_{1-x}Ge_x$/Si heterostructures," *Journal of Applied Physics*, vol. 70, pp. 2136-2151, 1991.

[5.5] D. J. Meyer, "Si-based alloys: SiGe and SiGeC," in Silicon Epitaxy, Semiconductors and Semimetals, vol. 72, pp. 345-395, Academic Press, 2001.

[5.6] N. Sato and T. Yonehara, "Hydrogen annealed silicon-on-insulator," *Applied Physics Letters*, vol. 65, pp. 1924-1926, 1994.

[5.7] J. P. Zhang et. al., "A comparison of the behaviour of $Si_{0.5}Ge_{0.5}$ alloy during dry and wet oxidation," *Thin Solid Films*, vol. 222, pp. 141-144, 1992.

[5.8] F. K. LeGoues, R. Rosenberg, and B. S. Meyerson, "Kinetics and mechanism of oxidation of SiGe: dry versus wet oxidation," *Applied Physics Letters*, vol. 54, pp. 644-646, 1989.

[5.9] D. K. Nayak, K. Kamjoo, J. S. Park, J. C. S. Woo, and K. L. Wang, "Wet oxidation of GeSi strained layers by rapid thermal processing," *Applied Physics Letters*, vol. 57, pp. 369-371, 1990.

[5.10] F. K. LeGoues, R. Rosenberg, T. Nguyen, F. Himpsel, and B. S. Meyerson, "Oxidation studies of SiGe," *Journal of Applied Physics*, vol. 65, pp. 1724-1728, 1989.

[5.11] T. K. Carns, M. O. Tanner, and K. L. Wang, "Chemical etching of $Si_{1-x}Ge_x$ in $HF:H_2O_2:CH_3COOH$," *Journal of the Electrochemical Society*, vol. 142, pp. 1260-1266, 1995.

[5.12] F. S. Johnson, D. S. Miles, D. T. Grider, and J. J. Wortman, "Selective chemical etching of polycrystalline SiGe alloys with respect to Si and SiO_2," *Journal of Electronic Materials*, vol, 21, pp. 805-810, 1992.

[5.13] A. H. Krist, D. J. Godbey, and N. P. Green, "Selective removal of a $Si_{0.7}Ge_{0.3}$ layer from Si (100)," *Applied Physics Letters*, vol. 58, pp. 1899-1901, 1991.

[5.14] S. Borel, C. Arvet, J. Bilde, S. Harrison, and D. Louis, "Isotropic etching of SiGe alloys with high selectivity to similar materials," *Microelectronic Engineering*, vol. 73-74, pp. 301-305, 2004.

[5.15] W. Kern and D. A. Poutinen, "Cleaning solutions based on hydrogen peroxide for use in semiconductor technology," *RCA Review*, vol. 31, pp. 187-206, 1970.

[5.16] D. B. Aubertine, "An x-ray diffraction study of concentration and strain dependent Si/SiGe interdiffusion", Ph. D. Thesis, Stanford University, 2003.

Chapter 5: Planar Double-Gate FET Process Development – Experimental Results

This page is intentionally left blank

Chapter 6

A Comparison Framework for Future Transistors

6.1 Introduction

In the preceding chapters, the double gate (DG) FET has been extensively discussed with respect to intrinsic device operation, extrinsic performance limiters, and a novel process to fabricate an ideal planar version of this transistor. The DG FET device design space can be explored by varying a number of material and structural parameters. Gate length (L_g), body thickness (T_{si}), effective gate oxide thickness (T_{ox}), channel material and gate dielectric material are some examples of such design variables. In the transistor design process, one is often confronted with the problem of how to compare the devices that result from such material and/or structural choices. In chapter 3, the I_{on}-I_{off} method was used to benchmark devices with different lateral doping profiles in the ultrathin extension regions. In this technique, the threshold voltage V_T is swept in order to generate a curve of the static off-state leakage current I_{off} as a function of the on-state saturated drive current I_{on}. At any given I_{off}, the curve that lies furthest to the right then represents the 'better' device. This method, which is widely used in the industry to

benchmark actual transistors that result from device design splits, has some shortcomings. First, the comparisons are made at a fixed supply voltage V_{dd}. There may be instances, such as comparisons of devices with channel materials differing in energy bandgap, such as Si and Ge, where it may not be fair to compare them at the same value of V_{dd}. Next, while the I_{on} value may be indicative of device speed, the actual inverter delay also depends upon the linear drive current [6.1]. This is not captured by the I_{on}-I_{off} method. Finally, this methodology is essentially only concerned about source-drain static power dissipation, through I_{off}. Other sources of static power such as gate leakage and dynamic power dissipation due to switching operations are neglected in using this technique.

This chapter takes a step back and looks at device optimization and comparison from a more holistic system performance point of view. Using total power and inverter delay as objective functions, a framework is developed to optimize a given transistor for minimizing the total power at a target inverter delay. Global comparisons of different transistor structures can then be made in a fair manner after each of them have been optimized using this methodology. In the next section, the motivation for this framework is explained using simple equations to describe power and delay. Next, there is a description of the methodology that was developed by a mixture of extensive device simulations and post-processing of the results. Finally, some of the key results obtained in the initial tests using this framework are discussed. The work in this chapter was done in collaboration with Dr. Pawan Kapur and Andy Chao, both from Stanford University.

6.2 CMOS Performance Metrics – Power and Delay

Neglecting manufacturing and design costs, from a pure device performance perspective, the two parameters that are most important in digital logic circuits are power and delay. These are now described in some more detail.

6.2.1 Transistor Delay

In a digital logic gate, the switching operation involves charge transfer between the output load capacitance and the supply/ground terminals through PMOS or NMOS transistors whose gate terminals are being driven by the input signal. The complementary nature of CMOS logic gates causes the load capacitance to typically swing through a voltage equal to V_{dd} in each switching operation. Therefore the charge transferred can be written as $C_{load}V_{dd}$. To a first order approximation, the transistor can be treated as a constant current source with a value equal to the drive current, I_{drive}. The intrinsic device delay, τ is therefore given by the expression,

$$\tau = \frac{C_{load}V_{dd}}{I_{drive}} \tag{6.1}$$

The current drive can be written as,

$$I_{drive} = WC_{ox}(V_{dd} - V_T)v_{inj} \tag{6.2}$$

where W is the device width, C_{ox} is the gate oxide capacitance per unit area, and v_{inj} is the effective injection velocity of carriers over the source barrier. In most digital logic circuits, the load capacitance at the output is typically the input capacitance of another transistor. In cases where the output drives an interconnect capacitance, buffer stages are inserted so that the output is still device-loaded. Neglecting parasitic capacitances, the output capacitance is equal to the gate oxide capacitance. For a transistor driving an identical device,

$$C_{load} = WL_gC_{ox} \tag{6.3}$$

The intrinsic device delay is therefore given by,

$$\tau = \frac{L_g}{v_{inj}(1 - \frac{V_T}{V_{dd}})} \tag{6.4}$$

This is an important equation since it explains some of the driving factors behind transistor scaling and the problems that arise as a result. First of all, scaling gate length directly

reduces the intrinsic device delay if all other parameters are kept constant. However, as L_g is reduced, V_{dd} is also reduced in order to keep the electric fields at manageable values from reliability considerations and also to reduce dynamic power dissipation. If V_T is also reduced in the same proportion as V_{dd}, the delay scales quite linearly with L_g. However, if the V_T becomes too low, the finite subthreshold swing causes the static power dissipation to grow exponentially. Therefore, in recent generations V_T has not been scaled down as fast than V_{dd}. This causes the denominator term in parentheses of eq (6.4) to reduce, increasing the delay. As a result, gate length scaling has been accelerated and it is now common to make L_g much smaller than the minimum pitch that is used to identify the particular technology node. For instance, at the '90 nm' node, the minimum gate length is around 50 nm [6.2]. This aggressive gate length scaling has put a greater emphasis on controlling short channel effects, driving the introduction and development of non-planar CMOS transistors such as ultrathin body fully depleted single-gate and multi-gate FETs. Another complementary approach for reducing the delay is to increase the v_{inj} term in the denominator. This is a driver for the introduction of high mobility channel materials such as strained-Si [6.2] and pure Ge [6.3].

It should be noted that the gate oxide capacitance C_{ox} gets cancelled out in eq. (6.4) and does not appear in the final delay approximation. The main reason for scaling the gate oxide thickness to increase C_{ox} has really been to improve gate control over the channel in order to suppress short channel effects.

6.2.2 Power Dissipation

There are two major components of power in digital CMOS circuits: dynamic and static power. Dynamic power P_{dyn} is dissipated each time a switching operation takes place. It is given by the expression,

$$P_{dyn} = \alpha C V_{dd}^2 f \tag{6.5}$$

where f is the clock frequency, C is the load capacitance being switched (mostly the device input capacitance C_{ox}), and α is the switching activity factor, which is a measure

of how frequently that load capacitance is switched. For latches and gates driven by the clock, it is high (typically 0.5), whereas for register files and cache transistors, it is low (typically < 0.01).

The static power P_{stat} is due to the source to drain subthreshold leakage current in the transistor off-state. It is equal to the product of the off-state leakage, $I_{SD,off}$ and the supply voltage.

$$P_{stat} = I_{SD,off} V_{dd} \tag{6.6}$$

The off-state leakage is approximately given by the equation for subthreshold MOS current,

$$I_{SD,off} = I_o \exp(\frac{-2.303 V_T^{'}}{S}) \tag{6.7}$$

where I_o is a constant, $V_T^{'}$ is the effective worst case threshold voltage that is lowered from the nominal V_T by the drain-induced barrier lowering, DIBL, and S is the sub-threshold swing. For a given delay, and hence specified L_{gate}, V_{dd}, and V_T, worsened short channel effects imply lower $V_T^{'}$ and higher S. These increases the static power exponentially.

The other sources of static power are gate leakage, drain junction leakage, and the band-to-band tunneling drain leakage. For well designed junctions, the latter two can be neglected. However, the gate leakage I_G can become very high in cases where the physical gate oxide thickness is small (<1.5 nm). With the use of high-k dielectric materials, the effective gate oxide thickness can be made very low for good short channel effect control without the accompanying penalty of increased gate leakage.

For a given transistor structure and a target delay, the choice of V_{dd}, V_T, and T_{ox} determines the total power dissipation. In the next section, a methodology is developed to evaluate and minimize the total power dissipation.

6.3 Methodology

For a given device, we first assume that all process and geometry parameters (such as L_{gate}, T_{ox}, silicon body thickness T_{si}, doping profile, etc.) are fixed. Assuming that the device has to switch at a certain target delay and with a certain switching activity factor, if we plot the static and dynamic power dissipation as a function of V_{dd}, the curves appear qualitatively as shown in Fig. 6.1. For the moment, the gate leakage power is neglected (i.e. the assumption of high-k dielectrics is made).

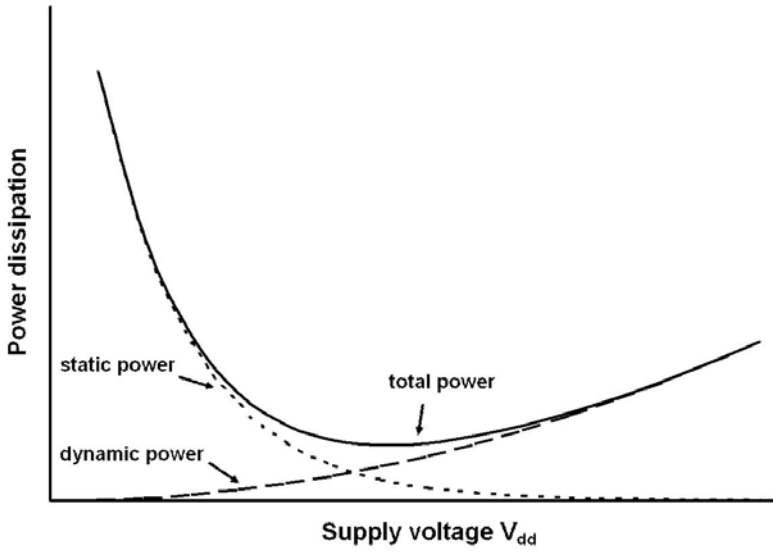


Fig. 6.1 Schematic illustration of the evolution of static and dynamic power components as a function of V_{dd} for a target delay. The dynamic power grows quadratically, while the static power reduces in an exponential fashion. Due to the opposing trends, the total power has a minimum at some V_{dd}.

Since the delay is fixed, from eq. (6.5), we see that the dynamic power increases as a quadratic function of V_{dd}. From eq. (6.4), it is apparent that the V_T has to increase linearly as V_{dd} is increased so as to keep the delay constant. From eq. (6.7), we therefore see that the static power reduces exponentially as V_{dd} is increased. Due to the opposite nature of these two trends, the total power has a minimum at some V_{dd} (and accompanying V_T).

The switching activity factor α essentially weights the dynamic power with respect to the static power in the summation. A low value of α puts a higher premium on static power.

If the gate oxide thickness T_{ox} is now lowered, still assuming no gate leakage, the dynamic power dissipation curve will increase while the static power curve will reduce. The latter is due to the implicit improvement in the short channel effect control which results in lower off-state current. For a device that already has good electrostatic gate control, this improvement will be marginal. The net result of changing T_{ox} is that a different optimum (minimum) power is reached and that happens at a different V_{dd}. If gate leakage is added to the static power, the curves change further due to the penalty posed by thin T_{ox} in terms of increased gate leakage power. At the end of this process, for a given structure, delay, and switching activity factor, we get a minimum total power value, and the corresponding values of V_{dd}, V_T, and T_{ox} that yield that minimum total power. By repeating this optimization for different values of target delay, we can generate a curve of the minimum power as a function of delay. Such a curve represents the best power-delay trade-off that can be achieved with a given device structure. Therefore, these curves can be used to compare various structures. At any given delay, the curve with lower power is the better one. Conversely, at any given power, the curve with lower delay is better.

6.4 Implementation

Although the aforementioned power-delay comparison/optimization technique is applicable to any futuristic transistor structure, we specifically use a nominally 18 nm gate length symmetric double-gate (DG) FET as a vehicle to exemplify the methodology and to implement a device simulation-based framework for the same. The DG FET structure, shown in Fig. 6.2, is the same as the one we have used in previous chapters for simulation-studies of DG FET electrostatics and extrinsic impedances. Such a device may be applicable for the 45 nm node high performance FET or 32 nm node low standby

power transistor. The body thickness T_{si} for this baseline structure is 7 nm. Thick 20 nm spacers separate the gates from the flared-out source/drain. As we have seen in chapter 3, such a geometry minimizes the impact of parasitic capacitance and resistance. The doping in the flared source/drain regions is assumed to be 10^{20} cm^{-3}. Unless otherwise mentioned, the lateral doping profile is assumed to roll off abruptly at the gate edge and contact resistance is neglected. The gates are assumed to be made of metal whose work-function Φ_m can be tuned arbitrarily within the Si bandgap in order to set the threshold voltage. Drift-diffusion transport models are used and quantum confinement effects, tunneling, and impact ionization are not turned on in the simulation models.

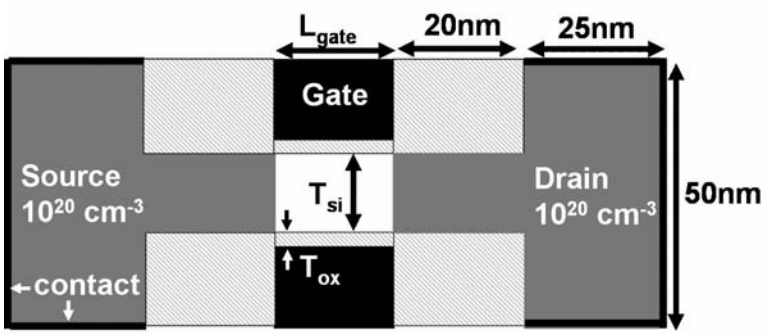


Fig. 6.2 Schematic of the Medici-simulated double-gate (DG) FET based on the ITRS 2003 45 nm HP/32 nm LSTP node. Baseline DG: L_{gate}=18 nm, T_{si}=7 nm. Back-gate (BG) FET: grounded bottom gate with midgap workfunction.

Medici simulations (dc and small signal ac analysis) are used to generate a family of curves plotting the drain current I_d and gate capacitance C_g as a function of gate voltage V_g and drain voltage V_d. Negative gate voltages are used to simulate the effect of gate workfunction. For each V_{dd}, the gate and drain voltages are swept from 0 to V_{dd} in steps of 0.05 V, while the gate workfunction is swept from 4.17 eV to 5.27 eV in steps of 0.05 V. The simulation outputs are stored in a look-up table for post-processing in MATLAB [6.4]. Fig. 6.3 shows a typical family of I_d-V_g curves for two values of T_{ox}. On the right-side axis, the discrepancy between the drift-diffusion and energy-balance

transport models is shown. The energy-balance modeling was also performed in Medici. The drift-diffusion models tend to underestimate the drive current by 30-40% compared to the more accurate energy balance models. However, we do not use the latter since they significantly increase the computational load. Moreover, there is no reliable calibration data for the 18 nm gate length DG FET energy balance models. The results from these simulations should not be taken to represent exact values of drive current (and hence delays). In spite of this inaccuracy, the trends in the power-delay trade-off and optimization are captured well.

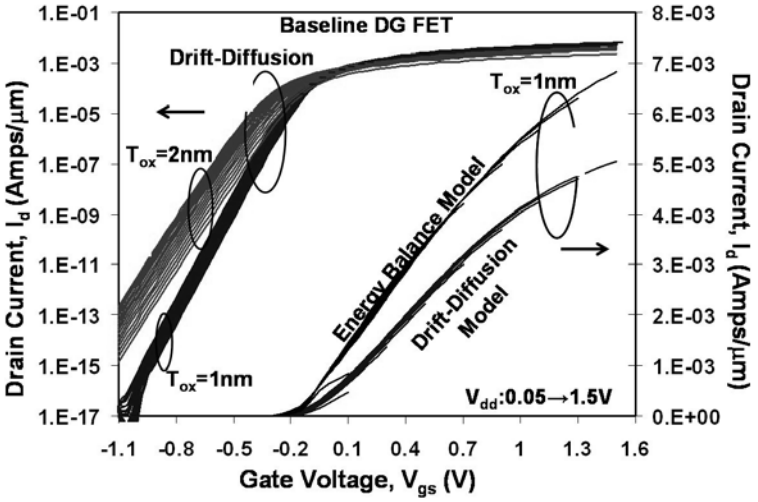


Fig. 6.3 Family of I_d-V_g curves with varying V_{dd} for 2 values of T_{ox}. Negative V_g is used to simulate the effective gate workfunction, which takes on all values within the Si bandgap. The right-side axis shows a 30-40% underestimate of the drive current using drift-diffusion as compared to the energy-balance transport model.

In order to calculate the delay from the contents of the I_d/C_g-$V_g/V_T/V_d$ look-up table, a two-step calculation is used to estimate the fanout of 1 (FO1) inverter delay. This delay calculation assumes a three-inverter configuration shown in Fig. 6.4. The input to the first inverter is assumed to be an abrupt step. The PMOS devices are assumed to be twice the width of the simulated NMOS devices and with symmetric values of V_T. The

temporal variations in currents and capacitances are calculated during the inverter switching transient. This is achieved by discretizing the time steps and using values from the simulated look-up table in conjunction with the Kirchoff's current law equations shown in Fig. 6.4. Such a numerical integration of the Medici I_d-V_g and C_g-V_g curves results in another column in the look-up table which has FO1 inverter delay for each point in V_g, V_T, and V_{dd} space.

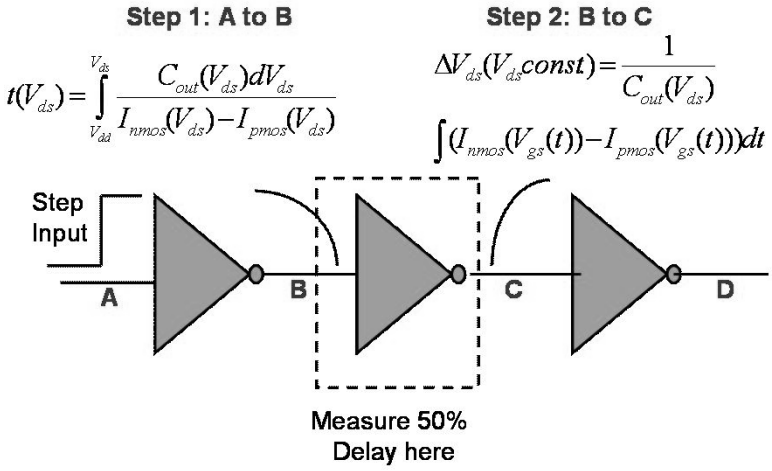


Fig. 6.4 Setup for FO1 inverter delay calculation by the numerical integration of Medici-simulated I_d-V_g and C_g-V_g curves.

Fig. 6.5 shows a comparison of the FO1 inverter delays calculated using the post-processing model with the more accurate (and computation-intensive) Medici mixed-mode transient simulations. When the feed-forward C_{gd} overshoot effect is accounted for, the calculated delays agree to within 30 % of the Medici-simulated values over a wide range of V_{dd}, V_t, and T_{ox} values. The 30 % underestimation is due to neglecting the Miller capacitance of the subsequent inverter stage in the post-processing calculations. Fortunately, this discrepancy is offset by the aforementioned 30-40 % discrepancy in the drive current due to the simpler drift-diffusion transport models. A key observation in Fig. 6.5 is that the use of the simple CV/I_{on} metric to calculate the FO1 inverter delay results in a

severe underestimation since it does not capture the dynamic nature of drive current and load capacitance during the switching transient.

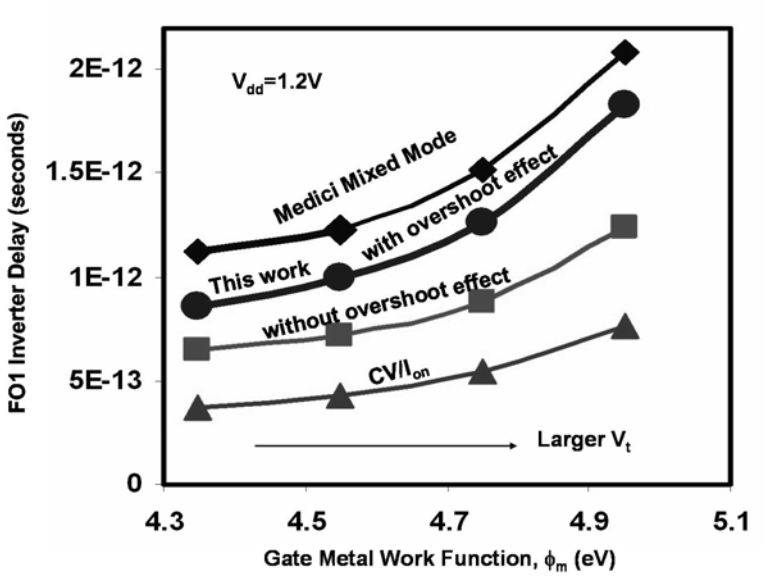


Fig. 6.5 Comparison of inverter FO1 delay calculation methods with mixed mode transient simulations in Medici. Use of the simple CV/I_{on} approximation to calculate FO1 delay causes a large underestimate.

With the delay model in place, the static power calculation due to source-drain leakage can be performed as shown in Fig. 6.6. This shows a plot of the source-drain leakage current as a function of the FO1 inverter delay for different values of V_{dd}. It should be noted that V_T adjustment is implicit to achieve the delay at a given V_{dd}. The curves shown in Fig. 6.6 are for a specific case of T_{ox} = 1 nm. With such plots, the source-drain leakage static power component as a function of V_{dd} (as shown in Fig. 6.1) can be calculated by choosing a point on the x-axis (FO1 delay) and picking off the leakage current values from curves with different V_{dd}. The leakage current is multiplied by V_{dd} to get the source-drain leakage static power. For a given V_{dd}, the leakage current decreases as the delay increases since the required V_T corresponding to that delay increases. At very low delay (i.e. low V_T), the leakage increases dramatically, more so as

the V_{dd} increases since the effect of DIBL further increases the off-state leakage. This explains the general shape of the curves and the crossover seen at low delay and high V_{dd}.

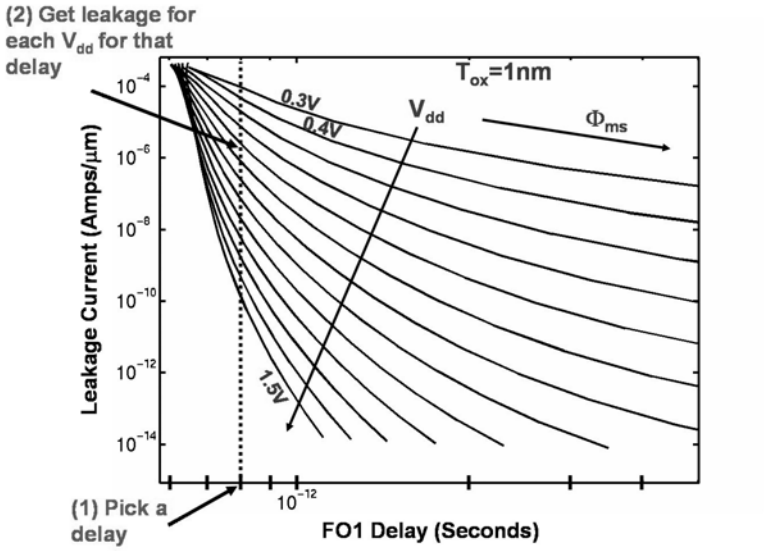


Fig. 6.6 Static power (due to source-drain leakage) calculation method. For every delay point, as V_{dd} is swept, implicit V_T adjustment sets the leakage current.

Next, the gate leakage is added in post-processing by combining a modified version of the analytical direct tunneling model of Lee et. al. [6.5] with an analytical DG FET charge model from Taur [6.6]. Fig. 6.7 shows excellent agreement between the post-processing model and computation-intensive direct tunneling simulations in Medici. The modeled gate leakage currents match simulations very well over 10 orders of magnitude under various V_{dd} and T_{ox} conditions. We assume that the only significant component of gate leakage is due to conduction band electron tunneling from the NMOS inversion layer into the gate electrode. The other components of tunneling (valence band hole tunneling, valence band electron tunneling, and electron tunneling from the metal gate into the substrate) are neglected since the larger barrier heights involved in those processes make them smaller in comparison. This approximation may not be valid in cases where there is a lot of gate to source/drain direct overlap or if the gate workfunctions

approach the band-edge. However in the baseline DG FET structure modeled here, the assumption is quite good.

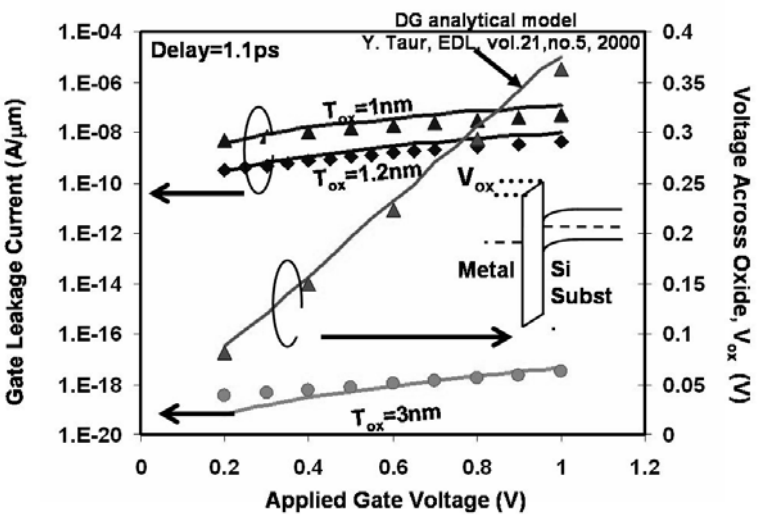


Fig. 6.7 Gate tunneling current and oxide voltage drop calculations by post-processing analytical models (solid lines) compared with Medici simulations (symbols).

Finally, in order to calculate the dynamic power, we assume some switching activity factor and a clock frequency that is given by a logic depth of 16 FO4 inverters. The weighted sum of the three power components is taken assuming that the '0' and '1' states of the inverter are equally likely.

6.5 Results – Total Power Optimization

Fig. 6.8 shows the NMOS total power-V_{dd} curves with the weighted contributions from each of the three individual components. Values of T_{ox} ranging from 1 nm to 2 nm in steps of 0.2 nm are used to generate a family of 6 curves. As expected (and depicted in Fig. 6.1), the total power does indeed show a global minimum (optimum) with respect to V_{dd} and T_{ox}. At this optimum point, the ratios of source-drain static leakage power and

gate leakage power to dynamic power are 20 % and 5 % respectively. The source-drain static leakage power has the strongest dependence on V_{dd}. The gate leakage power is highly dependent on the T_{ox} value due to exponentially increasing direct tunneling probability as T_{ox} is reduced.

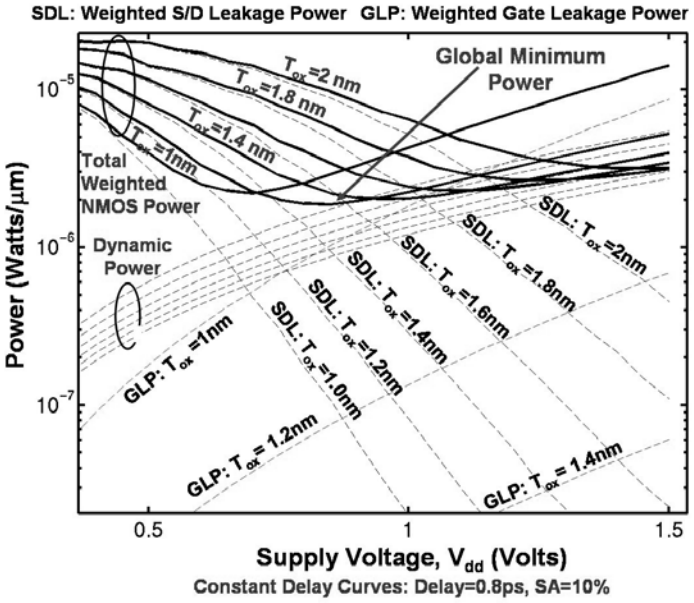


Fig. 6.8 Total power optimization curve for a specific delay and switching activity factor. The dashed curves show the weighted individual components of power – dynamic, source-drain leakage (SDL), and gate leakage power (GLP).

The optimized total power is dependent on the value of T_{ox} in the presence of gate leakage. Fig. 6.9 shows this dependence for the baseline device used with a moderate switching activity of 10 % (such as in the logic datapath). For the case of each delay, there is a optimum with respect to T_{ox}. At high values of T_{ox}, the source-drain leakage dominates due to degraded short channel effect control. At low values of T_{ox}, the gate leakage dominates. This can be clearly seen by comparing the SiO_2 curves with high-k dielectric curves for the same effective oxide thickness. In this analysis, the ideal high-k

curves have been obtained by simply turning off the gate leakage model in post-processing. No mobility degradation is assumed.

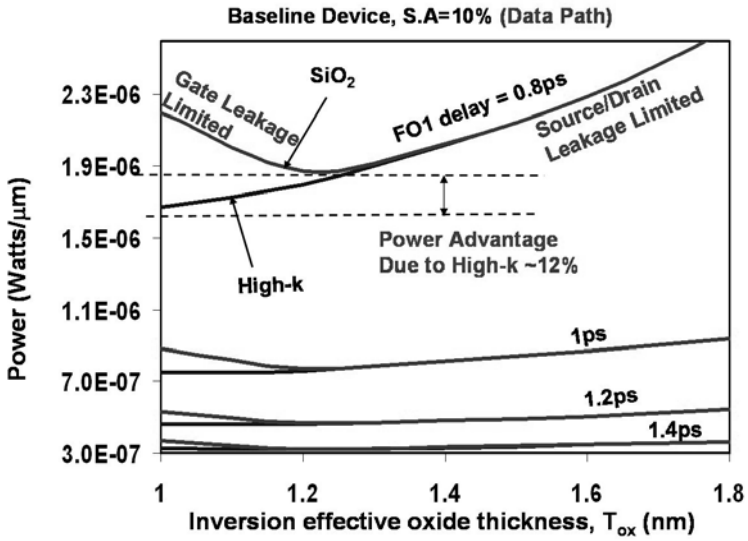


Fig. 6.9 Optimized total power as a function of the effective oxide thickness T_{ox} at inversion comparing the SiO_2 with perfect high-k dielectric for the baseline DG FET with moderate (10%) switching activity.

Interestingly, even such an optimistic high-k assumption gives only about 12 % reduction in the total power for the low delay datapath transistors. The high-k dielectric advantage diminishes at higher values of delay. This rather surprising result can be understood as being a result of the inherent electrostatic robustness of the ultrathin body DG FET. Therefore there is only an incremental advantage in scaling T_{ox} further. In addition, the absence (or lesser amount) of gate leakage in the wider PMOS transistors causes it to be weighted less as compared to the source-drain leakage and dynamic power in the total inverter power.

In the case of transistors used in lower switching activity applications (such as registers), the power advantage of high-k dielectrics over leaky SiO_2 increases. This is a result of the higher premium placed on static power dissipation compared to dynamic

power in such low switching activity devices. This is seen in Fig. 6.10 which is the same curve as shown in Fig. 6.9 except for the lower switching activity of 1 %

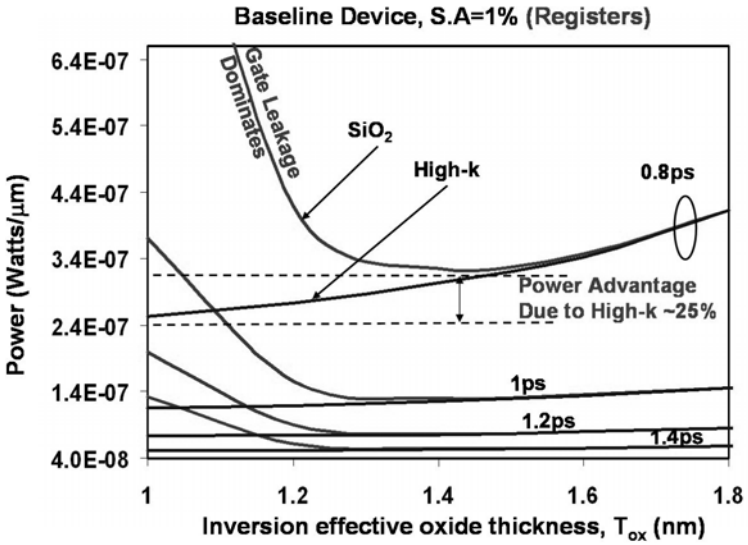


Fig. 6.10 Optimized total power as a function of the effective inversion oxide thickness comparing SiO_2 with and ideal high-k dielectric for a device with low (1 %) switching activity.

6.6 Results – Optimum Power-Delay Comparisons

Using the framework just described for optimizing total power with respect to V_{dd} and T_{ox} for a given structure, delay, and switching activity, we now use optimum power-delay curves to compare various device structures.

Fig. 6.11 shows the optimal power delay curves for 3 kinds of device structures – (a) the baseline DG FET with ideal (laterally abrupt) source/drain extensions aligned to the gate edge and zero contact resistance, (b) the baseline DG FET with realistic parasitic series resistance (3.5 nm/dec source/drain extensions which have an underlap of 7.5 nm from the gate edge – the underlap is chosen based on the optimization study in chapter 3), and (c) a back-gate (BG) FET which is essentially the DG FET operated as a single gate

FET with the second gate held at a fixed voltage. Such a structure is thought to be promising as a variable-threshold [6.7] for adaptive power control during chip operation. For each of these three devices, optimal power-delay curves are plotted assuming either leaky SiO_2 or a perfect high-k dielectric. In all cases, the switching activity is set to 10 % (datapath).

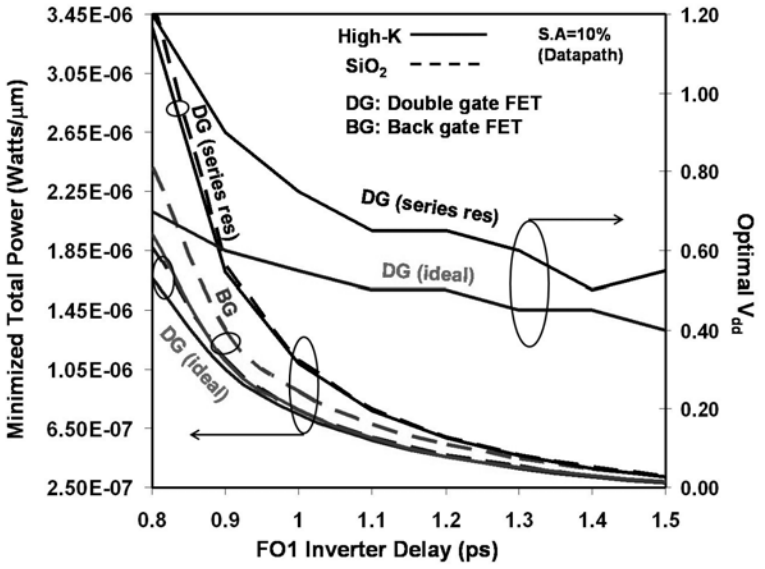


Fig. 6.11 Comparison of devices using optimal power-delay curves. The impact of material change (high-k vs. SiO_2), structure change (DG vs. BG), and parasitic resistance (ideal vs. realistic) is shown. The right-side axis shows the optimal V_{dd} for the DG FET with and without series resistance.

In all cases, there is a dramatic rise in total minimum power below a certain delay (> 100 % increase in power for 20 % reduction in delay below 1 ps.) Moreover, all three devices are rather comparable at large delays. At lower values of delay, the curves separate, enabling a clear comparison. The DG FET devices are better than the BG FETs assuming that both have negligible parasitic resistance. This is because of the degraded subthreshold swing in BG devices as a result of enhanced capacitive division of the gate voltage between the surface potential and the back gate. Also, in Fig. 6.11 we see that the

impact of reducing parasitic resistance is far greater than that of high-k dielectric insertion. Furthermore, from the right-side axis, it is evident that the presence of realistic series resistance increases the optimum V_{dd}. This is required in order to make up for the loss of gate overdrive due to the voltage drop in the source series resistance. Besides the lower parasitic resistance effects, in general, devices with better SCE control and/or mobility require lower optimum V_{dd}. This can be seen easily from eq. (6.4). In devices that have better electrostatic robustness, the V_T can be lowered, thus enabling lower V_{dd} for a given delay. Similarly, higher mobility (leading to higher injection velocity) can enable operation at a reduced overdrive (and hence reduced V_{dd}). The slight non-monotonicity of the V_{dd} curves is a result of the errors due to interpolation. Fig. 6.11 is an example of how the optimal power-delay comparison framework can be used to benchmark the effects of material and/or structure change, and parasitic elements in devices.

Fig. 6.12 shows the impact of switching activity on the power-delay curves. For any given delay, clock devices, which have the largest switching activity, dissipate the highest amount of power, and need to be operated at the lowest V_{dd}.

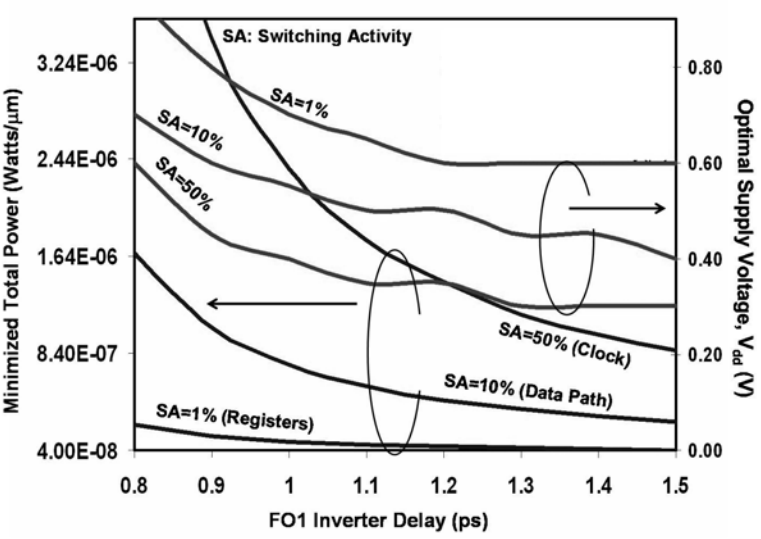


Fig. 6.12 Minimum total power-delay curves with corresponding optimal V_{dd} (shown on right-side axis) for devices with different switching activity.

On the other hand, devices used in register/cache with low switching activity dissipate the least power and need to be operated at high V_{dd} (in order to enable high V_T for static leakage power reduction).

In Fig. 6.13, we show the results of comparing 18 nm gate length DG FETs with different Si body thickness. In all cases, realistic parasitic series resistance is modeled (3.5 nm/dec lateral abruptness, 7.5 nm underlap, 5 x 10^{-8} ohm-cm^2 specific contact resistivity) and ideal high-k dielectric is assumed in order to neglect gate leakage. It is clear that reducing T_{si} results in better power-delay curves despite the increased series resistance in the thinner extensions. This means that at least in this T_{si} range, the improvement in short channel effect control due to thinner bodies is a stronger factor than the parasitic resistance degradation in dictating the optimal power-delay curves. From the points depicting optimal T_{ox} shown on the right-side axis, it is also seen that thinner bodies have a higher optimal $T_{ox.}$ The reason for this is that in the thin body devices, the electrostatic gate control is already so good that reducing T_{ox} degrades dynamic power (via increased C_{ox} at the load) more than the incremental improvement in short channel effects.

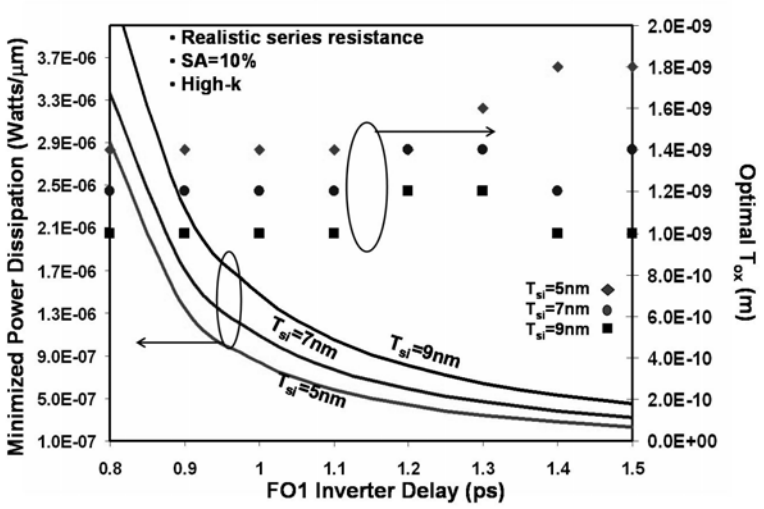


Fig. 6.13 Minimum total power-delay curves, along with the corresponding optimal T_{ox} for DG FETs with different Si body thickness (T_{si}).

The optimum values of T_{ox}, ranging between 1.4 to 1.8 nm are well within the reach of current technology, and even if conventional oxynitrides are used instead of high-k dielectrics, the impact of gate leakage on total power optimization is not significant at these values of thickness.

So far, all the comparisons were made assuming no variations in the nominal gate length. In reality, devices are over-designed so that they still meet power and performance specifications under the worst case process variations. Fig. 6.14 shows the impact of variations in L_{gate} (spatial) and V_{dd} (temporal). The optimum power in the presence of variations is calculated by using a gate workfunction Φ_m which ensures that even under the worst case conditions (sub-nominal V_{dd} and super-nominal L_{gate}), the delay specification is met in the critical path. Using this value of Φ_m, the static power is then calculated under worst case conditions (sub-nominal L_{gate} and super-nominal V_{dd}) and the entire optimization is repeated.

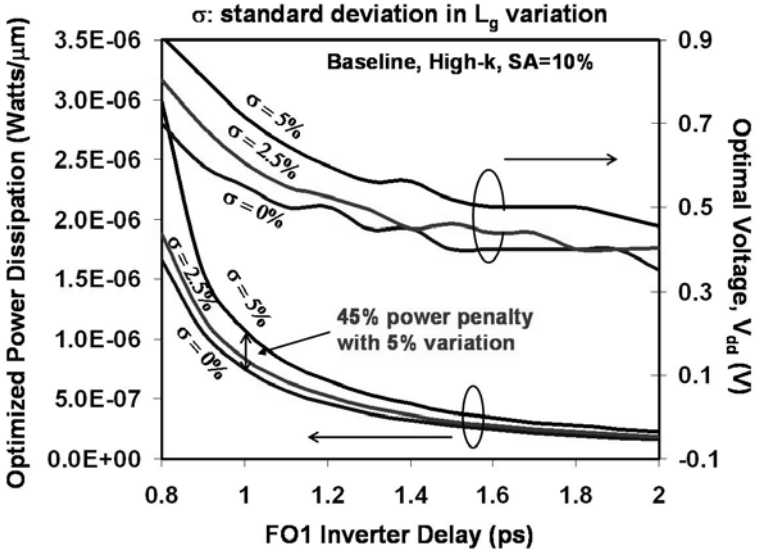


Fig. 6.14 Impact of process-induced gate length variations on the minimum power-delay curves and the corresponding optimum V_{dd}.

From the optimum power-delay curves in Fig. 6.14, it is seen that as the magnitude of variations increases, the minimum power increases. At a FO1 inverter delay of 1 ps, even a small 5 % variation in gate length requires over-design that yields a 45 % power penalty. In addition, the optimum V_{dd} required is higher when process variations are present.

Finally, at a target clock frequency and given body thickness T_{si}, Fig. 6.15 shows that there is actually an optimum gate length L_{gate} that minimizes the total power. The optimum arises since too large an L_{gate} gives good short channel effect control, but requires a low V_T. On the other hand, too small an L_{gate} can allow higher V_T operation, but degrades short channel effects. In Fig. 6.15, the optimal L_{gate} occurs at approximately 2-3 times the T_{si} if process-induced variations are neglected. The inclusion of process variations shifts the optimal L_{gate} to higher values with an accompanying power penalty.

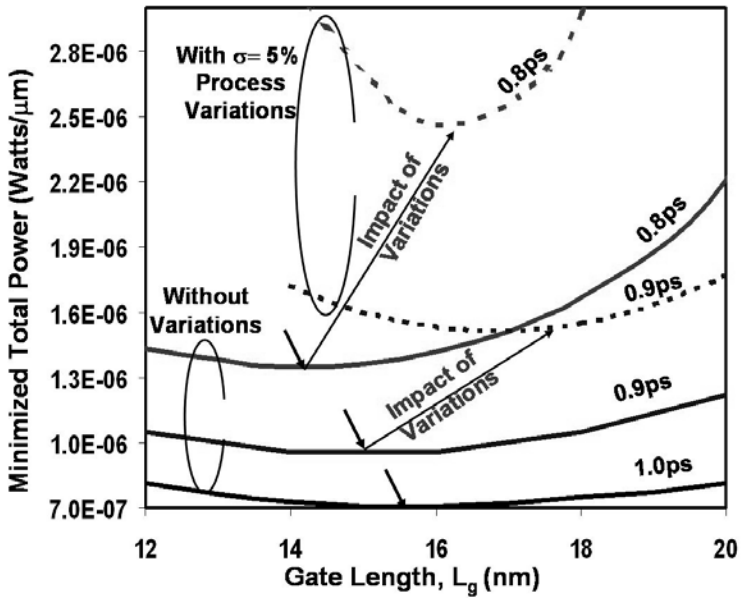


Fig. 6.15 Optimal gate length that minimizes total power for a given target delay and structure (baseline DG FET with T_{si} = 7 nm). For any delay, the incorporation of process-induced variations increases the optimal L_g and the optimized total power.

6.7 Summary

Simple I_{on}-I_{off} or CV/I_{on} comparisons of futuristic transistor structures are inadequate since they a) do not properly model delay, b) neglect dynamic power, and c) do not enable fair comparisons of structures which may require different supply voltages. In this chapter, we have developed a total power minimization framework and used it as a comparison standard to benchmark the impact of material and structural innovations on power and delay. Using an exemplary 18 nm gate length DG FET targeted for the 45 nm high performance logic node, the impact of high-k dielectric insertion and reduction of parasitic resistance in such a device is quantified. As an output of the power minimization, the optimum values of V_{dd}, V_T, and T_{ox} are obtained for a given device structure with a certain switching activity and target delay. For the devices in this study, we find that it is more important to reduce the parasitic resistance than to use high-k gate dielectrics. The latter may first find use in devices used in low switching activity circuits such as registers and cache memory. For higher switching activity devices, it is actually more beneficial to reverse scale the effective gate oxide thickness since the minimum total power occurs at higher values of T_{ox} in ultrathin body DG FETs. This obviates (or at least delays) the need for introducing high-k gate dielectrics in DG FETs. The inclusion of process-induced variations in this methodology invariably results in higher optimum V_{dd} and higher optimized total power. The specific conclusions mentioned above are used only as a vehicle to illustrate the versatility of the methodology. Given better current transport models in simulation tools, it should be possible to extend the scope to cover devices with novel channel materials such as Ge or C nanotubes in the comparison.

References

[6.1] M. H. Na, E. J. Nowak, W. Haensch, and J. Cai, "The effective drive current in CMOS inverters," in *IEDM Technical Digest*, 2002, pp. 121-124.

[6.2] T. Ghani et. al., "A 90nm high volume manufacturing logic technology featuring novel 45nm gate length strained silicon CMOS transistors," in *IEDM Technical Digest*, 2003, pp. 978-980.

[6.3] H. Shang et. al., "High mobility p-channel germanium MOSFETs with a thin Ge oxynitride gate dielectric," in *IEDM Technical Digest*, 2002, pp. 441-444.

[6.4] Mathworks Corporation, Natick, MA, MATLAB version 7.0.0, 2004.

[6.5] W.-C. Lee and C. Hu, "Modeling CMOS tunneling currents through ultrathin gate oxide due to conduction- and valence-band electron and hole tunneling," *IEEE Transactions on Electron Devices*, vol. 48, no. 7, pp. 1366-1373, Jul 2001.

[6.6] Y. Taur, "An analytical solution to a double-gate MOSFET with undoped body," *IEEE Electron Device Letters*, vol. 21, no. 5, pp. 245-247, May 2000.

[6.7] A. Khakifirooz and D. A. Antoniadis, "Effect of back-gate biasing on the performance and leakage control in deeply scaled SOI MOSFETs," in *2000 IEEE International SOI Conference Proceedings*, 2000, pp. 58-59.

This page is intentionally left blank

Chapter 7

Conclusions and Recommendations

7.1 Dissertation Summary

As the gate length of high performance MOS transistors is scaled down into the sub-50 nm regime, the suppression of short channel effects becomes very important. Left unchecked, these lead to a dramatic increase in static power dissipation during the off-state of the transistor. A number of fundamental and practical limits impede the traditional scaling of conventional bulk FETs. It becomes necessary to invoke innovations in materials and/or structures to prolong device scaling for technology nodes at the end of the roadmap [7.1]. The double-gate (DG) FET is a novel device structure that can potentially supplement or replace planar bulk FETs in future technology generations. The research performed towards this dissertation is aimed at studying some of the problems and solutions dealing with the technology and scaling of DG FETs.

From an intrinsic scalability viewpoint, it is very important to have an ultrathin (body thickness less than half of the gate length) undoped body in DG FETs. The ultrathin body, along with two gates, allows for better gate shielding of the drain electric field lines. As the body thickness (T_{si}) is reduced, keeping all other parameters unchanged, the electrostatic gate control is improved with the consequent amelioration of short channel

effects. The impact of T_{si} scaling can studied by analytical modeling using evanescent mode approximations as well as by 2-D device simulations. Since the magnitude of short channel effects is now exponentially related to T_{si}, it is important to minimize T_{si} variations across devices. One of the most important requirements for a DG FET is a uniformly ultrathin body.

As a result of the unconventional structure and ultrathin body, the DG FET has sources of extrinsic impedance that are different from those in planar bulk transistors. Unless minimized, these can limit the overall device performance even if the intrinsic device has very good characteristics. Parasitic capacitance in DG FETs arises mainly as a result of gates that are either misaligned to each other, differently sized, or with very thin sidewall spacers separating them from the source and drain regions. Mixed-mode (device and circuit) simulations have been used to study the impact of these non-optimal gate structures on the inverter delay and transistor off-state leakage current. The ideal DG FET needs to have gates that are perfectly self-aligned to each other and of the same size. In addition, they need to have low capacitance to the source/drain by using sidewall spacers that are several times thicker than the gate oxide. Parasitic resistance comes about due to contact resistance and series resistance in the ultrathin source/drain extension regions. Ideally, one would like to have zero specific contact resistivity and extremely abrupt lateral doping profiles. If practical considerations force the device designer to work with finite contact resistivity and lateral doping gradients, flared-out source/drain structures should be used to minimize the contact resistance, and the extension doping offset (underlap) from the gate edge should be optimized. The underlap optimization arises as the result of a trade-off between series resistance and short channel effect degradation. The optimal underlap depends upon the device electrostatic integrity and the maximum tolerable off-state leakage current. Metal source/drain DG FETs offer a path to circumventing the problems of dopant placement and profile control. However, in order for such Schottky-barrier DG FETs to be competitive with their optimized doped source/drain counterparts, the metal-semiconductor junctions need to have nearly zero barrier heights.

The DG FET can be implemented in a number of ways depending upon the orientation of the current carrying plane and the direction of current flow. Of these variants, the planar DG FET is interesting since the critical body thickness dimension is vertically directed and therefore can be defined, in principle, sub-lithographically by a film deposition step. The key challenge in building planar DG FETs is the placement and self-alignment of the bottom gate. Previous attempts at building planar DG FETs have suffered from problems of excessive gate capacitance, gate mis-sizing, or process complexity. In this work, we have proposed a novel process to build a planar DG FET that combines some of the main ideas from the prior art, while overcoming their shortcomings. The main idea inherent to this process is the use of disposable layers that act as placeholders for the top and bottom gates-to-be. SiGe is a suitable sacrificial material. Its lattice is close to that of Si making it possible to heteroepitaxially grow a single crystal Si body above it. At the same time, SiGe is chemically different enough from Si so as to allow its removal by selective etching with respect to Si. In addition, the enhanced oxidation rate of SiGe compared to Si can be utilized to form thick gate sidewall spacers for low parasitic capacitance. The in-situ doped source/drain regions which are deposited by epitaxy are ideally suited for the optimization of contact and extension resistance. The proposed structure therefore has all the desirable characteristics of an ideal planar DG FET: 1) deposition- controlled ultrathin and uniform body, 2) fully self-aligned top and bottom gates with same size and thick sidewall spacers for low parasitic capacitance, 3) flared-out source/drain for low parasitic resistance, and 4) the capability to integrate high-k dielectric/metal gates as a result of the gate-last nature of the process. In addition, due to the relative decoupling of the channel and source/drain formation steps, minor variations on the baseline process could be used to fabricate interesting device structures such as heterostructure-channel DG FETs and metal source/drain DG FETs. Such a process therefore also serves as a test-vehicle for novel device ideas.

The experimental portion of this work mainly deals with the process development of the unit steps and their integration in order to demonstrate double-gate transistors

using the proposed flow. High quality single-crystalline trilayers of SiGe/Si/SiGe have been grown by reduced pressure epitaxial CVD on both blanket as well as patterned substrates. Cross-sectional TEM has been extensively used for recipe development and to ascertain the absence of defects in such metastably strained films. Oxidation in steam has been carried out to confirm the 3-4X oxide growth rate enhancement on SiGe films with 25% Ge atomic fraction. A number of recipes, both wet and dry, have been used to isotropically etch SiGe selectively with respect to Si. High (> 50:1) selectivity has been obtained using a wet etch that removes SiGe by an oxidation/etch process. In-situ doped source/drain regions have been formed by chemical vapor deposition followed by low temperature annealing (650°C) to minimize the thermal budget while yielding high active doping concentrations. A slightly simplified version of the process has been successfully integrated to obtain functional double-gate transistors. As expected of DG FETs with those dimensions, electrical measurements show very good subthreshold turn-off characteristics, with nearly ideal (67 mV/dec.) subthreshold swing and almost zero DIBL. The behavior of the current-voltage curves as a function of the applied substrate back-bias has been used to distinguish the working DG FETs from parasitic planar bulk transistors.

A methodology has been developed for the optimization of the supply voltage, threshold voltage, and effective gate oxide thickness in order to minimize the total power dissipation in digital logic transistors with a given switching activity ratio and target delay. A framework incorporating device simulation and post-processing analysis has been created for fair comparisons of future transistor structures using such minimum total power-delay curves for benchmarking. While such a comparison technique is quite general in its scope, it has been applied to a 45 nm node high performance DG FET in order to exemplify its utility. One of the interesting results of such a study has been the relative lack of need for high-k gate dielectrics in such DG FETs where the short channel effects are already well controlled by the ultrathin body.

7.2 Contributions

- A systematic study of extrinsic resistance in ultrathin body DG FETs showing the importance of extension series resistance due to laterally non-abrupt doping profile. We have shown that it is necessary to tune the extension doping so as to have an off-set or underlap from the gate edge. This underlap is chosen by maximizing the drive current for a fixed off-state leakage current. The optimum arises as a balance between drive current degradation due to short channel effects and series resistance. The optimal underlap depends upon the inherent robustness of the device to short channel effects, and on the maximum tolerable off-state leakage current [7.2, 7.3].
- Proposal of a novel process flow to fabricate an ideal planar DG FET with a) deposition-controlled ultrathin and uniform body, b) fully self-aligned and identically-sized top and bottom gate with low parasitic capacitance sidewall spacers, c) flared-out source/drain structure for low extrinsic resistance, and d) gate-last feature allowing the integration of high-k/metal gate stack.
- Unit process development to verify the feasibility of the key steps in the proposed flow: a) high quality epitaxial CVD of trilayer SiGe/Si/SiGe stack, b) oxidation rate enhancement on SiGe compared to Si, c) isotropic etching of SiGe selective to Si, and d) in-situ doped source/drain formation by CVD.
- Process integration to build planar DG FETs. Functional double-gate transistors have been demonstrated with excellent subthreshold characteristics (nearly ideal subthreshold swing and very low DIBL) as proof of concept [7.4].
- Development (in collaboration with co-workers at Stanford University) of a new methodology to optimize supply voltage, threshold voltage, and effective oxide thickness for minimizing the total power dissipation in a digital logic transistor structure with a given switching activity and target delay. This methodology forms the basis of a framework using device simulation and post-processing for fair comparison of future generation transistor structures [7.5].

7.3 Recommendations for Future Work

The following is a list of ways in which the present work can be extended in future research.

- The extrinsic impedance simulation study carried out in this dissertation focused only on Si DG FETs. It may be interesting to extend this study to Ge DG FETs. One would expect that, due to the enhanced channel carrier transport, extrinsic resistance would be more of a performance bottleneck in Ge transistors. In addition, experiments have not been able to demonstrate very high active n-type doping in Ge. Given this fact, the examination of the applicability of metal source/drain DG FETs (even with finite barrier heights) in Ge channel systems is needed.
- The underlap optimization relies on the prior knowledge of the lateral doping profile. Experimental studies of dopant diffusion in ultrathin films and characterization of lateral dopant profiles in such layers will be a very useful contribution. One might expect that, similar to lateral solid phase epitaxial regrowth [7.6], the kinetics of dopant diffusion in ultrathin films will be markedly different from that in bulk systems.
- It will be useful to optimize the baseline planar DG FET process so as to completely eliminate the parasitic bulk FET (without the need to apply substrate back-bias). This can be done by starting on SOI substrates and using a SiGe etch that is selective to oxide as well as Si (such as $NH_4OH:H_2O_2:H_2O$ or a CF_4-based chemical dry etching scheme). By improving the basic DG FET process and device yield, a study of mobility in ultrathin body DG FETs will reveal valuable experimental information about quantum confinement and carrier transport in ultrathin body systems.
- Some of the variations on the baseline planar DG FET process may be useful projects to work on. In particular, the heterostructure channel DG FET should be relatively easy to build. Also, by carrying out the dual of this process to build SiGe channel DG FETs (using sacrificial Si layers and a Si wet etch such as TMAH that is selective to

SiGe), it will be possible to study carrier transport in DG FETs with ultrathin SiGe channels.

- The SiGe/Si/SiGe trilayer epitaxy combined with selective SiGe etching can be used for applications other than building DG FETs. One such structure is a $SiO_2/Si/SiO_2$ quantum well. There have been some simulation studies [7.7] of quantum wells formed in Si/SiO_2 systems, with an ultrathin Si layer sandwiched in between thin SiO_2 barriers. However, due to the difficulty in building perfect ultrathin (< 5 nm) single-crystal Si layers on direct-tunneling (< 3 nm) oxides, such quantum wells have not yet been experimentally realized to the best of our knowledge.
- The minimum power-delay comparison framework can be utilized to examine the impact of novel transistor channel materials such as Ge, strained-Si, diamond, etc. The inclusion of band-to-band tunneling models, either in the Medici simulation, or in post-processing, will be necessary to properly model the deleterious effects that accompany low energy band-gap materials. Also, the effects of process-induced variations and the applicability of multiple supply-threshold voltage schemes need to be examined more carefully.

References

[7.1] International Technology Roadmap for Semiconductors, 2003 edition, SIA, 2003.

[7.2] R. S. Shenoy and K. C. Saraswat, "Optimization of extrinsic source/drain resistance in ultrathin body double-gate FETs," in *Proceedings of the 2003 IEEE Silicon Nanoelectronics Workshop*, 2003, pp. 8-9, 2003.

[7.3] R. S. Shenoy and K. C. Saraswat, "Optimization of extrinsic source/drain resistance in ultrathin body double-gate FETs," *IEEE Transactions on Nanotechnology*, vol. 2, no. 4, pp. 265-270, Dec 2003.

[7.4] R. S. Shenoy and K. C. Saraswat, "Novel process for fully self-aligned planar ultrathin body double-gate FET," in *2004 IEEE International SOI Conference Proceedings*, 2004, pp. 190-191.

[7.5] P. Kapur, R. S. Shenoy, A. K. Chao, Y. Nishi, and K. C. Saraswat, "Power optimization of future transistors and a resulting global comparison standard," to be presented at IEDM 2004.

[7.6] P. Kalavade, "Novel device structures for CMOS scaling," Ph. D. Thesis, Stanford University, 2002.

[7.7] C.-H. Choi, Z. Yu, and R. W. Dutton, "Resonant gate tunneling current in double-gate SOI: A simulation study," *IEEE Transactions on Electron Devices*, vol. 50, no. 12, pp. 2579-2581, 2003.