

INTEGRATED CIRCUITS AND SYSTEMS

Colinge

Edited by
J.-P. Colinge

FinFETs and Other Multi-Gate Transistors



FinFETs and Other Mu

CBN BRNCH
GenColl

TK 7872
.T73
F56
2008
Copy 1

 Springer

FinFETs and Other Multi-Gate Transistors

Series on Integrated Circuits and Systems

Series Editor: Anantha Chandrakasan
Massachusetts Institute of Technology
Cambridge, Massachusetts

FinFETs and Other Multi-Gate Transistors

Jean-Pierre Colinge (Ed.)

ISBN 978-0-387-71751-7

Design for Manufacturability and Statistical Design: A Constructive Approach

Michael Orshansky, Sani R. Nassif, and Duane Boning

ISBN 978-0-387-30928-6

Low Power Methodology Manual: For System-on-Chip Design

Michael Keating, David Flynn, Rob Aitken, Alan Gibbons, and Kaijian Shi

ISBN 978-0-387-71818-7

Modern Circuit Placement: Best Practices and Results

Gi-Joon Nam and Jason Cong

ISBN 978-0-387-36837-5

CMOS Biotechnology

Hakho Lee, Donhee Ham and Robert M. Westervelt

ISBN 978-0-387-36836-8

SAT-Based Scalable Formal Verification Solutions

Malay Ganai and Aarti Gupta

ISBN 978-0-387-69166-4, 2007

Ultra-Low Voltage Nano-Scale Memories

Kiyoo Itoh, Masashi Horiguchi and Hitoshi Tanaka

ISBN 978-0-387-33398-4, 2007

Routing Congestion in VLSI Circuits: Estimation and Optimization

Prashant Saxena, Rupesh S. Shelar, Sachin Sapatnekar

ISBN 978-0-387-30037-5, 2007

Ultra-Low Power Wireless Technologies for Sensor Networks

Brian Otis and Jan Rabaey

ISBN 978-0-387-30930-9, 2007

Sub-Threshold Design for Ultra Low-Power Systems

Alice Wang, Benton H. Calhoun and Anantha Chandrakasan

ISBN 978-0-387-33515-5, 2006

High Performance Energy Efficient Microprocessor Design

Vojin Oklibdzija and Ram Krishnamurthy (Eds.)

ISBN 978-0-387-28594-8, 2006

Abstraction Refinement for Large Scale Model Checking

Chao Wang, Gary D. Hachtel, and Fabio Somenzi

ISBN 978-0-387-28594-2, 2006

A Practical Introduction to PSL

Cindy Eisner and Dana Fisman

ISBN 978-0-387-35313-5, 2006

Thermal and Power Management of Integrated Systems

Arman Vassighi and Manoj Sachdev

ISBN 978-0-387-25762-4, 2006

Continued after index

Jean-Pierre Colinge
Editor

FinFETs and Other Multi-Gate Transistors

 Springer

ed after index

Jean-Pierre Colinge
Tyndall National Institute
Cork, Ireland

Series Editor:
Anantha Chandrakasan
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139
USA

Library of Congress Control Number: 2007935904

ISBN 978-0-387-71751-7

e-ISBN 978-0-387-71752-4

Printed on acid-free paper.

© 2008 Springer Science+Business Media, LLC

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden. The use in this publication of trade names, trademarks, service marks and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

9 8 7 6 5 4 3 2 1

springer.com

Table of Content

Preface	v
Table of Content	ix
Contributors	xv
1 The SOI MOSFET: from Single Gate to Multigate.....	1
1.1 MOSFET scaling and Moore's law	1
1.2 Short-Channel Effects.....	2
1.3 Gate Geometry and Electrostatic Integrity	4
1.4 A Brief History of Multiple-Gate MOSFETs	8
1.4.1 Single-gate SOI MOSFETs	9
1.4.2 Double-gate SOI MOSFETs.....	11
1.4.3 Triple-gate SOI MOSFETs.....	12
1.4.4 Surrounding-gate (quadruple-gate) SOI MOSFETs.....	13
1.4.5 Other multigate MOSFET structures.....	14
1.4.6 Multigate MOSFET memory devices.....	16
1.5 Multigate MOSFET Physics.....	17
1.5.1 Classical physics.....	17
1.5.1.1 Natural length and short-channel effects	17
1.5.1.2 Current drive.....	23
1.5.1.3 Corner effect	26
1.5.2 Quantum effects.....	28
1.5.2.1 Volume inversion.....	28
1.5.2.2 Mobility effects.....	31
1.5.2.3 Threshold voltage	31
1.5.2.4 Inter-subband scattering.....	35
References	37

2 Multigate MOSFET Technology	49
2.1 Introduction	49
2.2 Active Area: Fins	52
2.2.1 Fin Width.....	52
2.2.2 Fin Height and Fin Pitch.....	57
2.2.3 Fin Surface Crystal Orientation.....	61
2.2.4 Fin Surface Preparation	64
2.2.5 Fins on Bulk Silicon	65
2.2.6 Nano-wires and Self-Assembled Wires	66
2.3 Gate Stack.....	70
2.3.1 Gate Patterning	70
2.3.2 Threshold Voltage and Gate Workfunction Requirements	71
2.3.2.1 Polysilicon Gate.....	74
2.3.2.2 Metal Gate	75
2.3.2.3 Tunable Workfunction Metal Gate	75
2.3.3 Gate EWF and Gate Induced Drain Leakage (GIDL)	79
2.3.4 Independently Controlled Gates	82
2.4 Source/Drain Resistance and Capacitance.....	85
2.4.1 Doping the Thin Fins.....	85
2.4.2 Junction Depth.....	87
2.4.3 Parasitic Resistance/Capacitance and Raised Source and Drain Structure	87
2.5 Mobility and Strain Engineering	91
2.5.1 Introduction	91
2.5.2 Wafer Bending Experiment.....	92
2.5.3 Nitride Stress Liners.....	93
2.5.4 Embedded SiGe and SiC Source and Drain	94
2.5.5 Local Strain from Gate Electrode.....	95
2.5.6 Substrate Strain: Strained Silicon on Insulator.....	97
2.6 Contacts to the Fins	98
2.6.1 Dumbbell source and drain contact	99
2.6.2 Saddle contact.....	99
2.6.3 Contact to merged fins.....	100
Acknowledgments	100
References	101

.....	49
.....	49
.....	52
.....	52
.....	57
.....	61
.....	64
.....	65
.....	66
.....	70
.....	70
nents	71
.....	74
.....	75
.....	75
.....	79
.....	82
.....	85
.....	85
.....	87
.....	87
.....	91
.....	91
.....	92
.....	93
.....	94
.....	95
.....	97
.....	98
.....	99
.....	99
.....	100
.....	100
.....	101

3 BSIM-CMG: A Compact Model for Multi-Gate Transistors	113
3.1 Introduction	113
3.2 Framework for Multigate FET Modeling	114
3.3 Multigate Models: BSIM-CMG and BSIM-IMG	115
3.3.1 The BSIM-CMG Model	116
3.3.2 The BSIM-IMG Model	116
3.4 BSIM-CMG	117
3.4.1 Core Model	117
3.4.1.1 Surface Potential Model	117
3.4.1.2 I-V Model	123
3.4.1.3 C-V Model	126
3.4.2 Modeling Physical Effects of Real Devices	130
3.4.2.1 Quantum Mechanical Effects (QME)	131
3.4.2.2 Short-channel Effects (SCE)	137
3.4.3 Experimental Verification	144
3.5 The BSIM-IMG Model	147
3.5.1 Surface Potential of independent DG-FET	147
3.5.2 BSIM-IMG features	150
3.6 Summary	151
References	151
 4 Physics of the Multigate MOS System	 155
4.1 Device electrostatics	155
4.2 Double gate MOS system	163
4.2.1 Modeling assumptions	163
4.2.2 Gate voltage effect	167
4.2.3 Semiconductor thickness effect	169
4.2.4 Asymmetry effects	174
4.2.5 Oxide thickness effect	178
4.2.6 Electron tunnel current	180
4.3 Two-dimensional confinement	184
References	185

5 Mobility in Multigate MOSFETs	191
5.1 Introduction	191
5.2 Double-Gate MOSFETs and FinFETs.....	192
5.2.1 Phonon-limited mobility.....	197
5.2.2 Confinement of acoustic phonons	202
5.2.3 Interface roughness scattering.	205
5.2.4 Coulomb scattering.....	209
5.2.5 Temperature Dependence of Mobility.....	212
5.2.6 Symmetrical and Asymmetrical Operation of DGSOI FETs	214
5.2.7 Crystallographic orientation	218
5.2.8 High-k dielectrics.	224
5.2.9 Strained DGSOI devices.....	226
5.2.10. Summary.....	232
5.3 Silicon multiple-gate nanowires	233
5.3.1 Introduction	233
5.3.2 Electrostatic description of Si nanowires.	235
5.3.3 Electron transport in Si nanowires.....	239
5.3.4 Surface roughness.....	245
5.3.5 Experimental results and conclusions.....	246
References	247
 6 Radiation Effects in Advanced Single- and Multi-Gate SOI	
MOSFETs.....	257
6.1. A brief history of radiation effects in SOI	257
6.2. Total Ionizing Dose Effects	259
6.2.1 A brief overview of Total Ionizing Dose effects.....	259
6.2.2 Advanced Single-Gate FDSOI devices	260
6.2.2.1 Description of Advanced FDSOI Devices.....	261
6.2.2.2 Front-gate threshold voltage shift.....	261
6.2.2.3 Single-transistor latch	264
6.2.3 Advanced Multi-Gate devices	266
6.2.3.1 Devices and process description.....	267
6.2.3.2 Front-gate threshold voltage shift.....	269
6.3 Single-Event Effects	271
6.3.1 Background.....	271
6.3.2 Effect of ion track diameter in nanoscale devices	274
6.3.3 Transient measurements on single-gate and FinFET SOI	
transistors.....	279
6.3.4 Scaling effects	285
References	287

.....	191
.....	191
.....	192
.....	197
.....	202
.....	205
.....	209
.....	212
I FETs	214
.....	218
.....	224
.....	226
.....	232
.....	233
.....	233
.....	235
.....	239
.....	245
.....	246
.....	247
II	
.....	257
.....	257
.....	259
.....	259
.....	260
.....	261
.....	261
.....	264
.....	266
.....	267
.....	269
.....	271
.....	271
.....	274
SOI	
.....	279
.....	285
.....	287

7 Multi-Gate MOSFET Circuit Design.....	293
7.1 Introduction	293
7.2 Digital Circuit Design.....	294
7.2.1 Impact of device performance on digital circuit design	294
7.2.2 Large-scale digital circuits.....	302
7.2.3 Leakage-performance trade off and energy dissipation.....	306
7.2.4 Multi- V_T devices and mixed- V_T circuits	311
7.2.5 High-temperature circuit operation	312
7.2.6 SRAM design	313
7.3 Analog Circuit Design.....	315
7.3.1 Device figures of merit and technology related design issues.....	315
7.3.1.1 Transconductance	315
7.3.1.2 Intrinsic transistor gain	316
7.3.1.3 Matching behavior.....	317
7.3.1.4 Flicker noise.....	318
7.3.1.5 Transit and maximum oscillation frequency	319
7.3.1.6 Self-heating.....	320
7.3.1.7 Charge trapping in high-k dielectrics.....	320
7.3.2 Design of analog building blocks	321
7.3.2.1 V_T -based current reference circuit	321
7.3.2.2 Bandgap voltage reference.....	322
7.3.2.3 Operational amplifier.....	324
7.3.2.4 Comparator	325
7.3.3 Mixed-signal aspects	326
7.3.3.1 Current steering DAC	327
7.3.3.2 Successive approximation ADC	327
7.3.4 RF circuit design.....	329
7.4 SoC Design and Technology Aspects.....	331
Acknowledgments	332
References	333
Index.....	336

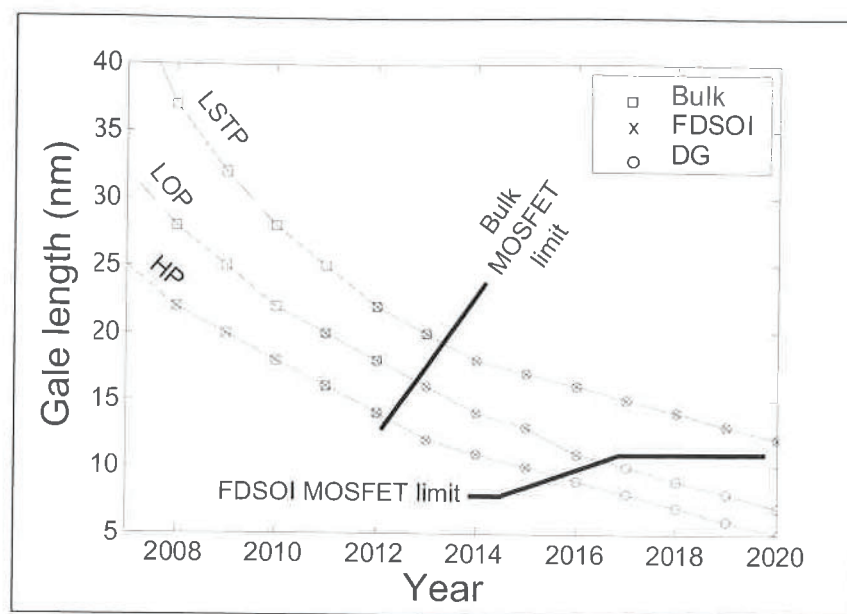


Fig. 1.5. Evolution of gate length predicted by the 2005 ITRS for high-performance (HP), low operating power (LOP), and low standby power (LSTP) digital circuits.

1.4 A Brief History of Multiple-Gate MOSFETs

In a continuous effort to increase current drive and better control short-channel effects, silicon-on-insulator MOS transistors have evolved from classical, planar, single-gate devices into three-dimensional devices with a multi-gate structure (double-, triple- or quadruple- gate devices). It is worth noting that, in most cases, the term “double gate” refers to a single gate electrode that is present on two opposite sides of the device. Similarly, the term “triple gate” is used for a single gate electrode that is folded over three sides of the transistor. One remarkable exception is the M1GFET (Multiple Independent Gate FET) where two separate gate electrodes can be biased with different potentials. It is also worth pointing out that one device may have several different names in the literature (Table 1.1).

Table 1.1. Device names found in the literature.

Acronym	Also known as
MuGFET (Multiple-Gate FET)	Multi-gate FET, Multigate FET
MIGFET (Multiple Independent Gate FET)	Four-terminal (4T) FinFET
Triple-gate FET	Trigate FET
Quadruple-gate FET	Wrapped-Around Gate FET Gate-All-Around FET Surrounding-Gate FET
FinFET	DELTA (fully DEpleted Lean channel TrAnsistor)
FDSOI (Fully Depleted SOI)	Depleted Silicon Substrate
PDSOI (Partially depleted SOI)	Non-Fully Depleted SOI
Volume Inversion	Bulk Inversion
DTMOS (Dual Threshold Voltage MOS)	VTMOS (Varied Threshold MOS) MTCMOS (Multiple Threshold CMOS) VCBM (Voltage-Controlled Bipolar MOS) Hybrid Bipolar-MOS Device

FETs

ontrol short-
olved from
ices with a
ices). It is
to a single
the device.
ode that is
ption is the
parate gate
rth pointing
ie literature

1.4.1 Single-gate SOI MOSFETs

Figure 1.6 shows the “Family Tree” of SOI MOSFETs and shows the evolution from partially depleted, single-gate devices to multi-gate, fully depleted structures. Partially depleted silicon MOSFETs are the successors of earlier SOS (Silicon-On-Sapphire) devices. PDSOI MOSFETs were first used for niche applications such as radiation-hardened or high-temperature electronics. At the turn of the century PDSOI technology became mainstream as major semiconductor manufacturers started to use it to fabricate high-performance microprocessors. The low-voltage performance of PDSOI devices can be enhanced by creating a contact between the gate electrode and the floating body of the device. Such a contact improves the subthreshold slope, body factor and current drive, but limits the device operation to sub-1V supply voltages. [15-23] Fully depleted SOI devices have a better electrostatic coupling between the gate and the channel. This results in a better linearity, subthreshold slope, body coefficient and current drive. FDSOI technology is used in a number of applications ranging from low-voltage, low-power to RF integrated

circuits. There exists a series of textbooks on SOI devices and technology, [24-26] as well as on SOI circuit design.[27-29]

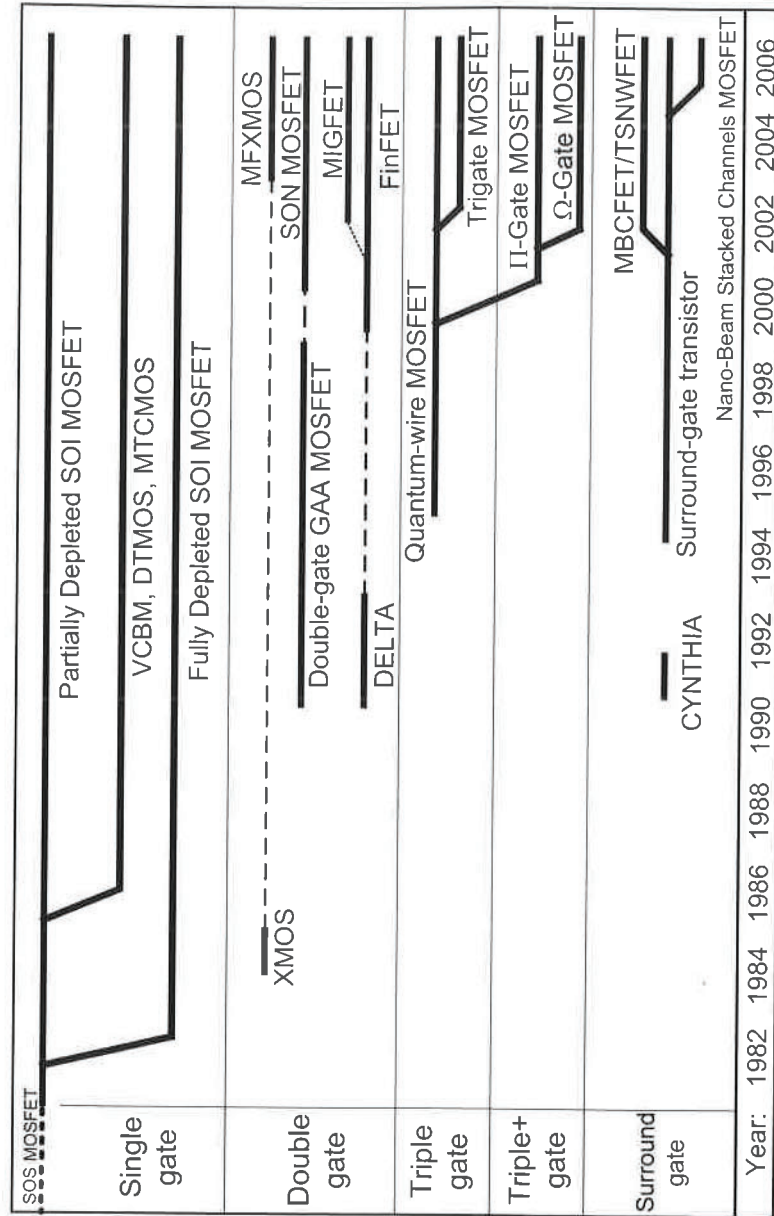


Fig. 1.6. "Family tree" of SOI and multigate MOSFETs.

chnology,

1.4.2 Double-gate SOI MOSFETs

The first article on the double-gate MOS (DGMOS) transistor was published by T. Sekigawa and Y. Hayashi 1984.[30] That paper shows that one can obtain significant reduction of short-channel effects by sandwiching a fully depleted SOI device between two gate electrodes connected together. The device was called XMOS because its cross section looks like the Greek letter Ξ (Ξ). Using this configuration, a better control of the channel depletion region is obtained than in a “regular” SOI MOSFET, and, in particular, the influence of the drain electric field on the channel is reduced, which reduces short-channel.[31] A more complete modeling that includes Monte-Carlo simulations, was published by Frank, Laux and Fischetti in 1992 in a paper that explores the ultimate scaling of the silicon MOSFET.[32] According to that article, the ultimate silicon device is a double-gate SOI MOSFET with a gate length of 30 nm, an oxide thickness of 3 nm, and a silicon film thickness of 5 to 20 nm. Such a (simulated) device shows no short-channel effects for gate lengths larger than 70 nm, and provides transconductance values up to 2300 mS/mm. The first fabricated double-gate SOI MOSFET was the “fully Depleted Lean-channel TrAnsistor (DELTA, 1989)”,[33] where the device is made in a tall and narrow silicon island called “finger”, “leg” or “fin” (Figure 1.7). The FinFET structure is similar to DELTA, except for the presence of a dielectric layer called the “hard mask” on top of the silicon fin. [34-38] The hard mask is used to prevent the formation of parasitic inversion channels at the top corners of the device.

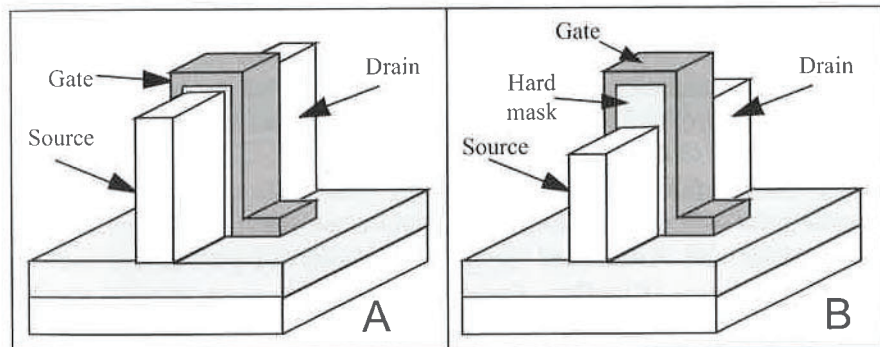


Fig. 1.7. Examples of double-gate MOS structure: A: DELTA MOSFET; B: FinFET.

Other implementations of vertical-channel, double-gate SOI MOSFETs include the “Gate-All-Around device” (GAA), which is a planar MOSFET with the gate electrode wrapped around the channel region (Figure

Nano-Beam Stacked Channels MOSFET													
Year:	1982	1984	1986	1988	1990	1992	1994	1996	1998	2000	2002	2004	2006

1.8.A)[39], the Silicon-On-Nothing (SON) MOSFET,[40-42] the Multi-Fin XMOS (MFXMOS) [43], the triangular-wire SOI MOSFET [44] and the Δ -channel SOI MOSFET.[45] It is worth noting that the original GAA device was a double-gate device, even though the gate was wrapped around all sides of the channel region, because the silicon island was much wider than thick. Nowadays, most people use the GAA acronym for quadruple-gate or surrounding-gate devices having a width-to-height ratio much closer to unity.[46-48]

The MGFET (Multiple Independent Gate FET) is a double-gate device in which the two gate electrodes are not connected together and can, therefore, be biased with different potentials.[49-52,160] The main feature of the MGFET is that the threshold voltage of one of the gates can be modulated by the bias applied to the other gate. This effect is similar to the body effect in FDSOI MOSFETs.[53] An application using MGFET is signal modulation. A simple square law mixer can be formed using a single MGFET by applying a small RF signal to one gate and a large low-frequency signal to the other gate. This single-device modulation is possible because the channel is fully depleted and the gates are perfectly symmetrical and aligned. This signal modulation circuit reduces transistor counts and rail-to-rail transistor stack, making it possible to design compact low-power mixers.[54]

1.4.3 Triple-gate SOI MOSFETs

The triple-gate MOSFET is a thin-film, narrow silicon island with a gate on three of its sides.[55] Implementations include the quantum-wire SOI MOSFET (Figure 1.8.B) [56-57] and the trigate MOSFET.[58-59]

The Electrostatic Integrity of triple-gate MOSFETs can be improved by extending the sidewall portions of the gate electrode to some depth in the buried oxide and underneath the channel region (Π -gate device [60-61] and Ω -gate device [62-64]). From an electrostatic point of view, the Π -gate and Ω -gate MOSFETs have an effective number of gates between three and four. The use of strained silicon, a metal gate and/or high-k dielectric as gate insulator can further enhance the current drive of the device.[65-68]

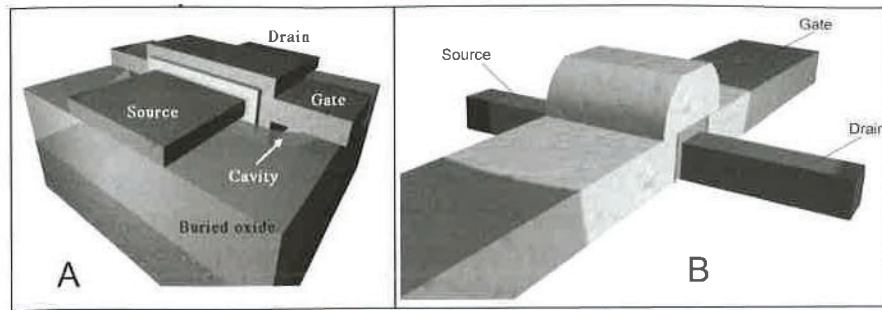


Fig. 1.8. A: Gate-All-Around (GAA) MOSFET; B: Triple-gate MOSFET.

1.4.4 Surrounding-gate (quadruple-gate) SOI MOSFETs

The structure that theoretically offers the best possible control of the channel region by the gate, and hence the best possible Electrostatic Integrity is the surrounding-gate MOSFET. The first surrounding-gate MOSFETs were fabricated by wrapping a gate electrode around a vertical silicon pillar. Such devices include the CYNTHIA device (circular-section device) [69-70] and the pillar surrounding-gate MOSFET (square-section device).[71] More recently, planar surrounding-gate devices with square or circular cross sections have reported.[72-73] Surrounding-gate SOI MOSFETs with a gate length as small as 5 nm and a diameter of 3 nm have shown to be fully functional.[74-75]

To increase the current drive per unit area, multiple surrounding-gate channels can be stacked on top of one another, while sharing common gate, source and drain. Such devices are called the Multi-Bridge Channel MOSFET (MBCFET), [76-77] the Twin-Silicon-Nanowire MOSFET (TSNWFET) [78], or the Nano-Beam Stacked Channels (GAA) MOSFET.[79] Analytical models for the electrical characteristics of cylindrical surrounding-gate MOSFETs can be found in the literature.[80-83]

Schematic cross sections corresponding to the different gate structures described in the previous sections are shown in Figure 1.9.

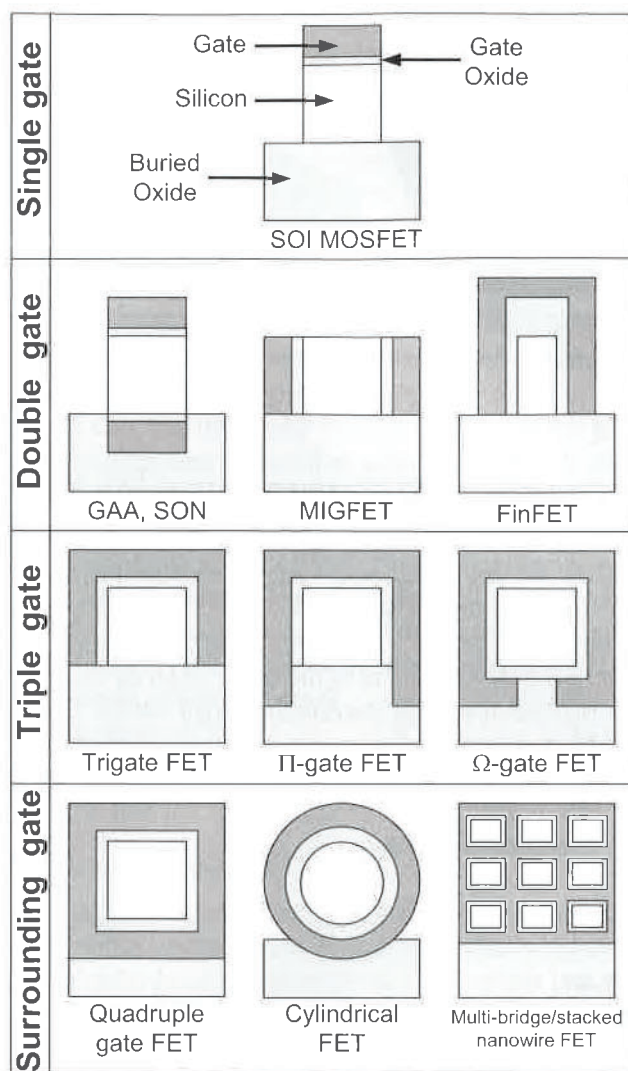


Fig. 1.9. Different gate structures.

1.4.5 Other multigate MOSFET structures

The Inverted T-channel FET (ITFET) combines a thin-film planar SOI device with a trigate transistor (Figure 1.10.A). [84-85] It comprises planar horizontal channels and vertical channels in a single device. The devices have multi-gate control around these channels. The Inverted T-gate

structure has several advantages: the large base helps the fins from falling over during processing; it also allows for transistor action in the space between the fins, which is left unused in other MuGFET configurations. These additional channels increase the current drive. Numerical simulation of an N-channel ITFET reveals different turn-on mechanisms in different parts of the device. The corners of the device turn on first, immediately followed by the surface of the planar regions and the vertical channel. Since each ITFET has about seven corner elements they constitute a significant current to each ITFET device and in a well-designed device can yield substantially more current than a planar device of equivalent area.

The bulk FinFET is a FinFET made on bulk silicon instead of an SOI wafer. Fins are etched on a bulk silicon wafer and trimmed using an oxidation step. Field oxide is deposited to avoid inversion between the fins (Figure 1.10.B). Device with fin width down to 10 nm have shown to have good punchthrough immunity down to the sub-20nm gate length regime.[86-87]

The multi-channel Field Effect Transistor (McFET) is a modified bulk FinFET where a trench is etched in the center of the fin.[88] The trench is filled by the growth of a gate oxide and the deposition of gate material. This process produces a device having two very thin “twin” fins running from source to drain (Figure 1.10.C).

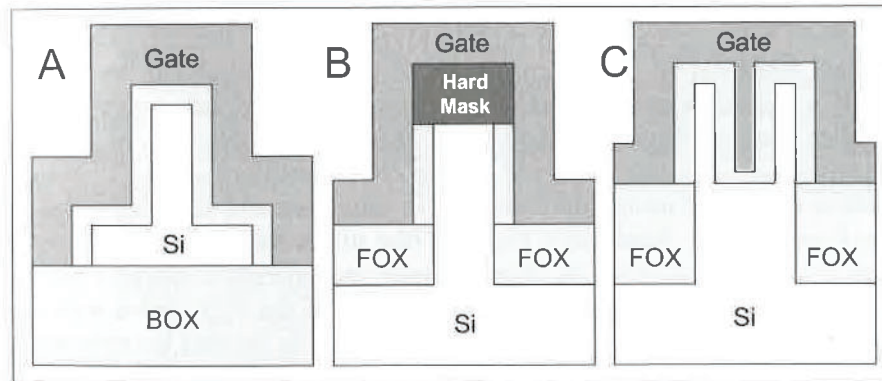


Fig. 1.10. Cross section of A: Inverted T channel FET; B: Bulk FinFET; C: Multi-channel Field Effect Transistor.

The 4-gate transistor (G^4 FET) has two (top and bottom) MOS gates and two (lateral) JFET gates. It is operated in accumulation mode and has the same structure as an inversion-mode partially depleted SOI MOSFET with

1 planar SOI
prises planar
The devices
erted T-gate

two independent body contacts. These lateral contacts play the role of source and drain in the G^4 FET, while the junctions are used as lateral gates. The current in the body of the device is controlled by the front and back MOS gates, and by the two lateral junctions. The height and width of the conductive path is modulated by a mix of MOS and JFET effects. Each gate has the ability of switching the transistor on and off.[89-91] A model for the G^4 FET has been published in the literature.[92]

1.4.6 Multigate MOSFET memory devices

SONOS (Silicon-Oxide-Nitride-Oxide-Silicon) devices are non-volatile flash memory devices. They are essentially used in mobile applications. Flash devices have a small form factor, high storage density and low power consumption. For logic applications FinFET type devices are known to have good scalability down to 10-nm gate length. The FinFET device architecture combined with an ONO trapping layer as gate dielectric enables memory cells with feature sizes well below 50-nm. SONOS FinFET cells are programmed and erased using Fowler–Nordheim tunneling. SONOS FinFET memory devices show excellent functionality down to 20 nm channel length. [93-94] As an alternative to the ONO layer, nanocrystals embedded within the gate dielectric can be used to trap charges and achieve a similar memory effect. FinFET flash memory devices with a V_{TH} window larger than 1 volt have been demonstrated using silicon nanocrystals embedded in the gate oxide.[95]

The SOI Zero-capacitor RAM (ZRAM) memory cell consists of a single PDSOI transistor in the floating body of which a charge can be stored. Such a capacitorless, one-transistor memory cell is, of course much smaller than a classical DRAM cell that requires the use of both a transistor and a storage capacitor. In the ZRAM memory cell a “1” binary state is stored by biasing the transistor in saturation and injecting holes in the floating body. Applying a negative bias to the bit line (device source) removes the charge from the floating body and, therefore, stores a binary “0” in the device. Since the threshold voltage of the FET varies with the charge stored in the body the logic state can easily be read by measuring the drain current when the device is turned on. Even though the storage of a charge requires a floating body, and, therefore, a partially depleted device, it is possible to realize ZRAM cells in fully depleted FinFETs of trigate devices. By applying a back-gate bias to the device, a floating body can be created at the bottom of the otherwise fully depleted fin and ZRAM operation can be activated.[96] Retention times of few tens of milliseconds are observed at room temperature on FinFETs with 300 nm channel length.

the role of
d as lateral
ie front and
nd width of
ffects. Each
1] A model

non-volatile
pplications.
ty and low
devices are
The FinFET
/er as gate
low 50-nm.
r-Nordheim
unctionality
ONO layer,
used to trap
ash memory
emonstrated

ts of a single
n be stored.
course much
e of both a
a "1" binary
ting holes in
vice source)
ores a binary
ries with the
y measuring
he storage of
ally depleted
1 FinFETs of
floating body
n and ZRAM
milliseconds
annel length.

These values suit well typical embedded DRAM specification. FinFETs provide better retention characteristics than trigate devices which have a retention time of approximately 2 ms. Retention time decreases at shorter channel length but even 100 nm devices still show retention time of few milliseconds.[97]

When a positive bias is applied to the underlying silicon substrate (back gate) and a negative bias is applied to the gate of a MuGFET, it is possible to create an inversion layer at the bottom of the device and accumulation at the other interfaces. In that case, the so-called "MetaStable Dip" (MSD) effect can be observed, in which the transconductance as a function of the gate voltage increases, then drops, and then increases again. This effect is time-dependent and presents an hysteresis effect which depends whether the gate voltage is scanned in the forward or reverse mode.[98] The hysteretic nature of the phenomenon makes it useful for potential single-transistor memory applications.[99]

1.5 Multigate MOSFET Physics

1.5.1 Classical physics

1.5.1.1 Natural length and short-channel effects

The Electrostatic Integrity (*EI*), which we know to be related to short channel effects, was compared between a bulk transistor, a FDSOI MOSFET, and a double-gate device in Figure 1.3. It was shown that the *EI* can be improved by reducing the gate oxide thickness and/or by decreasing the silicon film thickness (the junction depth in the case of a bulk device). We are now going to extend the analysis to all types of multigate MOSFETs.

The potential distribution in the channel of a fully depleted multigate MOSFET can be obtained by solving Poisson's equation using the depletion approximation:

$$\frac{d^2\Phi(x, y, z)}{dx^2} + \frac{d^2\Phi(x, y, z)}{dy^2} + \frac{d^2\Phi(x, y, z)}{dz^2} = \frac{qN_a}{\epsilon_{si}} \quad (1.7)$$

Let us first understand the meaning of this equation. It can be rewritten in the form:

iconductor-

Quantum-
sactions on

stigating the
FETs. Solid-

i: Quantum-
and carbon-

MOSFETs.
Nanoscaled

-dimensional

n, W. Xiong,
'atrino: Low-
EEE Electron

Schrüfer, K.
Effects in Pi-
(2006)

Agarwal, C.H.
reye, G.Q. Lo,
owire gate-all-
ation and low
M, paper 20.4

guyen, G. Van
, E. Suzuki, M.
ith asymmetric
7)

2 Multigate MOSFET Technology

Weize (Wade) Xiong

2.1 Introduction

Manufacturing a self-aligned double-gate MOSFET has been the holy grail of device engineers and researchers ever since it was proposed by Sekigawa and Hayashi in 1984.[1] The first modern self-aligned vertical multi-gate MOSFET was called DELTA (fully **DE**pleted **L**ean channel **Tr**Ansistor).[2] This device was proposed by D. Hisamoto *et al.* in 1989. Figure 2.1 shows a cross section of the DELTA MOSFET.

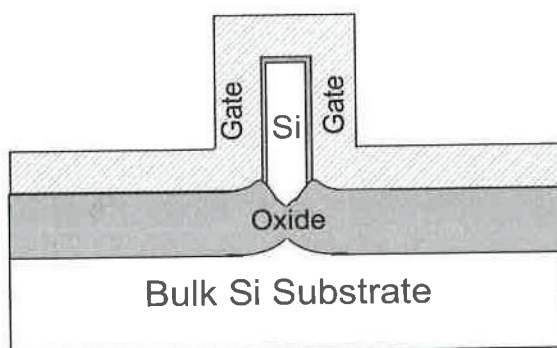


Fig. 2.1. DELTA MOSFET [1-2]. The oxide under the Si fin was formed through LOCOS oxidation while the Si fin was protected by the nitride hard mask and a nitride spacer prior to the oxidation process.

The two key features of the DELTA MOSFET are: 1) front and back MOSFET gates are inherently self-aligned and 2) the channels are on the sidewall of the silicon body/fin. The compatibility of the DELTA FET with existing planar CMOS processing is the main reason why it was

adopted as the mainstream multi-gate structure. FinFETs, Tri-gate and Π/Ω gate devices have all evolved from of this original concept [3-8].

The critical fabrication steps in the front-end processing of a multi-gate MOSFET include, sequentially: 1) fin formation, 2) gate stack formation, 3) source and drain extension implant, 4) spacer formation, 5) epitaxial raised source/drain formation, and 6) deep source/drain implantation and activation anneal.

The fabrication flow of a Tri-gate MOSFET on an SOI substrate is shown in Figure 2.2. The SOI silicon top layer (T_{si}) thickness defines the fin height (Fin_{HEIGHT}). The fin pattern and the critical dimension of fin width (Fin_{WIDTH}) can be defined by optical lithography or by spacer image transfer (SIT) [9-10], followed by plasma etching.

After fin etch, the fin sidewall surfaces are rough. Therefore, oxidation and H_2 annealing are often used to smooth the sidewalls.[11-12] Next, the gate dielectric is grown and metal gate is deposited. It is suitable to tune the threshold voltage (V_{th}) of the MOSFET by using a gate material that has the appropriate effective workfunction rather than by doping the channel. Sections 2.3, 2.5 and 2.6 will show that it is highly desirable to have intrinsic or lightly doped channels.

Since the gate stack is over the fin topography, a planarization step is desirable to flatten the gate surface, which reduces the burden on photolithography and gate etch. Significant overetch of the gate material is required to clear the bottom of the fins. As a result, the gate etch must have a high selectivity to the gate dielectric on top of the fin, if one wants to avoid damage to the fin during gate etch. Source and drain (S/D) extensions are formed after gate patterning using low-energy and large-tilt angled implants.[13-14] Next, S/D offset spacers are formed along the sidewalls of the gate and fin. The sidewall spacers on the fins are subsequently removed to expose the fin to grow raised source and drain using selective epitaxy.[13-15] The raised source and drain structure helps to reduce the parasitic resistance associated with thin fins.[13]

The rest of this chapter is divided into sections according to multi-gate device fabrication sequences. Section 2.2 covers fin formation, Section 2.3 describes gate stack structures, Section 2.4 describes source and drain formation and Section 2.5 highlights mobility and strain engineering.

gate and
3-8].

multi-gate
formation,
) epitaxial
ation and

substrate is
defines the
sion of fin
acer image

, oxidation
2] Next, the
ible to tune
aterial that
doping the
desirable to

ation step is
burden on
e material is
te etch must
if one wants
drain (S/D)
and large-tilt
ed along the
the fins are
ce and drain
ructure helps

to multi-gate
tion, Section
rce and drain
reering.

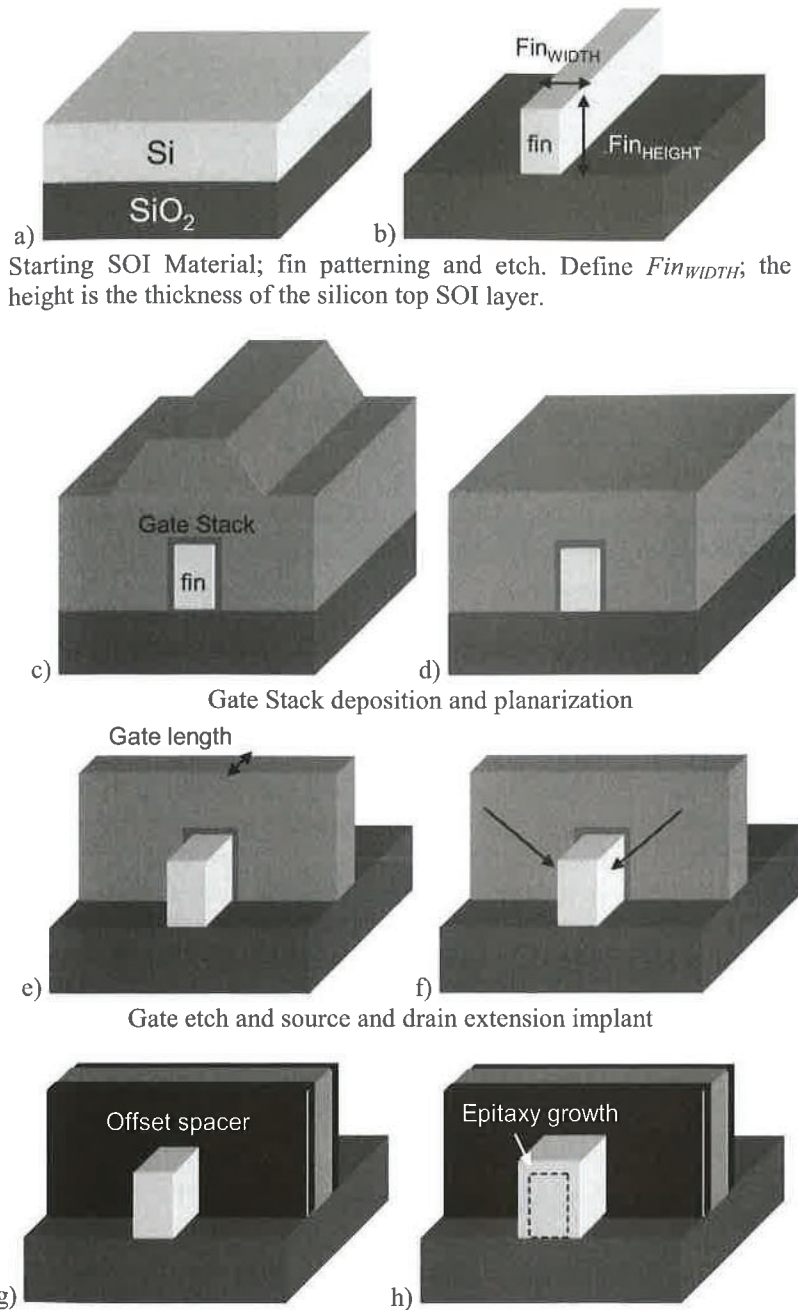


Fig. 2.2. General Fabrication Sequence of a Tri-gate MOSFET.

2.2 Active Area: Fins

2.2.1 Fin Width

As the MOSFET gate length, L_g , continues to shrink, year after year, the drain starts to compete with the gate electrode for control of the channel potential. The resulting “short-channel effects” (SCE) are: 1) higher sub-threshold leakage, 2) a threshold voltage (V_{th}) roll-off, and 3) a form of punch-through between the drain and source when V_{DS} is equal to the supply voltage.[16]

To reduce the drain influence in the channel, scaled planar bulk MOSFETs and PDSOI MOSFETs rely on: 1) gate oxide thickness reduction and 2) higher channel doping. The use of a thinner gate oxide increases the gate-to-channel capacitance. A higher channel doping concentration reduces charge sharing between the gate and drain in the channel, and it creates a large potential barrier between source and drain.

However, in modern MOSFETs, direct tunneling current through the gate dielectric prevents further scaling of the gate oxide thickness. The use of high channel doping concentrations reduces carrier mobility and increases Gate Induced Drain Leakage (GIDL).[17-18]

Another way to control SCEs is to use two or more gate electrodes and a thin, fully depleted semiconductor body. The idea is to create a large potential curvature in the source and drain direction by introducing a large electrical field gradient in the direction vertical to the gate. The key parameter in this approach is the thickness of the thin, fully depleted semiconductor body. In the case of the multi-gate technologies, the fin width is the thickness of the thin, fully depleted semiconductor body.

Chapter 1 introduced a powerful concept called natural length, λ , with:

$$\lambda = \sqrt{\frac{\epsilon_{si}}{2\epsilon_{ox}}} t_{si} t_{ox} \quad \text{for a double-gate MOSFET} \quad (2.1)$$

λ is a measure of the short-channel effect. It represents the distance of penetration of the drain electric field in the channel. A small value λ is desirable to minimize the SCE.

λ is proportional to the square root of the product of the device body thickness t_{si} , (or Fin_{WIDTH} in multi-gate technology) and gate oxide thickness t_{ox} . Thus, one can trade off t_{ox} scaling with Fin_{WIDTH} reduction. This approach is not possible for bulk or PDSOI MOSFETs.[17]

In a double-gate MOSFET, the center of the fin is the furthest away from the gate electrodes. Therefore, gate control is the weakest at this point. Should punchthrough occur, it would be in the center of the fin.[20] This implies using a slightly different boundary condition than Equation 2.1, which assumes the punch through is along the channel surfaces.

In this section, we will derive a new value for λ based on the assumption that punchthrough occurs in the center of the channel. The new λ has the same physical meaning as Equation 2.1, but it can better explain some of the experimental results shown later in this section. We will use the same coordinate system as in Figure 1.11. Starting with Poisson's equation, we can write:

$$\frac{d^2\Phi(x,y,z)}{dx^2} + \frac{d^2\Phi(x,y,z)}{dy^2} + \frac{d^2\Phi(x,y,z)}{dz^2} = \frac{qN_{sub}}{\epsilon_{si}} \quad (2.2)$$

For single-gate or double-gate devices $\frac{d\phi}{dz} = 0$, such that Equation 2.2 is reduced to:

$$\frac{d^2\Phi(x,y)}{dx^2} + \frac{d^2\Phi(x,y)}{dy^2} = \frac{qN_{sub}}{\epsilon_{si}} \quad (2.3)$$

Assuming a parabolic potential distribution in the y-direction [21] we can write:

$$\Phi(x,y) \approx c_0(x) + c_1(x)y + c_2(x)y^2 \quad (2.4)$$

We now apply the boundary condition $\left. \frac{d\Phi(x,y)}{dy} \right|_{y=0} = 0$ at $y = \frac{t_{si}}{2}$,

which yields:

$$\Phi(x,y) = \Phi_f(x) + \frac{\epsilon_{ox}}{\epsilon_{si}} \frac{\Phi_s(x) - \Phi_{gs}}{t_{ox}} y - \frac{1}{t_{si}} \frac{\epsilon_{ox}}{\epsilon_{si}} \frac{\Phi_s(x) - \Phi_{gs}}{t_{ox}} y^2 \quad (2.5)$$

where $\Phi_s(x)$ is the front- and back-surface potential and $\Phi_{gs} = V_{gs} - V_{FB}$ is the front gate voltage V_{gs} minus the flat-band voltage, V_{FB} . Since the potential at the center of the fin, $\Phi_c(x)$, is the most relevant to SCE, we extract a relationship between $\Phi_c(x)$ and $\Phi_s(x)$ from 2.5 by substituting $y = \frac{t_{si}}{2}$ as follows:

$$\Phi_s(x) = \frac{1}{1 + \frac{\epsilon_{ox}}{4\epsilon_{si}} \frac{t_{si}}{t_{ox}}} \left(\Phi_c(x) + \frac{\epsilon_{ox}}{4\epsilon_{si}} \frac{t_{si}}{t_{ox}} \Phi_{gs} \right) \quad (2.6)$$

Expressing $\Phi(x, y)$ as a function of $\Phi_c(x)$ we obtain:

$$\Phi(x, y) = \left(1 + \frac{\epsilon_{ox}}{\epsilon_{si}} \frac{y}{t_{ox}} - \frac{\epsilon_{ox}}{\epsilon_{si}} \frac{y^2}{t_{ox} t_{si}} \right) \left(\frac{\Phi_c(x) + \frac{\epsilon_{ox}}{4\epsilon_{si}} \frac{t_{si}}{t_{ox}} \Phi_{gs}}{1 + \frac{\epsilon_{ox}}{4\epsilon_{si}} \frac{t_{si}}{t_{ox}}} \right) - \left(\frac{\epsilon_{ox}}{\epsilon_{si}} \frac{y}{t_{ox}} \Phi_{gs} - \frac{\epsilon_{ox}}{\epsilon_{si}} \frac{y^2}{t_{ox} t_{si}} \Phi_{gs} \right) \quad (2.7)$$

Substituting 2.7 back in 2.3 we get:

$$\frac{d^2 \Phi_c(x)}{dx^2} + \frac{\Phi_{gs} - \Phi_c(x)}{\lambda^2} = \frac{qN_A}{\epsilon_{si}} \quad (2.8)$$

where natural length λ is now written:

$$\lambda = \sqrt{\frac{\epsilon_{si}}{2\epsilon_{ox}} \left(1 + \frac{\epsilon_{ox}}{4\epsilon_{si}} \frac{t_{si}}{t_{ox}} \right) t_{si} t_{ox}} \quad (2.9)$$

Numerical simulations in [17] show that for reasonable short-channel effect control, the gate length needs to be 5-10 times greater than λ .

We have established a theoretical relationship between L_g and λ . For a device designer, it would be more relevant to establish a guideline for the L_g to t_{si} ratio to control short-channel effects. Equation 2.9 does not separate the effects of t_{si} and t_{ox} . To walk around the problem, let us

assume a good SCE control can be achieved with $L_g = n\lambda$, where n is some arbitrary number. We now have:

$$\frac{L_g}{n} = \lambda = \sqrt{\frac{\epsilon_{si}}{2\epsilon_{ox}} \left(1 + \frac{\epsilon_{ox}}{4\epsilon_{si}} \frac{t_{si}}{t_{ox}} \right) t_{si} t_{ox}} \quad (2.10)$$

In the limiting case when $L_g = 0$, we can write:

$$\frac{\epsilon_{si}}{2\epsilon_{ox}} \left(1 + \frac{\epsilon_{ox}}{4\epsilon_{si}} \frac{t_{si}}{t_{ox}} \right) t_{si} t_{ox} = 0 \quad (2.11A)$$

which implies:

$$t_{si} = -4 \frac{\epsilon_{si}}{\epsilon_{ox}} t_{ox} \quad (2.11B)$$

The right-hand term of Equation 2.11B represents the intercept point of the L_g vs. Fin_{WIDTH} plot for $L_g = 0$. Thus, one can define an effective fin width, $FinW_{eff}$, as:

$$FinW_{eff} = Fin_{WIDTH} + 4 \frac{\epsilon_{si}}{\epsilon_{ox}} t_{ox} \quad (2.12)$$

Equation 2.12 shows that the impact of gate oxide can be considered equivalent to a constant offset of physical fin width. The validity of Equation 2.12 was demonstrated in reference [22]. Fig 2.3 shows contours of subthreshold swing on a L_g vs. Fin_{WIDTH} plot with 48 devices having 6 different Fin_{WIDTH} values and 8 different gate lengths. Each line represents a fixed short channel effect control. Since all devices have the same gate oxide (1.6nm oxynitride), they all intercept $L_g = 0$ at the same point. This

point is approximately equal to $-4 \frac{\epsilon_{si}}{\epsilon_{ox}} t_{ox}$.

Figure 2.4 shows the experimental double-gate FinFET results of $L_g/FinW_{eff}$ vs. subthreshold swing (or Subthreshold Slope, SS).[22] If we assume that a SS value of 90mV/dec is acceptable, then $L_g/FinW_{eff}$ is ~ 1.5 . A triple-gate MOSFET with enhanced gate electrostatic control of the channel provides a lower L_g/Fin_{WIDTH} ratio than a double-gate FinFET for a same SCE control.[23] L_g/Fin_{WIDTH} as low as 1.0 was reported for Ω -Gate devices in [8]. *It is worthwhile noting that the minimum feature size in multi-gate technology is the fin width (Fin_{WIDTH}), and not the gate length.*

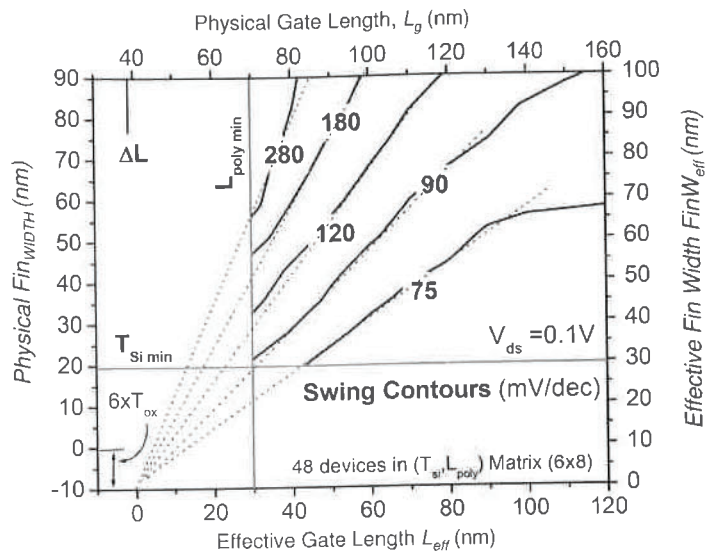


Fig. 2.3. Subthreshold swing contours for NMOS FinFET devices, $V_d=0.1V$. Fin_{WIDTH} varies from 20nm to 82.5nm in steps of 12.5nm, physical gate length; L_g varies from 70nm to 157.5nm in steps of 12.5nm. Contours are extrapolated with dotted lines to a common origin.[22] Copyright© 2001 IEEE.

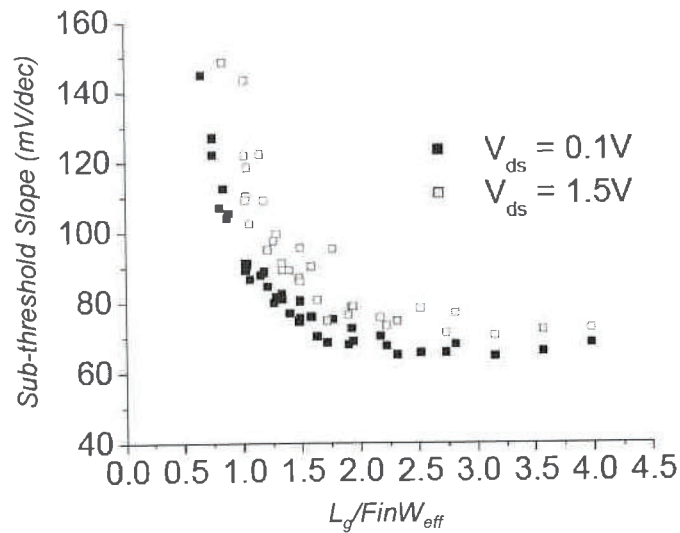


Fig. 2.4. Measured subthreshold swing vs. $L_g/FinW_{eff}$. Fin_{WIDTH} varies from 20nm to 90nm; L_g varies from 30nm to 120nm. The subthreshold slope is shown for both for $V_d=0.1V$ and 1.5V.[22] Copyright© 2001 IEEE.

2.2.2 Fin Height and Fin Pitch

Having understood the requirements on Fin_{width} for SCE control, let us look at how fin pitch and fin height affect the drive current. For a double-gate FinFET, the channels are on the sidewalls of the fins. The effective channel width is thus equal to: [24]

$$W_{eff} = 2n Fin_{HEIGHT} \quad (2.13)$$

Where n is an integer number that is equal to the number of fins in a device. W_{eff} of multi-gate devices can only be increased or decreased by a discrete amount equal to $2 \times Fin_{HEIGHT}$. [25]

The fin pitch, P_{FIN} , is the spacing between the fins plus the fin width. P_{FIN} is limited by the lithography pattern capability. Exactly one fin can be placed in one P_{FIN} . Therefore, the effective channel width per pitch is $2 \times Fin_{HEIGHT}$. Figure 2.5 shows the layout comparison between a planar device and a FinFET. For the planar MOSFET, the effective device channel width, W_{eff} , is equal to the footprint on the substrate, W_{foot} . For FinFET, W_{eff} is related to W_{foot} through:

$$W_{eff} = 2Fin_{HEIGHT} \frac{W_{foot}}{P_{FIN}} \quad (2.14)$$

If we want the FinFET layout to be "competitive", we need $W_{eff} \geq W_{foot}$, or:

$$2Fin_{HEIGHT} \geq P_{FIN} \quad (2.15)$$

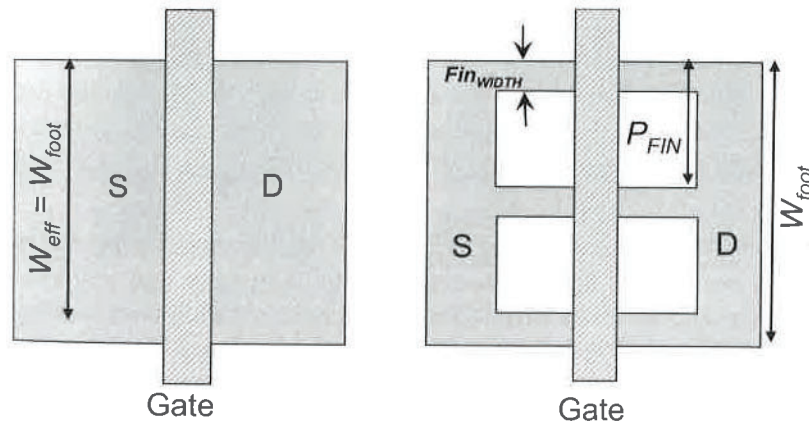


Fig. 2.5. Layout comparison between a planar MOSFET and a typical FinFET.

The analysis of the tri-gate channel width is further complicated by the “volume inversion” or “bulk inversion” effect reported in [26]. For thin fins, the inversion charge is not limited to the device surfaces. A large portion of the inversion charge is in the bulk of the fins, even at high gate voltages. Thus, without activating the top gate, multi-gate devices already have a significant charge at the top surface of the fins. Figure 2.6 shows the two-dimensional numerical simulation of carrier density on the top surface of 1) a tri-gate, 2) a double-gate and 3) a FinFET with a hard mask on top of the fin.

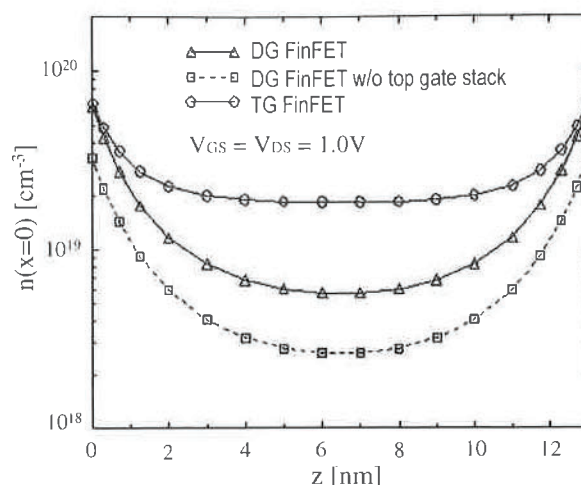


Fig. 2.6. Davinci prediction of on-state electron density along the top fin surface at the center of the channel ($x = 1/2 L_{eff}$). DG stands for double gate. TG stands for tri-gate. Due to bulk inversion, the electron density on the top fin surface for DG FinFET is comparable to that in the TG device [26]. Copyright© 2005 IEEE.

Due to this “bulk inversion”, the effective width, ΔW_{eff} , of the top gate in a tri-gate MOSFET is only a fraction of the physical top gate width. Numerical simulation in [26] shows ΔW_{eff} is a function of fin height and fin width.

For tri-gate devices, Equation 2.12 becomes: $2Fin_{HEIGHT} + \Delta W_{eff} \geq P_{FIN}$

Fins can be defined by conventional photolithography and reactive ion etch. Fins with width less than 20nm have been reported with 193nm dry photolithography processing.[27-28] Figure 2.7 shows the TEM cross section of some of these fins.[27-31] Continuing advances in lithography pitch and line dimension, including emerging 193nm photolithography and Extreme UV (EUV) lithography pave the way for further Fin_{WIDTH} and P_{FIN} reduction.

ated by the
5]. For thin
es. A large
at high gate
ices already
e 2.6 shows
on the top
a hard mask

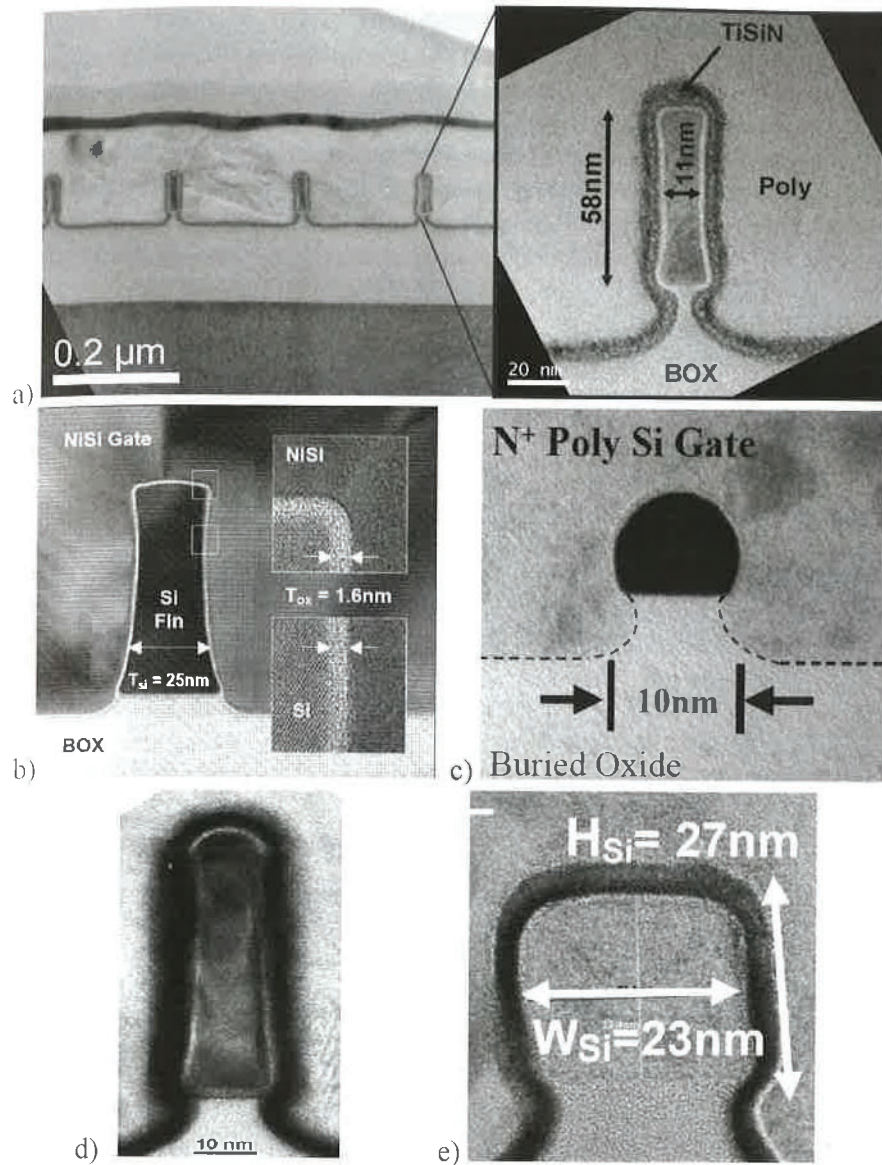


Fig. 2.7. Cross-sectional TEM of fins under gates. a) [27] Copyright© 2006 IEEE, b) [28] Copyright© 2002 IEEE, c) [29] Copyright© 2004 IEEE, d) [30] Copyright© 2005 IEEE, e) [31] Copyright© 2006 IEEE.

To increase the current drive per unit W_{foot} , we can increase Fin_{HEIGHT} , or reduce P_{FIN} . Both options have practical limitations. Increasing Fin_{HEIGHT} undermines the fin stability and increases the difficulty of gate patterning.

top fin surface
. TG stands for
surface for DG
05 IEEE.

the top gate in
p gate width.
height and fin

$W_{eff} \geq P_{FIN}$
d reactive ion
with 193nm dry
ie TEM cross
in lithography
ithography and
 n_{WIDTH} and P_{FIN}

P_{FIN} is limited by the optical lithography capability. In order to improve P_{FIN} beyond optical lithography capability, spacer-defined fin patterning has been proposed.[32] Figure 2.8 shows the general concept. First a sacrificial pattern is defined with conventional optical lithography. Then spacers are formed on the sidewalls of the sacrificial pattern through thin film deposition and anisotropic reactive ion etch (RIE). The sacrificial pattern is subsequently selectively removed leaving the remaining spacers to serve as hard mask for fin etch.

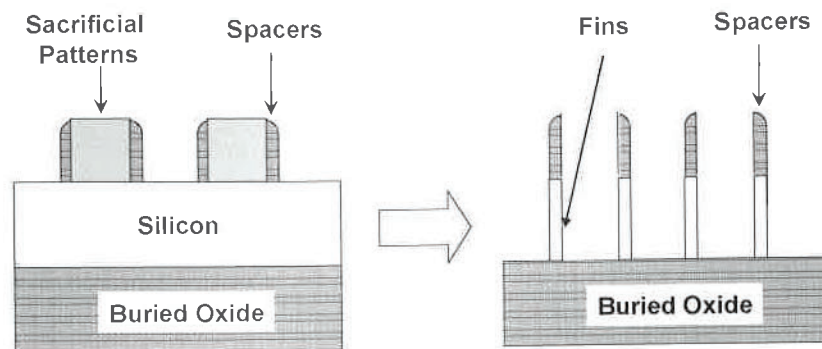


Fig. 2.8. Illustration of a spacer-defined fin formation process.

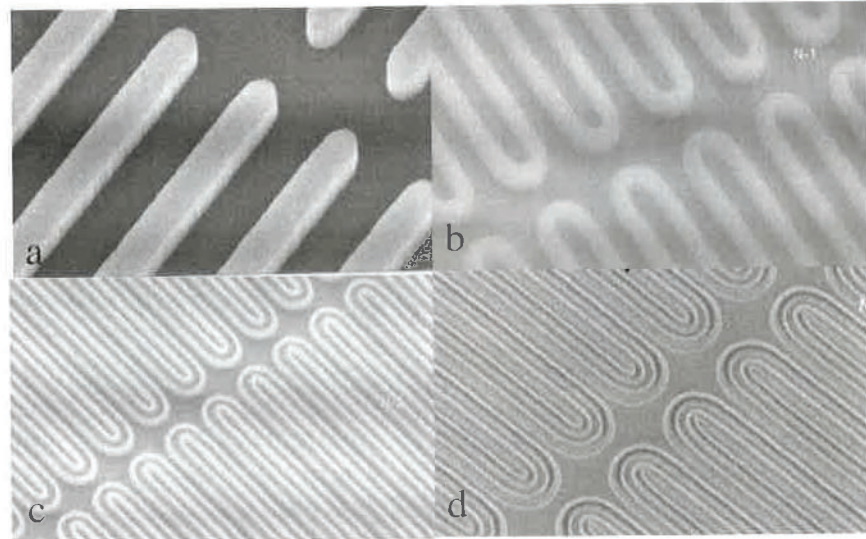


Fig. 2.9. SEM pictures of three progressive spacer-defined fin formation processes. a) initial optical lithography of sacrificial pattern. b) after first spacer imaging; c) after second spacer imaging; and d) after third spacer imaging. The final pattern pitch is 6 times smaller than the optical lithography pitch. REPRINTED WITH PERMISSION FROM [33]. COPYRIGHT 2003 AMERICAN INSTITUTE OF PHYSICS.

to improve
patterning
pt. First a
lithy. Then
rough thin
sacrificial
ing spacers

spacers



ess,



fin formation
after first spacer
r imaging. The
itch. REPRINTED
JTE OF PHYSICS.

There are two advantages to spacer-defined fin formation. First, it doubles the number of fins for a given lithography pitch, *i.e.*: it reduces the fin pitch to half of the lithography pitch. In fact, when the spacer process is used twice, the pitch can be reduced by a factor of four. Figure 2.9 shows the SEM images of spacer imaging technique progressively being used three times.[33] Second, the fin width is defined by the thickness of deposited film, which has the potential to 1) be narrower than a lithography-defined fin, 2) have better CD control and 3) have reduced line-edge roughness.[32] The disadvantage of spacer-defined fin is that a clean-up mask is required to remove the fins where designers do not want them.[25] CMOS FinFET devices have been successfully fabricated using this imaging technique. Successful conversion of a planar microprocessor design into a spacer-defined FinFET microprocessor was also demonstrated in [25].

2.2.3 Fin Surface Crystal Orientation

Electrons and holes mobility depends on the orientation of the silicon surface as well as on the current flow direction.[34-36] Figure 2.10 shows the measured surface mobility of electrons and holes as a function of the inversion layer charge density.[37] For electrons, the highest mobility surface is (100) and the lowest mobility surface is (110). For holes, the (110) surface has the highest mobility and the (100) surface has the lowest mobility. Electrons and holes at a (111) surface have mobilities laying somewhere between those at (100) and (110) surfaces. The origins of the above differences are due to the anisotropy of effective mass and surface scattering.[35,36]

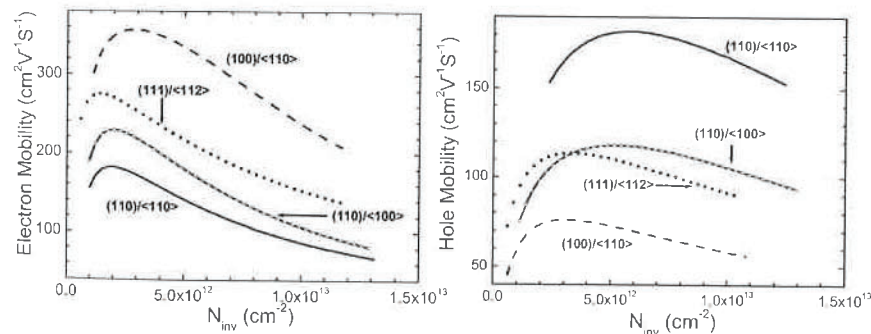


Fig. 2.10. Effective surface mobility of electrons and holes as a function of inversion layer charge density, for different surface orientations.[37] Copyright© 2003 IEEE.

Multi-gate MOSFETs have been fabricated on all 3 major Si surfaces [27,32,19]. The directions of the current flow for the respective surfaces are $\langle 100 \rangle$ for (100), $\langle 110 \rangle$ for (110) and $\langle 112 \rangle$ for (111). (110) Fin sidewall surfaces are made by patterning fins at 0 or 90 degrees with respect to the $\langle 110 \rangle$ notch on silicon wafers with a (001) surface orientation. This layout style is the same as for conventional planar devices. (100) fin surfaces are made by rotating the fin layout by a 45 degree angle with respect to the $\langle 110 \rangle$ notch direction. Figure 2.11 shows the layout directions of (100) fin surfaces and (110) fin surfaces.[38] Alternately, the Si top layer of the SOI wafers can be rotated by a 45 degree angle with respect to the substrate during bonding, to have the (100) surface fin layout at 0 and 90 degrees and the (110) surface at a 45 degree rotation.[39]

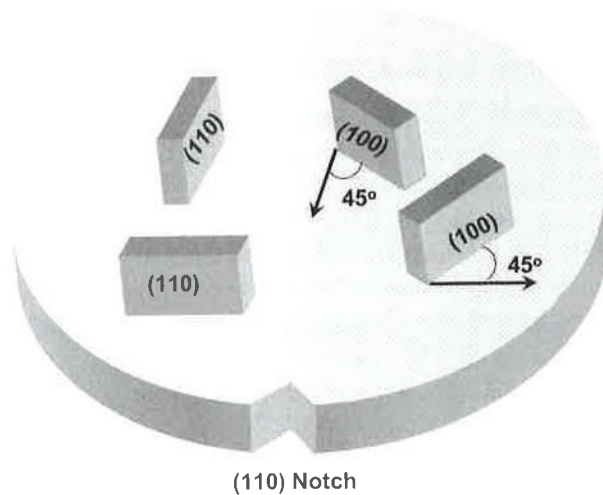


Fig. 2.11. Fins with (110)/ $\langle 110 \rangle$ and (100)/ $\langle 100 \rangle$ surfaces/directions can be readily made using standard Si substrate by rotating the fin layout by 45 degrees.

Having fins with both (110) and (100) surface orientation on the same chip allows one to optimize electron (NMOS) and hole (PMOS) mobility simultaneously. The price to pay for such devices is an increased layout area [41] and a higher photolithography process complexity. Optimizing optical lithography lines at 0, 90 and 45 degree angles is challenging. Figure 2.12 shows the comparison of a 6-transistor SRAM cell layout containing both (110) and (100) fins directions (mixed layout) with a standard layout with only one fin orientation. The mixed layout results in a 16.6% larger cell area.[41]

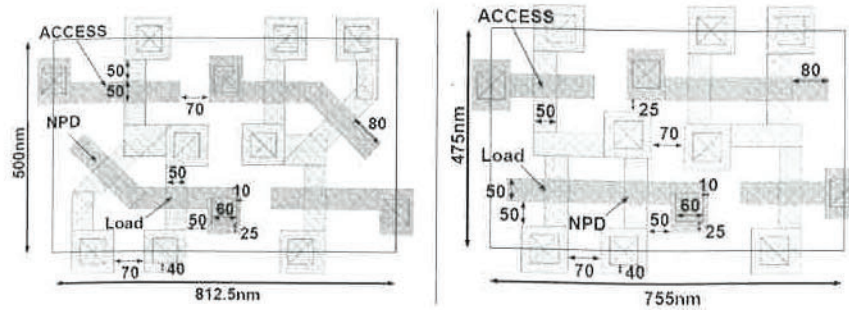


Fig. 2.12. 6-transistor SRAM cell layout comparison: standard layout (right) and layout with NMOS pull-down transistors rotated by 45 degrees (left).[41] Copyright© 2005 IEEE.

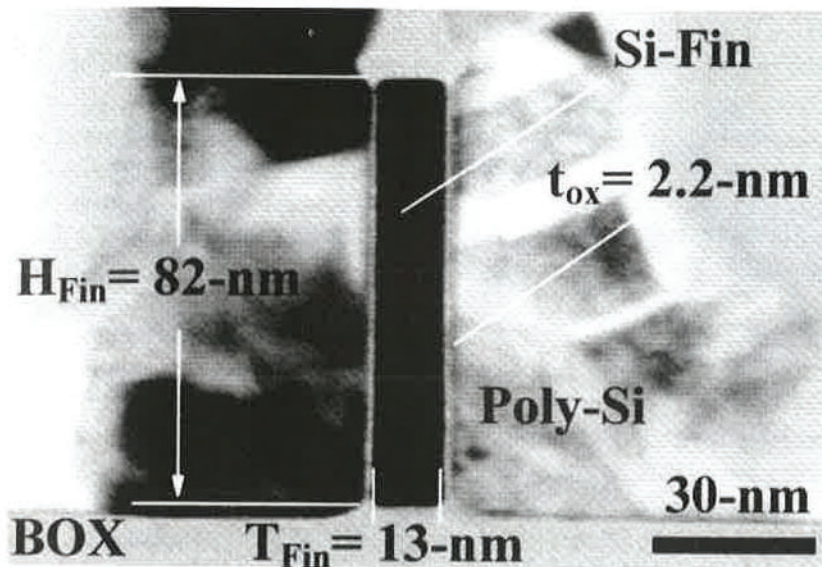


Fig. 2.13. Cross-sectional TEM of a (111)-surface sidewall fin. The smooth rectangular shaped fins were formed using surface-orientation dependent wet etch chemistry.[19] Copyright© 2003 IEEE.

The formation of (111) surface fins on (110) surface SOI substrates has been demonstrated.[19,42] In this experiment, fin patterns were aligned parallel to the $\langle 112 \rangle$ direction. The advantage of the (111) surface fins is that very smooth and straight fins can be formed by a wet etch process. The (111) surface silicon has an extremely low etch rate in tetramethylammonium hydroxide (TMAH) solution, compared with other crystallographic planes. The etch rate ratio at 50°C for (110) planes

si surfaces
re surfaces
(110) Fin
grees with
1) surface
nal planar
ut by a 45
2.11 shows
rfaces.[38]
ed by a 45
o have the
face at a 45

ections can be
y 45 degrees.

n on the same
40S) mobility
creased layout
y. Optimizing
s challenging.
M cell layout
layout) with a
yout results in

versus (111) planes is 20:1. Thus, after initial plasma fin etch, an additional wet etch in TMAH can ensure very straight and smooth (111)-surface fins. Figure 2.13 shows the TEM of a smooth and vertical (111)-sidewall surface fin. One can appreciate the nearly perfect rectangular shape of the fin. The smoothness of the (111) surface fin using surface orientation dependent wet etch chemistry is very attractive. However, the drawback of this technique is that (111) surface carriers mobility is rather low. Furthermore, the mobility enhancement with strain engineering is not currently as well characterized on (111) as on the other two surfaces. Strain engineering will be further discussed in Section 2.5.

2.2.4 Fin Surface Preparation

The fin surfaces exhibit significant roughness right after plasma etch. In addition, the fin corners are sharp. Gate oxide grown on these surfaces is prone to defects and localized thinning. Without proper treatments, silicon crystal defects such as twin formation in the fin corners could happen.[40] Sharp corners also result in a kink in the subthreshold region if the fin doping level is high.[43] Such kinks are caused by parasitic corner devices, which have a lower threshold voltage than the sidewalls due to the higher electrical field in the corners. The fin corner problem is similar to the parasitic STI edge device effect that can be found in planar CMOS technologies.

Carrier mobility along the fin sidewalls without treatment is rather poor.[44] The simplest way to repair surface roughness and damage from the plasma etch is through oxidation. The requirement for fin sidewall smoothing oxidation is similar to STI liner oxidation. It is desirable to have conformal oxidation with a growth rate that is insensitive to the fin surface orientation. In-Situ-Steam-Generated (ISSG) rapid thermal oxidation has such desired properties.[45-46]. ISSG typically uses pressure above 5 Torr. Hydrogen and oxygen are introduced simultaneously into the chamber. The reactants remain as free species (O_2 and H_2) until they reach the wafer surface, where heat induces a local reaction. O_2 and H_2 react to form steam and oxygen radicals. The unstable oxygen radicals promote further oxidation. As a result, the oxidation is conformal and has no surface orientation preference. In addition, ISSG has the advantage of being insensitive to the doping level in the silicon and it provides better corner rounding than conventional oxidation.

The sacrificial smoothing oxide is typically removed using a hydrofluoride (HF)-based isotropic wet etch step.[40] Some of the buried oxide

etch, an
oth (111)-
ical (111)-
rectangular
ng surface
wever, the
ty is rather
ering is not
o surfaces.

asma etch. In
e surfaces is
ents, silicon
happen.[40]
on if the fin
rner devices,
to the higher
milar to the
anar CMOS

ent is rather
damage from
fin sidewall
s desirable to
ive to the fin
apid thermal
uses pressure
aneously into
H₂) until they
n. O₂ and H₂
xygen radicals
ormal and has
e advantage of
rovides better

using a hydro-
e buried oxide

in the SOI substrate is etched at the same time as the sacrificial oxide. As a result, buried oxide recesses below the bottom fin surface and creates the Π/Ω -shaped gate (See Fig. 2.8a). Π/Ω -Gate MOSFETs take advantage of such buried oxide recess to improve the gate electro-static control of the channel.[7-8] However, such a recess can undermine the fin stability if the fin width is scaled below 10nm.

H₂ annealing is another way to smooth the fin surfaces and round the corners. H₂ annealing has been shown to: 1) improve electron mobility, 2) reduce gate leakage and 3) lower 1/f noise.[11,12,47] The H₂ annealing process conditions have a great impact on the final shape of the fins. If the temperature is too high and/or the pressure is too low, the Si fins can agglomerate and break. The narrower the fins, the more sensitive they are to the H₂ annealing conditions.

2.2.5 Fins on Bulk Silicon

The fins we described so far are fabricated on SOI wafers. MuGFETs can also be made on bulk silicon wafers.[48-53] In some references they are called body-tied FinFETs or bulk FinFETs. The main advantages of using bulk silicon over SOI substrates are 1) lower wafer cost and 2) better substrate heat transfer rate. However, bulk multi-gate MOSFET require additional isolation steps, which increase the cycle time and cost. Furthermore, the fin height of bulk-based multi-gate devices are entirely determined by the fin etch step. This puts more pressure on etch variation control, since any fin height variation translates into transistor width variation.

The control of the electrical potential by the gate in bulk fins is as good as in their SOI counterparts. But underneath the fin, in bulk devices, the source and drain have to be separated by heavy channel stop implants to prevent sub-surface punchthrough. As a result, multi-gate devices made on bulk silicon suffer from higher source and drain diode leakage and higher junction capacitance than devices made in SOI.

The bulk silicon under the fin can be accessed by a body contact. But the body factor of a bulk FinFET is very low, since the electrostatic potential inside the fins is dominated by the gate, not by the body. Therefore, body bias is not effective in changing the threshold voltage of a bulk multi-gate MOSFET. Figure 2.14 illustrates the general fabrication process of a bulk MuGFET.

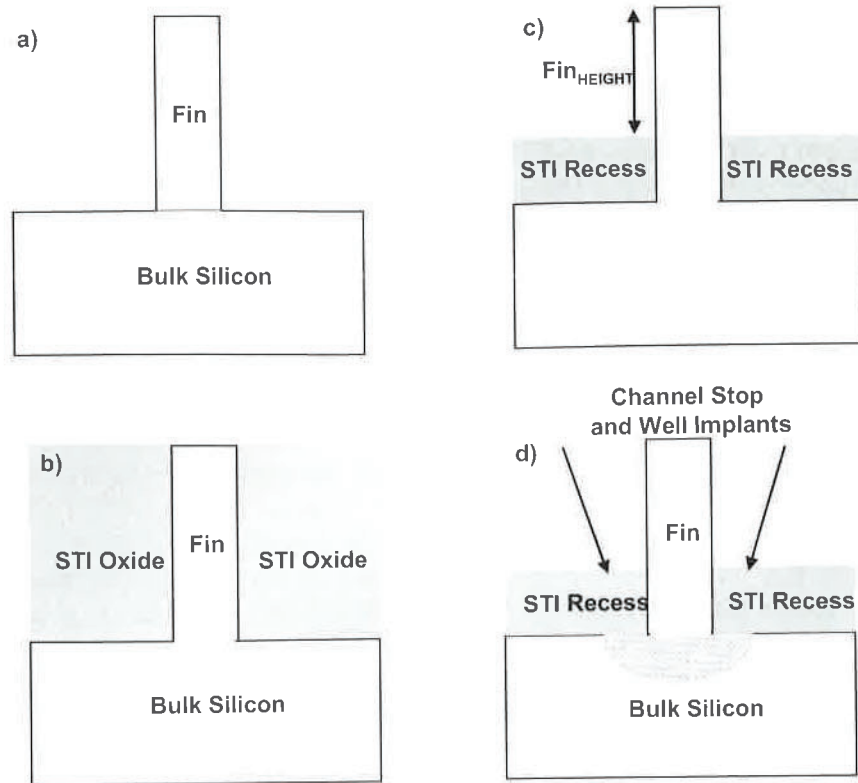


Fig. 2.14. General process flow for making a FinFET on a bulk silicon bulk wafer. a) Fin etched in bulk silicon, the initial Fin_{HEIGHT} is greater than final device Fin_{HEIGHT} to allow some STI (Shallow trench Isolation) oxide. b) STI oxide deposition and planarization. c) STI oxide recess to expose silicon fin with desired Fin_{HEIGHT} . d) Channel stop and Well implants. Channel stop is needed to prevent the source and drain punch through below the fin in the bulk silicon.

2.2.6 Nano-wires and Self-Assembled Wires

The ultimate MOSFET structure is a cylindrical wire-like gate all-around device, as show in Figure 2.15. This structure provides the best electrostatic control of the channel by the gate [54]. The natural length, λ , for such a device is given by Equation 2.16.

$$\lambda = \sqrt{\frac{2\varepsilon_{si}t_{si}^2 \ln(1 + \frac{2t_{ox}}{t_{si}}) + \varepsilon_{ox}t_{si}^2}{16\varepsilon_{ox}}} \quad (2.16)$$

where t_{si} is the diameter of the wire. To keep short-channel effects under control, t_{si} shrinks with gate length. When t_{si} is below 10nm, the device structure is typically called a “nanowire MOSFET”.

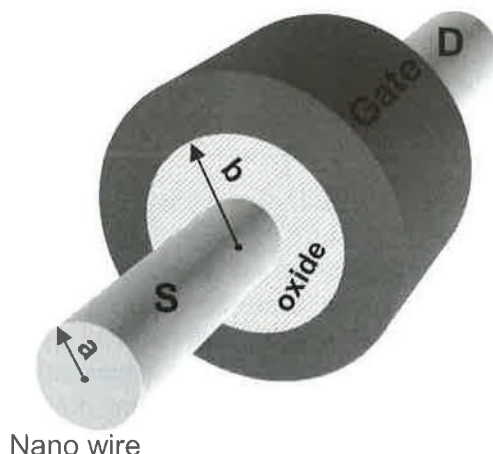


Fig. 2.15. Illustration of a silicon nanowire transistor. The gate stack is wrapped around the semiconductor nanowire. a is the radius of the nanowire.

Because the nanowire transistor has a cylindrical cross section, its capacitance can no longer be calculated using a simple parallel-plate model. Considering Figure 2.15, the inversion charge in the nanowire can be approximated by a charge sheet on the outer surface of the wire. The charge on the gate is on the inner wall of the gate electrode. The capacitance can be obtained by calculating the voltage drop across the gate oxide. Applying Gauss’s law to an infinitely long cylinder one can find the electrical field outside the cylindrical wire:

$$E = \frac{Q}{2\pi\epsilon_{ox}r} \quad (2.17)$$

Where Q is the charge per unit length at the surface of the wire, and r is the radius of the wire. The voltage drop across the gate oxide can be found by integrating the electrical field along a radial line:

$$\Delta V = \frac{Q}{2\pi\epsilon_0 r} \int_a^b \frac{1}{r} dr = \frac{Q}{2\pi\epsilon_{ox} r} \ln\left(\frac{b}{a}\right) \quad (2.18)$$

where a is the nanowire radius, and $b = a + t_{ox}$. The capacitance per unit gate length is:

$$\frac{C}{L_g} = \frac{Q}{\Delta V} = \frac{2\pi\epsilon_{ox}}{\ln\left(\frac{b}{a}\right)} \quad (2.19)$$

Converting Equation 2.19 to a capacitance per unit area of the nanowire surface, we obtain:

$$\frac{C}{A} = \frac{2\pi\epsilon_{ox}}{\pi a \ln\left(\frac{b}{a}\right)} = \frac{2\epsilon_{ox}}{a \ln\left(\frac{b}{a}\right)} \quad (2.20)$$

Equation 2.20 indicates the effective gate capacitance of the nanowire transistor is a function of the nanowire radius as well as the gate oxide thickness. This relationship can be exploited to reduce the effective gate oxide thickness without using higher k dielectric. As an example, a nanowire MOSFET with $a=4\text{nm}$, $t_{ox} = 2\text{nm}$, the effective gate oxide thickness is 1.62nm. The above calculation does not take bulk inversion or quantum-induced volume inversion effects into account. In those cases, the effective nanowire radius is decreased, since the charges are not limited to the nanowire surface.

Silicon nanowires can be made with conventional patterning described in section 2.2.2. A circular cross section can be made by first removing the buried oxide below the silicon, and then rounding of the wires to an approximately circular cross section can be achieved using H_2 annealing.[29]

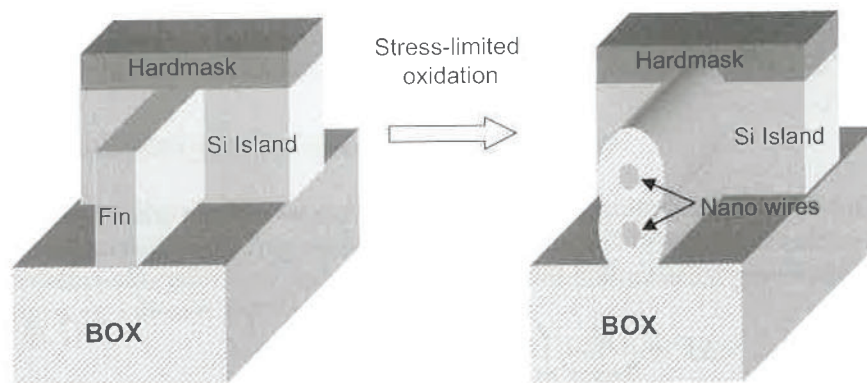


Fig. 2.16: Twin Si nanowires formed through stress-limited oxidation. A hard mask is needed to keep the Si island region from oxidizing. The presence of Si island at the end of the wires is needed to anchor the wires.

(2.19)

nanowire

(2.20)

e nanowire
gate oxide
fective gate
example, a
gate oxide
inversion or
those cases,
ges are not

g described
st removing
wires to an
using H_2



dation. A hard
presence of Si

A twin-wire structure has been made using stress-limited oxidation of tall fins. The stress-limited oxidation oxidizes the wire slower than planar silicon, which produces a final structure comprising two Si wires, one on the top of the fin, and the one other at the bottom of the fin (Figure 2.16). It is important to protect the source and drain region during the oxidation. After removing the oxide, two silicon wires are formed.[55] Twin nanowires can also be formed by oxidation using doping-dependent oxidation rate effects.[126]

The formation of high-density, self-assembled nanowires has been demonstrated using cylindrical-phase diblock copolymers.[56] The polymer has the ability to divide a lithographic pattern (W) into integer number of sub-unit with a distance, l , between cylinders. l is intrinsic length of the polymer which determine the number of sub-units (n) fits in W . The relationship between n and l is defined by Equation 2.21.

$$n = \frac{2}{\sqrt{3}} \left(\frac{W}{l} \right) \quad (2.21)$$

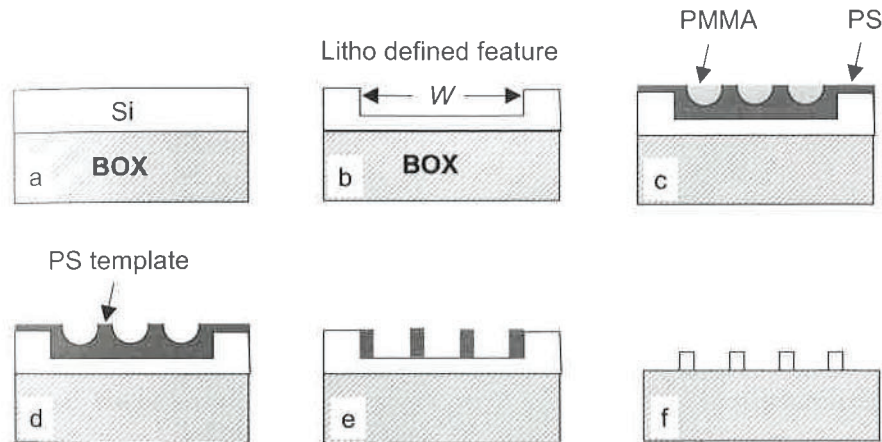


Fig. 2.17: Self-assembled nanowire formation. The initial lithography-defined pattern on an SOI wafer (a,b) is divided into self-aligned patterns by cylindrical-phase diblock copolymers (c). d) PS template formation for plasma etching (e). The remaining PS features are used as mask to etch the silicon nanowires (f).

The self-assembly process is illustrated in Figure 2.17. The sidewalls of the lithography-defined trench initiate the assembly of the polymer domain. In Reference 56, diblock copolymers composed of 70% polystyrene (PS) and 30% poly (methyl-methacrylate) (PMMA) (70:30 PS:PMMA) with total molecular weight $M_n=64$ kg/mol are shown to form

hexagonal lattices of 20 nm diameter cylindrical PMMA domains ($l = 35$ nm) in a matrix of PS. After film assembly, PMMA is removed from the half-cylinder domains by immersing the substrate in acetic acid. Plasma etch is then used to define the Si nanowires.

2.3 Gate Stack

2.3.1 Gate Patterning

This section presents the process steps used in the gate formation of a multi-gate MOSFET. The topography created by the fins presents many challenges. One of these challenges is illustrated in Figure 2.18: photoresist has a surface planarization effect when deposited on a non-planar topography. The wavy gate material stepping over the fins results in different photoresist thickness along the width of the transistor. The photoresist critical dimension (CD) after exposure is a function of the resist thickness. This effect is known as “swing curves”. Without gate planarization, gate CD will vary periodically through the entire width of the device, with a period equal to the fin pitch (Figure 2.18). Therefore, the use of a planarized gate is highly desirable.

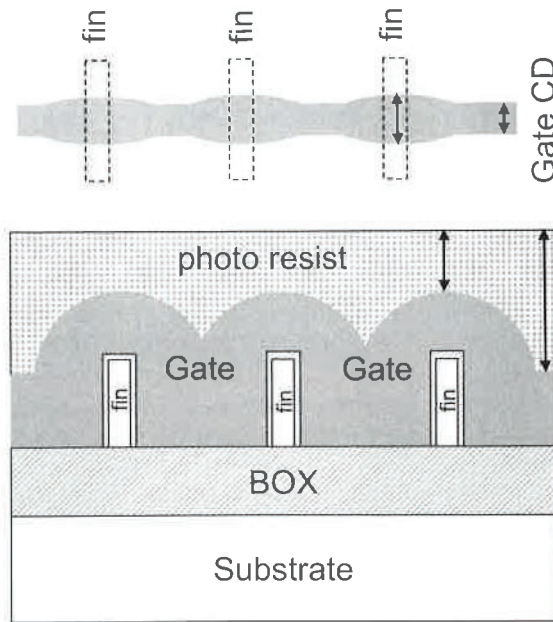


Fig. 2.18. Patterning gate on an irregular topography results in wavy gate CD due to the variation in resist thickness. A planarization step is required.