Simulation Study of Scaling Design, Performance Characterization, Statistical Variability and Reliability of decananometer MOSFETs

Xingsheng Wang

Submitted to University of Glasgow,
Department of Electronics and Electrical Engineering,
in fulfilment of requirements for the degree of
Doctor of Philosophy

May 2010

Dedicated to

my parents for their endless support…

Abstract

This thesis describes a comprehensive, simulation based scaling study – including device design, performance characterization, and the impact of statistical variability – on deca-nanometer bulk MOSFETs. After careful calibration of fabrication processes and electrical characteristics for n- and p-MOSFETs with 35 nm physical gate length, 1 nm EOT and stress engineering, the simulated devices closely match the performance of contemporary 45 nm CMOS technologies. Scaling to 25 nm, 18 nm and 13 nm gate length n and p devices follows generalized scaling rules, augmented by physically realistic constraints and the introduction of high-k/metal-gate stacks. The scaled devices attain the performance stipulated by the ITRS. Device a.c. performance is analyzed, at device and circuit level. Extrinsic parasitics become critical to nano-CMOS device performance. The thesis describes device capacitance components, analyzes the CMOS inverter, and obtains new insights into the inverter propagation delay in nano-CMOS. The projection of a.c. performance of scaled devices is obtained.

The statistical variability of electrical characteristics, due to intrinsic parameter fluctuation sources, in contemporary and scaled decananometer MOSFETs is systematically investigated for the first time. The statistical variability sources: random discrete dopants, gate line edge roughness and poly-silicon granularity are simulated, in combination, in an ensemble of microscopically different devices. An increasing trend in the standard deviation of the threshold voltage as a function of scaling is observed. The introduction of high-k/metal gates improves electrostatic integrity and slows this trend. Statistical evaluations of variability in I_{on} and I_{off} as a function of scaling are also performed.

For the first time, the impact of strain on statistical variability is studied. Gate line edge roughness results in areas of local channel shortening, accompanied by locally increased strain, both effects increasing the local current. Variations are observed in both the drive current, and in the drive current enhancement normally expected from the application of strain. In addition, the effects of shallow trench isolation (STI) on MOSFET performance and on its statistical variability are investigated for the first time. The inverse-narrow-width effect of STI enhances the current density adjacent to it. This leads to a local enhancement of the influence of junction shapes adjacent to the STI. There is also a statistical impact on the threshold voltage due to random STI induced traps at the silicon/oxide interface.

Acknowledgment

First of all, I would like to express my most sincere gratitude to my supervisors Professor Asen Asenov and Dr. Scott Roy. They guided my PhD researches through more than 3 years. Professor Asenov inspired me with his knowledge, enthusiasm and wisdom in semiconductors, and he gave me countless suggestions and advice on the research fields, and also answered those academic questions once obsessed in my mind. He devoted a lot of time and patience to the reading and correction of this thesis.

I am grateful to Dr. Scott Roy for his valuable comments, advice, encouragement and thesis reading. I thank him for his help to solve those problems when I appeared as a fresh device modeller. His pleasant and optimistic appearance greatly impresses me.

I would also like to thank those Doctors Binjie Cheng, Gareth Roy, and Stanislav Markov for their helpful thesis reading and suggestions. My sincere thanks go to Binjie Cheng for his discussions and suggestions about my research, and also for plenty of personal suggestions from the beginning. Gareth taught me the basics of Linux. Stanislav and I had fruitful discussions. I feel grateful to Andrew R. Brown for his countless patient answers about the ‘atomistic’ simulator and others. Thanks also go to Dr. Campbell Millar for helpful discussions about programming.

I appreciate the discussions with Dr. Karol Kalna about whatever has been involved, and Dr. Jeremy Watling about high-k gate stack. I thank Dr. Antonio Martinez for discussions reminding me of theoretical studies in physics.

This work is supported partitially by the EPSRC project (EP/E003125/1) “Meeting the design challenges of the nano-CMOS electronics,” and partially by the Overseas Research Students Awards Scheme (ORSAS). I acknowledge their support for my PhD studies.

Last but not least, I want to thank my parents. They teach me the life principles. My father wished me to be a scientist expert. Their understanding encourages me, and their sacrificial giving and support drive me.

Publications

X. Wang, S. Roy, A. R. Brown, and A. Asenov, “Impact of STI on statistical variability and reliability of decananometre MOSFETs,” submitted to *IEEE Electron Device Letters.*

B. Cheng, D. Dideban, N. Moezi, C. Millar, G. Roy, **X. Wang**, S. Roy and A. Asenov, “Statistical-variability compact-modeling strategies for BSIM4 and PSP,” *IEEE Design & Test of Computers*, Vol. 27 No. 2, pp.26-35, March/April 2010.

B. Bindu, B. Cheng, G. Roy, **X. Wang**, S. Roy, A. Asenov, “Parameter set and data sampling strategy for accurate yet efficient statistical MOSFET compact model extraction,” *Solid-State Electronics*, Vol.54 No.3, pp.307-315, March 2010.

B. Benbakhti, K. Kalna, **X. Wang**, B. Cheng, A. Asenov, G. Hellings, G. Eneman, K. D. Meyer and M. Meuris, “Impact of raised source/drain in the $In_{53}Ga_{47}As$ channel implant-free quantum-well transistor,” in *Proc. 11th ULIS*, pp.129-132, March 18-19, 2010.

N. A. Kamsani, B. Cheng, C. Millar, N. Moezi, **X. Wang**, S. Roy and A. Asenov, “Impact of slew rate definition on the accuracy of nanoCMOS inverter timing simulations,” in *Proc. 11th ULIS*, pp.53-56, March 18-19, 2010.

B. Cheng, D. Dideban, N. Moezi, C. Millar, G. Roy, **X. Wang**, S. Roy and A. Asenov, “Capturing intrinsic parameter fluctuations using the PSP compact model,” in *Proc. Design, Automation and Test in Europe*, pp.650-653, Germany, March 8-12, 2010.

X. Wang, S. Roy, and A. Asenov, "Impact of strain on the performance of high-k/metal replacement gate MOSFETs," in *Proc. IEEE 10th Ultimate Integration on Silicon (ULIS 2009)*, pp.289-292, Aachen Germany, March 18-20, 2009.

A. Asenov, S. Roy, A. R. Brown, G. Roy, C. Alexander, C. Riddet, C. Millar, B. Cheng, A. Martinez, N. Seoane, D. Reid, M. F. Bukhori, **X. Wang**, U. Kovac, "Advanced simulation of statistical variability and reliability in nano CMOS transistors," in *IEDM Tech. Dig.*, pp. 421, December 2008.

X. Wang, S. Roy, and A. Asenov, "High performance MOSFET scaling study from bulk 45 nm technology generation," in *Proc. of IEEE 9th International Conference on Solid-State and Integrated-Circuit Technology (ICSICT 2008)*, pp. 484-487, Beijing China, October 20-23, 2008.

B. Bindu, B. Cheng, G. Roy, **X. Wang**, S. Roy, and A. Asenov, "An efficient data sampling strategy for statistical parameter extraction of nano-MOSFETs," in *IEEE Workshop on Compact Modeling*, pp.55-59, Japan, September 8, 2008.

X. Wang, S. Roy, and A. Asenov, "Impact of strain on LER variability in bulk MOSFETs," in *Proc. of IEEE 38th European Solid-State Device Research Conference (ESSDERC 2008)*, pp.190-193, Edinburgh Scotland, September 15-19, 2008.

X. Wang, B. Cheng, S. Roy, and A. Asenov, "Simulation of strain enhanced variability in nMOSFETs," in *Proc. IEEE 9th Ultimate Integration on Silicon (ULIS 2008)*, pp.89-93, Udine Italy, March 12-14, 2008.

Contents

List of tables

List of figures

List of symbols

Symbols	Descriptions	Units
α	Scaling factor for electric field	-
β	Gain factor of a MOS transistor	A/V^2
Δ	Root mean square (rms) for LER	nm
ΔE_{strain}	Energy band split due to strain	eV
Δz	Distance of inversion carrier centroid away the interface	nm
ε	Permittivity	F/cm
ε_{highK}	High-k material permittivity	F/cm
ε_{ox}	Silicon dioxide permittivity	F/cm
ε_{si}	Silicon permittivity	F/cm
$\bar{\varepsilon}$, ε_{ij}	Strain tensor, strain component	-
ϕ	Potential	Volt (V)
ϕ_i	Potential defined by $\phi_i = -E_i/q$	Volt (V)
Φ	Quasi-Fermi potential	Volt (V)
φ_m	Metal work function	Volt (V)
κ	Scaling factor	-
Λ	Correlation length for LER	nm
λ	Wave length	nm
μ	Mobility	cm^2/Vs
μ_n	Electron mobility	cm^2/Vs
μ_p	Hole mobility	cm^2/Vs
ρ	Charge density	C/cm^3
ρ_{trap}	Trap charge density	C/cm^3
$\bar{\sigma}$, σ_{ij}	Stress tensor, stress component	N/cm^2
τ	Inverter propagation delay	s
τ_{int}	Intrinsic inverter delay	s
ω	Angular frequency ($\omega = 2\pi f$)	radians/s
ψ_B	Difference between intrinsic and extrinsic Fermi leverls	Volt (V)
A	Area	cm^2
C	Capacitance	F
C_{gc}	Gate-to-channel capacitance (per unit width)	F (F/μm)
C_{gg}	Total gate capacitance (per unit width)	F (F/μm)
C_j	P-N junction capacitance (per unit width)	F (F/μm)
C_{ox}	Gate oxide capacitance per unit area	F/cm^2
C_1, C_2, C_3	Fitting coefficients	-
D, D_n, D_p	Diffusion coefficient, electron and hole diffusion coefficient	cm^2/s
E	Electric field	V/cm
E_a	Activation energy in NBTI	eV
E_C, E_V	Conduction, valence band edge energy	eV
E_F	Fermi level	eV
E_i	Intrinsic Fermi level	eV

f	Circuit frequency	Hz
f_T	Cut-off frequency	Hz
G, G_{ij}	Small-signal conductance matrix, small-signal conductance	siemens
g_m	Small-signal transconductance	siemens
h	Planck's constant (6.626×10^{-34})	Js
$\hbar$	Reduced planck's constant ($\hbar=h/2\pi$)	Js
I_d	Drain current	A
I_{dsat}	Drain saturation current	A
I_{eff}	Effective drive current	A
I_{off}, I_{on}	Off-current (per unit width), on-current (per unit width)	A (A/μm)
J, J_n, J_p	Current density, electron and hole current densities	A/cm^2
k	Boltzmann's constant (1.38×10^{-23})	J/K
l	Mean free path	nm
L	Channel length	nm
L_{gate}	Physical gate length	nm
L_{met}	Metallurgical channel length	nm
n	Electron density	cm^{-3}
n_i	Intrinsic electron density	cm^{-3}
m	MOSFET body-effect coefficient	-
m_l	Carrier longitudinal mass	kg
m_t	Carrier transverse mass	kg
N_a	Acceptor doping concentration	cm^{-3}
N_a^-	Ionized acceptor doping concentration	cm^{-3}
N_C	Effective density of states of conduction band	cm^{-3}
N_d	Donor doping concentration	cm^{-3}
N_d^+	Ionized donor doping concentration	cm^{-3}
N_p	Poly-silicon doping concentration	cm^{-3}
N_{sub}	Substrate doping concentration	cm^{-3}
N_t	Trap sheet density	cm^{-2}
p	Hole desntiy	cm^{-3}
P	Power dissipation	W
P_{active}	Active power dissipation	W
$P_{passive}$	Standby power dissipation	W
P_d	Possibility to travel d without suffering from collisions	-
P_{E_b}	Possibility to obtain energy E_b (eV)	-
q	Electronic charge (1.6×10^{-19})	C
Q_{dm}	Maximum depletion-layer charge per unit area	C/cm^2
Q_{inv}	Inversion charge per unit area	C/cm^2
R_{net}	Net recombination rate	1/s
S	Sub-threshold slope	mV/dec
t	Time	s
t_{highK}	High-k material thickness	nm
t_{ox}	Silicon oxide thickness	nm
T	Temperature	Kelvin (K)

v	Carrier velocity	cm/s
v_{sat}	Carrier saturation velocity	cm/s
V	Voltage	Volt (V)
V_d	Drain voltage	Volt (V)
V_{dd}	Supply voltage	Volt (V)
V_{fb}	Flat-band voltage	Volt (V)
V_g	Gate voltage	Volt (V)
V_{in}	Input node voltage of a logic gate	Volt (V)
V_{out}	Output node voltage of a logic gate	Volt (V)
$V_{overshoot}$	Inverter output overshoot voltage during input transition	Volt (V)
V_{th}	Threshold voltage	Volt (V)
$V_{th,sat}$	Saturation threshold voltage	Volt (V)
W	Channel width	μm
W_d	Depletion-layer width	nm
W_{dm}	Maximum depletion-layer width	nm
x_j	Junction depth	nm
Y, Y_{ij}	Admittance matrix, admittance component	siemens

Chapter I

1. Introduction

1.1. Motivation

The progressive down-scaling of bulk metal-oxide-semiconductor field effect transistors (MOSFETs) has been the driving force behind the integrated circuit (IC) industry for several decades, continuously delivering higher component densities and greater chip functionality, while reducing the cost per function from one CMOS technology generation to the next. Moore's law boosts IC industry profits by constantly releasing high-quality and inexpensive electronic applications into the market using new technologies. From the 1 μm gate lengths of the eighties to the 35 nm gate lengths of contemporary 45 nm technology, the industry successfully achieved its scaling goals, not only miniaturizing devices but also improving device performance.

However, the years of 'happy scaling' are over. Several challenges are facing the further miniaturization of the transistors in Si chips. First, there are the process challenges of continued scaling including, among others, sub-wavelength patterning and the formation of ultra-shallow junctions. Continual process innovation is needed, including immersion optical lithography, and the flash lamp and laser annealing implemented at the 45 nm CMOS technology to achieve high-fidelity patterns and high-activation/low dopant diffusion. Secondly, the scaling of contemporary deep-decananometer gate length transistors results in the deterioration of electrical characteristics due to short-channel, quantum mechanical and transport effects. Countermeasures like halo-doping can suppress the short-channel effect in conventional MOSFETs but at the expense of increased quantum effects and reduced mobility. Novel MOSFET structures such as fully-depleted silicon-on-insulator devices are required to achieve a fundamental improvement in electrostatic integrity. Finally, intrinsic parameter fluctuations, resulting from the discreteness of charge and the granular nature of matter in real decanano devices, hamper the integration of scaled transistors. Statistical variability unavoidably increases drastically

with further scaling. It requires the introduction of novel device architectures and the adaptation of smarter circuit and system design strategies.

It is natural to ask the questions how far the scaling of the conventional bulk MOSFET can continue in the presence of the above technological or physical limitations. This project tries to shine some light on these questions through aggressively scaling contemporary bulk MOSFETs using advanced commercial TCAD tools, and simulating their statistical variability using the Glasgow 'atomistic' simulator.

1.2. Aims and objectives

The aim of this PhD is to study the realistic scaling of bulk MOSFETs and their statistical variability. In order to accomplish this aim, several key research objectives have been identified:

- To study the physical effects governing the operation of aggressively scaled bulk MOSFETs, and the associated challenges for contemporary CMOS technology;
- To master TCAD process and device simulations, and 'atomistic' simulation techniques;
- To calibrate the simulations in respect of state-of-the-art 45 nm technology n- and p-MOSFETs, and to perform careful scaling to smaller technologies according to the prescriptions and the requirements of the ITRS;
- To perform a statistical variability study on these scaled bulk MOSFETs;
- To investigate the impact of realistic structures such as stressors and shallow trench isolation on the statistical variability and reliability;
- To evaluate the performance of scaled devices, including their impact on timing and speed of simple circuits.

1.3. Thesis outline

This PhD thesis includes eight chapters, formulated in a general structure of introduction-main body-conclusion.

Chapter 1 describes the motivation and the aim and objectives of this PhD study.

Chapter 2 starts with the bulk MOSFET scaling rules and the projections of the ITRS. It then describes several scaling challenges including: feature patterning, power control, poly

depletion effects, quantum effects, short-channel effects, reliability and variability. Finally it explains technology boosters, such as induced strain and high-k/metal gates, that enable further device scaling.

Chapter 3 focuses on the simulation tools and methodology used in this study. It selectively describes the simulation of key process steps, the physics and numerical methods behind the device simulation, and finally the ‘atomistic’ simulation techniques.

Chapter 4 presents the scaling study based on bulk 45 nm technology MOSFETs. It starts with the calibrations of both the process and the device simulators against the doping and the characteristics of published Toshiba MOSFETs. Then the devices are redesigned adopting the latest technology features including stress engineering to achieve the equivalent performance to the contemporary 45 nm CMOS technology. The careful scaling informed by ITRS proceeds to 25 nm, 18 nm, 13 nm physical gate lengths.

Chapter 5 presents the predictive study of statistical variability of the scaled MOSFETs. The statistical distribution of threshold voltage, on-current, and off-current are obtained for 35 nm, 25 nm and 18 nm gate length MOSFETs. The expected increase in the statistical variability with scaling is observed. This includes the detailed statistical distributions of these key parameters.

Chapter 6 investigates the impact of strain and shallow trench isolation on the statistical variability. The strain variability in channel due to gate line edge roughness is studied. The stress enhances the drive current variation. The shallow trench isolation (STI) enhances the current density near the isolation edge, inducing a threshold voltage lowering in narrow transistors. The impact of STI-adjacent junction shapes on the random dopant induced threshold voltage fluctuations is also investigated.

Chapter 7 presents a systematic study of the a.c. and dynamic performance of the scaled devices. It analyzes the MOSFET capacitances using the split C-V approach. Mixed-mode inverter simulation is performed to analyze the intrinsic and extrinsic inverter delay.

Chapter 8 draws conclusions and proposes future avenues for related research.

Chapter II

2. MOSFET technology scaling, challenges and boosters

The advance of the IC industry has been driven by the continuous down-scaling of MOSFETs for several decades. A consistent evolution of MOSFET technologies has been essential in overcoming many perceived ‘insurmountable’ barriers. Recently, novel materials and processes have been introduced, and new device structures proposed, to ensure further benefits from device miniaturization.

2.1. MOSFET scaling

2.1.1. Past scaling trends

The scaling of the MOSFET devices has continued from its first introduction in integrated circuits four decades ago. This has resulted in doubling the component density on a single chip by proportionally scaling of the transistor dimensions over a period of time. This reduces the cost per function and delivers more functions at the same time, which is the essence of the famous *Moore’s law* [1]. At the same time the scaling leads to improved performance while controlling the power consumption by reducing the supply voltage and carefully tuning the design. Schematically shown in Figure 2.1, a MOSFET consists of two back-to-back connected *p-n* junctions. The gate voltage applied across metal-oxide-semiconductor (MOS) capacitor creates an inversion channel connecting the source and the drain, and controls the carrier density in it. From an operational point of view, the MOSFET has two critical structural parameters, namely gate length and gate dielectric thickness. MOSFET scaling affects both lateral and vertical device dimensions. While the reduction of the lateral dimensions increases the transistor density in a chip, the reduction of the oxide thickness is needed to ensure good electrostatic integrity.

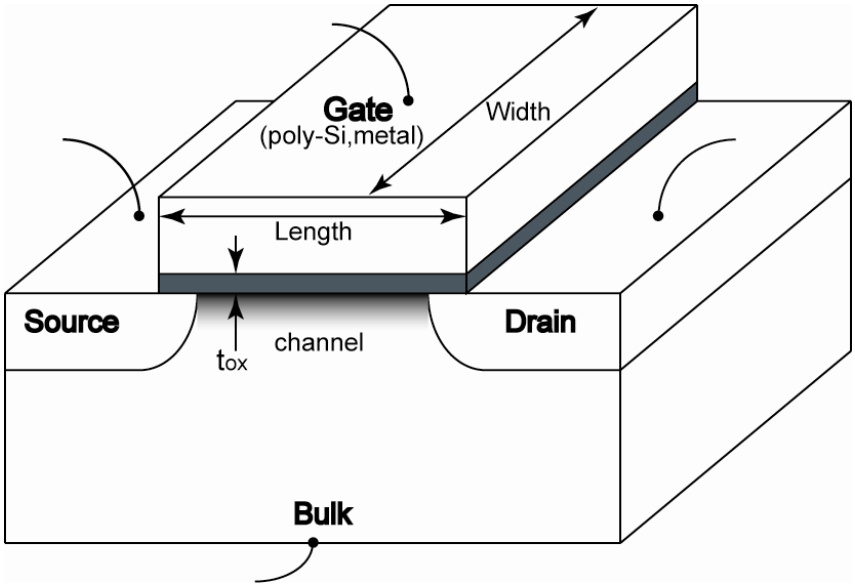


Figure 2.1 Schematic view of a surface-channel MOSFET device indicating physical gate length, channel width and physical gate dielectric oxide thickness (t_{ox}).

2.1.1.1. Transistor dimensions

With the scaling of the lateral dimensions by a factor of $1/\sqrt{2}$ each technology generation, the bulk MOSFET area is reduced by one half, doubling the circuit density and enhancing functionality. Since the publication of the famous paper on scaling by Dennard *et al.* in 1974 [2], the gate length of the transistor in modern chips has been reduced by more than two orders featuring 35 nm gate length MOSFETs in the 45 nm technology generation [3]. The equivalent oxide thickness has been reduced from 100 nm to around 1.0 nm in contemporary technology. In addition, the number of transistors in a microprocessor chip, has increased 100,000 times from the first microprocessor Intel 4004 (about 2,300 transistors) released in 1971, by transistor scaling and additional efforts of compacting the physical layout [4][5].

In contemporary MOSFETs the drain current I_d is determined by [6]

$$I_d / W = C_{ox}(V_g - V_{th})v, \tag{2.1}$$

where W is the channel width, C_{ox} is the gate capacitance per unit area, and v is the source end carrier velocity. The saturation transconductance g_m may be obtained by

$$g_m / W = \frac{\partial I_d}{\partial V_g} / W = C_{ox} \times v = \frac{\varepsilon_{ox}}{t_{ox}} \times v. \tag{2.2}$$

where ε_{ox} is the oxide permittivity. The carrier velocity is usually saturated at short-channel MOSFETs, thus g_m/W is an index of gate oxide thickness t_{ox}. Since gate capacitance per unit area is inversely proportional to oxide thickness, both the device current and the

saturation transconductance are closely related to oxide thickness. As demonstrated in Figure 2.2 for several past technology generations [7][8][9][10], saturation transconductance has been increasing linearly with the reduction of gate length. This indicates that in the past the oxide thickness has been scaled at an identical pace with gate length.

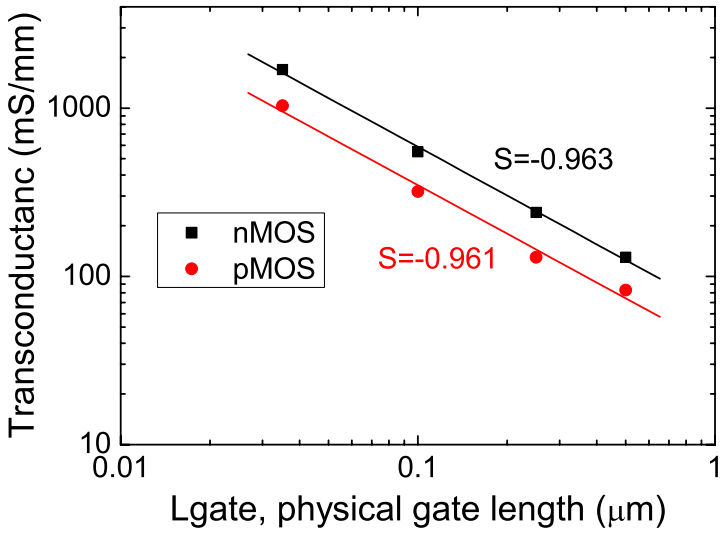


Figure 2.2 Saturation transconductance increases with reduction of gate length, indicating the uniform decreasing of gate oxide thickness

2.1.1.2. Supply and threshold voltage scaling

As illustrated in Figure 2.3 the supply voltage has been continuously reduced with the advance of CMOS technology.

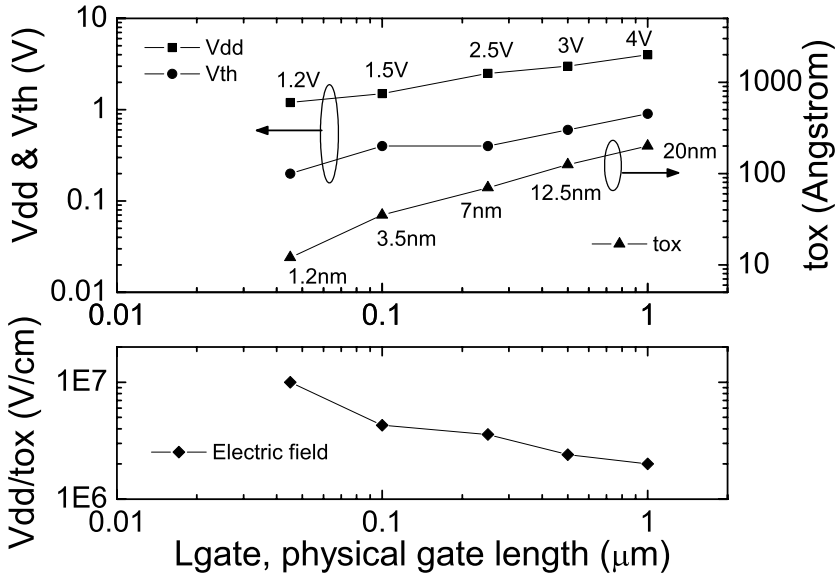


Figure 2.3 Supply voltage (V_{dd}), threshold voltage (V_{th}) and gate oxide thickness were scaled down with gate length in the past. Meanwhile the vertical gate electric field increased rapidly.

In addition, the threshold voltage has also been reduced to maintain appropriate gate voltage overdrive according to equation (2.1). However, the reduction of supply voltage and threshold voltage has lagged the gate length and gate dielectric thickness scaling. The slower reduction of the threshold voltage is due to a combination of factors including increasing channel doping needed to control short-channel effects. The sub-linear threshold voltage scaling in turn has retarded the scaling of the supply voltage.

2.1.1.3. Doping profiles

The MOSFET substrate doping, and especially the channel doping concentration, have been continuously increasing since the beginning of scaling. From the original doping density of approximately 2.5×10^{16} cm^{-3} in 1 μm gate length transistors [2], it has already reached more than 2×10^{18} cm^{-3} in contemporary 35 nm gate length MOSFETs [11]. In addition to the average increase of the channel doping, the design of the vertical doping profile has also changed. Retrograde doping has been introduced to short channel devices to control short-channel effects and to achieve simultaneously low threshold voltage and elevate mobility by reducing impurity scattering [11]. Halo doping implanted using a small tilt angle effectively blocks lateral field penetration from source/drain, reducing further short-channel effects without affecting adversely the threshold voltage [12]. Figure 2.4 shows the retrograde doping and halo doping.

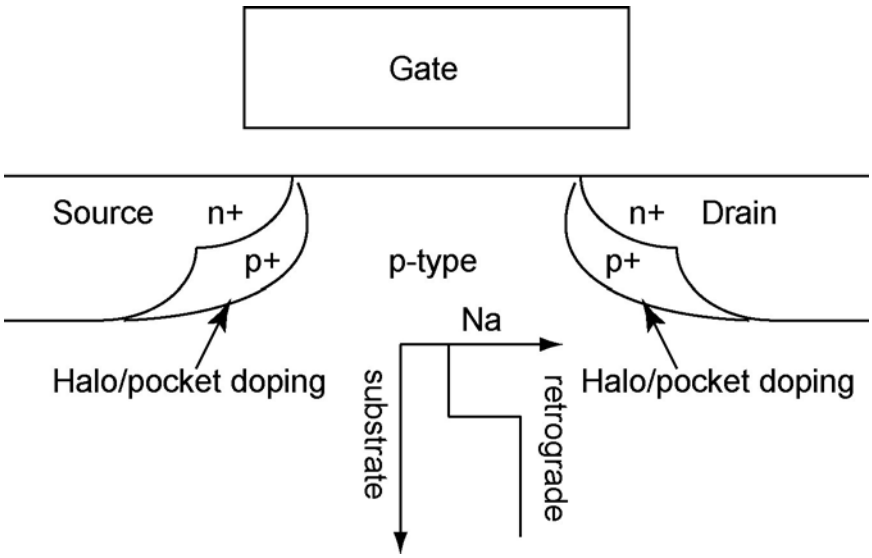


Figure 2.4 Schematic view of the doping profile in an n-channel MOSFET, including retrograde doping along substrate depth and halo/pocket doping close to source/drain.

In conjunction with the changes in the vertical doping design, structure modifications and changes have been made to the design of the source/drain regions. Lightly doped drain (LDD) was introduced to alleviate reliability concerns during the 5V supply voltage period. High field related impact ionization near the drain junction caused serious reliability

concerns at this stage [13]. The n^- regions introduced near n^+ source/drain diffusion regions by additional implantation reduces the high field into n^- region near drain [14] at the expense of additional series resistance [15]. Finally, shallow source/drain extensions (SDE) have been introduced to manage short-channel effects in successive technologies reducing the coupling between the drain voltage and the carrier concentration in the channel. Meanwhile the need of low sheet resistance of SDE requires high doping activation without significant additional diffusion.

At 0.5~0.25 micron technology in 1990s [16], the integration requirements for abrupt transition between the active transistor area and the insulation and better planarity prompted the replacement of the LOCal Oxide of Silicon (LOCOS) horizontal isolation with the vertical Shallow Trench Isolation (STI). Compared with LOCOS's shortfalls, such as lateral extension of the bird's beak, boron encroachment and oxide thinning, STI achieved high scalability and planarity [17][18]. Inverse-narrow-width effect has created some problems in STI MOSFETs due to enhanced edge conduction by fringing field at the edge of the STI, leading to the threshold voltage lowering with the decrease of MOSFET width [19]. However, improvements, such as sidewall implantation and corner rounding, were adopted to suppress parasitic conduction effect at the corner [20][21].

2.1.2.Scaling principles

2.1.2.1. Constant-field scaling

CMOS technology emerged in 1960's and started to replace bipolar devices in integrated circuits logic and memory applications, where chip density and cost are primary concerns. The applications of high resolution lithographic capability, and the introduction of controllable doping profiling using ion implantation [2] paved the way of device miniaturization achieving the reduction of cost/performance towards large scale integration in early 1970's [2][22][23][24]. At the origin of these efforts a seminal paper by Dennard *et al.* established the rules of so-called constant-field scaling [2].

The essence of constant-field scaling is to maintain a constant electric field while simultaneously reducing dimensions and supply voltage in concert by the same scaling factor κ (>1). This requires an appropriate increase of substrate doping concentration to scale the depletion layer width. This is schematically illustrated in Figure 2.5. This is based on the Poisson's equation (2.3)

$$\nabla_r \cdot (-\varepsilon \nabla_r \phi) = \rho = q(p - n + N_d^+ - N_a^-), \qquad (2.3)$$

where N_d^+, N_a^- are ionized donor and acceptor concentrations respectively, ρ is charge density and ε is material permittivity. In sub-threshold regime where charge density is mainly determined by ionized impurity concentration and carrier density contributions in the channel are negligible, the current continuity equation can be decoupled from the Poisson equation. If the scaled dimensions is $\boldsymbol{r}'=\boldsymbol{r}/\kappa$, we apply Poisson's equation to the scaled device,

$$\nabla_{r'} \cdot (-\varepsilon \nabla_{r'} \phi') = \rho' . \tag{2.4}$$

In both equations ∇ is the gradient operator on dimensional position $\boldsymbol{r}$ and $\boldsymbol{r}'$, ϕ and ϕ' is electrostatic potential, ρ and ρ' is space charge density for sub-threshold region respectively for original and scaled devices. The maintenance of constant field requires $-\nabla_{r'} \phi' = -\nabla_r \phi$. This results in $\phi' = \phi/\kappa$ and $\rho' = \kappa\rho$ in the light of $\nabla_{r'} = \kappa \nabla_r$. This could be achieved by scaling down supply voltage V_{dd} by the factor κ, and by increasing substrate doping N_{sub} (N_a for p-type or N_d for n-type) by the same factor κ.

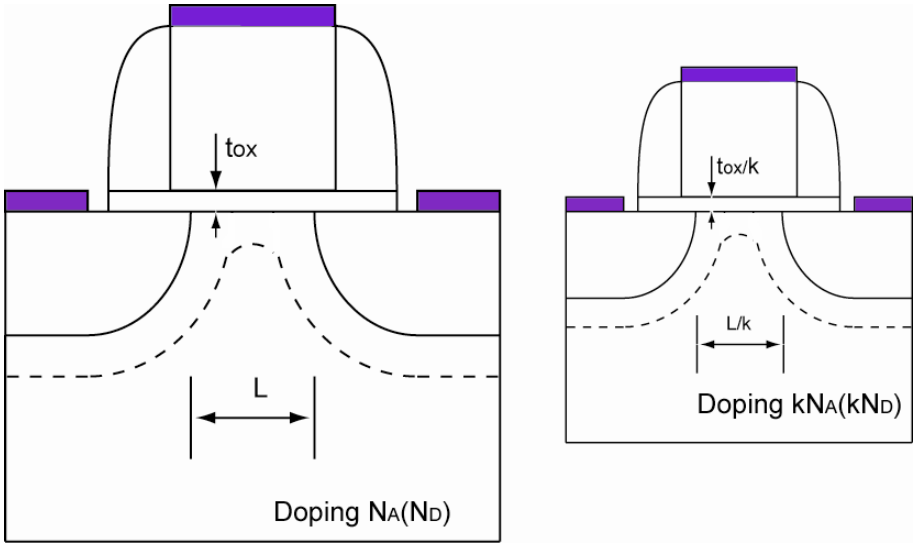


Figure 2.5 Schematic view of constant-field scaling principle shows proportional dimension reduction and doping increase.

The rules of constant-field scaling for other device parameters listed in Table 2.1 can be deduced in a relatively easy way [25]. The current will be reduced by a factor κ according to equation (2.1), therefore the power consumption per circuit will reduced by a factor $1/\kappa^2$. It enables the constant power consumption on a chip.

2.1.2.2. Generalized scaling principle

Constant-field scaling principles had been an elemental strategy in designs of MOSFET devices and circuits, and worked as a successful guide for the design down to 1-μm gate

length MOSFET. However, the difficulties in reduction of threshold voltage and junction built-in potential exposed the limited flexibility of the constant-field scaling scenario in the design of quarter-micron MOSFET technology. As a result Baccarani *et al.* proposed a generalized set of scaling rules in 1984, allowing for further device miniaturization under the above mentioned constraints [26].

The fundamental novelty in the generalized scaling rules is to relax the scaling pace of voltage. Assuming that scaled dimensions of MOSFETs are $\boldsymbol{r}' = \boldsymbol{r}/\kappa$ (κ>1), and introducing an additional scaling factor α >1, the applied potential in the scaled device is $\phi' = (\alpha/\kappa)\phi$. Accordingly the scaled device electric field is $-\nabla_{r'}\phi' = \alpha(-\nabla_r\phi)$. Applying to the Poisson's equations (2.5), in terms of equation (2.3), results in

$$\nabla_{r'} \cdot (-\varepsilon \nabla_{r'} \phi') = \rho' = \alpha\kappa\rho \,. \tag{2.5}$$

It means the channel impurity concentration has to increase by a factor $\alpha\kappa$.

Table 2.1 Scaling principles for MOSFET device and circuit parameters.

Scaled parameters	Constant-field scaling	Generalized scaling	
Dimensions (L, W, t_{ox}, x_j)	$1/\kappa$	$1/\kappa$	
Voltage (V)	$1/\kappa$	α/κ	
Doping concentration (N_a, N_d)	κ	$\alpha\kappa$	
Electric field (E)	1	α	
Depletion-layer width (W_d)	$1/\kappa$	$1/\kappa$	
Capacitance ($C=\varepsilon A/t_{ox}$)	$1/\kappa$	$1/\kappa$	
Inversion charge density (Q_{inv})	1	α	
Carrier velocity (v)	1	α (Long ch.)	1 (Vel. sat.)
Current, drift (I)	$1/\kappa$	α^2/κ	α/κ
Delay time/circuit ($\tau \sim CV/I$)	$1/\kappa$	$1/\alpha\kappa$	$1/\kappa$
Power dissipation/circuit ($P \sim VI$)	$1/\kappa^2$	α^3/κ^2	α^2/κ^2
Power-delay product/circuit ($P\tau$)	$1/\kappa^3$	α^2/κ^3	
Circuit density ($\propto 1/A$)	κ^2	κ^2	
Power density (P/A)	1	α^3	α^2

Generalized scaling principle slows down the scaling pace of supply voltage by a factor α >1, which has been confirmed by the historical data presented in subsection 2.1.1. The substrate doping increase trend is also reflected in reported device scaling scenarios. This increase is more aggressive than dimensional reduction. Scaling parameters in scaled

device are presented in Table 2.1. It should be noticed that the carrier velocity is treated differently in long channel devices, where linear relationship between the velocity and parallel electric field is assumed, and in short channel devices where velocity saturation at the source end of the channel is assumed [25].

Compared to the constant power consumption of constant-field scaling, the power dissipation using generalized scaling significantly increases. While both the current and voltage in short-channel devices increase by a factor α compared with those using constant-field scaling, the power consumption on a chip boosts by a factor α^2 for each generation, which brings integration problems.

2.1.3.Recent scaling trends and ITRS

2.1.3.1. Recent MOSFET technologies

Further advancement of the MOSFET technology becomes complicated, as gate length enters within sub-100 nm nanometre regime at the 130nm technology generation, a commonly accepted critical dimension for nanoelectronics. At such dimensions, performance deterioration issues associated with transport property limitations of Si started to emerge. From the technological point of view imaging and patterning has become extremely challenging. As optical lithography is the ‘work horse’ of the semiconductor sector, sub-wavelength patterning is pushed to adopt resolution enhancement techniques (RET), such as optical proximity correction (OPC), to print finer lines and increase fidelity of pattern transfer [27][28][29]. Due to extreme oxide thickness scaling, the gate oxide direct tunnelling through oxides thinner than 20Å started to dominate the gate leakage [30][31], and poses a challenge for further gate oxide scaling. Simultaneously intrinsic parameter fluctuations due to the discreteness of charge and matter, including random discrete dopants, line edge roughness, and poly-silicon granularity, introduce statistical variability in the electrical characteristics of nanoscale MOSFETs, influencing device performance and circuit yield [32][33][34][35].

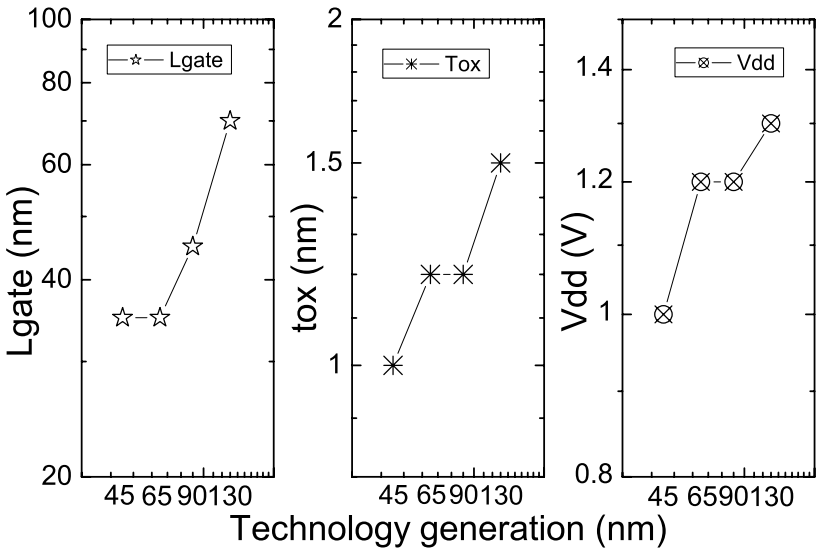


Figure 2.6 Evolutions of gate length, gate oxide thickness (EOT) and supply voltage in recent MOSFET technologies by Intel.

A reflection of these challenges is the reduction in the scaling pace for gate lengths and vertical dimensions in the generations subsequent to the 90 nm node [5][10] as shown in Figure 2.6. The down-scaling of supply voltage and gate oxide in fact stopped somewhere between the 90 nm and the 65 nm technology generations. The scaling of gate length has not happened in the Intel transition from 65 nm to 45 nm technology generation. Notably the oxide thickness (EOT) at the 45 nm technology generation sharply decreased to 1 nm, which was achieved by the introduction of high-k/metal gate stack in order to prevent excessive gate leakage while reducing the EOT. In addition, the supply voltage was decreased to 1V to avoid the severe power issue, although the threshold voltage in these recent technologies has stalled [3][10][36][37][38].

2.1.3.2. ITRS prospect

The latest published 2008 update of International Technology Roadmap for Semiconductors (ITRS) relaxed the pace of introducing new technology generations [39]. A technology generation is defined by the half pitch of Dynamic Random Access Memory (DRAM) or the half pitch of metal 1 (M1) interconnects of Microprocessor unit (MPU)/Application-specific integrated circuit (ASIC) depending on the relevant technology driver. The 2008 ITRS update maintains the trend of half pitch reduction to ensure a steady increase in transistor density, but the projection for the physical gate length scaling is relaxed to 0.71 every 3.8 years until the year 2022, far slower than previous scaling that followed 0.71 every 2 years, as depicted in Figure 2.7.

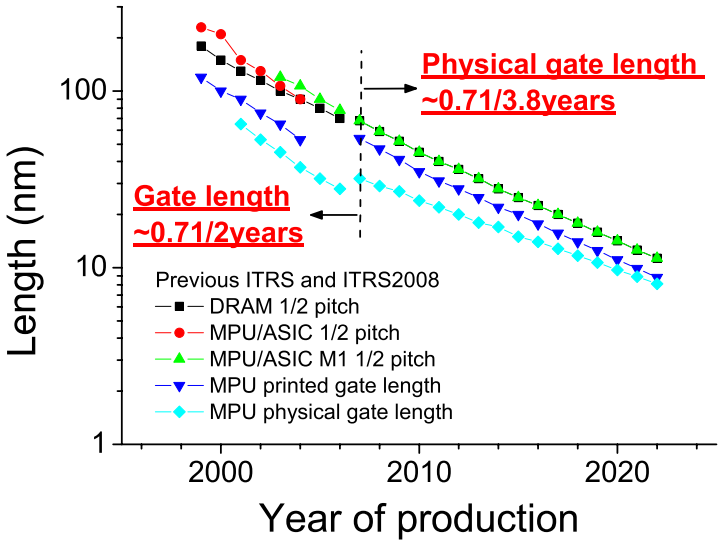


Figure 2.7 Summary of feature sizes in the evolution of technology in terms of recording and projection of ITRS editions.

Note that the projected usage of planar bulk MOSFETs has been extended until it reaches physical gate length of 14 nm in the year 2016. Accordingly the introduction of fully depleted silicon-on-insulator (FD SOI) MOSFET has been delayed to the year 2013, and double gate MOSFETs, such as FinFETs, to the year 2015. To maintain the same performance with scaling as before, the slower reduction of gate length will be compensated by so-called 'Equivalent Scaling', namely through enhancement of interconnects, mobility enhancement of strained silicon, novel gate stack, novel channel materials and alternative device architectures.

Table 2.2 lists the key projected parameters for high performance logic planar bulk MOSFETs. EOT is the to equivalent oxide thickness in terms of the equal capacitance of high permittivity ε_{highK} material with a physically thinness t_{highK}. The gate stack usually also includes an interfacial silicon dioxide layer with thickness t_{ox} between high-k dielectrics and silicon in order to improve interface quality and obtain high barrier. Therefore the EOT is calculated by,

$$\frac{EOT}{\varepsilon_{ox}} = \frac{t_{highK}}{\varepsilon_{highK}} + \frac{t_{ox}}{\varepsilon_{ox}}. \tag{2.6}$$

EOT_{elec} is the sum of EOT and additional equivalent layer due to the poly-silicon depletion effect and quantum confinement effect in both inversion layer and poly-depletion layer. Note that the additional equivalent layer occupies a considerable portion of the electrical oxide thickness for present dielectric stacks. I_{off} and I_{on} are the drain current at the drain bias of V_{dd} and gate bias of zero and V_{dd} respectively, i.e. off-state current and the drive

current. C_{gg} is the total gate capacitance including overlap and fringe parasitic capacitances. τ_{int} here is intrinsic gate delay due to total gate capacitance.

Table 2.2 High-performance logic technology projection for extended planar bulk MOSFETs in terms of ITRS.

Year	2008	2009	2010	2011	2012	2013	2014	2015	2016
M1 1/2 pitch (nm)	59	52	45	40	36	32	28	25	22.5
Physical L_{gate}(nm)	29	27	24	22	20	18	17	15	14
EOT (Å)	12	10	9.5	8.8	7.5	6.5	6	5.3	5
EOT_{elec} (Å)	19.4	13.3	12.7	11.9	10.4	9.3	8.75	7.95	7.6
V_{dd} (V)	1.1	1.1	1.07	1	1	1	0.97	0.92	0.9
$V_{th,sat}$ (mV)	225	196	175	168	94	103	102	107	112
I_{off} (μA/μm)	0.13	0.17	0.46	0.71	0.7	0.64	0.69	0.71	0.68
I_{on} (μA/μm)	1006	1317	1370	1333	1639	1807	1816	1793	1762
C_{gg} (fF/μm)	0.72	0.87	0.81	0.79	0.84	0.84	0.83	0.81	0.79
$\tau_{int}= C_{gg} V/I$ (ps)	0.79	0.73	0.64	0.60	0.51	0.46	0.45	0.42	0.4

2.2. Scaling challenges of MOSFETs

2.2.1. Feature patterning

From the viewpoint of circuit integration, the possibility to pattern ever smaller features on silicon has driven the integration of millions or billions of transistors on a chip. Among various patterning technologies, optical lithography is by far the mostly widely used by the semiconductor industry. Since 1998, it has entered the sub-wavelength imaging era. With improvements of light sources and applications of resolution enhancement techniques (RET) it is still the most cost-efficient and powerful patterning tool for further increase in chip density. The krypton fluoride (KrF) excimer laser introduced at the 180 nm CMOS technology has a wavelength of 248 nm. The argon fluoride laser (ArF) introduced at the 130 nm CMOS node has a wavelength of 193 nm. In sub-wavelength optical lithography [40], light diffraction and interference from sub wavelength pattern features causes image disorder. Therefore, resolution enhancement techniques are widely implemented. These include, among others, optical proximity correction (OPC) and phase-shifting masks (PSM). The numerical aperture (NA), which represents the light-gathering ability of optical lens, is proportional to the refraction index n of the media light travel through and the sine of the half-angle θ of the maximum light cone of the projection lens. Usually the

resolution is proportional to the wavelength, and inversely proportional to the NA. In optics, in order to distinguish two fuzzy Airy disc images (light spots), a separation distance of these two disc images has to be [41]

$$separation = k_1 \frac{\lambda}{NA}. \tag{2.7}$$

Rayleigh proposed k_1=0.61 to locate one disc centre at the minimum intensity point of the other disc. Alternative resolution limit is proposed by Sparrow using k_1=0.5 [41]. As a result, the minimum feature (half pitch) that a lithography process can achieve is determined by

$$R = k_2 \frac{\lambda}{NA}, \tag{2.8}$$

with k_2=0.3 or k_2=0.25.

Optical proximity corrections (OPC) adjust local exposure by correcting mask layouts. Proximity effects cause feature biasing, line-end pullback, and corner rounding [41]. By adding serifs at corners and slender line ends of mask layouts, lithography exposure is locally adjusted, and corner rounding and line-end pullback are compensated. Phase-shifting mask has phase shifter which changes the electric field of a neighbouring image, effectively causing boundary darkness by neighbour compensation [42]. These techniques are schematically illustrated in Figure 2.8.

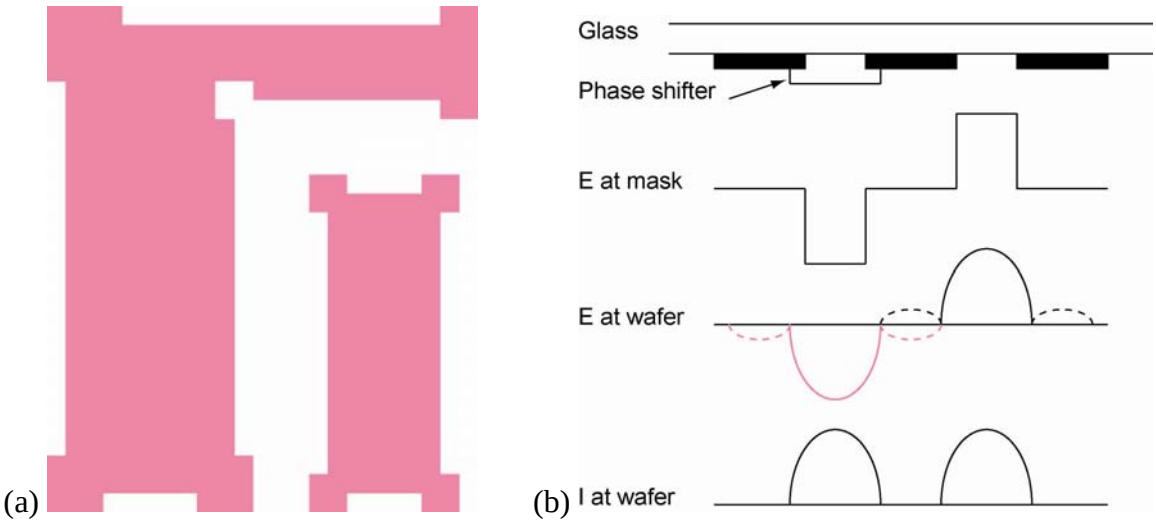


Figure 2.8 (a) Schematic view of OPC mask layout shows that serifs are added at corners and line-ends; (b) Phase-shifting mask operates to improve image boundary quality. Ultimately the intensity of image boundary at wafer tends to vanish.

2.2.2. Power and performance management

A major challenge for further device scaling is the control of the circuit power consumption within acceptable limits. In CMOS circuits, the total consumed power can be split into mainly active switching power P_{active} and standby leakage power $P_{passive}$ [43][44][45]. P_{active} is the active switching power, proportional to the number of switching circuits N_{active}, switching frequency f, load capacitance per circuit C_{load}, and supply voltage,

$$P_{active} = N_{active} C_{load} V_{dd}^2 f \,. \tag{2.9}$$

$P_{passive}$ is the standby leakage power, proportional to the number of passive, non-switching circuits $N_{passive}$, supply voltage, and off-state leakage current,

$$P_{passive} = N_{passive} V_{dd} I_{off} = N_{passive} V_{dd} I_0 \exp\left(-\frac{V_{th}}{S}\right), \tag{2.10}$$

where I_0 is the drain current at threshold voltage, and S is the device sub-threshold slope. The best value of S is limited to 60 mV/dec at room temperature T = 300K, and depends on the so-called body effect coefficient m,

$$S = \ln 10 \frac{kT}{q} m = \ln 10 \frac{kT}{q}\left(1 + \frac{\varepsilon_{si} / W_{dm}}{C_{ox}}\right), \tag{2.11}$$

where k is the Boltzmann constant, ε_{si} is the silicon permittivity, W_{dm} is the maximum depletion width, and C_{ox} is the gate oxide capacitance per unit area.

Since the down-scaling of the device dimensions increases circuit integration density, N_{active} typically increases as a result. f is in general inversely proportional to the transistor switching delay, and hence increases with scaling, but this is compensated by the reduction of C_{load} by the factor of dimension scaling, as seen from Table 2.1. Therefore, a reduction in supply voltage V_{dd} is critical to limit the increase in the active power dissipation of a system. The same rate of reduction in V_{th} would be necessary to maintain drive current, but the lowering of V_{th} results in an exponential increase in standby power according to equation (2.11). In addition, Joule heating and the resultant temperature increase of the chip during circuit operation leads to higher sub-threshold slope S and lower mobility, again, exponential increase in the passive power. This imposes a limitation on the minimum V_{th} design, which in turn limits the switching speed.

As scaling proceeds beyond the 90 nm technology, supply voltage scaling has practically stopped with the threshold voltage stalling at approximately 0.2V, closely approaching the

trade-off limits between leakage and drive currents due to sub-threshold slope limit. Managing the power dissipation to attain an optimal balance between active and passive power dissipation has become extremely complicated [46]. The leakage power of the worst-case lowest threshold voltage transistors becomes an important fraction of total power. Moreover, power and performance trade-off management leads to the emergence of new power scaling strategies, like adaptive body bias control, aimed at mitigating leakage while allowing for sufficient gate voltage overdrive.

2.2.3. Vertical scaling

2.2.3.1. Polysilicon depletion effect

While the channel doping concentration increases with scaling to maintain electrostatic integrity, the polysilicon doping concentration remains limited to 10^{19}~10^{20}cm^{-3} due to doping solid solubility limits. This, in combination with the extreme scaling of the gate oxide thickness, results in the degradation of the gate capacitance and transconductance [47][48][49]. This degradation is due to the increase of the effective oxide thickness, resulting from the polysilicon depletion layer when the device is operated at inversion. This polysilicon depletion effect is more pronounced at low polysilicon gate doping densities. The reduction in device performance is complemented by a threshold voltage shift. An analytical expression for the threshold voltage shift ΔV_{th}^{poly} due to the polysilicon depletion effect can be obtained from the Poisson equation, by including the polysilicon region [50][51],

$$\Delta V_{th}^{poly} = 2\psi_B \frac{N_{sub}}{N_p}. \qquad (2.12)$$

Here ψ_B is the difference between intrinsic and extrinsic Fermi levels in the substrate, and N_{sub} and N_p are the doping concentrations in the substrate and polysilicon, respectively. As stated earlier, an increase of substrate doping, or a decrease of polysilicon doping, both lead to an increase in threshold voltage.

2.2.3.2. Quantum effects

The increasingly strong surface electric field near the silicon/oxide interface creates a potential well, as the energy bands bend to form an inversion channel. This leads to quantum confinement of the inversion carriers, giving rise to discrete sub-bands for motion in the direction perpendicular to the interface and shifting the peak of the inversion charge

centroid away from the interface (although retaining free continuum motion in the plane parallel to the interface) [52]. The quantum mechanical confinement increases the effective oxide thickness, decreasing the inversion charge density at a given bias, and, in combination with the ground state shift, increases the threshold voltage [53]. Comprehensive 1D and 2D Schrödinger-Poisson solutions demonstrate the impact of quantum confinement on sub-threshold slope [54], drain induced barrier lowering (DIBL), and short-channel effects (SCE) [55]. In Si, the peak of the inversion carrier concentration is located around 1.2nm away from interface [53][56]. As a result, the effective oxide thickness under inversion bias conditions can be expressed as [53],

$$t_{ox}^{QM} = t_{ox} + \frac{\varepsilon_{ox}}{\varepsilon_{si}} \Delta z \,, \tag{2.13}$$

where Δz is the distance of the inversion charge centroid away from interface. Accordingly, the correction to the threshold voltage due to quantum confinement effects can be written in the form [53],

$$\Delta V_{th}^{QM} = qN_{sub}\left(\frac{\Delta z}{2\varepsilon_{si}} + \frac{t_{ox}}{\varepsilon_{ox}}\right)\Delta z \,. \tag{2.14}$$

In scaled devices with ultra thin-gate oxide, significant performance degradation is attributed to the quantum confinement effects, because of the increasing weight of Δz in the total effective oxide thickness. Moreover, the threshold voltage shift increases with the increase of substrate doping, and poses constraints on the design of substrate doping profiles in scaled devices.

2.2.3.3. Gate tunnelling

Gate tunnelling current has become a major contributor to static power dissipation, making it comparable to the dynamic power dissipation for sub-65 nm technology generations with pure SiO_2 or SiON dielectric [57]. While several mechanisms of gate leakage exist, the most important one in contemporary technology is direct tunnelling, where carriers tunnel through the entire width of the potential barrier formed by the gate dielectric, as is schematically illustrated in Figure 2.9 [58]. Direct tunnelling is exponentially sensitive to the physical thickness of the gate dielectric, and for sub-2 nm SiO_2 dominates by orders of magnitude the leakage due to Fowler-Nordheim tunnelling (also shown in Figure 2.9) or trap assisted tunnelling [59][60][61][62]. For a 2 nm SiO_2, the direct tunnelling current density exceeds 1 A/cm^2, while for a 1 nm SiO_2, the tunnelling current density exceeds 10^4 A/cm^2, (for 1 V bias) [62].

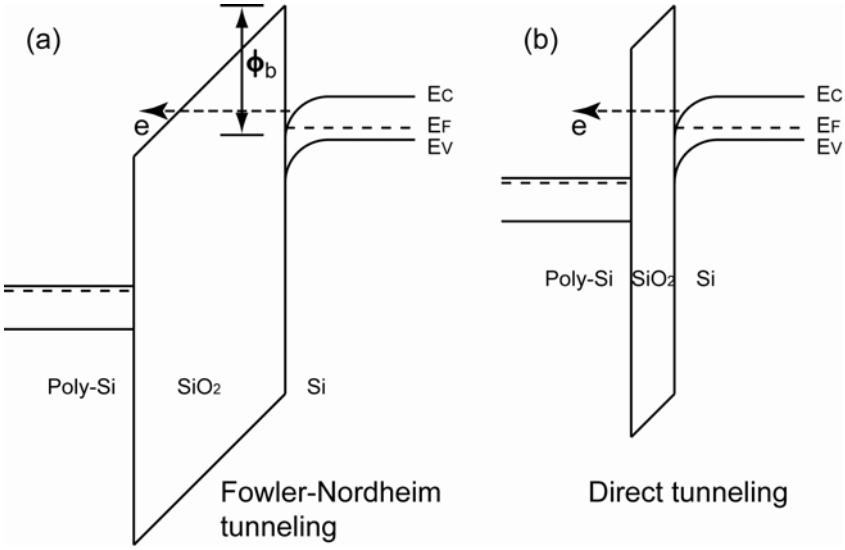


Figure 2.9 Band diagram view of Fowler-Nordheim tunnelling (a) and direct tunnelling (b) in MOS capacitors.

This excessively high gate leakage imposes a fundamental limit on the scaling of the silicon dioxide as a gate dielectric. Continuous reduction of the EOT to 1 nm is realised by introducing alternative gate dielectric materials with higher permittivity, thus maintaining a sufficiently large physical thickness to suppress direct tunnelling. A natural choice of alternative oxide is SiON (with a dielectric constant of about 6), since the presence of nitrogen also improves oxide reliability by mitigating boron penetration in p-channel transistors [63]. However, further EOT scaling below 1 nm is only possible with the introduction of transitional-metal oxides [e.g. HfO_2] with dielectric constants beyond 20, referred to as high-k materials [64]. Introduction of such novel dielectrics entails significant technological innovation, including the replacement of the poly-Si gate with a metal gate (helping to eliminate poly-depletion effects), and represents the state of the art in gate stack engineering, allowing the 45 nm technology to achieve 20 times smaller gate leakage compared to its predecessor [3].

It should be noted that high-k gate stacks inevitably exhibit a thin, mostly sub-stoichiometric, SiO_x interfacial layer above the Si substrate, which are difficult to reduce below 0.5 nm [65]. Additionally, oxides of transitional metals have a smaller band-gap [66]. This means the potential barrier for tunnelling is substantially reduced (e.g. to less than 2 eV in HfO_2, compared to 3.15 eV in SiO_2), so that Fowler-Nordheim tunnelling becomes an appreciable component of the gate leakage.

2.2.4.Lateral effects

2.2.4.1. Threshold voltage roll-off and DIBL

Short channel effects (SCE) relate to the loss of electrostatic control of the gate over the charge in the channel of the transistor. They are associated with the enhanced electrostatic influence of the drain, as the channel length shrinks. This influence is due to the relative enlargement of the depletion layer of the source/drain p-n junction, with respect to the channel length. One measurable manifestation of SCE is the threshold voltage roll-off, illustrated in Figure 2.10 (a). It consists of a rapid reduction in V_{th} as the gate length is reduced, while maintaining the same vertical doping profile. This is due to the reduction of the lateral potential barrier with gate length scaling, as shown in Figure 2.10 (b).

The V_{th} roll-off is more dramatic when the drain bias is high. This is expected, since an increase in drain voltage leads to further penetration of the drain-induced field into the channel of the transistor, reducing the lateral potential barrier that is typically controlled by the gate. This effect is termed drain induced barrier lowering (DIBL) and is illustrated in Figure 2.11 (a). As in the case of V_{th} roll-off, the lowering of the lateral potential barrier is reflected in lower V_{th}.

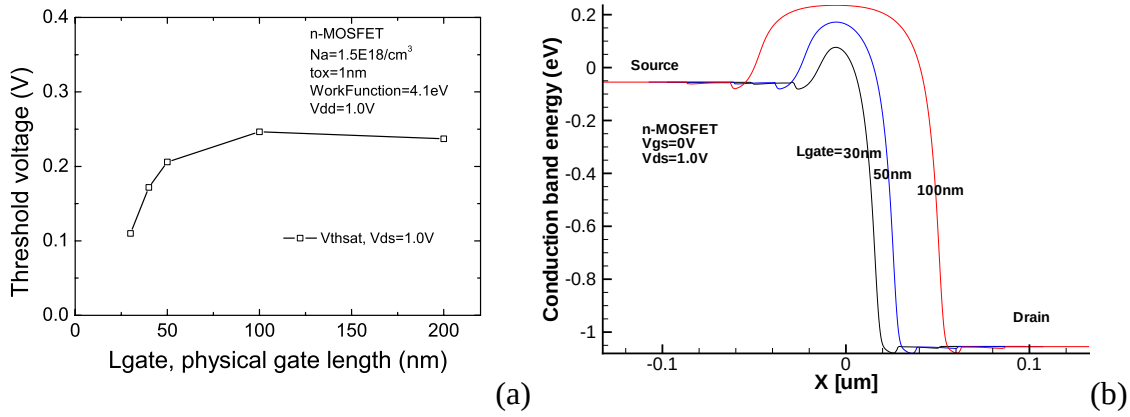


Figure 2.10 (a) A simulation shows that threshold voltage roll-off characteristics. The threshold voltage is determined at the drain current 5×10^{-8}W/L ampere from *Id-Vg*; (b) Gate barrier lowering due to gate length shrinking.

V_{th} lowering due to DIBL can be qualitatively explained by a semi-empirical 'charge sharing' model [67]. The idea is schematically illustrated in Figure 2.11(b), and considers the splitting of the depletion charge under the gate into two parts – one controlled by the gate, the other controlled by the source and drain. This introduces a correction of the maximum depletion charge controlled by the gate which determines the threshold voltage.

For an n-channel MOSFET the correction leads to the following expression for the threshold voltage:

$$V_{th} = V_{fb} + 2\psi_B - \frac{Q'_{dm}}{C_{ox}}, \tag{2.15}$$

with $Q'_{dm} = -qN_a^- \left(\frac{L + L'}{2L} \right) W_{dm}$, where L, L' and W_{dm} are defined in Figure 2.11 (b).

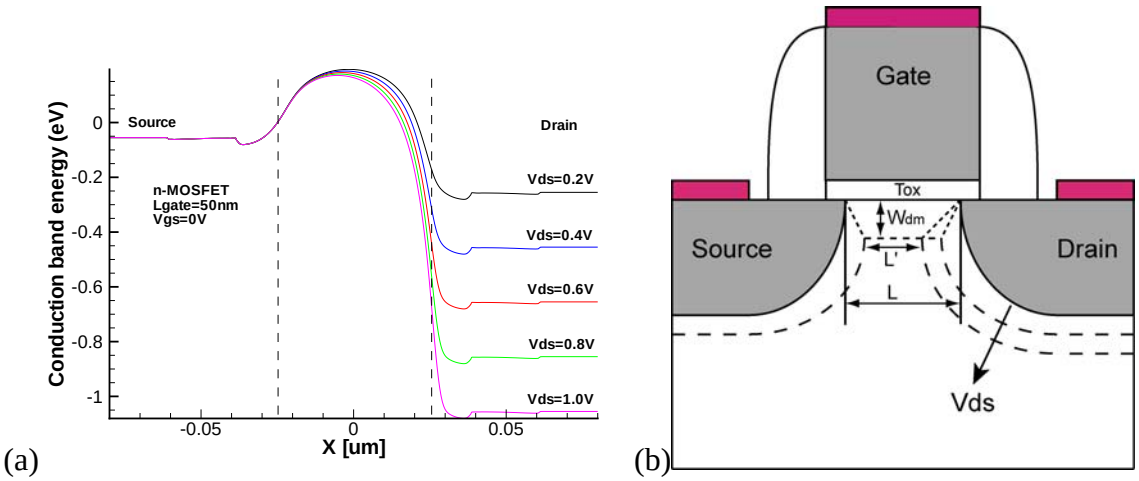


Figure 2.11 A simulation of a MOSFET demonstrating drain induced barrier lowering (a); The schematic view of charge-sharing model of short-channel effect (b).

Mitigating SCE without increasing the vertical channel doping requires the incorporation of '*halo*'-doping around the p-n junctions. The halo implants are of the same type as the substrate doping, but with a much higher concentration, in order to reduce the depletion layer of the source-channel and drain-channel *p-n* junctions.

In ultra-short channel devices, a reverse V_{th} roll-off is observed [68]. This is a result of the overlapping of the source and drain halos, leading to a net increase of the doping in the channel, hence an increase in V_{th}.

2.2.5. Hot-carrier degradation and BTI

High field induced hot-carrier (HC) degradation affects reliability and causes long-term instability [69][70], manifested by a threshold voltage increase and drive current reduction. Hot carriers generated by the high electric field near the drain are injected into the oxide with enough energy to create defect states (traps) in the oxide near the silicon/oxide interface [71]. It is found that only hot electrons having energy of 0.6eV larger than the Si-SiO_2 conduction band discontinuity can cause SiO_2 degradation in n-channel MOSFETs. The degradation is attributed to the breaking of the ≡SiH bond at the interface [70][72].

Phenomenological models are used to describe HC degradation, of which the most widely adopted is the 'lucky electron' model, proposed by Shockley [73]. It is based on the assumption that only the hot electrons accelerated by the electric field and not suffering collisions are most likely to cause impact ionization.

The probability of an electron obtaining enough energy E_b to surmount oxide barrier, given an electric field E is [70][74] of the form $P_{E_b} = \exp(-E_b / qEl)$. The probability of a hot electron travelling a distance d arriving at the interface Si/SiO2 without suffering from energy loss collisions is [75] of the form $P_d = \exp(-d / l)$, where l is the hot electron mean free path. Consequently, the injection probability is proportional to the product of these two probabilities [76][77],

$$I_g = C_1 I_d P_{E_b} P_d, \tag{2.16}$$

where C_1 is a fitting coefficient and I_d is the drain current.

Performance degradation can become severe at elevated temperatures even at low fields. Negative bias temperature instability (NBTI) is another very important reliability concern for contemporary p-channel MOSFETs. The threshold voltage shift primarily depends on stress bias condition, stress time and stress temperature. The observed power-law dependence on stress time [78][79] is explained by the Si-SiO_2 interface diffusion-reaction model [72]. This can be used to estimate the threshold voltage shift due to NBTI [80],

$$\Delta V_{th} = C_2 \exp\left(\gamma V_g\right) \exp\left(-E_a / kT\right) t^{0.25}, \tag{2.17}$$

where E_a is a fitting parameter (the NBTI activation energy) and C_2 and γ are fitting coefficients.

The wide use of nitrogen in sub-2nm gate oxides is found to enhance NBTI [81]. NO gas annealing leads to lower activation energy E_a and nitride oxides acquire more positively charged traps during stress compared to the pure SiO_2, which degrades the transistor lifetime by 2 to 3 decades. From a device design point of view, the suppression of the impact of nitrogen on NBTI is an important concern.

It is worth noting that the NBTI degradation is recoverable. While static measurement of NBTI shows continuous degradation, pMOSFETs under dynamic stress conditions undergo passivation/relaxation stages between stresses, and the threshold voltage accordingly recovers after stress removal [82][83].

Greater HC degradation is observed in narrow-width MOSFETs with STI [84], although initial impact ionization rate is not bigger than that of large channel width fresh devices. This effect is ascribed to the increase of impact ionization rate and injection rate in narrow n-MOSFETs [84] and a higher oxide electron trapping efficiency in narrow p-MOSFETs [85].

2.2.6.Statistical variability

In contemporary MOSFETs with sub-50 nm channel length, the number of dopants in the channel depletion is of the order of a hundred, and the number of interface traps is of order of ten. The exact number and location of the discrete dopants and traps fluctuate from device to device. In addition, resist-defined gate line edge roughness is unavoidable. The gate material granularity and the oxide thickness fluctuation of 1 interatomic layer of the Si crystal lattice, also contribute to the microscopic differences in devices with identical macroscopic parameters. These microscopic fluctuations originate either from uncontrollable process aspects or from intrinsic material granularity. All these factors result in uncontrollable statistical variations in the electrical characteristics of nominally identical devices. Therefore, the characteristics of an idealised ‘uniform’ transistor lose their representative meaning. Instead, a statistically significant ensemble of device characteristics has to be obtained to aid the design of circuits and systems, based on devices with statistical variability in mind.

2.2.6.1. Random discrete dopants

Random discrete dopants (RDD) introduced by ion implantation are a major source of statistical variability in sub-0.1 μm CMOS transistors [86][32]. Random dopant fluctuations cause threshold voltage variation, partially due to the fluctuation in depletion charge, due to dopant number fluctuation [87]. Accordingly, analytical models attempting to estimate the threshold voltage standard deviation (σV_{th}) are based on the variance in dopant number in the channel region of the transistors, which is known to follow a Poisson distribution. The variance in depletion layer charge of ionized dopants is [87][88][89]

$$\sigma\left(Q_{dm}LW\right) = C_3 q\sqrt{\bar{N}_a LWW_{dm}}\,, \tag{2.18}$$

and therefore,

$$\sigma V_{th} = \frac{\sigma Q_{dm}}{C_{ox}} = C_3 \frac{qt_{ox}}{\varepsilon_{ox}}\sqrt{\frac{\bar{N}_a W_{dm}}{LW}}\,, \tag{2.19}$$

where C_3 is the fitting coefficient. In [87][89] it is 1/2, and in [88][45] it is $1/\sqrt{3}$.

However, equation (2.19) underestimates σV_{th}. Full scale numerical simulations have shown [86][32] that the dopant number fluctuation is insufficient to explain the true magnitude of the threshold voltage fluctuation. The randomness in the dopant arrangement in devices with having identical number of dopants in the depletion region contributes further to V_{th} variability due to the spatial inhomogeneities in the potential and corresponding current percolations.

However, the analytical models capture essential aspects of threshold voltage variability. It is inversely proportional to the oxide capacitance, and increases with the substrate doping. Therefore, further reduction of EOT reduces V_{th} variability, but the benefits of this scaling trend are cancelled by the faster increase in channel doping that is needed to maintain the electrostatic integrity of a planar bulk MOSFET. In addition, the area reduction of a bulk MOSFET with scaling also contributes to increasing V_{th} variability.

2.2.6.2. Line edge roughness

In the past, the line edge roughness (LER) of the lithography patterns in the photoresist was negligible compared to the device dimensions, and was rarely taken into account in the device simulation and analysis. However, as the size of the printed features in contemporary circuits has reached decananometer dimensions, the number of molecules contained in the patterned resist dramatically decreases. Molecule aggregates of varying size produce non-uniformities on the resist edge, which are not smoothed by subsequent etching steps. This introduces microscopic fluctuations in the gate edge of the transistors, which are also transferred to the metallurgical *p-n* junction of the source and the drain. Since the resist polymer molecules are bigger than 1.0 nm [90], it is difficult to reduce LER below the state of the art $3\sigma \approx 4$ nm [91], which represents a significant fraction of the gate length in 45 nm technology devices.

2.2.6.3. Poly-silicon granularity related variability

The polycrystalline-silicon, used as gate material causes threshold voltage increases due to Fermi level pinning at the grain boundaries [92]. This is due to the defect states at grain boundaries energetically located in the polySi bandgap [94][95]. The stochastic nature of polysilicon grain shape and orientation make the corresponding Fermi level pinning an important source of device variability [34].

2.3. Technology boosters

The scaling challenges of recent MOSFET technologies, described in the previous section, demand the introduction of technology inventions and new materials. Remarkable advancements have already been achieved in channel and gate stack engineering. These innovative technology boosters bring about so-called *equivalent scaling* where the performance improvement and the aggressive pitch reduction continue to deliver the previously established performance trends, while the actual scaling of certain dimensions or electrical parameters (e.g. V_{dd} and V_{th}) has stalled.

2.3.1. Strained channel

2.3.1.1. Stress engineering

The effect of mechanical stress in semiconductor devices is not a new phenomenon. Smith [96] studied the relationship between resistivity change and tension in silicon and germanium in 1954, establishing a linear dependence. During MOSFET fabrication, thermal processes are common. Thermal cycling generates mechanical stresses between materials with different thermal expansion coefficients. In addition, silicon oxidation may cause compressive stress due to volumetric expansion while it consumes silicon. Therefore, shallow trench isolation around active regions can produce significant compressive stress by sidewall oxidation. As the area of a MOSFET gets smaller, the STI-induced stress in the active region becomes significant. The compressive stress from STI can lead to rapidly increased junction leakage [97] and, depending on layout conditions, can degrade the nMOSFET drive current [98].

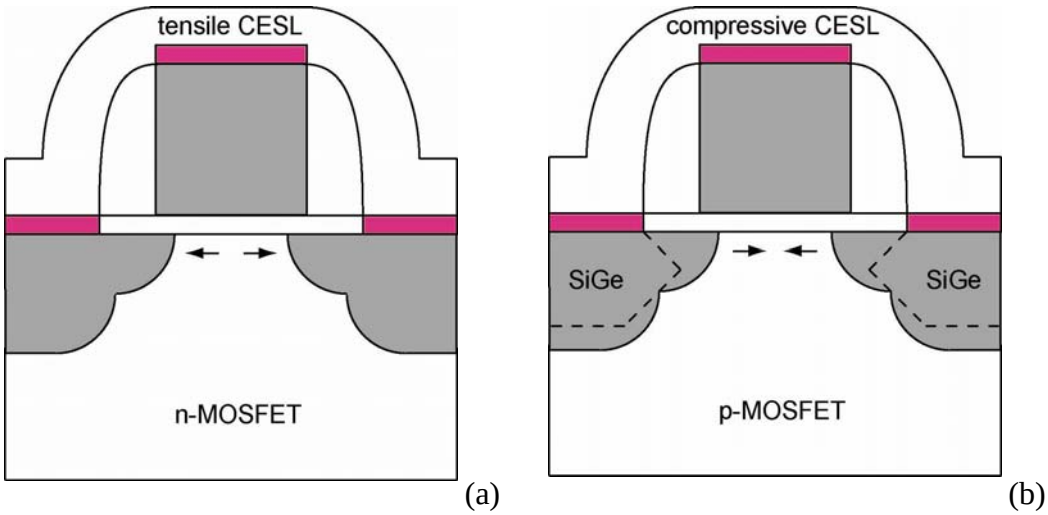


Figure 2.12 Schematic demonstration of strained silicon schemes for (a) nMOSFETs and (b) pMOSFETs.

Intentional use of mechanical stress to enhance the MOSFET performance started at the 90 nm CMOS technology. However, initial trials involved in-plane biaxial tensile stress in standard (001) wafers with a heterogeneous structure composed of epitaxial Si layer grown on a relaxed SiGe 'virtual substrate' [99][100][101]. Due to the larger lattice constant of the relaxed SiGe alloy, the silicon layer is stretched in both directions parallel to interface, leading to tensile strained silicon.

In commercial CMOS technology, alternative methods of stress engineering are used. A highly tensile nitride contact-etch-stop layer (CESL) is deposited on the top of n-MOSFETs, while SiGe is selectively, epitaxially grown in recessed source/drain regions. A compressive CESL is deposited for p-MOSFETs [37]. These methods introduce longitudinal uniaxial stress along <110> channels in contrast to the biaxial stress in an epitaxially grown Si layer on a SiGe virtual substrate. As schematically illustrated in Figure 2.12, tensile CESL causes tension in n-channel MOSFETs, and compressive CESL and SiGe source/drain areas introduce compressive stress in p-channel MOSFETs [102][103][104].

2.3.1.2. Performance enhancements due to strain

The biaxial tensile strain splits silicon's six-fold degenerate electron conduction band minima into two-fold Δ2 and four-fold Δ4 minima, and the split energy (in electron volts) due to strain is proportional to the Ge fraction x in $Si_{1-x}Ge_x$ [105][99].

$$\Delta E_{strain} = 0.67x\,. \tag{2.20}$$

Since the minima of the Δ4 valleys are barely affected, the minima of the Δ2 valleys are reduced by ΔE_{strain}. When the substrate is inverted, ΔE_{strain} adds to the quantisation induced splitting between the Δ2 and Δ4 subbands, making it energetically unfavourable for carriers to populate the Δ4 valleys. Hence all the inversion charge populates the Δ2 subbands that have smaller effective mass in the transport direction along the channel, compared to the transport mass for Δ4 carriers [99][106]. Additionally, the increased energy split between the Δ2 and Δ4 valleys reduces intra-valley scattering, thus further enhancing mobility [99][106].

The application of uniaxial strain along the channel has an impact on the electronic structure of the Si channel in a similar way, and therefore enhances electron mobility through the same mechanism as biaxial strain. Theoretical calculation shows that uniaxial strain offers more advantages over biaxial strain such as less band gap narrowing [107]. In

the case of holes, compressive stress splits the degenerated valence subbands, increasing the hole population in the subband with smaller transport effective mass [100].

2.3.2. High-k/metal gate

High-k/metal gates were introduced into mass production in 2007 by Intel in the 45 nm CMOS technology generation [3][38]. This is the first time that traditional oxides or oxynitrides have been replaced in gate stacks, to enable continuous scaling of the EOT.

2.3.2.1. High permittivity gate dielectrics

The replacement of SiO_2 by a high-k dielectric stack must satisfy a series of material constraints and process integration conditions. Although there are many potential high-k materials, based on their permittivity, a strict selection rules out many candidates. First of all, from a gate leakage perspective, a suitable conduction band offset is necessary to provide a sufficient barrier. For example, tantalum oxide has an adequately high permittivity of around 25, but the ~0.36eV conduction band barrier is not sufficient to provide any overall advantage over SiO_2. The narrower band gap of high-k materials cancels the benefit of the high dielectric constant. Thereby, a suitable trade-off between the dielectric constant and the conduction band offset is the first criterion for high-k dielectric candidates [108]. A few high-k dielectrics show the promise to replace silicon oxide, and some of their fundamental parameters are listed in Table 2.3 along with the corresponding parameters of SiO_2 [108][109][110].

Table 2.3 Some essential parameters for selected high-k materials and SiO_2.

Material	Bandgap (eV)	Relative dielectric constant	Conduction band offset (eV)	Leakage current reduction (ref SiO_2)	Thermal stability, T_{max} (°C)
SiO_2	9	3.9	3.15		
Al_2O_3	8.8	9.5-12	2.8	10^2-10^3	~1000
ZrO_2	5.7-5.8	12-16	1.4-1.5	10^4-10^5	~900
HfO_2	4.5-6	16-30	1.5	10^4-10^5	~430-600
$ZrSiO_4$	~6	10-12	1.5		
$HfSiO_4$	~6	~10	1.5		

Thermal stability is another critical property in application of high-k dielectrics to CMOS technology. Front end processes involve high temperature thermal annealing, usually

peaking above the 1,000°C necessary for dopant activation. Most high-k materials undergo crystallization at such high temperatures, as indicated in Table 2.3. An amorphous HfO_2 layer grown initially by atomic layer deposition (ALD) exhibits a poly crystalline structure after 500°C activation annealing [110]. Heterogeneous poly-crystalline orientation degrades the mobility by spatially varying the electric field – inducing additional scattering. Therefore, amorphous high-k materials are preferred. The incorporation of Si into a high-k material can markedly increase the thermal stability [64]. Nitrogen also can promote thermal stability and effectively prevent dopant penetration [111].

2.3.2.2. Metal gate

Initially, poly-Si/high-k combination gate stack was considered as a route to improving gate leakage. However theoretical studies and experimental data show a mobility degradation compared to the use of metal gates [112][113][114]. Table 2.4 lists the work functions (WF) of some commonly studied metals for MOSFETs [115]. Depending on the gate dielectric, the work function varies due to differing band alignments.

Table 2.4 Experimental vacuum (effective) work functions of selected metals on various dielectrics [115][116][117][118].

Metal/dielectric	Work function (eV)
Al/Al_2O_3	3.9
Al/SiO_2	4.14
Al/ZrO_2	4.25
W/SiO_2	4.6-4.7
Mo/SiO_2	5.05
Mo/HfO_2	4.95
Pt/SiO_2	5.59
Pt/HfO_2	5.23
Pt/ZrO_2	5.05
Ni/Al_2O_3	4.5
Ni/ZrO_2	4.75
TiN/SiO_2 [116]	4.2-4.9
TiN/HfO_2 [117]	~4.33-4.58 (effective)
TiN/HfO_2 [118]	3.6-5.1 (effective)

Metal gate implementation, however, has difficulties. Similar to the high-k dielectrics, the first issue is thermal stability of the work function. For integration with SiO_2 and Al_2O_3

dielectrics, the greatest difficulty is in finding a metal with a low work function (φ_m = 4.1-4.3 eV), although most midgap and high work function (φ_m = 4.9-5.2 eV) metals remain stable under high temperature in the range of 800-1,000°C [64]. Thermal process leads to a drift in the low work function of low WF metals towards the midgap due to the Fermi level pinning dependence of annealing temperature [119].

Gate-last or replacement-gate processes appear as a remedy for thermal stability problems. The high-k/metal gate stack is realized after all the front end processes, avoiding critical high temperature treatments. This requires the selective removal of a sacrificial poly-Si and oxide layer after planarization of the top of transistor by chemical mechanical polishing (CMP). Then the new high-k gate dielectric material is deposited (usually by ALD), followed by metal gate deposition [120][121]. An oxide interfacial layer (IL) is formed intentionally (or unintentionally in most cases of ALD deposition) at moderate temperature. The high-k/metal gate is thereafter subject to back end processes, such as interconnect deposition and packaging, the thermal budgets of which are usually lower than 500 ˚C.

Although experimentally controversial [122], an additional advantage of metal gates could be their ability to screen the soft phonon modes in high-k dielectrics from coupling to the inversion layer, giving rise to partially restored surface mobility, compared to the case of poly-Si/high-k gate stacks [113].

2.4. Summary

To sum up, this chapter first presented a comprehensive overview of the scaling of key parameters of bulk MOSFETs. It described the scaling rules: constant-field scaling and generalized scaling, and it explored the new scaling features beyond the 90 nm CMOS technology and the ITRS projections of design and performance over the next generations of devices. Secondly, the scaling challenges facing CMOS were described in detail, including: optical patterning difficulties; the trade-off between power dissipation and performance; the vertical and lateral scaling challenges such as gate direct tunnelling and short-channel effects; reliability and statistical variability. Finally, the technology boosters, such as stress engineering and high-k/metal gates, all employed to enable continued scaling, were presented. This chapter provides the basis of understanding needed for the next chapters to discuss research objectives, methods and results.

Chapter III

3. Simulation tools and methodology

With the rapid development of the IC industry, technology computer aided design (TCAD) has undergone a continuous evolution. Its importance, in assisting the design and performance evaluation of both devices and circuits, has been indispensable. Advanced CMOS technology process simulators and device simulators, which reflect the latest technology features and underlying physics, can reduce, by up to 40%, new technology development costs. TCAD has become crucially important in achieving the optimal design of modern semiconductor devices. The TCAD tools used in this work are Sentaurus Process and Device [123]. In addition, the Glasgow drift-diffusion 'atomistic' simulator, specially tailored to solve problems in statistical variability and reliability has been used, and is described in detail further below [124][125].

3.1. Process simulation

One main role of process simulation is to replicate real fabrication steps in simulation, obtaining relevant information about device structures and doping profiles which can then be used in device simulation. Usually one of two approaches is used to generate the semiconductor device structure needed before electrical device simulations can be performed. The first employs simple analytical approximations to mimic device geometries and doping distributions using, for example, empirically fitted Gaussian or complementary error functions. The other approach is to perform numerical process simulations, using accurate physical models of the sequential process steps. The choice between the two approaches depends on the goals, and the required accuracy, of the final numerical device simulations. Numerical process simulation is in the heart of this project and its use is determined by the nature of the project.

3.1.1. Ion implantation

Ion implantation is an effective approach to precisely introduce dopants into semiconductor materials, which has been widely employed in semiconductor

manufacturing since 1970's [126]. After ions of a certain charge/mass have been selected and accelerated by magnetic and electrostatic fields, these ionised dopants are focused forwards and scanned across the semiconductor substrate. Depending on the gained energy, the type of dopant, the incident direction etc., the average depth of penetration of the impurity dopant into the substrate material can be accurately determined, although a scattered distribution of dopants reflects the randomness of the ion energy transfer to the semiconductor lattice. An advantage of ion implantation is the low temperature at which it proceeds. The accurate doping profile control available allows for multiple implants to be used in the design of a particular doping profile, providing flexibility in the semiconductor device design process.

In process simulation, the doping profile resulting from ion implantation may be simulated statistically using a Monte Carlo approach. This approach simulates the dopant distribution by following the scattering of individual dopants within a lattice of host atoms and electrons, correctly accounting for ion energy losses, to estimate the final doping distribution. Analytical models can capture the doping probability distribution by functions parameterised by the moments of the statistical distributions. In multi-dimensional doping distributions a point-response distribution is used. The semiconductor device structure is illustrated in Figure 3.1. Energetic dopants bombard the substrate surface at point (*u, v*), and a dopant entering from this incident point will scatter with host atoms or electrons until it loses its energy and stops. Its stopping location is described by a distribution *F* in the 2D or 3D semiconductor space. In the 2D structure depicted in this figure, the dopant concentration at point (*x, y*) can be calculate by

$$C(x,y) = N_{dose} \int_{\Omega_{gas}} F\left(x, y, u(s), v(s)\right) ds \tag{3.1}$$

where N_{dose} is the total dose, and Ω_{gas} is incidence point set. The dopant concentration of a point (*x, y*) is the superposition of all possible entry dopant distributions.

The distribution function assumes independence of the primary and lateral distributions F_p and F_l. Therefore the distribution function *F* may be written as

$$F(x,y,u,v) = F_p\left(x-u(s)\right)\cdot F_l\left(y-v(s)\right), \tag{3.2}$$

$$F(x,y,z,u,v,w) = F_p\left(x-u(s)\right)\cdot F_l\left(y-v(s)\right)\cdot F_l\left(z-w(s)\right), \tag{3.3}$$

for 2D and 3D simulations respectively. The primary distribution F_p can have different order moments associated with entry point location: projected range R_p is the mean value

of stopping location. The standard deviation σ, skew γ, and kurtosis β are the second, third and fourth central moments of the distribution function. The primary distribution may be a Gaussian, a Pearson, a Pearson distribution with a linear exponential tail, and a dual Pearson distribution functions. The lateral distribution usually is a Gaussian function.

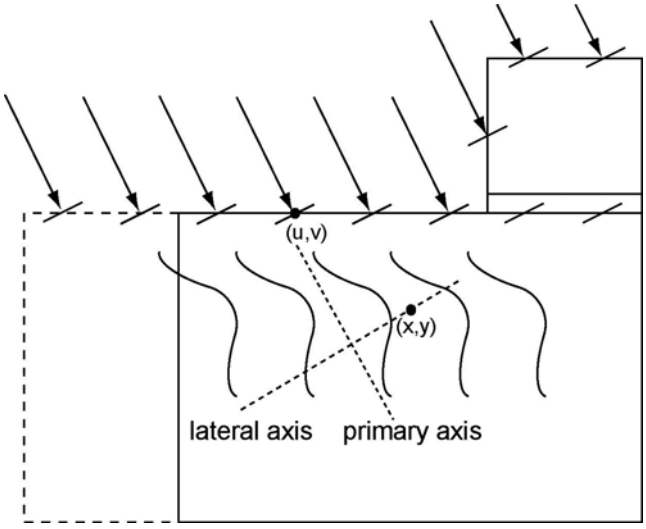


Figure 3.1 Schematic view shows the point-response distribution for ion implantation in Sentaurus process simulation. The arrow indicates the incident direction of the implanted ions, which impacts the solid surface. The curves inside the substrate indicate the primary distribution of dopants.

The ion stopping of ionized impurity dopants is due to energy loss in nuclear and electronic collisions with host materials. Electronic stopping power is proportional to the incident ion velocity. Therefore in high energy region, the energy is mainly transferred to electrons in electronic collision. Near the end of the energy loss process of low energy, the nuclear collision dominates causing host atom displacement damage. Heavy dopants have larger nuclear energy loss per unit distance and therefore small projected range but result in more damage [127].

3.1.2. Thermal annealing

Thermal processing is crucial to semiconductor device fabrication, in particular for: impurity dopant activation, damage annealing, film growth and strain engineering. For example, damage due to ion implantation partially renders the single crystalline substrate to be amorphous, with dopants in interstitial, rather than substitutional, sites. Proper thermal cycling can restore lattice crystallinity and improve device charge transport.

In a thermal process, the temperature changes with time and reactions may happen among impurity dopants, lattice atoms, and defects (interstitials and vacancies). Generally

speaking, the particle flux of a species A with charge c as a result of diffusion can be written as

$$\vec{J}_{A^c} = -d_{A^c}\left(\frac{n}{n_i}\right)^{-c}\nabla\left(A^c\left(\frac{n}{n_i}\right)^{c}\right). \tag{3.4}$$

Where A^c is the particle density, $d_A{}^c$ is the diffusivity, n is the electron concentration, and n_i the intrinsic electron density. In addition the continuity equation for species A with charge c is

$$\frac{\partial A^c}{\partial t} = -\nabla\cdot\vec{J}_{A^c} + R^{trans}_{A^c} - R^{cluster}_{A^c}. \tag{3.5}$$

$R^{trans}_{A^c}$ is the net reaction rate of the diffusing species A with charge c; $R^{cluster}_{A^c}$ is the negative contribution related to the clustering of the species. Equation (3.4) and (3.5) jointly construct the diffusion model.

Different boundary conditions can be applied, such as Dirichlet, Neumann or mixed boundary conditions. Considering the different solid solubility of dopants in different materials, segregation can be included in the boundary conditions. In the process of segregation the flux is given by

$$\vec{j}\cdot\vec{n} = k_{transfer}\left(C^1_A - \frac{C^2_A}{k_{segregation}}\right). \tag{3.6}$$

Where $\vec{n}$ is the interface standard normal vector; C^1_A is the concentration of dopant A in material 1; C^2_A is the concentration of A in material 2; $k_{transfer}$ is the transfer rate; and $k_{segregation}$ is segregation coefficient of A in second material relative to the first material; namely the impurity concentration ratio under equilibrium.

3.1.3. Film formation

Thermal oxidation is an important way to form thin gate silicon dioxide (SiO_2) films. Compared to chemical-vapour deposition (CVD) the thermal oxide has high density and low porosity. Thermal oxidation kinetics for thick oxide (30 to 2,000 nm) includes two processes [128]: a diffusion of oxidant species such as H_2O and O_2 through already formed oxide to the oxide/silicon interface, and reaction with the silicon at the interface. The growing amorphous oxide expands above the moving interface due to volumetric increase. The oxidant transport is governed by the diffusion equation. This process is illustrated in

Figure 3.2. At the reaction front at the interface the following reactions describe the growth kinetics.

$$\begin{aligned} Si(solid) + O_2(gas) &\rightarrow SiO_2(solid) \\ Si(solid) + 2H_2O(gas) &\rightarrow SiO_2(solid) + 2H_2(gas) \end{aligned} . \quad (3.7)$$

Consuming one Si atom generates one SiO_2 molecule but with almost the same mass density (2.33-2.2g/cm^3), that leads to SiO_2 expanding more than 125%.

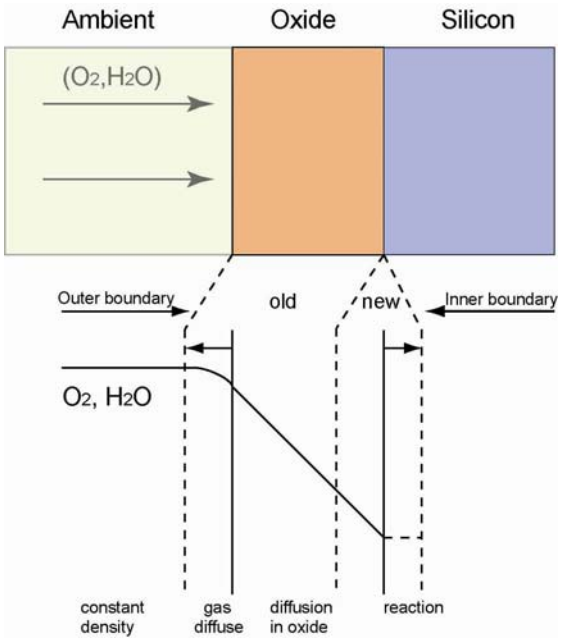


Figure 3.2 Illustration of thermal oxidation process on Si wafers. It indicates oxidant diffusion through oxide and its reaction with Si atoms at the interface with oxide.

The oxide thickness grows faster at higher temperature, and water vapour oxidation has higher reaction rate compared to elemental oxygen. The choice depends on the required oxide quality. Slow dry O_2 is usually used to grow thin high quality gate oxide while water steam is used to grow oxides for isolation.

Chemical vapour deposition (CVD) SiO_2 at low temperature of usually several hundreds of degrees Celsius is deposited as a result of reactions of external species such as silane SiH_4 with oxidants. Similarly, silicon nitride (Si_3N_4) is also deposited by CVD through the reaction of dichlorosilane $SiCl_2H_2$ with ammonia (NH_3). Silicon nitride can be used as gate spacer. Deposition in process modelling is usually modelled by geometrical deposition of a film layer in target structure rather than the precise modelling of the reaction processes.

3.1.4. Process induced stress

Stress is important in many aspects of modern process flows; the use of strain to enhance device performance has become the norm in sub-90 nm technologies. Stress may be caused by stressor layers, thermal mismatch and lattice mismatch. In modelling [123], the strain tensor is split into *deviatoric* and *dilatational* parts, where the deviatoric part is related to deformation without volume change while the dilatational part corresponds to pure volume change.

$$\varepsilon_{jk} = \overbrace{\varepsilon'_{jk}}^{deviatoric} + \overbrace{\frac{1}{3}\sum_l \varepsilon_{ll}}^{dilatational}. \tag{3.8}$$

The visco-elastic model is used to model the relationship between stress and strain. Dirichlet and Neumann boundary conditions apply. The model is decomposed in terms of volumetric and shear components of the stress tensor. The dilatational part of stress tensor σ actually uses a purely elastic equation neglecting the visco-elastic term in equation (3.9)

$$\frac{\dot{\sigma}_v}{K} + \frac{\sigma_v}{\bar{\eta}(T,\sigma_s)} = 3\dot{\varepsilon}_v \tag{3.9}$$

$$\sum_k \sigma_{kk} = 3K\sum_k \varepsilon_{kk} = -3p = 3\sigma_v \tag{3.10}$$

where K is the bulk modulus related to Young's modulus, and p is the hydrostatic pressure. The deviatoric part of stress tensor σ is described by

$$\frac{\dot{\sigma}'_{jk}}{G} + \frac{\sigma'_{jk}}{\eta(T,\sigma_s)} = 2\dot{\varepsilon}'_{jk} \tag{3.11}$$

where G is shear modulus and $\bar{\eta}$ and η are all viscosity related functions of temperature and shear stress σ_s. Usually, as applying equations (3.9)-(3.11), oxide and nitride are treated as visco-elastic materials, and silicon as an elastic material for which the viscosity terms are negligible.

3.1.5. Three dimensional process simulation methodology

Due to the extreme scaling of contemporary MOSFETs, accurate modelling of their three-dimensional (3D) structure is necessary in order to obtain proper physical insights into their detailed operation. In addition, novel candidates for future field-effect transistors, such as the double gate finFETs currently under investigation as replacements for

conventional planar bulk MOSFETs, are essentially 3D in nature. Current TCAD tools have 3D process simulation capabilities for many physical process steps (such as implantation and diffusion), but not for all. Therefore a well thought-through 3D process simulation methodology must be considered and constructed to investigate MOSFET scaling. Figure 3.3 demonstrates a strategy to integrate 3D process capabilities using several structure modifications.

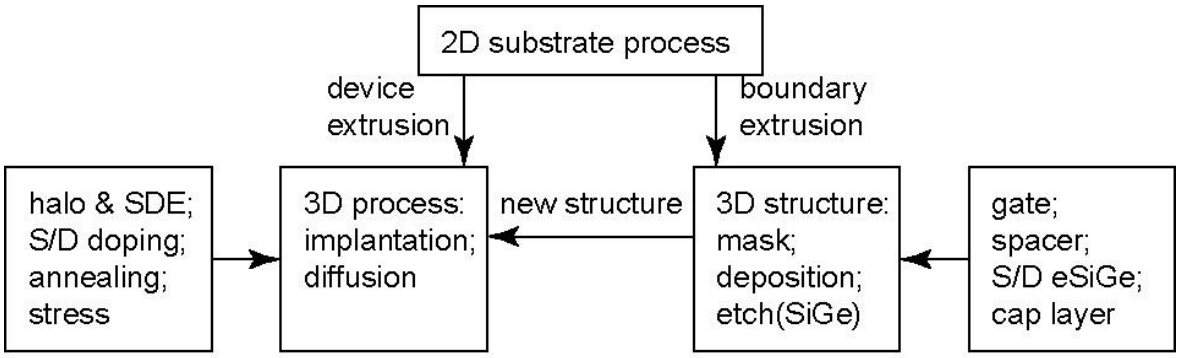


Figure 3.3 Schematic view of 3D process strategy. 3D structural changes such as deposition and etching which are used in 3D implantations and thermal processes .

Structures such as shallow trench isolation (STI) should be included into the simulation of narrow-width MOSFETs; such structures again require 3D simulation. In the 3D strategy illustrated above, STI formation and the substrate doping process is accomplished in 2D mode, then the device is extruded into the third dimension. Meanwhile its 2D boundary structure is imported into a structure editor. A 3D extrusion is followed by gate patterning, spacer formation, embedded SiGe and cap layer formation, with the previous structure saved, like a continuous snapshot, when a new structure is formed. These saved structures are fed back to 3D process simulator where previous data field are loaded and the structure undergoes new implantation and diffusion until a new structure is again imported if needed.

3.2. Device simulation

Modelling and simulating the physical phenomena and electrical characteristics of semiconductor devices provides comprehensive insights into the complex details of device operation, and allows reliable assessment of the potential impacts of physical effects and design optimisations.

3.2.1. Transport equations

The governing equations of semiconductor carrier transport include the Poisson equation and electron and hole continuity equations. The Poisson equation derives from Maxwell's equations and can be written in the form

$$\nabla \cdot (-\varepsilon \nabla \phi) = \rho = q\left(p - n + N_d^+ - N_a^-\right) + \rho_{trap}\,, \tag{3.12}$$

where ε is the permittivity; ϕ is the electrostatic potential; ρ is the charge density, q is the elementary electronic charge; n and p are electron and hole concentration; N_d^+ and N_a^- are the ionized donor and acceptor concentrations; and ρ_{trap} is the trap and fixed charge density. The Poisson equation is solved self-consistently with the current continuity equations for electrons and holes.

$$\begin{aligned} \nabla \cdot \vec{J}_n &= q(R_{net} + \frac{\partial n}{\partial t}) \\ \nabla \cdot \vec{J}_p &= -q(R_{net} + \frac{\partial p}{\partial t}) \end{aligned}, \tag{3.13}$$

which are deduced from $\nabla \cdot \vec{J} + \frac{\partial \rho}{\partial t} = 0$ obtained from Ampere's circuital law. In equations (3.13) $\vec{J}_n$ and $\vec{J}_p$ are current densities for electrons and holes and R_{net} is the net electron-hole recombination rate.

Depending on the complexity of semiconductor structure and the carrier transport behaviour, different transport models can be chosen, usually based on various approximations to the full Boltzmann transport equation. Models describe the current density with different degrees of complexity and may include self-heating and carrier energy transport. The Drift-diffusion model is one of the most simple approaches. It describes the current as a sum of two components describing the drift of carriers under the influence of the electric field and diffusion driven by concentration gradients.

$$\begin{aligned} \vec{J}_n &= qD_n \nabla n - q\mu_n n \nabla \phi = -qn\mu_n \nabla \Phi_n \\ \vec{J}_p &= -qD_p \nabla p - q\mu_p p \nabla \phi = -qp\mu_p \nabla \Phi_p \end{aligned} \tag{3.14}$$

where μ_n and μ_p is the electron and hole mobility; D is the diffusion coefficient which obeys *Einstein relation* when the system is close to thermal equilibrium namely $D = \mu \frac{kT}{q}$. Φ_n and Φ_p are electron and hole quasi-Fermi potentials described as,

$$\begin{aligned}\Phi_n &= \phi_i - \frac{kT}{q}\ln\left(\frac{n}{n_i}\right)\\ \Phi_p &= \phi_i + \frac{kT}{q}\ln\left(\frac{p}{n_i}\right)\end{aligned}. \tag{3.15}$$

Where n_i is the intrinsic electron density, and $\phi_i = -\frac{E_i}{q}$ is the electrostatic potential defined in terms of the intrinsic Fermi level E_i.

3.2.1.1. Numerical methods

Numerical methods are used to obtain solutions of the semiconductor equations described in the previous section for devices with realistic geometries and doping concentrations. The equations are discretised over a mesh or grid covering the device simulation domain using finite difference or finite element techniques, resulting in large system of algebraic equations. In the case of the box integration approach, which is a modification of the finite element approach, Gauss's theorem is used to transform the governing equations into the following integral form

$$\begin{aligned}&\varepsilon\int_{\partial V}\nabla\phi\cdot dA + \int_V \rho dV = 0\\ &\int_V\left(\frac{\partial n}{\partial t} + R_{net}\right)dV - \frac{1}{q}\int_{\partial V}\vec{J}_n\cdot dA = 0\\ &\int_V\left(\frac{\partial p}{\partial t} + R_{net}\right)dV + \frac{1}{q}\int_{\partial V}\vec{J}_p\cdot dA = 0\end{aligned}. \tag{3.16}$$

The discretised equations at grid element i of a volume V_i are finally written as in terms of electric field on current fluxes.

$$\begin{aligned}F_{\phi,i} &= \varepsilon\sum_j \frac{\phi_j - \phi_i}{d_{ij}}A_{ij} + \rho_i V_i = 0\\ F_{n,i} &= \left(\frac{\partial n}{\partial t} + R_{net}\right)V_i - \frac{1}{q}\sum_j J_{n,ij}A_{ij} = 0\\ F_{p,i} &= \left(\frac{\partial p}{\partial t} + R_{net}\right)V_i + \frac{1}{q}\sum_j J_{p,ij}A_{ij} = 0\end{aligned}. \tag{3.17}$$

The current densities are similarly discretised according to the neighbour grids. A single box in the mesh for the set of discretized equations is shown in Figure 3.4.

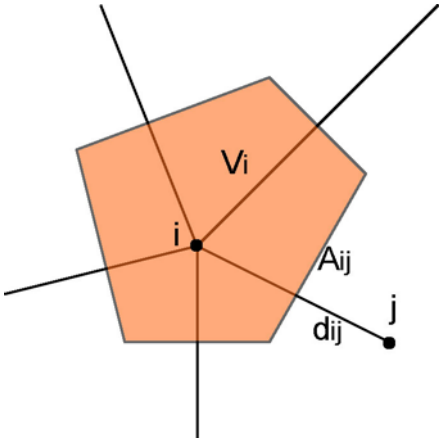


Figure 3.4 Schematic view of box discretization method in 2D. V_i is the box volume associated with gird i, A_{ij} is the interface area between boxes of i and j, and d_{ij} is the distance between i and j.

The set of algebraic equations obtained is nonlinear, and Newton or Gummel procedures are used for their linearization. The Gummel iteration method is well known for its stability [129]. The nonlinear Poisson's equation is solved first using an initial guess for the quasi-Fermi potential distribution, and the solution for the potential is substituted into continuity equations which are solved to obtain the carrier concentration. A new approximation for the quasi-Fermi levels is then substituted back to Poisson's equation. This is repeated until certain convergence criteria for the current are met. The Newton iteration method based on fixed point theory has a faster convergence [130]. Assuming that the vector of algebraic equations is $F=[F_\phi, F_n, F_p]^T$ and the solution variables are $W=[\phi, n, p]^T$, an approximation around the exact solution is obtained by a Taylor expansion [131],

$$F(W^k)+\frac{\partial F}{\partial W}\Delta W^k=0. \tag{3.18}$$

This results in a linear set of equations, solved using direct or iterative techniques. The nonlinear problem is solved by the following iterations.

$$W^{k+1}=W^k-J^{-1}F(W^k), \tag{3.19}$$

where the Jacobian matrix $J=\frac{\partial F}{\partial W}=\begin{pmatrix} \frac{\partial F_\phi}{\partial \phi} & \frac{\partial F_\phi}{\partial n} & \frac{\partial F_\phi}{\partial p} \\ \frac{\partial F_n}{\partial \phi} & \frac{\partial F_n}{\partial n} & \frac{\partial F_n}{\partial p} \\ \frac{\partial F_p}{\partial \phi} & \frac{\partial F_p}{\partial n} & \frac{\partial F_p}{\partial p} \end{pmatrix}$. The *Dirichlet* boundary condition reflects fixed values of the variable on the boundary grids, and the *Neumann* condition handles reflective boundaries with $\vec{J}\cdot\vec{n}=0$.

3.2.1.2. Density gradient quantum corrections

As discussed in preceded chapter, strong quantum confinement effects in the channel result in band splitting, and affect the carrier distribution in decananometer MOSFETs. The use of quantum corrections is required to accurately model carrier concentration, threshold voltage and gate capacitance. To calculate the quantum confinement effects a potential-like modification term Λ_n (for example for electrons) is added into classical electron density expression.

$$n = N_C F_{1/2}\left(\frac{E_{F,n} - E_C - \Lambda_n}{kT_n}\right) \tag{3.20}$$

where N_C is the effective density of electronic states of conduction band; $F_{1/2}$ is the complete Fermi-Dirac integral of order 1/2 here assuming Fermi-Dirac statistics rather than Boltzmann statistics; $E_{F,n}$ is the electron quasi-Fermi potential energy and E_C is conduction band edge energy. In the density gradient approximation Λ_n is expressed as [132][133],

$$\Lambda_n = -\frac{\gamma\hbar^2}{12m_n}\left\{\nabla^2 \ln n + \frac{1}{2}\left(\nabla \ln n\right)^2\right\} = -\frac{\gamma\hbar^2}{6m_n}\frac{\nabla^2\sqrt{n}}{\sqrt{n}} \tag{3.21}$$

where γ is a fitting parameter with default value 3.6 for electron; $\hbar$ is the reduced Planck constant; m_n is the electron effective mass; and $\nabla^2 = \nabla\cdot\nabla$ is a Laplacian operator. The similar quantum correction formalism is applied for holes with default value 5.6 for γ.

3.2.2.Mobility models

Different scattering mechanisms contribute to the carrier mobility. The carrier mobility may be obtained by combining different bulk mobilities μ_{b1}, μ_{b2}…and different surface mobilities μ_{s1}, μ_{s2}… according to their scattering contributions. This is done by applying Mathiessen's rule although which is not necessarily equal to the mobility [134].

$$\frac{1}{\mu} = \frac{1}{\mu_{b1}} + \frac{1}{\mu_{b2}} + \cdots \frac{1}{\mu_{s1}} + \frac{1}{\mu_{s2}} + \cdots . \tag{3.22}$$

At high field, the carrier velocity saturates and high-field mobility model needs to be taken into account based on the calculated low field mobility according to formula (3.22).

In an undoped lattice, the lattice vibrations dominate the scattering, and the phonon scattering related mobility model has the following form

$$\mu_{const} = \mu_L \left(\frac{T}{300K} \right)^{-\zeta} \tag{3.23}$$

indicating strong dependence on lattice temperature. Related parameters for electrons and holes are listed in Table 3.1. In doped bulk semiconductor, carrier scattering with ionized impurities further reduces the mobility. Masetti *et al.* proposed an empirical mobility model in doped semiconductors widely used in semiconductor device simulations [135],

$$\mu_{dope} = \mu_{\min 1} \exp\left(-\frac{P_c}{N_{tot}} \right) + \frac{\mu_{const} - \mu_{\min 2}}{1 + (N_{tot} / C_r)^{\alpha}} - \frac{\mu_1}{1 + (C_s / N_{tot})^{\beta}}, \tag{3.24}$$

here μ_{min1}, μ_{min2} and μ_1 are reference doping mobilities; P_c, C_r and C_s are reference doping concentrations; N_{tot} is the total donor and acceptor concentrations; α and β are fitting parameters. Their default values are listed in Table 3.2.

Table 3.1 Constant mobility model with default parameter values for Si.

Symbol	Electrons	Holes	Unit
μ_L	1417	470.5	cm^2/Vs
ζ	2.5	2.2	1

Table 3.2 Masetti doping-dependent mobility model with default parameter values for Si.

Symbol	Electrons	Holes	Unit
μ_{min1}	52.2	44.9	cm^2/Vs
μ_{min2}	52.2	0	cm^2/Vs
μ_1	43.4	29.0	cm^2/Vs
P_c	0	9.23×10^{16}	cm^{-3}
C_r	9.68×10^{16}	2.23×10^{17}	cm^{-3}
C_s	3.34×10^{20}	6.10×10^{20}	cm^{-3}
α	0.680	0.719	1
β	2.0	2.0	1

Another empirical expression for the doping-dependence form of mobility was proposed by Arora *et al.* [136] in which

$$\mu_{dope} = \mu_{\min} + \frac{\mu_d}{1 + (N_{tot} / N_0)^{A^*}}. \tag{3.25}$$

With $\mu_{\min} = A_{\min} \cdot (\frac{T}{300K})^{\alpha_m}$, $\mu_d = A_d \cdot (\frac{T}{300K})^{\alpha_d}$, and $N_0 = A_N \cdot (\frac{T}{300K})^{\alpha_N}$, $A^* = A_a \cdot (\frac{T}{300K})^{\alpha_a}$. The default values of fitting parameters are listed in Table 3.3.

Table 3.3 Arora doping-dependent mobility model with default parameter values for Si.

Symbol	Electrons	Holes	Unit
A_{min}	88	54.3	cm^2/Vs
α_m	-0.57	-0.57	1
A_d	1252	407	cm^2/Vs
α_d	-2.33	-2.23	1
A_N	1.25×10^{17}	2.35×10^{17}	cm^{-3}
α_N	2.4	2.4	1
A_a	0.88	0.88	1
α_a	-0.146	-0.146	1

The major difference in the Masetti model is the addition of the third term to the 'min-max' terms in Arora model to represent further mobility reduction at doping beyond 5×10^{19} (cm^{-3}). The mobility reduction rate with doping concentration is also somewhat different in the Masetti model.

Carriers in surface channel MOSFETs undergo scattering due to surface acoustic phonons and interface roughness. Lombardi *et al.* proposed a corresponding mobility model [137]

$$\mu_{ac} = \frac{B}{F_\perp} + \frac{C(N_{tot}/N_0)^\lambda}{F_\perp^{1/3}(T/300K)^k}, \tag{3.26}$$

$$\mu_{sr} = \left(\frac{(F_\perp / F_{ref})^{A^*}}{\delta} + \frac{F_\perp^3}{\eta} \right)^{-1}, \tag{3.27}$$

where the reference electric field F_{ref} =1V/cm and $F_\perp$ is the transverse electric field normal to semiconductor/insulator interface. Improved fitting A^* was obtained in [138] using

$$A^* = A + \frac{\alpha_\perp (n+p) N_{ref}^v}{(N_{tot}+N_1)^v}. \tag{3.28}$$

The surface mobility is combined with bulk mobility according to Mathiessen's rule as $\frac{1}{\mu} = \frac{1}{\mu_b} + \frac{D}{\mu_{ac}} + \frac{D}{\mu_{sr}}$ with $D = \exp(-x/l_{crit})$ where x is the distance from the interface. With N_{ref} =1 cm^{-3}, the reference and fitting parameter default values are listed in Table 3.4.

Table 3.4 Lombardi interface mobility model with default parameter values for Si.

Symbol	Electrons	Holes	Unit
B	4.75×10^{7}	9.925×10^{6}	cm/s
C	5.80×10^{2}	2.947×10^{3}	$cm^{5/3}V^{-2/3}s^{-1}$
N_0	1	1	cm^{-3}
λ	0.1250	0.0317	1
k	1	1	1
δ	5.82×10^{14}	2.0546×10^{14}	cm^2/Vs
A	2	2	1
$\alpha_{\perp}$	0	0	cm^3
N_1	1	1	cm^{-3}
v	1	1	1
η	5.82×10^{30}	2.0546×10^{30}	$V^2\ cm^{-1}s^{-1}$
l_{crit}	1×10^{-6}	1×10^{-6}	cm

A high-field mobility model encapsulates the low field model, a saturation velocity model and driving force model. Based on the original Caughey and Thomas model [139], Canali *et al.* proposed a high-field mobility model [140] in the following form,

$$\mu(F)=\frac{(\alpha+1)\mu_{low}}{\alpha+\left[1+\left(\frac{(\alpha+1)\mu_{low}F_{hfs}}{v_{sat}}\right)^{\beta}\right]^{1/\beta}} \tag{3.29}$$

where μ_{low} is the low-field mobility calculated by Mathiessen's rule, and $\beta=\beta_0(T/300K)^{\beta_{\exp}}$ is a factor exponentially dependent on temperature. v_{sat} and $F_{hfs,n/p}$ are saturation velocity and driving force electric field for electrons or holes, calculated by following formulas, in which $\Phi_{n/p}$ is electron/hole quasi-Fermi potential,

$$v_{sat}=v_{sat,0}\left(\frac{300K}{T}\right)^{v_{sat,\exp}}, \tag{3.30}$$

$$F_{hfs,n/p}=|\nabla\Phi_{n/p}|. \tag{3.31}$$

The reference and fitting parameter default values are listed in Table 3.5.

Table 3.5 Canali high-field mobility model with default parameter values for Si.

Symbol	Electrons	Holes	Unit
β_0	1.109	1.213	1
β_{exp}	0.66	0.17	1
α	0	0	1
$v_{sat,0}$	1.07×10^{7}	8.37×10^{6}	cm/s
$v_{sat,exp}$	0.87	0.52	1

3.2.3.Modelling stress-dependent mobility

Mechanical stress deforms the semiconductor crystalline lattice, resulting in alternations to energy band structure and carrier mobility. Bardeen and Shockely modelled this effect in terms of deformation potentials [141], and Smith proposed the piezoresistance model to express stress related conductance changes [96]. Accurate modelling of the impact of stress on mobility is critically important when stress engineering is intentionally employed in small geometry CMOS devices.

Generally, the stress tensor $\bar{\sigma}$ is a 3×3 matrix, and so is the strain $\bar{\varepsilon}$. The independent six components of stress may be written as a vector applying the transformation: $\sigma_{11}\rightarrow\sigma_1$, $\sigma_{22}\rightarrow\sigma_2$, $\sigma_{33}\rightarrow\sigma_3$ $\sigma_{23}\rightarrow\sigma_4/2$, $\sigma_{13}\rightarrow\sigma_5/2$, $\sigma_{12}\rightarrow\sigma_6/2$. The same applies to the strain.

Focusing on the conductivity changes due to small stresses, the piezoresistance mobility model suggests a linear relationship between electron or hole currents for strained and unstrained materials [142],

$$\begin{aligned} \bar{\mu}_\alpha &= \mu_\alpha^0(\bar{I} - \bar{\Pi}^\alpha \cdot \bar{\sigma}) \\ \vec{J}_\alpha &= \bar{\mu}_\alpha \cdot \left(\frac{\vec{J}_\alpha^0}{\mu_\alpha^0}\right) \end{aligned} \tag{3.32}$$

where $\bar{\mu}_\alpha$ is the mobility tensor of rank 2 (3×3 matrix) in the presence of stress while μ_α^0 is the isotropic scalar mobility without stress effects; $\bar{\Pi}^\alpha$ is the piezoresistance coefficient tensor and $\bar{I}$ is identity tensor; $\vec{J}_\alpha^0$ is the carrier current density without stress. α differentiates between electrons and holes. In cubic symmetry crystals such as Si the coefficient tensor can be simplified by rotating the coordinate system to become parallel to crystal axes, resulting in only three independent components. In addition, the piezoresistance coefficients are dependent on doping concentration and temperature which is related to anisotropic scattering [142][143], represented by a multiplicative doping and temperature dependent factor $P_\alpha(N,T)$, compared to the effective mass change which is represented by a constant term.

$$\bar{\Pi}^\alpha = \begin{pmatrix} \Pi_{11}^\alpha & \Pi_{12}^\alpha & \Pi_{12}^\alpha & 0 & 0 & 0 \\ \Pi_{12}^\alpha & \Pi_{11}^\alpha & \Pi_{12}^\alpha & 0 & 0 & 0 \\ \Pi_{12}^\alpha & \Pi_{12}^\alpha & \Pi_{11}^\alpha & 0 & 0 & 0 \\ 0 & 0 & 0 & \Pi_{44}^\alpha & 0 & 0 \\ 0 & 0 & 0 & 0 & \Pi_{44}^\alpha & 0 \\ 0 & 0 & 0 & 0 & 0 & \Pi_{44}^\alpha \end{pmatrix}, \tag{3.33}$$

$$\Pi_{ij}^{\alpha} = \Pi_{ij,\mathrm{var}}^{\alpha} P_{\alpha}(N,T) + \Pi_{ij,const}^{\alpha}$$
$$P_{\alpha}(N,T) = \frac{300K}{T} \frac{F_0'(E_{F,\alpha}/kT)}{F_0(E_{F,\alpha}/kT)}. \quad (3.34)$$

where $F_0(x)$ and $F_0'(x)$ are the Fermi integral of order 0 and its derivative; $E_{F,\alpha}$ is the energy of $E_F - E_C$ for electrons or $E_V - E_F$ for holes. The default values for these parameters are listed in Table 3.6.

Table 3.6 Piezoresistance coefficients for holes and electrons.

$\Pi_{11,const}^{p}$	$\Pi_{12,const}^{p}$	$\Pi_{44,const}^{p}$	$\Pi_{11,\mathrm{var}}^{p}$	$\Pi_{12,\mathrm{var}}^{p}$	$\Pi_{44,\mathrm{var}}^{p}$	Unit
5.1×10^{-11}	-2.6×10^{-11}	2.8×10^{-10}	1.5×10^{-11}	1.5×10^{-11}	1.1×10^{-9}	Pa^{-1}
$\Pi_{11,const}^{n}$	$\Pi_{12,const}^{n}$	$\Pi_{44,const}^{n}$	$\Pi_{11,\mathrm{var}}^{n}$	$\Pi_{12,\mathrm{var}}^{n}$	$\Pi_{44,\mathrm{var}}^{n}$	Unit
0	0	0	-1.026×10^{-9}	5.34×10^{-10}	-1.36×10^{-10}	Pa^{-1}

Intel proposed a physically-based analytic model for the hole mobility in the presence of stress [146]. It focuses on heavy hole valence band changes, which accounts for ~75% of whole hole concentration, in the 2D k-space, due to stress. The heavy hole valence band without stress is modelled by two degenerate ellipsoids aligned to the <110> and <-110> directions as illustrated in Figure 3.5 (top view).

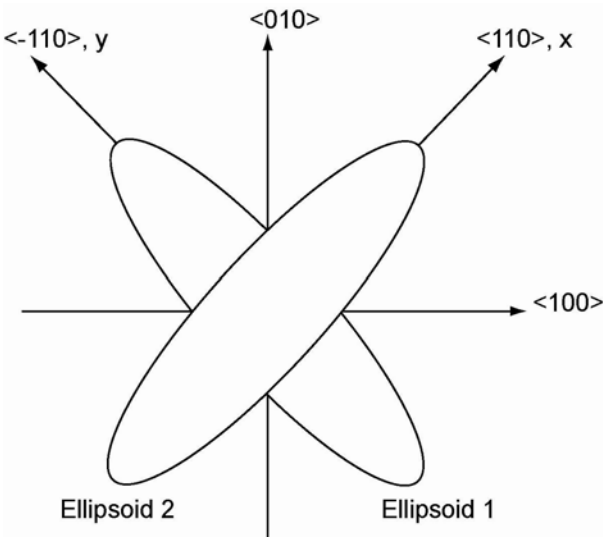


Figure 3.5 The simplified heavy hole 2D band structure without stress. Two equivalent ellipsoids are depicted in 2D treatment.

Without stress, the fraction of hole occupation in each ellipsoid is 0.5, furthermore due to equivalence of the two ellipsoids, the anisotropic scalar mobility corresponding to an applied electric field is given by the following expression

$$\mu_0 = q\langle\tau\rangle\left(\frac{0.5}{m_{t0}}+\frac{0.5}{m_{l0}}\right). \tag{3.35}$$

Where $\langle\tau\rangle$ is the mean-free time of scattering and m_{t0} and m_{l0} are the transverse and longitudinal masses in a system free of stress. In the presence of stress, these two ellipsoids will separated by energy ΔE_{strain}, given by equation (3.40) below. The ellipsoids are expressed in equations (3.36). For example if a uniaxial compressive stress is assumed along <110>, the ellipsoid 1 will expand and the ellipsoid 2 will shrink, leading to a higher population of holes redistributed into ellipsoid 1[144][145]. The mobility component along <110> will increase due to the rich population of holes with small transverse effective mass in ellipsoid 1 [146]. If a field is applied along <110> the result will be a current enhancement.

$$\begin{aligned} E_2 &= \frac{\hbar^2}{2}\left(\frac{k_x^2}{m_{l2}}+\frac{k_y^2}{m_{t2}}\right)+\frac{\Delta E_{strain}}{2} \\ E_1 &= \frac{\hbar^2}{2}\left(\frac{k_x^2}{m_{t1}}+\frac{k_y^2}{m_{l1}}\right)-\frac{\Delta E_{strain}}{2} \end{aligned}. \tag{3.36}$$

In equation (3.36), m_{t1} and m_{l1} are the transverse and longitudinal mass of ellipsoid 1 in the presence of stress, and m_{t2} and m_{l2} are the corresponding masses in ellipsoid 2. Therefore the principal axis of mobility is readily calculated according to formula in equation (3.37) assuming hole redistribution with fraction f_1 in ellipsoid 1 and $f_2 = 1 - f_1$ in ellipsoid 2. For an in-plane electric field applied at angle θ from the principal axis <110>, the scalar in-plane mobility is given in formula (3.38).

$$\begin{bmatrix}\mu_{110}\\ \mu_{-110}\end{bmatrix} = 2\mu_0\frac{m_{t0}m_{l0}}{m_{t0}+m_{l0}}\begin{bmatrix}\frac{f_1}{m_{t1}}+\frac{f_2}{m_{l2}}\\ \frac{f_2}{m_{l1}}+\frac{f_2}{m_{t2}}\end{bmatrix}, \tag{3.37}$$

$$\mu = 2\mu_0\frac{m_{t0}m_{l0}}{m_{t0}+m_{l0}}\left[\cos^2\theta\left(\frac{f_1}{m_{t1}}+\frac{f_2}{m_{l2}}\right)+\sin^2\theta\left(\frac{f_1}{m_{l1}}+\frac{f_2}{m_{t2}}\right)\right]. \tag{3.38}$$

The hole population may be obtained using Maxwell-Boltzmann statistics. An approximation is used of the form

$$f_1 = \frac{\exp\left(\frac{\Delta E_{strain}}{2kT}\right)}{\exp\left(\frac{\Delta E_{strain}}{2kT}\right)+\exp\left(-\frac{\Delta E_{strain}}{2kT}\right)} = \frac{1}{1+\exp\left(-\frac{\Delta E_{strain}}{kT}\right)}. \tag{3.39}$$

For semiconductor devices with in-plane stress of the form $\bar{\varepsilon} = \begin{pmatrix} b+a & s \\ s & b-a \end{pmatrix}$ with b the biaxial stress, a the asymmetric stress and s the shear stress, the band structure deformation can be formulated by the powers of stress tensor components.

$$\begin{aligned} \Delta E_{strain} &= d_1 s \\ \frac{1}{m_{t1}} &= \frac{1}{m_{t0}}(1 - s_{t1}s + s_{t2}s^2 + b_{t1}b + b_{t2}b^2) \\ \frac{1}{m_{t2}} &= \frac{1}{m_{t0}}(1 + s_{t1}s + s_{t2}s^2 + b_{t1}b + b_{t2}b^2), \\ \frac{1}{m_{t(001)}} &= \frac{1}{m_{t0}}(1 + b_{tt}b) \end{aligned} \tag{3.40}$$

The longitudinal masses of ellipsoids are assumed to be independent of stress, namely $m_{l1} = m_{l2} = m_{l0}$. The fitting parameters for equation (3.40) are listed in Table 3.7. $m_{l0} = 0.48$ and $m_{t0} = 0.15$ relative to hole mass.

Table 3.7 Intel physically-based stress-dependent hole mobility model parameters.

d_1[eV/Pa]	s_{t1}[1/Pa]	s_{t2}[1/Pa^2]	b_{t1}[1/Pa]	b_{t2}[1/Pa^2]	b_{tt}[1/Pa]
-6e-11	-9.442e-10	4.3066e-19	-1.0086e-10	6.5886e-21	1.2e-10

The Sentaurus TCAD tool generalizes the original Intel 2D plane hole mobility model to be applicable to 3D cases, with doping dependence and carrier redistribution considered among more than 2 ellipsoids. In 3D, three planes with six ellipsoids are considered independently. The stress tensor corresponding to each plane is selected from the 3D stress tensor. The mobility in the direction of [001] perpendicular to the plane has also been modified by $\mu_{001} = \mu_0 m_{t0}/m_{t(001)}$ according to the $m_{t(001)}$ calculated in the formula (3.40). If the number of ellipsoids considered in the simulation is larger than the default value 2, the distribution in each ellipsoid is adjusted.

3.3. The Glasgow 'atomistic' simulator

The aggressive scaling of MOSFETs exposes the inherently granular nature of materials and individually discrete nature of atoms which are usually as averaged in any large-scale device treatment. When the individual behaviour of atoms, molecules, and charges becomes prominent in determining particular semiconductor device behaviour, 3D statistical device modelling is required to reflect the corresponding variability in the device characteristics. The role of the device modelling changes from obtaining the behaviour of

individual devices to the prediction of the statistical behaviour of an ensemble of microscopically different transistors.

The discrete random variation of doping, gate length roughness, and random granularity of gate stack and work function are main sources of variability in MOSFET characteristics. The predictive modelling of these effects is of great importance for both device and circuit design.

3.3.1. Random discrete dopants (RDD)

The average number of dopants in the active channel region of nanoscale MOSFETs is less than hundred and fluctuations in this number give rise to variations in MOSFET characteristics. Incident impurity dopants implanted into a semiconductor substrate undertake a series of random collisions with host lattice atoms until coming to rest, thus resulting in random positions of the individual discrete dopants in the transistors. Ionized impurity dopants positioned at these lattice sites create discrete Coulomb potential peaks in the channel of small dimension MOSFET. The corresponding potential fluctuations affect the current flow and they are the origin of further variations in the electrical characteristics from device to device.

In numerical simulations, a semiconductor device is discretized by a mesh. How to assign random discrete dopants to the mesh and how to model discrete dopants related device characteristics are important issues. DD simulations deal with charge densities rather than point charges. Therefore, the point charge of each dopant has to be converted into a charge density and appropriately assigned to the mesh. Two methods are used in the introduction of random discrete dopants into DD simulations. The first employs Monte Carlo (MC) implantation and atomistic kinetic Monte Carlo (KMC) diffusion from the initial process simulation of the transistor [123][147]. As already mentioned in the section 2, the implanted ions undergo nuclear and electronic collisions until they lose their extra kinetic energy and come to rest. The MC simulation of the implantation process follows the random collisions and trajectories of the individual implanted dopants. The reactions between charged or neutral dopants, point defects, clusters occur during the thermal diffusion. Kinetic MC diffusion models the interactions between dopants and defects in the diffusion process in Sentaurus Process KMC simulator DADOS. The results of the atomistic scale process simulation are individual random positions of dopants in the device. The dopants are assigned to the mesh using the Sano approximation [148].

The second method of introducing random discrete dopants is based on the continuous doping profile obtained from conventional process simulation, and this method is widely used in treating random discrete dopant induced intrinsic parameter fluctuations such as threshold voltage variation [86][32][149]. A standard method to assign discrete dopants according to the continuous concentration involves visiting, one by one, each sites of the Si lattice covering the device [149]. The probability of a dopant being located in a small volume cell ΔV_i is determined by the local doping concentration $N(x_i,y_i,z_i)$ with the form of

$$p_i = N(x_i, y_i, z_i)\Delta V \tag{3.41}$$

where $\Delta V=a_{si}^3/8$ (=0.020022 nm^3) is the volume associated with each atom. A Von Neumann rejection technique is used to decide whether to introduce a dopant at the Si site. If an evenly distributed random number between 0 and 1 is less than probability calculated as formula (3.41), then a dopant is placed in the lattice site. Finally, the discrete dopants are transformed into doping density assigned to the neighbouring mesh nodes using the charge assignment scheme. The schematic view of the mesh cell containing a discrete dopant is illustrated in Figure 3.6.

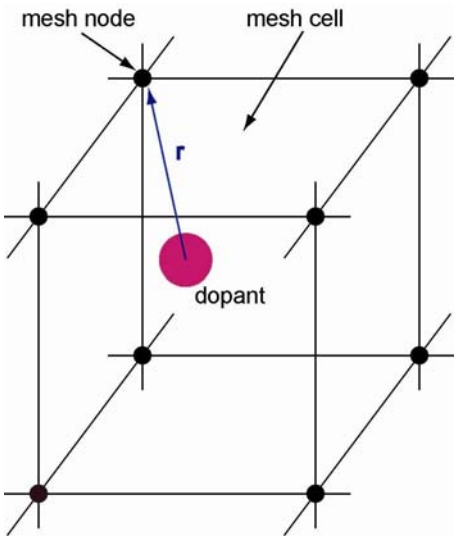


Figure 3.6 Schematic view of a mesh cell containing a discrete dopant. The point charge of dopant is assigned to those neighbouring nodes according to various charge assignment schemes.

Many charge assignment schemes are available. For example, the dopant charge can be assigned to the nearest grid point; the cloud-in-cell (CIC) linearly spreads point charge to mesh nodes based on the coordinate distances between the charge and the nodes; an assignment based on Gaussian distribution smears the charge across many grid nodes according to the chosen standard deviation [33].

Sharp Coulomb potentials appear in the solution of the Poisson equation due to the highly localized doping concentrations related to discrete dopant assignments. In DD simulations this results in artificial charge trapping wells which in physical reality would be prevented due to quantisation. To avoid this problem, attempts are made to split the Coulomb potential into long-range and short-range parts [150][151]. Density gradient quantum corrections are the physical way to avoid artificial carrier trapping. At mesh sizes below 1.0 nm, the effective quantum potential results in approximately 50mV conduction band edge lowering and the corresponding carrier density becomes independent of mesh spacing [33].

3.3.2.Line edge roughness (LER)

The granularity of the photoresist molecules defines the photoresist edges in a lithographic process. The corresponding line edge roughness is transferred to gate line edge by gate etching when the photoresist is used as mask. LER occurs due to slower dissolution rate of large polymer aggregates in the removal process of undesirable photo-resist [152][153]. The rough edge of photoresist is transferred to physical patterned lines on the semiconductor, such as polysilicon gate patterning. This process is illustrated in Figure 3.7. Therefore, poly-silicon gates have inherent line edge roughness (LER) and line width roughness (LWR), which today is on the same scale of the actual device dimensions.

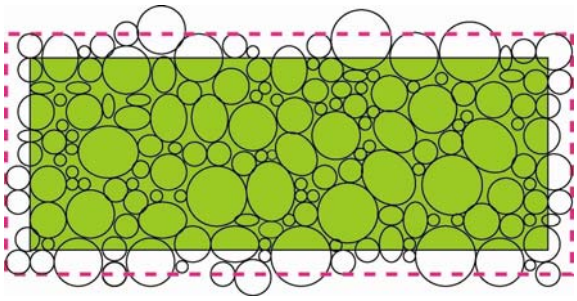


Figure 3.7 Schematic description of photoresist defined line edge roughness in lithography and etching. Dashed and solid lines represent the printed and physical lines respectively; the circles indicate the polymer segregates of photoresists.

As described in detail in Chapter 2, the LER can be characterized by a correlation function with two critical parameters: standard deviation and correlation length. In our simulations random gate edges are generated using the methodology described in [154][155]. This is based on a Fourier synthesis approach using the power spectrum of a Gaussian or exponential autocorrelation function.

$$S_G(k) = \sqrt{\pi}\Delta^2\Lambda \exp\left(-k^2\Lambda^2/4\right), \tag{3.42}$$

$$S_E(k) = \frac{2\Delta^2\Lambda}{1+k^2\Lambda^2}, \tag{3.43}$$

where Δ and Λ are RMS amplitude and correlation length respectively. In the inverse Fourier synthesis process that generates the random edge shape, the magnitude of the complex array is generated based on equation (3.42) or (3.43) and the phases are generated randomly.

3.3.3. Poly-silicon granularity (PSG)

Poly-silicon has been used as a gate material in MOSFETs. Grain boundaries in the polysilicon enhance dopant diffusion relieving poly-depletion effect [156]. Fermi level pinning in the Si band-gap occurring at poly-silicon grain boundaries locally alters the threshold voltage and lowers drive current, adding another source of statistical variability.

Poly-silicon Fermi level pinning has been investigated in [93][94][95] and is illustrated in Figure 3.8. Due to mismatch of neighbouring grain lattices, interface states are created at the grain boundary, resulting in carrier trapping. A depletion space charge region is created on either side of the boundary, enabling charge neutrality. The potential barrier induced by interface trapped charge causes surface potential fluctuations in the channel affecting the local threshold voltage. However, the lack of donor-type traps at boundary interfaces of p+ doping poly-silicon leads to negligible Fermi pinning in p-channel MOSFETs [157].

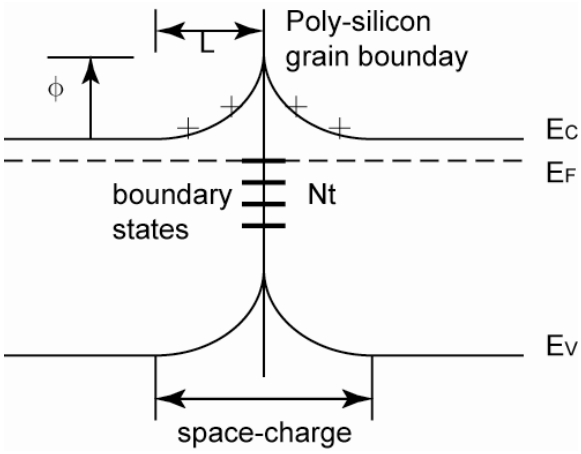


Figure 3.8 Band diagram view of poly-silicon grain boundary states induced Fermi pinning.

Assuming that the trapped charge sheet density is N_t (cm^{-2}) with consistent polysilicon doping concentration of N_p (cm^{-3}) on both sides of boundary, the Fermi pinning potential height can be obtained by applying Poisson's equation

$$\phi = -\frac{qN_pL^2}{2\varepsilon} = -\frac{qN_t^2}{8\varepsilon N_p}. \tag{3.44}$$

In ‘atomistic’ simulations the poly-silicon grain boundaries are generated from a large experimental AFM image of the poly-silicon structure, and the simulator randomly selects poly-silicon pieces trimmed out of the AFM image for simulated gates [34]. The Fermi pinning level at boundary is set 0.3eV below the Si conduction band edge, (then N_t is about 4×10^{13} cm^{-2} while N_p is 1×10^{20} cm^{-3}). The simulation results reveal that the polysilicon boundary results in an increase in the average threshold voltage, and a higher pinning barrier results in larger threshold voltage shift. In addition, the randomness of grain boundaries adds to the statistical variability of device parameters, and the threshold voltage variation is increased [34].

In summary, the salient details of the TCAD process and device simulators used in this work have been discussed. The Glasgow ‘atomistic’ simulator has also been described, with emphasis on its ability to accurately model statistical variability sources. The process steps replicated in these simulation tools have the capability to accurately model the relevant physics of the latest CMOS technology features and therefore enhance the optimal design of modern and scaled CMOS devices. The drift-diffusion device simulation with density gradient quantum corrections can provide the fundamental understanding of the scaling behaviour of electrical characteristics in scaled devices. The advanced statistical variability simulator enables the investigation of the in-depth microscopic behaviour of individual devices and the extraction of statistical information on the device parameters of an ensemble of devices.

Chapter IV

4. CMOS device design and characterization

MOSFET scaling has always been the driving force of the semiconductor industry delivering more functionality, higher speed, and lower cost per function. The study of CMOS scaling is critically important in predicting trends in performance and exploring potential technical barriers of future technology generations, which essentially would promote the foreseen feasible solutions and pave the road for next generation technologies. The scaling study is a foundation for the research conducted in the rest of this thesis.

4.1. Calibration

The scaling work starts with careful calibrations in respect of real fabricated deep deca-nanometer CMOS transistors, including physical dimensions, doping profiles and ultimately electrical characteristics. The test bed semiconductor devices are 35 nm gate length transistors published by Toshiba in 2002 [11]. At the time of publication the devices had the best reported I_{off} and I_{on} values. The devices have 35 nm physical gate length, NO gas annealed gate oxide, retrograde and super-halo doping profile, and low thermal process S/D extension shallow junctions. This calibration aims to closely replicate these features in TCAD process simulation and to match consistently the measured electrical characteristics using device simulation. Initial calibration was already carefully carried out using a now deprecated TCAD tool [158]. However, the calibration was updated employing the latest Sentaurus TCAD simulator in order to embed more capabilities and to reflect the latest technology features of modern CMOS devices.

4.1.1. Extraction of the real device structure

The benchmark MOSFETs described in this subsection will be examined in detail in terms of critical physical dimensions, process steps and essential electrical properties. The published transmission electron microscopy (TEM) photograph of the fabricated 35 nm gate length MOSFET structure is illustrated in Figure 4.1. The poly-gate thickness and lateral distance of S/D contacts to gate are measured. In terms of gate oxide thickness, an

initial 1.0–1.2 nm gate oxide was formed, and then annealed by nitric oxide (NO) ambient to form a silicon oxynitride SiO_xN_y gate dielectric. The 35 nm poly-Si gate is selectively etched by RIE. Shallow S/D junction extensions are formed by low energy implantation (1.0~1.5 keV for arsenic, less than 0.3 keV for boron) and followed by rapid thermal annealing (RTA). The structural data obtained from direct indication and measurement in the publication is summarized in Table 4.1.

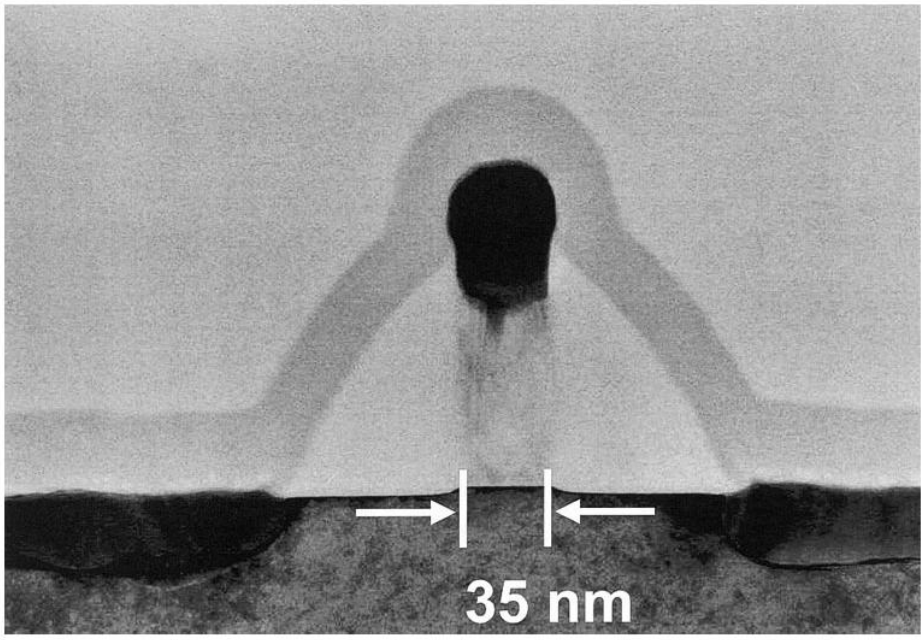


Figure 4.1 TEM photograph of 35 nm gate length Toshiba MOSFET. Reprinted with permission from Inaba *et al.*, "High performance 35 nm gate length CMOS with NO oxynitride gate dielectric and Ni salicide," *IEEE Trans. Electron Devices*, Vol.49 No.12, (© 2002 IEEE).

Table 4.1 Physical dimensions of Toshiba 35 nm gate length MOSFET (* measured)

	Gate length L_{gate} [nm]	Oxide thickness t_{ox} [nm]	Junction depth x_j [nm]	PolySilicon thickness* [nm]	Distance of S/D to gate* [nm]
nFET	35	1-1.2	20	140~150	~55
pFET	35	1-1.2	33	140~150	~55

The process features a retrograde channel doping profile and super-halo doping to suppress punch-through and SCE effects. The processing split experiments were carried out to achieve the optimal balancing between low threshold voltage V_{th} and suppressed SCE, using a wide range of implantation doses and energies. In general the application of a well designed halo process can significantly improve V_{th} roll-off characteristics. The

information of split tests about doping dose and energy is presented in Table 4.2 including channel doping, halo doping and S/D doping.

Table 4.2 Process doping implantation information of 35 nm gate length Toshiba MOSFETs.

	Retrograde doping dose [ions/cm^2] /energy [keV]	Halo doping dose [ions/cm^2] /energy [keV]	S/D extension doping dose [ions/cm^2] /energy [keV]
n-MOSFET	~1-5×10^{13} / 60-150	<3×10^{13} / unavailable	Unavailable / 1.0-1.5
p-MOSFET	~1-4×10^{13} / 80-120	~1-3×10^{13} / 40-50	Unavailable / <0.3

4.1.2. Calculations of doping profiles

The components of the complex 2D non-uniform doping profile includes vertical retrograde channel doping, lateral abrupt super-halo doping, and ultra-shallow S/D extensions formed by extremely low implantation energy and RTA. Figure 4.2 presents the measured experimental data of S/D extension arsenic doping concentration and channel indium doping concentration in the n-MOSFET. Focusing firstly on the channel doping profile, the total dose of indium dopants within depth of $x_0 = 0.15$ μm can be calculated according to the measured doping concentration by integration.

$$D = \int_0^{x_0} N(x)dx . \qquad (4.1)$$

The calculated total indium dose is 3.55×10^{13} ions/cm^2. The probability density can be described as $p(x)=\Delta D(x)/D$, allowing the calculation of the doping centroid, which is closely related to ion implantation projected range but also affected by followed thermal diffusions. We use the definition of the projected range to describe this mean value.

$$R_p = \int xp(x)dx . \qquad (4.2)$$

The calculated centroid of indium dopants is positioned at depth 68.6 nm from substrate interface.

It is clear that the channel doping profile is formed by multiple implantations due to the flat tails on both sides of the maximum. However it possesses a primary implantation to determine the projected range and initial portrait. Separating the near-surface doping profile and the prolonged deep tail, the main fraction of the doping profile is located in the range from 35 nm to 100 nm. Although affected by implantation and diffusion from both

sides, the initial implantation guess can be based on this central region of main doping distribution. Calculations based on this region yield a projected range 69 nm, the dose within this region of 2.66×10^{13} ions/cm^2, standard deviation of 14.6 nm, skew of 63 nm, and kurtosis of 2.41. The extracted parameters provide an initial guess for replicating the channel doping profile at process simulation.

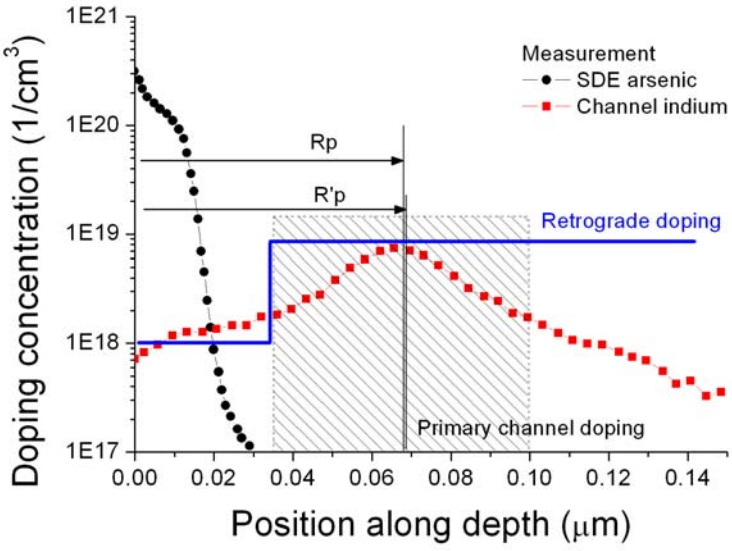


Figure 4.2 Experimental doping profiles of 35 nm gate length Toshiba n-MOSFET.

The features of the retrograded doping profile are outlined in Figure 4.2, having a low doping concentration near surface and high doping concentration beyond the depth of S/D extension junction. The low doping region is within the channel depletion region, providing low threshold voltage. The high doping region controls the deep S/D related punch-through leakage current. The retrograde doping profile is formed by the primary implantation, in which projected range tunes the retrograde position while dose and standard deviation are used to control retrograde step height.

4.1.3. Calibration methodology

A systematic simulation calibration is carried in respect of the test bed Toshiba MOSFETs to provide a realistic device structure and current-voltage characteristics as a starting point for the design of the scaled devices in this chapter. The systematic calibration includes the following steps.

Process simulation

1 Introduce the structures of MOSFETs and match retrograde channel doping profile and SDE doping profile in process simulation;

2 Tune the well doping and the halo doping to reduce benchmarking error;

Device simulation

3 Tune mobility models;

4 Match electrical characteristics: I_d-V_g and I_d-V_d in device simulation.

The flowchart of the calibration process is summarized in Figure 4.3. Firstly the simulation project matches the published process details of real structure as accurately as possible. It includes the structural parameters listed in Table 4.1, process flow steps such as ion implantation with specific energy and dose indicated in Table 4.2 and spike RTA. Closely matching retrograde channel doping, the calibration process uses the margin of well doping and halo doping for process simulation flexibility depending on the device simulation error.

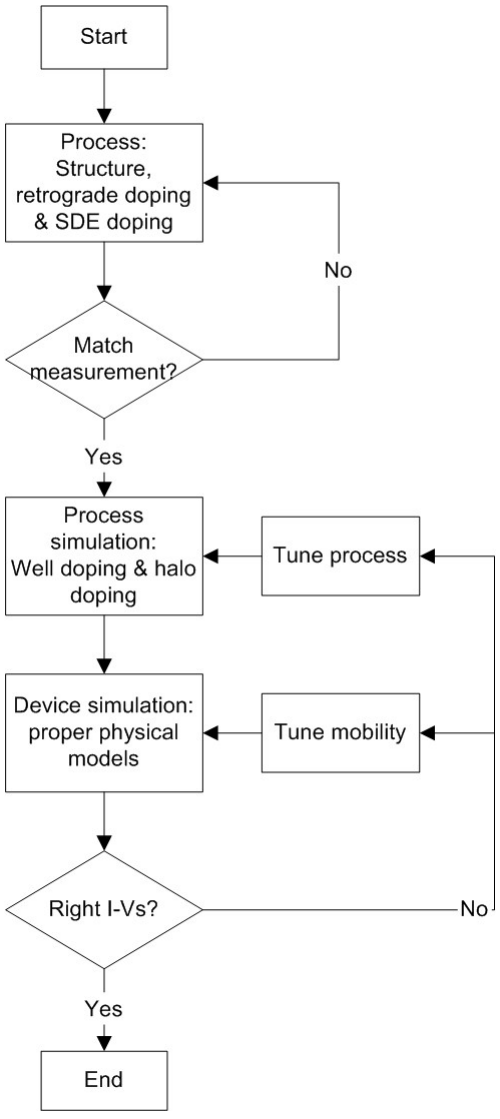


Figure 4.3 Simplified flowchart of systematic simulation calibration methodology

After the first stage, the calibration process focuses on tuning mobility parameters. First the device electrostatics are adjusted using low and high drain voltage biased I_d-V_g characteristics in the sub-threshold regime. Thus, low field mobility is tuned to match the low drain voltage I_d-V_g characteristics. Secondly high field mobility is tuned to the high

drain voltage I_d-V_d characteristics. In simulation the low field mobility tuning is primarily based on phonon scattering and impurity scattering mobility models. The high field mobility tuning adjusts the saturation velocity and the critical field in the field dependent mobility models.

All following simulations are carried out with the commercial TCAD package Sentaurus from Synopsys [Version A-2007.12].

4.1.4. Calibration results

4.1.4.1. Process calibration

The calibration starts with the design of an appropriate Sentaurus input process simulation file reflecting the structural and processing data from subsection 4.1.1. The gate dielectric is silicon oxynitride SiO_xN_y with 1.2nm thickness and relative permittivity 5.45 according to mole fractions of SiO_2 and Si_3N_4. The poly-Si thickness is 150 nm while the distance of S/D contacts to gate is 52 nm. In simulations, the source and drain electrodes are treated as Ohmic contacts. The resulting ultra-shallow SDE junction depth is 20 nm for the n-MOSFET and 28 nm for the p-MOSFET. The calibrated structure is illustrated in Figure 4.4.

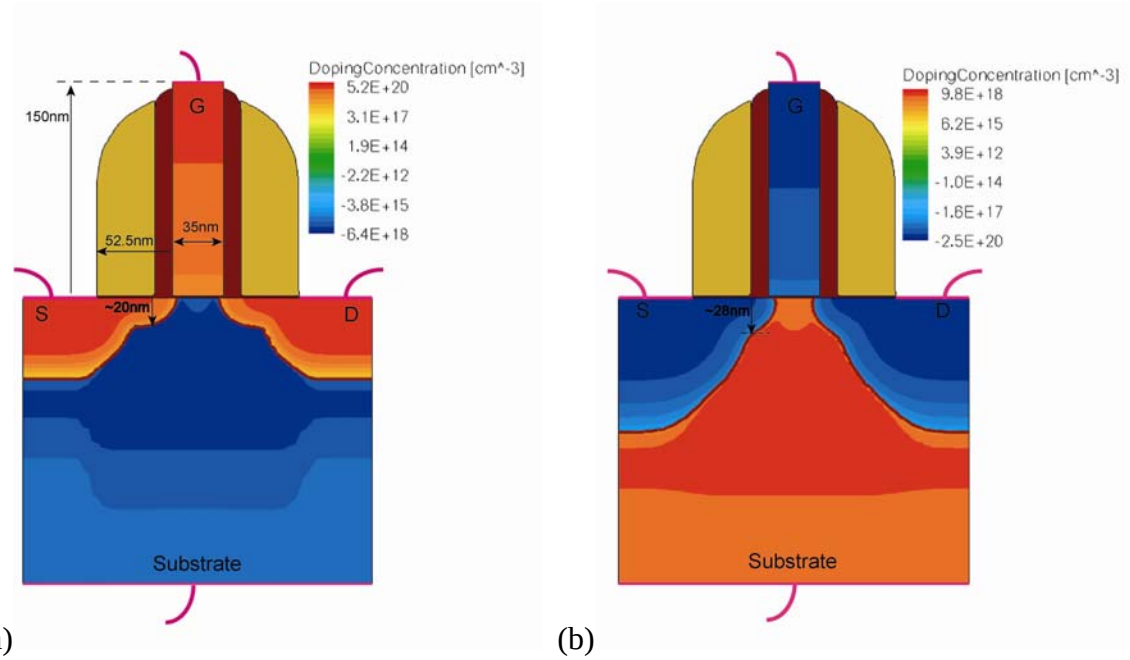


Figure 4.4 Simulation structures of 35 nm gate length n-MOSFET (a) and p-MOSFET (b) based on Toshiba experimental data.

Relying on the analysis of the retrograde doping profile of the n-MOSFET, the primary Indium implantation is applied with dose 3.2×10^{13} ions/cm^2 and energy 115 keV. Two

additional implantations with small doses are used to achieve the final doping distribution. The source/drain extensions are formed by extremely low implantation energy 1 keV and low thermal budget of 4 seconds RTA of 900°C. The RTA thermal process reduces the halo doping diffusion away from source or drain corners. The comparison of simulation and experimental data in terms of channel indium and SDE arsenic is shown in Figure 4.5. The retrograde indium and the ultra-shallow SDE doping profiles accurately match the measured data. The abruptness of SDE ~2.9 nm/dec is achieved.

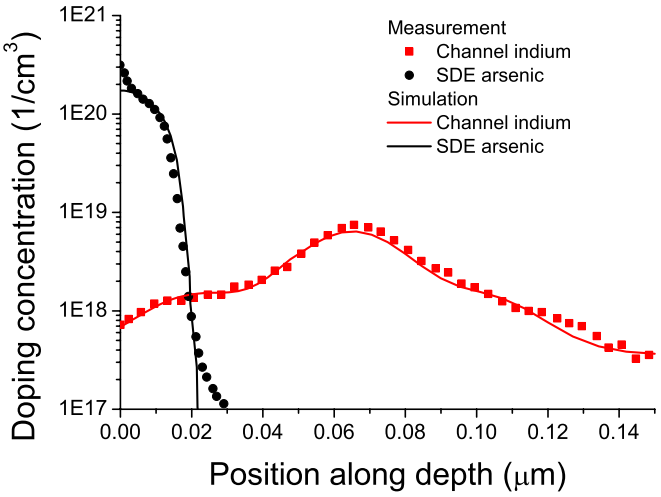


Figure 4.5 Calibrated Channel retrograde indium and SDE abrupt arsenic doping profiles in n-MOSFETs.

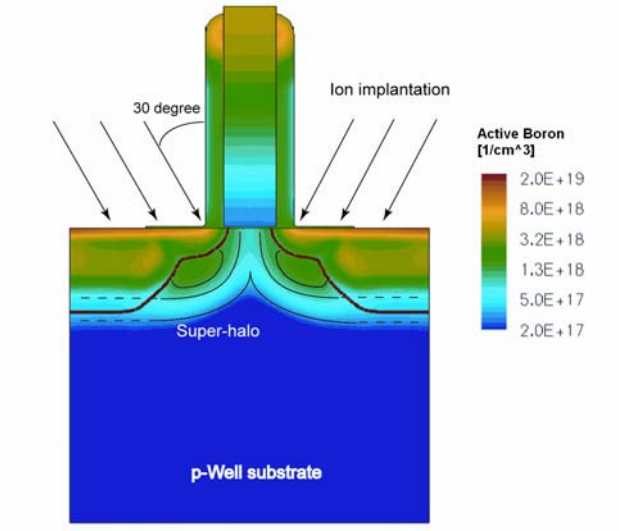


Figure 4.6 Halo process and final boron distribution in 35 nm gate length n-MOSFETs showing only two rotation directions of multiple implantations.

The halo doping of BF_2 is implanted with dose 2×10^{13} ions/cm^2 and energy 21keV carried out from quadric-frontal rotation directions and 30° tilt. The tilt enables the halo dopants to penetrate into the substrate under the gate in front of the S/D junctions. Thus the halo doping provides an effective barrier against lateral source/drain field penetration. The halo

process is illustrated in Figure 4.6, with the halo sitting next to the corner of SDE and deep S/D. Steep halo is desirable although the thermal process smears the halo doping upwards towards the channel surface and laterally towards the centre of the gate. Therefore a very limited thermal budget is required, especially in small dimension MOSFETs. The halo doping in this simulation is moderately successful in achieving the above objectives.

The similar simulation steps are carried out for p-MOSFET calibrations. The channel retrograde doping is constructed using multiple implantations, and halo doping is applied as in the n-MOSFET following the guidelines of Table 4.2. The final simulation structure of p-MOSFET is illustrated in (b) of Figure 4.4.

4.1.4.2. Electrical characteristics calibration

At the electrical calibration stage the mobility models implemented in the TCAD device simulation tool have been tuned according to experimental data in conjunction with the published literature deriving each mobility model. The Masetti and Arora models, which are described in Chapter 3, are adjusted to fit the doping dependent mobility, although the doping concentration dependences of the mobility are different. As seen in Figure 4.7, the phonon scattering in the Arora model is abnormally low compared with experimental data [159], and the mobility decreases much faster when the total doping concentration increases from 1×10^{16} to 1×10^{18} cm^{-3}. In addition, as shown in the insert figure of Figure 4.7, the Masetti model better represents the mobility at extremely high doping concentrations.

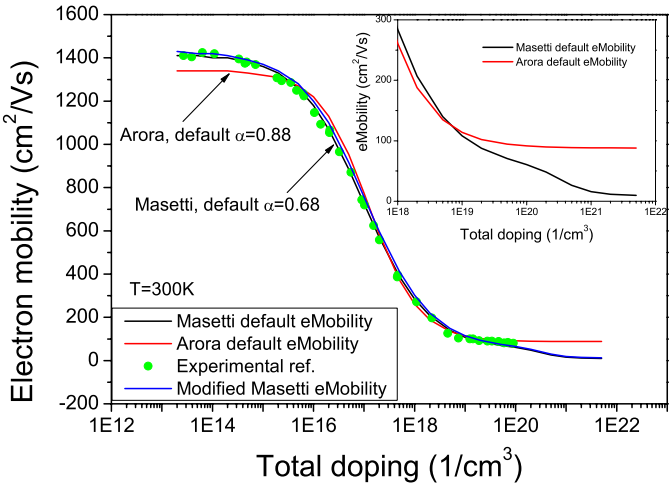


Figure 4.7 Mobility model choice and modification in device simulation calibration including experimental reference.

Choosing the Masetti model provides an advantage in better characterization of mobility in contemporary MOSFETs with high doping concentration. Small adjustments are made to

the mobility modification in calibration process related to 'min-max' terms by slightly increasing the phonon mobility and 'min' mobility. The third term remains relatively constant, close to its default. The final modified curve is shown in Figure 4.7. The default parameters of the interface mobility model and high-field mobility model are used in the simulations.

The thicker poly-Si in the Toshiba devices gives rise to a low doping concentration near gate insulator, exacerbating the poly depletion effect. High channel doping concentrations in deca-nanometer MOSFETs (above 1×10^{18}cm^{-3}) leads to inversion carriers being restricted in a narrow triangle-like potential well, and the confinement effects affect the spatial and energy distribution of the carriers, resulting in threshold voltage shift and gate capacitance reduction. The poly depletion effect and the channel quantum effects are included in the simulations as shown in Figure 4.8. The conduction band edge in poly-Si is bent up 0.16eV resulting in 2.5nm poly depletion due to low poly doping of approximately 3.6×10^{19}/cm^3 at the bottom of the gate. The inversion carrier distribution has a maximum at a depth 0.9 nm below the silicon interface, which is modelled using density gradient quantum correction. The quantum effects also reduce the gate capacitance by adding a quantum capacitance in series with the gate oxide and poly-depletion capacitances.

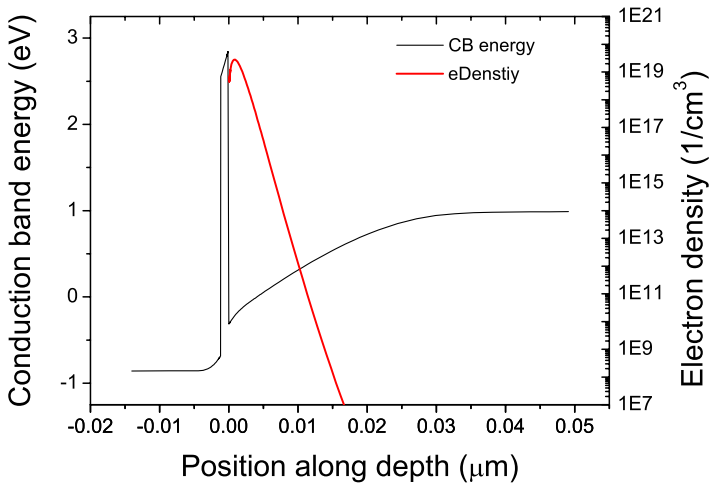


Figure 4.8 Inversion carrier profile in the channel and the conduction band edge in n-MOSFET at on-state.

The drain current *vs.* gate voltage (I_d-V_g) and drain current *vs.* drain voltage (I_d-V_d) are the primary targets of calibration in the device simulations. Related to these *I-V* characteristics are key figures of merit including: the threshold voltage and sub-threshold slope (SS) which affect both I_{on} and I_{off}; and DIBL characterizing the drain voltage influence on the threshold voltage and determined by the 2D lateral doping profile. The calibration starts by adjusting the electrostatics with respect to the low and high voltage I_d-V_g characteristics,

aiming to match the same SS and DIBL. Then the mobility is tuned to match the current magnitude. Finally a check is made that the I_d-V_d curves match well. For p-MOSFETs the mobility models adopted in simulations are the default ones. Figure 4.9 and Figure 4.10 illustrate the calibrated I-V characteristics of the n- and p-channel MOSFETs respectively. Due to more available data the calibration of the n-MOSFET is more accurate.

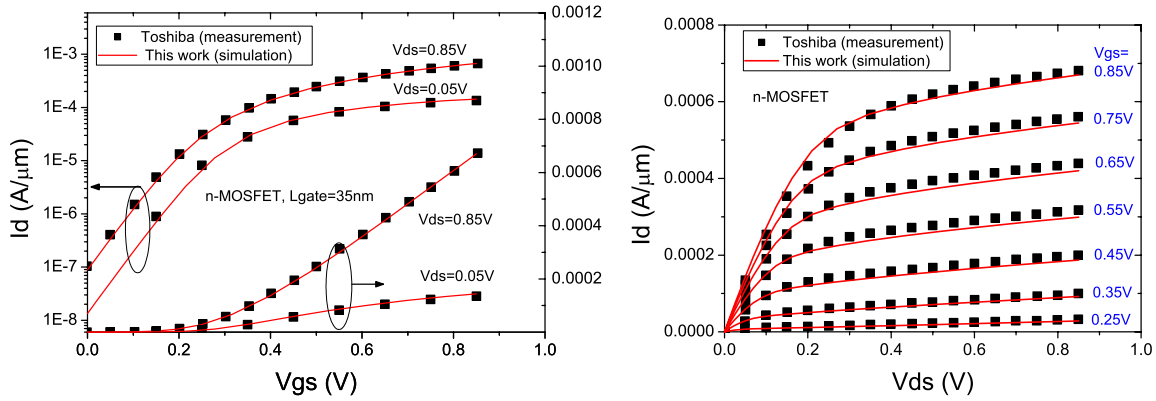


Figure 4.9 *Id-Vg* and *Id-Vd* characteristics calibrations of Toshiba 35 nm gate length n-channel MOSFETs with supply voltage 0.85V.

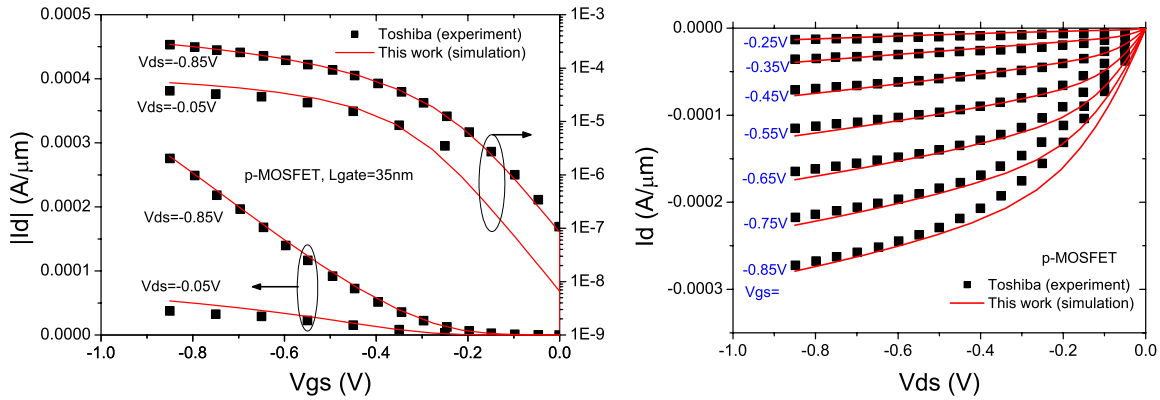


Figure 4.10 *Id-Vg* and *Id-Vd* characteristics calibrations of Toshiba 35 nm gate length p-channel MOSFETs with supply voltage 0.85V.

Table 4.3 Simulation and experiment performance parameters.

	Toshiba (experiment)		This work (simulation)	
	n-MOSFET	p-MOSFET	n-MOSFET	p-MOSFET
I_{on} (μA/μm)	676	272	670	279
I_{off} (nA/μm)	100	100	88.4	84.2
SS (mV/dec)	86.1	92.3	86.6	97.3
DIBL (mV/1V)	unavailable	unavailable	90	114

The key figures of merit extracted from simulations are compared with experimental data in Table 4.3. The results identify the quality of the calibration process. This is a starting point of the following improvement of the 35 nm gate length MOSFET design and its scaling to 25, 18 and 13 nm physical gate lengths.

4.2. Modernization of CMOS devices

4.2.1.45 nm CMOS technology

The 45 nm CMOS technology generation has already been developed and reported in several IEDM papers and foundries with several technology innovations and new materials introduced to overcome existing challenges [3][38][160][161][162]. To accurately represent the latest CMOS device trends, the Toshiba devices developed in 2002 are inadequate in providing physical insights for contemporary devices. Therefore, an update of this old 35 nm design, to incorporate the latest technology features and modernize the process flow, is carried out based on current 45 nm CMOS technologies. In the next few subsections, important technology features of the contemporary 45 nm technology generation are highlighted.

4.2.1.1. High-k / metal gate

For the first time, in the 45 nm technology reported by Intel, grown silicon oxide has been replaced by ALD hafnium-based high-k dielectric as the gate insulator. The poly-Si gate electrode has also been superseded by optimal dual band edge work function metal gates. Based on these innovations the Intel 45 nm CMOS technology has demonstrated advantages in reducing gate leakage by ×25 for nMOSFETs and ×1000 for pMOSFETs compared to the 65 nm technology. Drive currents of 1.36 mA/μm for nMOSFETs and 1.07 mA/μm for pMOSFETs at off-currents of 100nA/μm have been achieved [3]. The structure of these Intel transistors is illustrated in Figure 4.11, with the high-k dielectric and the metal gate clearly visible [38].

The high-k gate stack has enabled an ×0.7 equivalent scaling of the electrical gate oxide thickness compared to the 65 nm CMOS technology. The gate oxide thickness scaling was one of problems holding back scaling trends, due to atomic scaling limits of the SiO_2 thickness and excessive gate leakage. An equivalent oxide thickness (EOT) of 1.0 nm has been achieved by using hafnium-based high-k gate stacks. This does not only limit the gate leakage, but also improves the gate control over the channel. A major benefit of the metal

gate is the elimination of the poly-Si depletion effect, which causes serious degradation of the gate capacitance in scaled device due to limitations in the poly-silicon doping concentration. Although controversial, the metal gate is also said to screen the high-k soft optical (SO) phonon scattering, therefore limiting the detrimental effect of SO phonons on carrier mobility.

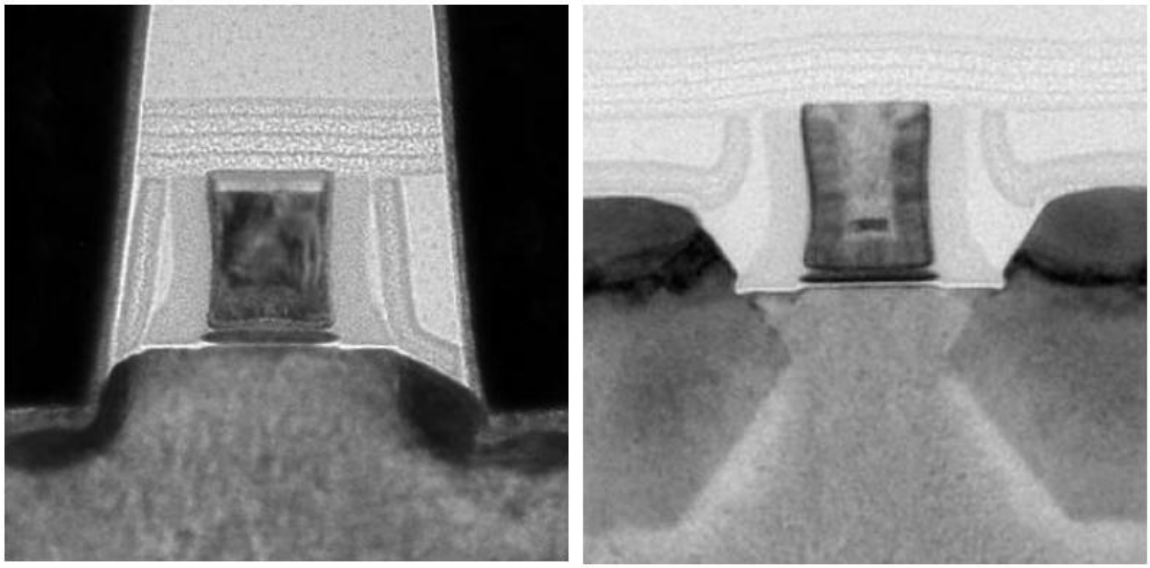


Figure 4.11 Intel 45 nm technology CMOS transistors. The left one is nMOS and the other is pMOS. Reprinted with permission from Auth *et al.*, "45 nm high-k + metal gate strain-enhanced transistors," in *Symp. VLSI Tech. Dig.*, (© 2008 IEEE).

A process simulation closely reflecting the 45 nm Intel damascene or replacement gate process flow was developed. Atomic layer deposition of high-k material is carried out instead of gate SiO_2 growth in process flow. After S/D salicidation and interlayer dielectric (ILD0) deposition, a CMP step is introduced and then remove dummy poly-Si. Relevant band edge work function metal is deposited for the n- and p-MOSFETs, then Al is filled into gate trenches for low resistance gate [3], and a metal CMP is finally carried out [163].

4.2.1.2. Stress engineering

Stress enhancement of carrier mobility as an essential technology booster was first introduced by Intel in the 90 nm node while gate oxide thickness and supply voltage scaling stagnated. Continued performance improvements were achieved by the use of rich-Ge epitaxy Si (eSiGe) in etched sources and drains, compressive contact etch stop layers (CESL) in p-channel MOSFETs and tensile CESL in n-channel MOSFETs, combined with various stress memorization techniques (SMT) – since the scaling limits eSiGe volume and stress liner thickness leading to challenges in maintaining strain enhancement levels from one generation to the next.

Stress enhancement has been realized by increasing Ge content of embedded SiGe in Intel CMOS technology from 17% in the 90 nm technology and 23% in the 65 nm technology, to 30% in the 45 nm technology. The proximity of eSiGe to channel is a key parameter in strain enhancement. Closer proximity can ensure an enhancement of carrier mobility. SiGe in the Intel pMOSFET has the proximity of around 8 nm while the eSiGe of the Fujitsu pMOSFET is located around 16 nm away from the channel. The S/D recess in the pMOSFET is usually created after spacer formation, ensuring a high quality SDE. Therefore the eSiGe proximity is restricted by the spacer thickness and recess etching facet shape. Stress liners for both nMOSFETs and pMOSFETs are limited to a small space in minimum contact pitch devices, and reducing the spacer is an efficient way to improve space for dual stress liners (DSL) such as an L-shaped spacer [160]. In addition, stress memorization techniques such as multiple-stressor technology (MST) are applied in Fujitsu CMOS devices. Poly-gate stressor liners are multiplicatively inserted into the process flow; deposited, annealed and thereafter removed from process flow such as in the insert of offset spacer formation and halo and extension doping, and the insert of S/D implantation and S/D salicide which all include thermal processes [162]. The stress is transferred into the channel and memorized during the annealing process and lattice repair. Stress enhancement benefits from a gate-last process in the high-k + metal gate technology. The dummy polySi removal followed by eSiGe S/D formation leads to additional increase in the compressive stress by 50% in channel [38].

4.2.1.3. Other features in 45 nm CMOS technology

193 nm dry lithography is used by Intel in the 45 nm CMOS technology; however a double patterning scheme is adopted in the printing of critical layers. Using this scheme, first step lithography patterning is applied to define parallel continuous lines. Only a series of discrete pitches are formed within the minimum pitch that determines the technology generation. A second lithography patterning is carried to cut the lines. The separate orthogonal patterns ensure a sharp poly line endcap [38].

Significant improvements were introduced in 193 nm lithography at the 45 nm technology level. The numerical aperture (NA) was increased to 1.2 in order to improve resolution by adding liquid with refraction index bigger than 1 (such as water of ~1.33) between the wafer and optical lithography projection lens. Immersion lithography enables a resolution of 40 nm half pitch.

It is important to have ultra-shallow SDE junctions and low S/D series resistances at 45 nm CMOS technology and beyond. Consequently laser annealing (LA) was introduced to activate SDE dopants. This ensures low sheet density but shallow SDE junction depth. In addition it also promotes the gate capacitance due to high activation of poly-silicon dopants. Hence it improves the short channel effect control and enhances the drive current. Table 4.4 collects the principal features applied in 45 nm CMOS technology.

Table 4.4 45 nm CMOS technology features of various foundries.

Foundry	Intel	TSMC	IBM	Fujitsu
L_{gate} (nm)	35	30	35	35
Contact pitch (nm)	160	126	140	140
Gate stack (thickness)	Metal/high-k (EOT 1.0nm)	PolySi/SiON (EOT 1.25 nm)	PolySi/SiON (1.2nm)	PolySi/SiON (unknown)
Stress enhancement	eSiGe, CESL, SMT, Gate-last	eSiGe, CESL, SMT	eSiGe, CESL, SMT	eSiGe, CESL, SMT
Critical layer patterning	0.92NA/193nm, dry litho, double patterning	1.2NA/193nm, immersion litho	1.2NA/193nm, immersion litho	193nm, immersion litho
Innovative annealing	none	LA	LA	unavailable
V_{dd} (V)	1.0	1.0	1.0	1.0
I_{on} (mA/μm) @ I_{off}=100nA/μm	1.36/1.07 (nFET/pFET)	1.2/0.75 (nFET/pFET)	1.15/0.785 (nFET/pFET)	1.22/0.765 (nFET/pFET)

4.2.2. Simulating the 45 nm technology CMOS

In this subsection, we introduce into the 35 nm gate length Toshiba device structure the latest features of 45 nm CMOS technology. The aim is to design a contemporary set of 35 nm gate length CMOS that replicates the leading 45 nm technology device performance. The redesign includes minimum pitch dimensions and stressors carefully developed using the latest TCAD simulators. The simulated electrical performance is verified against the equivalent performance of 45 nm CMOS technology reported by Intel and TSMC [3][161].

4.2.2.1. Process simulation

The Intel 45 nm technology has already allowed the SRAM cell size to be scaled down to 0.346 μm^2, and the dense SRAM cell area of IBM approaches 0.249μm^2. Therefore in our

simulations, the physical transistor dimensions are chosen to match requirements of minimum contact pitch of 45 nm CMOS technology. The contacted gate pitch is selected to be 160 nm, which is identical to the Intel process. The simulated devices include complete Source/Drain regions and as a result the lateral dimension of the simulation domain is 200 nm. The physical gate length in 45 nm node is kept the same as in the 65 nm technology by the majority of foundries, and therefore a typical value of 35 nm is used. Poly-silicon gate and nitrided oxide insulator are still maintained in simulations considering that the majority of foundries have not yet introduced high-k/metal gate. EOT is chosen to be 1.0 nm corresponding to 1.4 nm SiON thickness. As mentioned above, the poly-silicon thickness is a factor affecting polysilicon depletion. Reducing thickness can increase the doping concentration near the interface and decrease the poly-silicon depletion effect. The risk of boron penetration in ultra-thin poly-silicon can be effectively suppressed by nitrogen in the gate oxide. The introduced realistic thickness of poly-silicon used in simulations is approximately 60 nm. Values of the gate to S/D contact distance of 30~37 nm are selected with reference to realistic choices made by foundries.

The doping profile is essentially transferred from the Toshiba device calibration. Process steps follow those of the Toshiba devices. They include well and retrograde implantation doping, gate patterning, SDE low energy implantation and halo implantation followed by low thermal budget and silicon nitride spacer formation, and finally S/D formation. An introduction of stressors is made in process flow. In p-channel MOSFETs, source/drain silicon is etched and Ge-rich epitaxy is performed in the recessed regions. The compressive stress is released into channel after the following thermal processing. After source/drain formation, and selective stressed contact etch stop layer is deposited over the transistor, introducing tensile stress in n-channel and compressive stress in p-channel MOSFETs. The process flow is demonstrated in Figure 4.12, with the stress engineering steps highlighted in red.

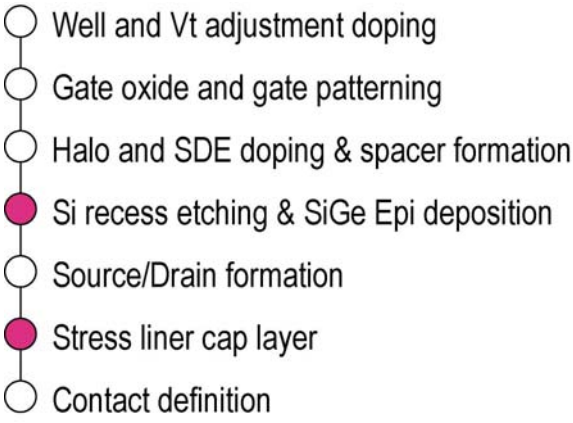


Figure 4.12 Process flow in simulations of 45 nm CMOS technology

Continued scaling exacerbates the short channel effect. Halo doping is used not only to effectively suppress SCE, but as a result of the doping diffusion towards channel to adjust the threshold voltage, compensating the threshold voltage decrease due to the improved poly depletion performance of the redesigned devices. The retrograde doping is maintained exactly as for the calibrated transistors. Spike RTA is applied following SDE/halo doping implantations, which activates dopants and reduces diffusion. Laser annealing is applied with a one millisecond exposure as a final activation step for S/D doping to repair implantation damage and to activate dopants with negligible diffusion.

Stress engineering is introduced at the TCAD simulation level. The first introduction of a stressed contact etch stop layer and SiGe source/drain in the 90 nm technology node results in a 10% increase in drive current for nMOS, and 25% for pMOS. This was achieved using an approximately 75 nm thick highly tensile nitride cap layer stressed to 20G dynes/cm^2 (=2.0G Pascal) [37]. Moderately thick 30 nm cap layers stressed to 2G Pascal are introduced in our simulations and Ge content in eSiGe is selected to be 30%. The Si recess etching is Σ-shaped ensuring SiGe stressor proximity to the channel. The proximity to the channel is chosen to be 12.5 nm and the Si recess depth is 75 nm, matching that of Intel 45 nm technology pMOSFET. The final structure of the simulated 35 nm gate length n-channel and p-channel MOSFETs are illustrated in Figure 4.16 and Figure 4.17 respectively. The SiGe expitaxy includes in-situ p-type doping, and p-type dopants of source/drain diffuse laterally towards channel and vertically towards deep substrate, extruding the S/D regions laterally and weakening the SCE control.

4.2.2.2. Electrical characteristics of the redesigned 35 nm MOSFETs

At a supply voltage of 1.0 V, on-currents of mass production high performance 45 nm technology transistors are presently above 1.2 mA/μm. Therefore in the following simulations the supply voltage was selected to be 1.0 V. In the electrical device simulation the drift-diffusion approach is used, including self-consistent solutions of Poisson equation and current continuity equations. The quantum effects are included through density gradient quantum corrections. The mobility models take into account impurity scattering, surface acoustic phonon and surface roughness scattering, and high field mobility dependence. The stress enhancement is separately included via a piezo-resistance mobility model for moderately strained n-MOSFETs and band deformation model for highly strained p-MOSFETs.

The nMOSFET and pMOSFET drive currents obtained by simulation at $V_{dd} = 1$ V are found to be 1275 μA/μm and 551 μA/μm for the n- and p-MOSFETs respectively, with off-currents of around 100 nA/μm. The unloaded RC delay is within 2 ps. The n-channel MOSFET performance is illustrated in Figure 4.13, including the I_d-V_g and the I_d-V_d characteristics. The sub-threshold slope is 88 mV/dec. The SCE are also well controlled by halo doping, resulting in a DIBL of 94 mV/V. The equivalent p-MOSFET characteristics are illustrated in Figure 4.14. The SS of the p-MOSFET is 90 mV/dec. The short channel effect control is not as good as that if the n-MOSFET, due to the SiGe stressor proximate to the channel, yielding a DIBL of 115 mV/V.

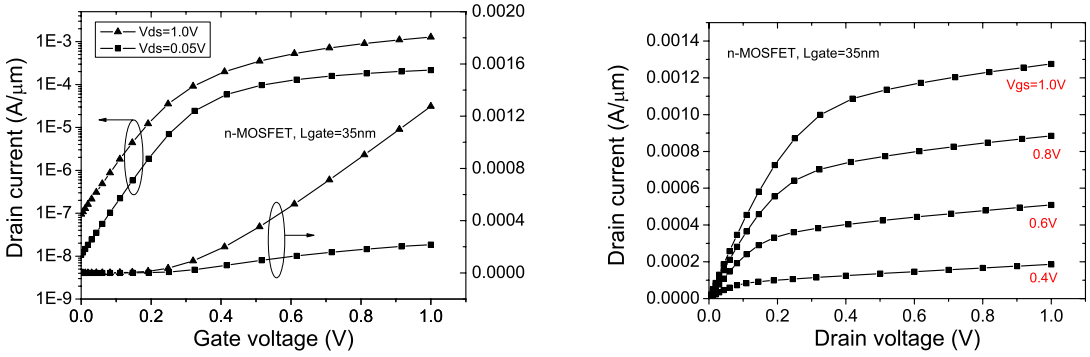


Figure 4.13 High/low drain voltage *Id-Vg* characteristics and *Id-Vd* characteristics of redesigned 35 nm gate length n-MOSFET.

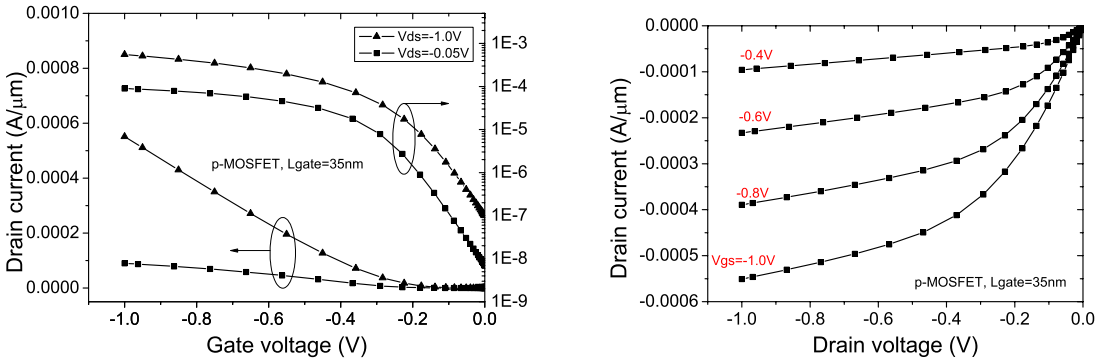


Figure 4.14 High/low drain voltage *Id-Vg* characteristics and *Id-Vd* characteristics of redesigned 35 nm gate length p-MOSFET.

In our simulation, a 20.8% enhancement of the drive current for nMOSFETs, and 14.6% enhancement of the drive current for pMOSFETs, are obtained due to careful stress engineering. The investigations show that the thickness of the stress liners, the stress in the stress liners and the spacer thickness are all critical parameters determining cap layer induced stress. In addition, the proximity and Ge content of the stressor SiGe in the source and drain are critically important. Corresponding negative effects should also be

considered. For example, closer stressor proximity increases SCE. In addition, the stressor space limits the minimum MOSFET contact pitch.

4.3. Scaled CMOS development

The redesigned 35 nm gate length MOSFETs, corresponding to the 45 nm CMOS technology, will be used in the following chapters to study specific statistical variability and statistical reliability issues. It is also the starting point for the design of a set of scaled devices with 25 nm, 18 nm and 13 nm physical gate lengths in this chapter.

4.3.1. CMOS scaling design

Scaling down of conventional bulk MOSFETs is still the driving force of CMOS technology. The 2008 update of the International Technology Roadmap for Semiconductors (ITRS) extended the 'lifetime' of conventional bulk MOSFETs to the implementations of 14 nm gate length devices expected to be in production around the year 2016. Therefore it is important to obtain predictive information about the behaviour of future generations of MOSFETs, and a viable strategy to do this is to scale down the current 45 nm technology CMOS, choosing reasonable scenarios.

The application of the generalized scaling principles mentioned in previous chapter is now limited by short channel effect and diminishing return in terms of device performance. However, the generalized scaling rule is adopted as the major guideline which sets the requirements for performance figures of merit, such as the drive-current/off-current ratio and electrostatic integrity. To continue Moore's law scaling, the MOSFET size should be scaled accordingly. The lateral dimension scaling scenarios ensure a corresponding SRAM cell size reduction. In simulations, the gate length, the gate width, contact pitch and the stress liner thickness follow the same scaling trend, as indicated in Table 4.5. However, in reality, the cap layer, for example, may not scale to the same proportions due to the need for drive current enhancement. The EOT plays a key role in determining key figures of merit such as threshold voltage, SS, DIBL and drive current. However EOT scaling has to take other limiting factors into account. One of such considerations is gate direct tunnelling leakage, which dramatically increases the overall leakage, overtaking the sub-threshold leakage in SiON devices below 2.0 nm thickness. Thus the EOT for scaled device is selected partially by realistic considerations, partially with a reference to the ITRS following the scaling details in Chapter 2. Similar considerations guide the supply voltage scaling. The supply voltage scaling slows down because the threshold voltage scaling has

in fact almost stopped in ultra-small MOSFETs, but mainly due to the ever increasing variability.

Table 4.5 Simulation specifications for scaled CMOS

L_{gate} (nm)	35	25	18	13
EOT (nm)	1.0	0.9	0.7	0.5
Stress liner thickness (nm)	30	22	15	11
Lateral length (nm)	200	130	100	70
V_{dd} (V)	1.0	1.0	1.0	0.9

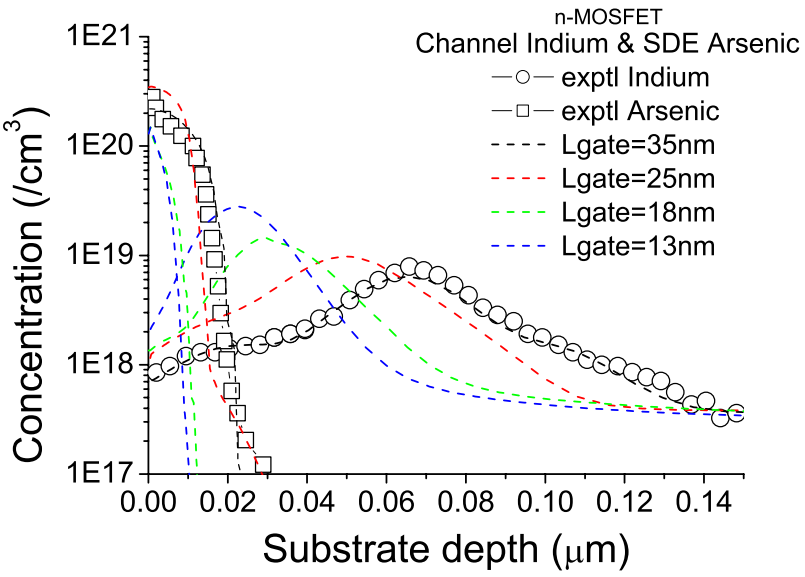


Figure 4.15 Doping concentration scaling of n-MOSFETs

The doping concentration has to be increased in scaled devices to suppress SCE. The scaling factor of the doping concentration is around 1.6~1.8; a requirement to appropriately control the sub-threshold slope and DIBL. A retrograde channel doping design is used in the scaled devices and the retrograde depth is scaled in the same proportion as the device dimension. SDE doping increases, but not in the cases of high-k/metal gate MOSFETs. SDE junction depth is reduced by decreasing the energy of implantation and lowering the thermal budget through steeper spike RTA and LA. The final doping profiles for the scaled devices are shown in Figure 4.15 in the case of n-MOSFETs.

The scaling of the complementary MOSFETs follows a similar scenario. The outcome of the scaling is illustrated in Figure 4.16 for n-channel MOSFETs and in Figure 4.17 for p-

channel MOSFETs respectively. The scaling not only reduces the dimension vertically and laterally, but also achieves the desirable doping features such as retrograde doping and halo doping. The stress in the CESL layer is maintained at 2.0 G Pascal in scaled devices, and the Ge content is maintained at 30%, however the stressor dimensions are proportionally scaled including the stress liner thickness and SiGe recess aspect.

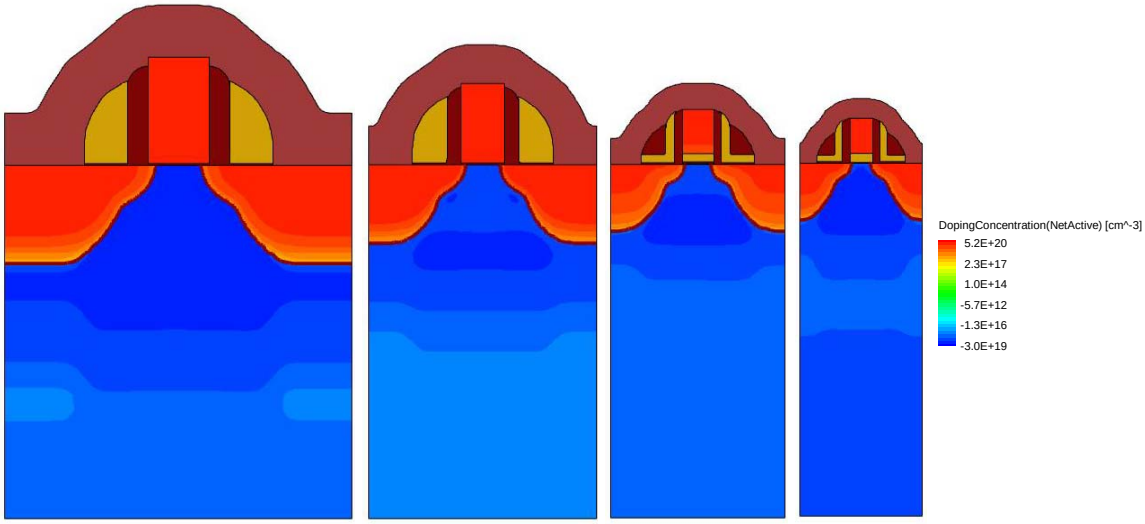


Figure 4.16 The n-channel MOSFET structures and doping profiles respectively with 35, 25, 18 and 13 nm physical gate length.

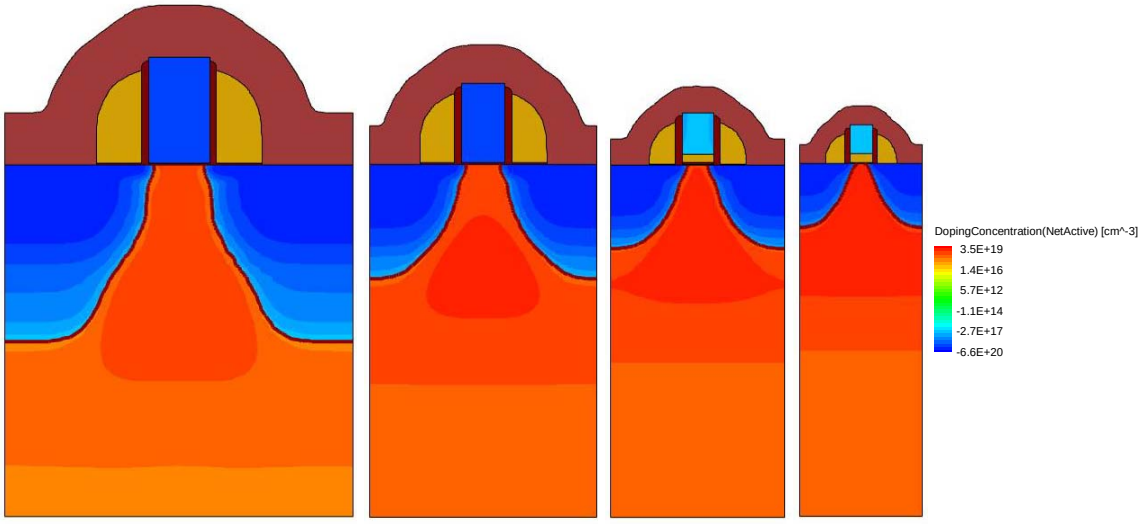


Figure 4.17 The p-channel MOSFET structures and doping profiles respectively with 35, 25, 18 and 13 nm physical gate length.

4.3.2.Scaled CMOS characterization

This subsection describes in detail the scaling of the complementary MOSFETs including the proper process flow and the simulated electrical characteristics.

4.3.2.1. 25 nm gate length CMOS

First we start with the poly-Si/SiON gate version of the 25 nm gate length CMOS based on the scaling of the redesigned 35 nm gate length CMOS devices from the previous section. It is extremely challenging to simultaneously achieve good SCE control and high performance due to the following reasons: Firstly the electrical EOT scaling has practically halted due to deteriorating poly-silicon depletion effect with the increase of channel doping; secondly, benefit from stress enhancement in the high-k/metal gate replacement process is unavailable. However, a successful process design has been achieved through careful doping design, including super-halo, and due to aggressive stress engineering. The halo implantation dose is 3.6-9.6×10^{13} cm^{-2} and the implantation energy is 15-25 keV.

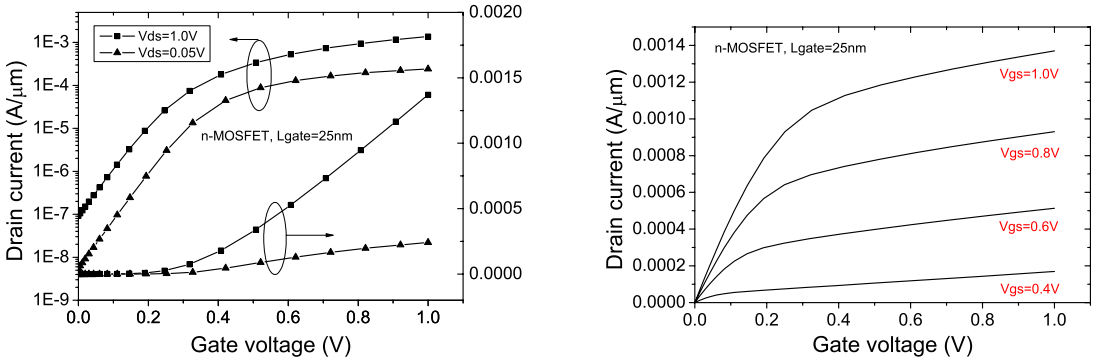


Figure 4.18 High/low drain voltage *Id-Vg* characteristics and *Id-Vd* characteristics of poly-silicon 25 nm gate length n-channel MOSFET.

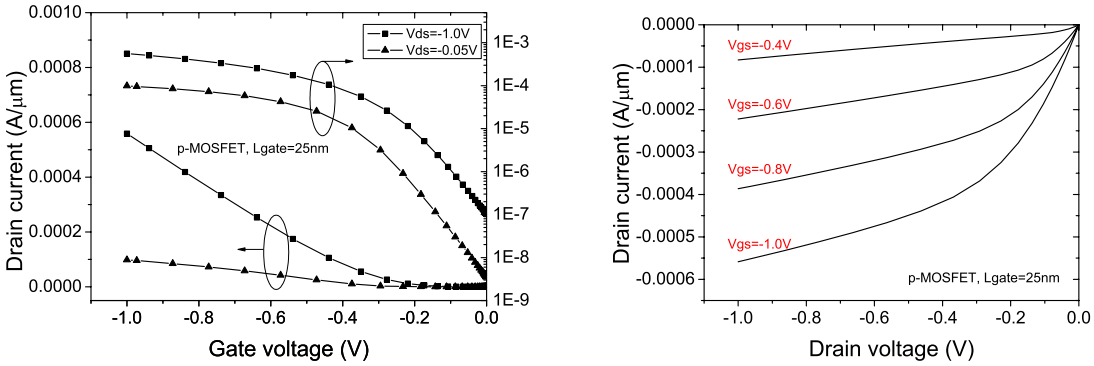


Figure 4.19 High/low drain voltage *Id-Vg* characteristics and *Id-Vd* characteristics of poly-silicon 25 nm gate length p-channel MOSFET.

The electrical characteristics of the 25 nm n- and p-channel MOSFETs are illustrated in Figure 4.18 and Figure 4.19 respectively. High drive currents of 1370 μA/μm and 559 μA/μm in the n- and p-channel cases respectively are achieved at around 100 nA/μm off-current. The sub-threshold slopes (SS) are 93 mV/dec and 100 mV/dec respectively,

while DIBL values are 115 mV/V and 147 mV/V respectively for the n- and p-channel transistors. The exacerbated SCE of the pMOSFET is a combined result of weakened gate control and the S/D proximity of the SiGe stressor. A high-k/metal gate version of the 25 nm p-MOSFET, with significant reduction of SCE by the improved gate control, yields 99 mV/V DIBL.

4.3.2.2. 18 nm gate length CMOS

High-k/metal gate is introduced in the 18 nm gate length CMOS devices. A high-k dielectric of 1.0 nm thick HfO_2, which has relative dielectric constant 20 in the simulations, is deposited on the top of an 0.5 nm interfacial oxide layer. A conduction band edge metal is deposited on the top of high-k dielectric in the n-MOSFET regions. A gate-last process is applied to p-MOSFET regions with the original metal removed, and substituted by valence band edge metal deposited for the p-MOSFETs. Spike RTA is ramped to 900°C for 4 s for SDE/halo activation, followed by a 0.7 ms LA applied to S/D doping activation. The poly-silicon depletion effect is eliminated, improving SCE and aiding the doping design. Retrograde channel doping is continuously formed and super-halo doping is implanted with dose 4-12×10^{13} cm^{-2} and energy 10.5-18 keV. The final structure and doping concentration of the 18 nm transistors are illustrated in Figure 4.16 and Figure 4.17 respectively for n- and p-MOSFETs.

The *I-V* characteristics of the complementary 18 nm transistors are shown in Figure 4.20 and Figure 4.21. High performance is achieved for high-k/metal gate 18 nm gate length MOSFETs. The drive current is 1755 μA/μm at an off-current of 75 nA/μm for the n-MOSFET and 1268 μA/μm at an off-current of 95 nA/μm for the p-MOSFET. pMOSFET drive current is particularly enhanced by the combined effect of metal gate and the gate-last stress enhancement. The SS for nMOSFETs and pMOSFETs is 87 mV/dec and 91 mV/dec respectively, better than in previous generation polysilicon/SiON gate devices. In addition, DIBL is also improved compared to the 25 nm MOSFETs, with DIBL values of 65 mV/V and 123 mV/V respectively.

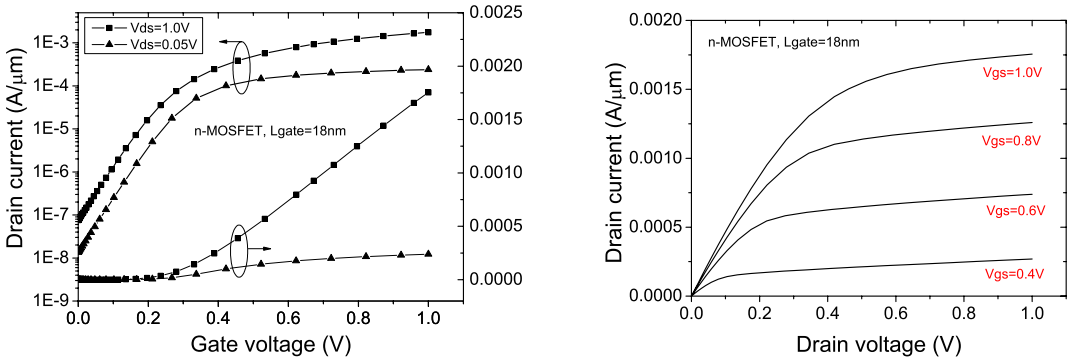


Figure 4.20 High/low drain voltage *Id-Vg* characteristics and *Id-Vd* characteristics of high-k/metal gate 18 nm gate length n-channel MOSFET.

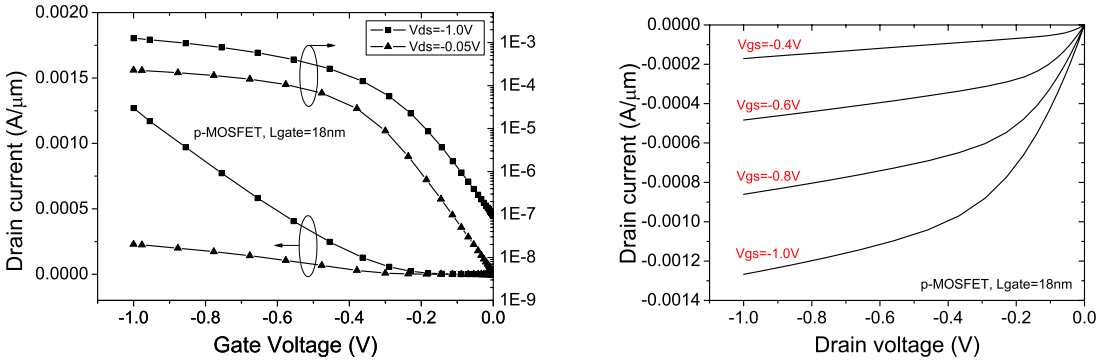


Figure 4.21 High/low drain voltage *Id-Vg* characteristics and *Id-Vd* characteristics of high-k/metal gate 18 nm gate length p-channel MOSFET.

4.3.2.3. 13 nm gate length CMOS

The ultra-short, 13 nm physical gate length CMOS devices are designed featuring high-k/metal gates, optimal stress, spike RTA/LA annealing and carefully crafted doping profiles. HfO_2/IL oxide gate dielectric (HfO_2 has relative dielectric constant 20) and dual band edge metal gates are utilized. The final structure and doping profiles of the 13 nm gate length MOSFETs are illustrated in Figure 4.16 and Figure 4.17. Retrograde channel doping and halo doping are implanted. Ultra-shallow SDE junctions are formed by an extremely low implantation energy of 0.35~0.4 keV, followed by a spike RTA ramped to 900°C over 2 s. The LA is applied for only 0.5 ms.

The electrical characteristics of the 13 nm gate length transistors are presented in Figure 4.22 and Figure 4.23. The supply voltage is scaled to 0.9V and the V_{th} at drain current 1×10^{-5}A/μm for nMOS and pMOS devices is 185 mV and 186 mV respectively. Drive currents of 1782 μA/μm at an off-current of 65 nA/μm has been achieved for the nMOSFET and 1298 μA/μm for the pMOSFET. The SS values are 91 mV/dec and

90 mV/dec respectively. The SCE is well controlled yielding DIBL values of 64 mV/V and 120 mV/V for the two devices.

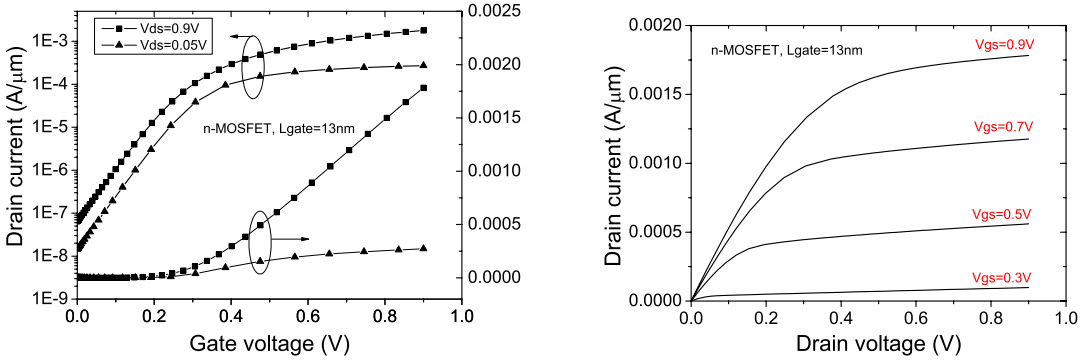


Figure 4.22 High/low drain voltage *Id-Vg* characteristics and *Id-Vd* characteristics of high-k/metal gate 13 nm gate length n-channel MOSFET.

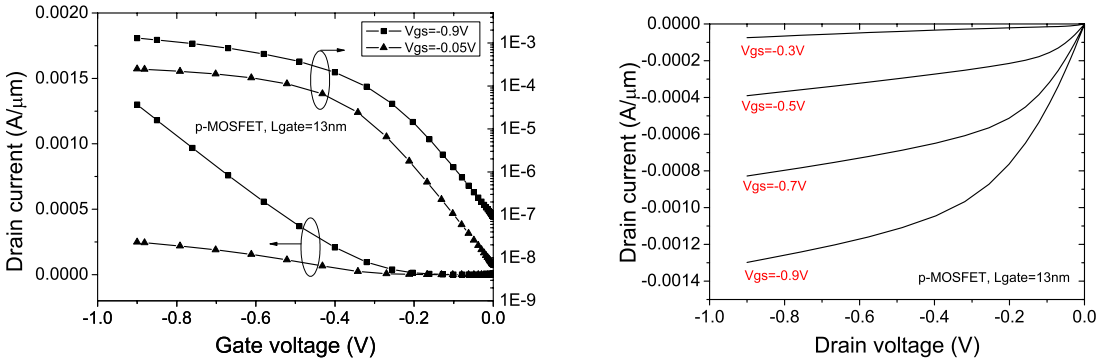


Figure 4.23 High/low drain voltage *Id-Vg* characteristics and *Id-Vd* characteristics of high-k/metal gate 13 nm gate length p-channel MOSFET.

4.4. Strain scaling

4.4.1. Gate-last benefit

It has been reported that the metal gate gate-last process, which removes the rigid poly-silicon gate after S/D formation and before the deposition of the high-k/metal gate stack, further increases channel strain and carrier mobility [164][165][166]. However, there are no systematic studies that differentiate the various aspects of the metal gate related performance enhancement and consider its behaviour with further scaling.

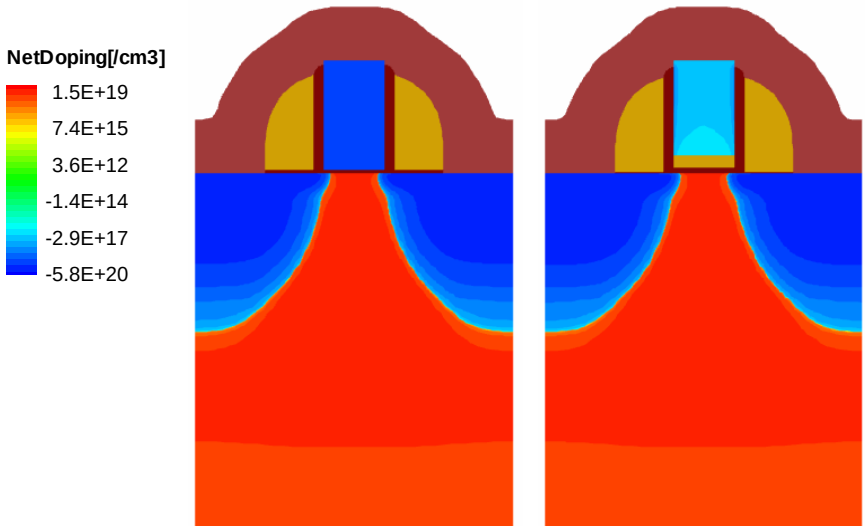


Figure 4.24 Simulation structures and doping profiles of both poly-silicon gate (left) and high-*k*/metal gate (right) 25 nm gate length pMOSFETs.

The 25 nm gate length p-MOSFET is selected as the test device. To highlight the benefit of the high-*k*/metal gate gate-last process, two p-MOSFETs are simulated (Figure 4.24): (i) the first is poly gate device fabricated using traditional flow; (ii) for the second device the poly-silicon gate and the gate oxide are removed after S/D formation and high-*k* gate dielectric is deposited over a grown thin oxide, followed by the deposition of a metal gate. Figure 4.25 illustrates the stress distribution in the two simulated devices. The main difference between the two devices is in the channel regions. The gate-last process increases the compressive stress from -1.28 GPa in the poly-silicon gate transistor to -2.81 GPa in the metal gate transistor.

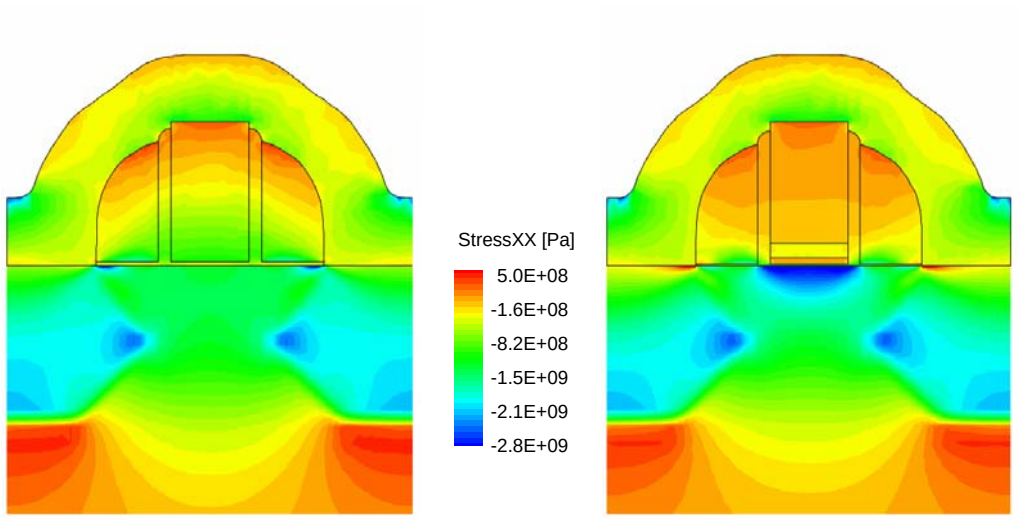


Figure 4.25 The comparison of channel direction normal stress distribution over the control poly-silicon gate device and the high-k/metal replacement gate device.

To explore the strain enhancement mechanism, ‘measurements’ of strain/stress just before poly-silicon gate removal and after high-*k*/metal gate formation in the same device are

taken and the results are shown in Figure 4.26. Due to the removal of the rigid poly gate, more strain is introduced in the channel region, changing the stress transfer path and achieving strain crowding into a shallow region in the channel. Because the deposited high-*k*/metal conforms to the available space without inserting significant stress of its own, the increased stress is memorized by the channel. This is confirmed by the almost zero strain/stress in the high-*k*/metal gate after the deposition process, and sharp strain peak at the gate edge in the gate-last process device illustrated in Figure 4.26.

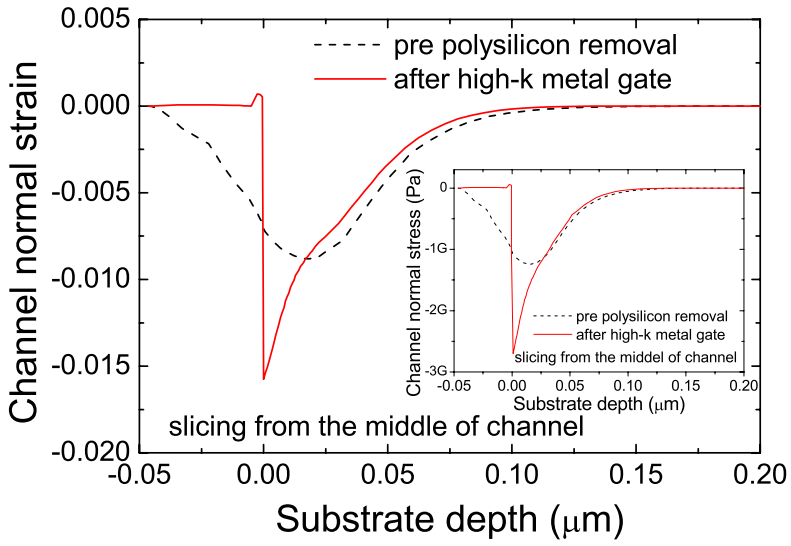


Figure 4.26 Channel direction normal strain/stress is compared between before poly gate removal and after the formation of high-k/metal gate.

The high-*k*/metal gate gate-last process significantly improves transistor performance. The dependence of the drain-current *vs.* gate-voltage (I_d-V_g) at high drain voltage is illustrated in Figure 4.27 in comparison to the control poly-gate transistor. A 65% increase in drive current is achieved for the same level of leakage current. The sub-threshold slope is also improved from 99 mV/dec to 89 mV/dec. From the I_d-V_d curves in the insert of Figure 4.27, it is clear that significant increase in the current is achieved in both the linear region and in the saturation region both contributing to higher switching speed [168].

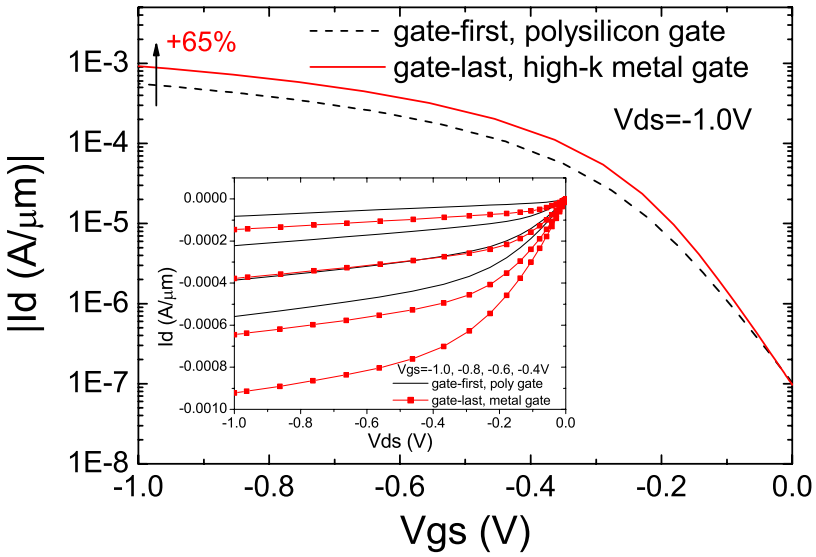


Figure 4.27 *Id-Vg* and *Id-Vd* electrical characteristics curves simulated for gate-first and gate-last processed devices.

Table 4.6 The relationship between on-currents of strained/unstrained gate-first/gate-last devices.

I_{on} [μA/μm]	Gate-first, poly-silicon gate	Gate-last, high-*k* metal gate	Enhancement
Unstrained	496.8	745.8	1.5012
Strained	558.7	922.6	1.6513
Enhancement	1.125	1.237	1.10

NB: Gate-first poly strained I_{off} = 103 nA/μm; Gate-last high-*k*/metal strained I_{off} = 97 nA/μm.

Since the models used in the DD simulation do not reflect the impact of the metal gate on the high-*k* related SO phonons, the main improvement in the simulated device performance comes from the hole mobility enhancement due to increased stress, and from the decrease of the electrical EOT associated with the metal gate. In order to differentiate their contributions, device simulations which do not account for the impact of the process induced strain on mobility are also performed for both the poly-gate and the gate-last devices. Table 4.6 compares the results.

It is clear that the electrical EOT reduction in the high-*k*/metal gate case results in an approximately 50% performance increase compared to the performance of the unstrained devices. Considering that the difference of gate over-drive voltages of the poly and metal gate devices is approximately 20 mV, this performance increase is almost entirely related to an increase of the inversion capacitance. When strain is considered in the simulations,

the drive current enhancement associated with the gate-last process increases to more than 65%. The drive current is given by

$$I_d = Q_{inv} \times W \times v \,. \tag{4.3}$$

Where Q_{inv} is the inversion charge per unit area at the source end of the channel, W is the channel width, and v is the corresponding carrier velocity. Therefore, the additional strain related increase in device performance in the gate-last transistor is associated with a more than 10% increase in the average carrier velocity at the source end of the channel.

A major advantage of the gate-last process is the additional strain and mobility enhancement available, compared to the poly-silicon case. It is essential to explore how this strain enhancement depends on stressor design. The $Si_{1-x}Ge_x$ stressor design parameters explored in this study include the proximity of the stressor to the channel and the germanium content, x, in the $Si_{1-x}Ge_x$.

The impact of the stressor proximity is illustrated in Figure 4.28 and shows that the closeness of the stressor to the channel not only increases the strain/stress in both the poly-silicon and the metal gate cases but also results in a stronger enhancement from the gate-last process technique. A closer stressor focuses more effectively the strain/stress into the channel, particularly in the absence of a rigid polysilicon gate. This is facilitated by the sharpening of the intrusion point of the Σ-like eSiGe. Increasing the Ge content of the eSiGe stressor increases the strain/stress as illustrated in Figure 4.29. However the gate-last related enhancement saturates at approximately 30% Ge content. This is probably related to the fact that at high Ge contents the stress is large enough to overcome poly-silicon rigidity and the extra gain from poly-silicon removal is therefore reduced.

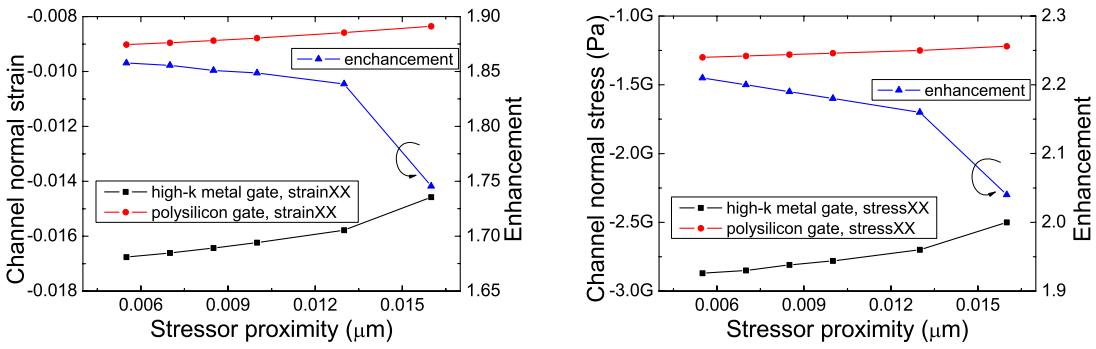


Figure 4.28 Stressor proximity dependency at Ge 30%. The closer the stressor, the bigger strain/stress, and the more enhancement from the gate-last process.

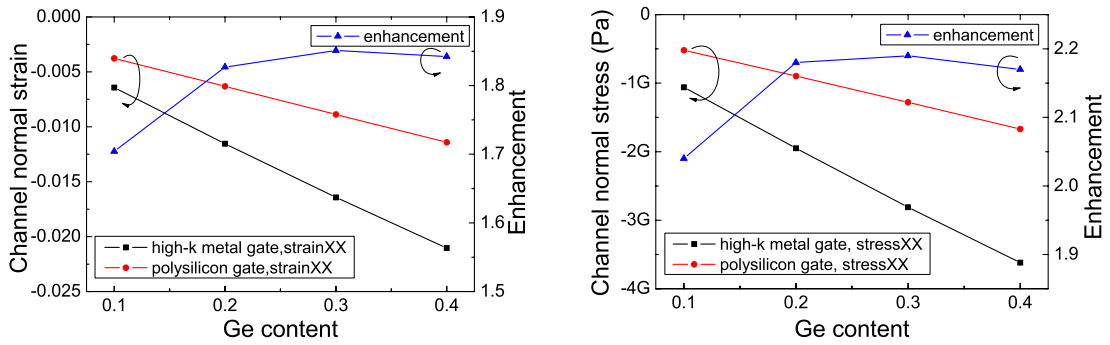


Figure 4.29 Ge content dependency at stressor proximity length 8.5 nm. The more Ge content, the bigger the strain/stress. Gate-last strain enhancement at first increases then saturates with increase of Ge content.

High performance has been achieved in the TCAD design of 25 nm gate length high-*k*/metal gate pMOSFETs employing the gate-last strain enhancement technique. The main advantage of the metal gate is a drastic reduction in electrical EOT, resulting in an almost 50% improvement in drive current. The strain enhancement of the gate-last process adds another 10% to the performance enhancement. The gate-last associated strain improvement will increase with further scaling of stressor proximity to channel but saturates with increasing Ge fraction in the stressor.

4.4.2. Scaling of strain enhancement

Strain fortification of drive current is a well-known fact in modern CMOS devices. The enhancement magnitude, however, is limited by the available space for stressors. Due to the reduction of gate gap in minimum pitch CMOS, the stress liner thickness and SiGe volume is limited. Too thick a stress liner will not increase stress any more [167], and SCE also affects the proximity of SiGe to channel. The enhancement of drive current for each device is listed in Table 4.7. Stressors in our simulations increase the carrier mobility and as a consequence enhance the on-current from 12.5% to 35.5%. Generally the strain enhancement decreases with device scaling due to stress liner thickness reduction and SiGe volume decrease.

Table 4.7 Strain enhancement of Id,sat in both NMOS and PMOS.

Gate length [nm]	35	25	18	13
Idsat enhancement, NMOS	20.8%	35.5%	13.9%	11.5%
Idsat enhancement, PMOS	14.6%	12.5%	23.5%	21.9%

4.5. Scaling summary

The increase of gate capacitance with scaling leads to the stronger control of the gate over the channel carriers, increasing the transconductance. This is illustrated in the Figure 4.30. The transconductance of nMOSFETs extracted at on-state smoothly increases with the reduction of gate length. The linear relationship reflects the inverse proportion of g_m to the gate oxide thickness. However in the pMOSFETs, the transconductance has a sharp increase at 18 nm gate length devices. The high-k/metal replacement gate process not only eliminates the poly depletion effect, improving the gate capacitance, but also increases the carrier mobility due to additional stress delivered in the gate-last technique.

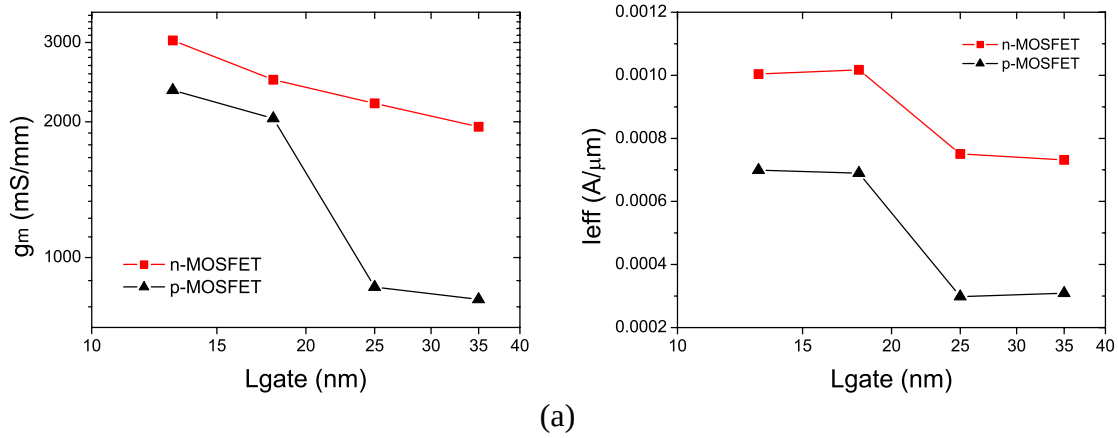


Figure 4.30 Transconductance trend (a) and effective drive current trend (b) with scaled MOSFETs.

The drive current is the key performance indicator for digital applications. High drive current leads to high switching speed in circuits. $C_{load}V_{dd}/I_{dsat}$ is the performance metric characterizing the circuit speed. However, other current criteria have been proposed to accurately represent circuit switching characteristics in terms of device characteristics instead of saturation on-current [168][169][170][171]. Na *et al.*, for example, averages the switching current trajectory by integration in terms of inverter 50% to 50% delay, and proposes a simple form of effective drive current

$$I_{eff} = \frac{I_H + I_L}{2}. \tag{4.4}$$

Where $I_H = I_{ds}(V_{gs} = V_{dd}, V_{ds} = V_{dd}/2)$, and $I_L = I_{ds}(V_{gs} = V_{dd}/2, V_{ds} = V_{dd})$.

A positive trend in the effective drive current scaling is obtained from simulations, as demonstrated in Figure 4.30 (b). For example, the effective drive current of the 35 nm nMOSFETs is 731 μA/μm. One way to promote effective current is to increase both I_H and

I_L by decreasing threshold voltage, but this will result in large leakage current. Another way is to increase I_L by increasing transconductance. Owing to the high-k/metal gate introduction, effective current is significantly improved through increase of I_H and I_L, both benefiting from gate capacitance increase.

In summary, a full set of scaled CMOS devices has been designed, starting from the calibrated, and then the improved 35 nm MOSFET devices, and adopting realistic physical scaling scenarios. The simulated 45 nm technology MOSFETs have achieved performance equivalent to that of industry leading technologies. The scaled devices, with careful and realistic process design, have also achieved the performance prescribed by the ITRS. The scaling behaviour of key performance indicators have been reported as meeting the requirements of future CMOS technology generations.

Chapter V

5. Scaling study of statistical variability

Although parameter fluctuations of MOSFETs due to process variation is not a new topic, due to the drastic scaling of device dimensions localised within-die variability is becoming dominant. The introduction of counter-measures (such as layout regularity) has been used to some extent mitigate the layout and strain induced variability [172][173]. However, the statistical variability due to the discreteness of charge and granularity of matter cannot be overcome by layout regularity and improved processing steps. The statistical variability already affects the yield of contemporary SRAM intensive CMOS circuits and systems and the problems will increase in future technology generations [174]. The MOSFET statistical variability due to sources including random discrete dopants (RDD), line edge roughness (LER) and poly-silicon granularity (PSG) has been carefully investigated via extensive numerical simulations [124][125]. In this chapter, a comprehensive simulation study of the statistical variability of the scaled bulk MOSFETs from Chapter 4 has been carried out using the Glasgow 3D 'atomistic' simulator.

5.1. Simulation methodology

The design and scaling of the set of bulk MOSFETs using the Synopsys TCAD suite has been presented in Chapter 4. The output from the TCAD process simulation including the mesh, material properties, doping profiles and other data, can be imported into the Glasgow 'atomistic' simulator to carry out the variability simulations. To allow the data transfer, the output from the process simulator must first be pre-processed, inactive regions are trimmed from the output, and generated doping profiles are re-assigned to a rectilinear simulation mesh used for device simulation. Secondly, the 'atomistic' device simulator is carefully calibrated in respect of the electrical I_d-V_g and I_d-V_d characteristics produced by the Sentaurus device simulator. This extraction and calibration results in a nominal device with characteristics matching those obtained from the TCAD simulation.

The continuously doped nominal device can be used to create ensembles of devices containing statistical variability sources such as RDD, LER and PSG. The transistor

samples used in the statistical simulation are macroscopically identical but microscopically different. RDD are the most important source of statistical variability in nano-scale bulk MOSFETs. LER induced gate length variation is always present due to the molecular nature of the resist used in patterning the gate. It is remarkably difficult to reduce LER beyond the 65 nm technology generation. Fermi-level pinning at the poly-silicon grain boundaries leads to variations in devices with poly-silicon gates. The details of modelling RDD, LER and PSG are formulated in Chapter 3.

In this simulation study of statistical variability, LER is modelled using an rms magnitude of $\Delta = 1.33$ nm and a correlation length of $\Lambda = 30$ nm. An average grain size of 40 nm is used to model PSG, assuming Fermi level pinning of 0.3 eV at below the conduction band edge. Due to the lack of donor-type traps along the poly-silicon grain boundaries in p-MOSFETs [157], the simulation of p-MOSFETs excludes PSG. The generation of RDD is based on the continuous doping profile using a rejection technique.

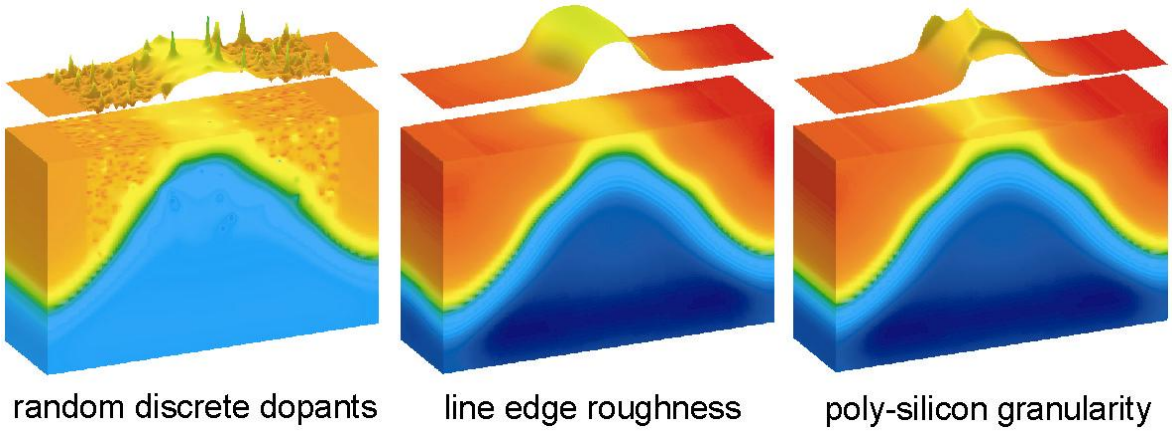


Figure 5.1 Statistical variability sources from random discrete dopants, line edge roughness and poly-Si granularity in a 35×35 nm^2 physical gate area n-MOSFET [125].

Each of the simulated sources of statistical variability illustrated in Figure 5.1 has a different impact on the operation of a MOSFET. Variations in the characteristics of n-channel MOSFETs due to RDD are caused by ionized acceptors in the channel, creating current percolation paths which determine the threshold voltage, and allowing some of devices to turn on earlier. LER alters the lateral doping profile and the metallurgical channel length fluctuates accordingly, resulting in current density variation along the channel width. Fermi level pinning along the crystalline boundaries of the poly-silicon gate results in localized changes in the threshold voltage. The magnitude of the corresponding threshold voltage variation depends on the size and the orientation of the grains. In nano-scale MOSFETs all of these sources of variability act in concert. To understand the overall effect of RDD, LER and PSG on device operation it is necessary to simulate an ensemble

of microscopically different transistors in order to obtain a evaluation of statistical distributions of key parameters. The validity of this approach to the simulation of statistical variability has been proved via comparisons of statistical simulations and measurements [175][157].

5.2. Statistical variability of 35 nm poly-gate MOSFETs

As described in Chapter 4, the 35 nm gate length CMOS transistor has been carefully designed using extensive TCAD simulations to achieve high performance similar to performance of contemporary 45 nm CMOS technology transistors introduced by Intel and TSMC. However, no systematic comprehensive simulations of statistical variability of contemporary 45 nm CMOS transistors in mass production have been published. This chapter presents a comprehensive simulation study of statistical variability in 35 nm, 25 nm and 18 nm gate length MOSFETs. To study the intrinsic parameter variation in the 35 nm gate length device an ensemble of 200 microscopically different MOSFETs are simulated including statistical variability introduced by RDD, LER for both n-channel and p-channel MOSFETs and also PSG for n-MOSFETs.

The results of this simulation are illustrated in Figure 5.2 and Figure 5.3 showing drain current *vs.* gate voltage (I_d-V_g) electrical transfer characteristics of 200 microscopically different devices, plotted together with the characteristics of a continuously doped transistor. These devices have been simulated at a high drain voltage of 1.0 V and low drain voltage of 0.05 V for both n-/p-MOSFETs. The distribution of I_d-V_g traces around the nominal current shows the dramatic impact of the statistical variability sources. It can be observed that the variation of the leakage current covers more than two orders of magnitude. This wide range of variation could have a significant impact on the leakage power lost in corresponding circuit and system. Threshold voltage variation as explicit as current variation leads to the matching problem particularly in SRAM application.

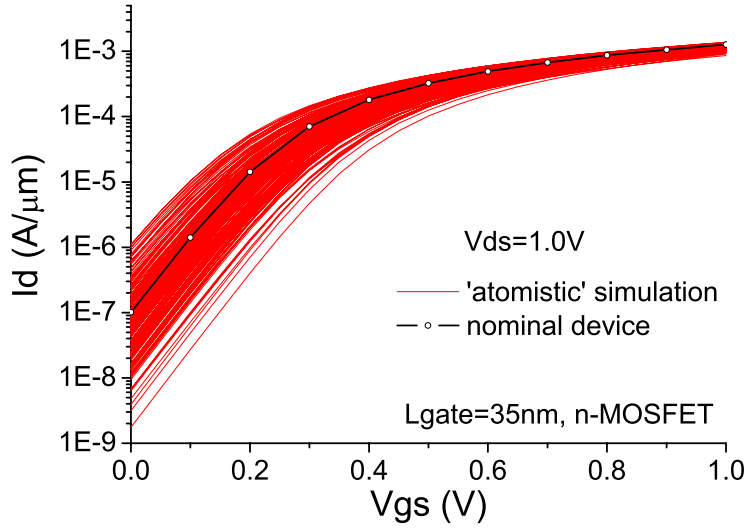


Figure 5.2 Id-Vg characteristics for 35 nm gate length n-channel MOSFETs subject to RDD, LER and PSG induced statistical variability.

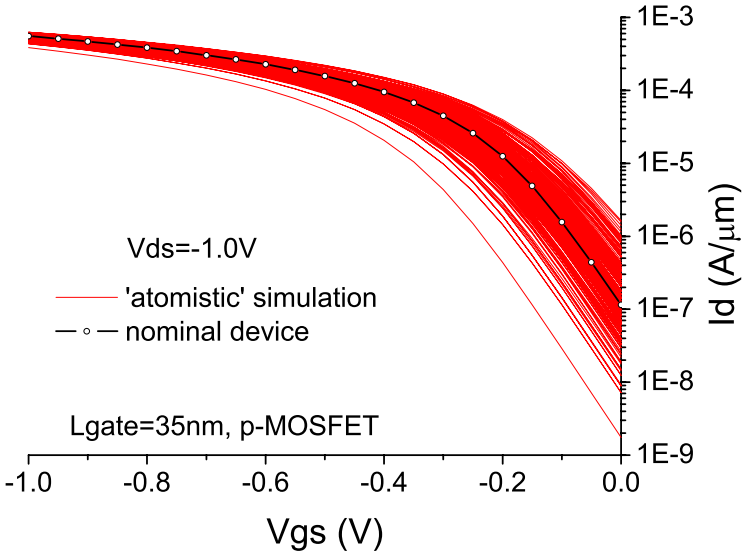


Figure 5.3 Id-Vg characteristics for 35 nm gate length p-channel MOSFETs subject to RDD and LER statistical variability.

Next we discuss the detailed statistics of important MOSFET parameters such as threshold voltage, DIBL and sub-threshold slope. The threshold voltage is extracted from the I_d-V_g characteristics using a current criterion. The threshold current in Amperes is typically defined by:

$$I_d = \sim 4\times10^{-7}\times\frac{W}{L}, \qquad (5.1)$$

where W and L are the channel width and channel length both in microns. For the high performance transistors studied here, due to increased leakage current, the threshold

current is altered to the current in the uniform device at V_{gs} = 0.2 V when V_{ds} = 1.0 V. For a square 35 nm gate length transistor, the threshold currents are 1.43×10^{-5} A/μm and 1.25×10^{-5} A/μm for n-MOSFETs and p-MOSFETs respectively.

Histograms of distributions of the threshold voltage (V_{th}), the off-current (I_{off}) and drive current (I_{on}) due to combined sources of fluctuations are shown in Figure 5.4 for the n- and p-MOSFETs respectively. The threshold voltage standard deviation (σV_{th}) is 45.2mV and the mean value is 0.220V for nMOSFETs including the impact of RDD, LER and PSG. The 6σ value of threshold is approximately 0.27V implying a wide range of threshold voltages which will lead to problems in power management, switching times and timing closure. For the p-channel transistors, the variability sources exclude PSG, yielding threshold voltage standard deviation of 44.2mV and an average of -0.206V. It can be noted that the average threshold voltage of the n-channel MOSFETs is larger than that of p-MOSFETs, due to poly-silicon Fermi pinning and the associated threshold voltage shift.

The distribution of the off-current shown in Figure 5.4 determines the standby power and the increase of the worst-case leakage is the largest concern associated with off-current variation. One standard deviation of log(I_{off}) is around 0.53~0.54 for 35 nm gate length MOSFETs which is ~3.4 times of average I_{off}.

The drive current variation in Figure 1.4 has a standard deviation σI_{on} of 97.3μA/μm for the n-channel MOSFET and 43.4μA/μm for the p-channel MOSFET. One standard deviation corresponds to 8.5% and 8.2% of average drive currents respectively. The approximations inherent to the Drift-Diffusion approach used in these simulations do not allow the accurate capturing of transport variation effects and therefore underestimates the actual variations in drive current. Comprehensive Monte Carlo simulations are required to obtain a true magnitude of on-current variation [176].

However even this underestimated variation in on-current it will create issues in driving the next stage circuit leading to timing variations. It is noticeable that the average drive currents are lower than nominal values. The reason for this is the artificial trapping of carriers in potential wells associated with ionized dopants in the source/drain which increases the access resistances [33].

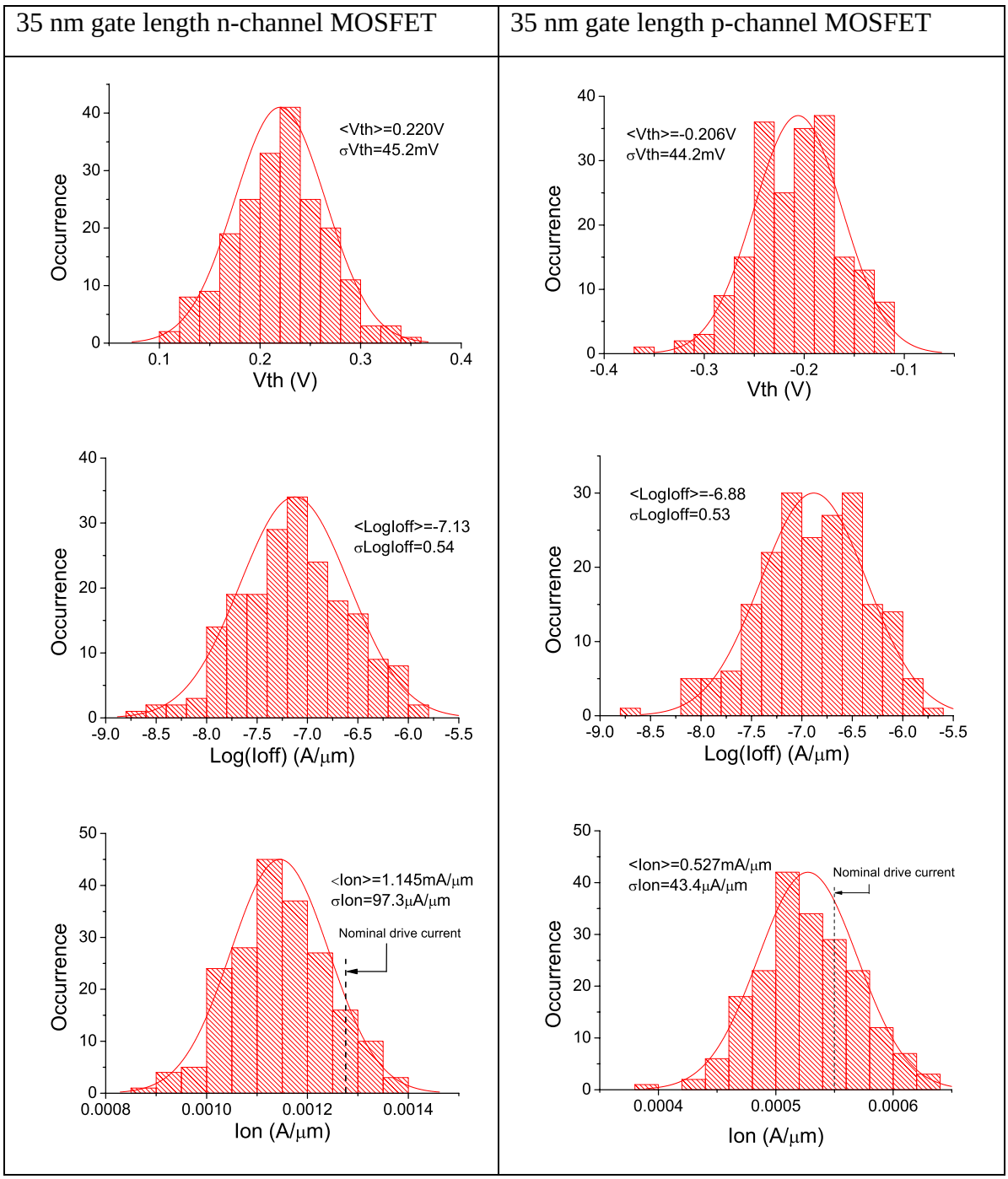


Figure 5.4 Histograms of intrinsic parameter fluctuation for 35 nm poly-gate MOSFETs.

Figure 5.5 illustrates the impact of the variability sources on the sub-threshold slope distribution. The average sub-threshold slope is 88mV/dec and 90mV/dec for n-/p-MOSFETs respectively. Under the thermal limit at T=300K, the major part of the distribution spreads to the right side, resulting in an elongated tail on the right side. The short-channel effects are also influenced by the variability sources, and this is demonstrated by the DIBL variations in Figure 5.5. The average DIBL is 95.5mV/1V and 124.8mV/1V for nMOS and pMOS respectively while the standard deviations are both around 15mV/1V. The variations in the drain induced barrier lowering are introduced by

both random dopant variation close to drain affecting the drain field penetration in the channel end and also by the gate LER.

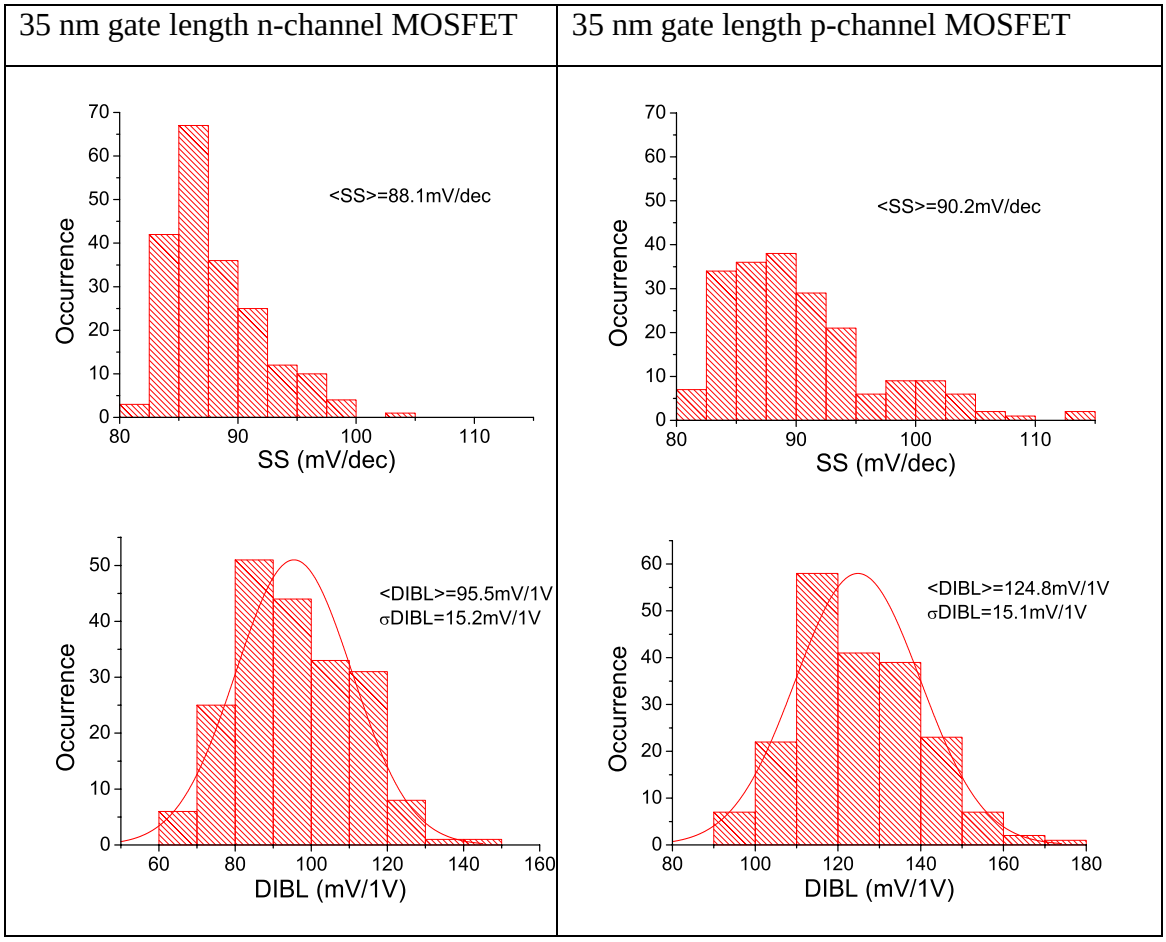
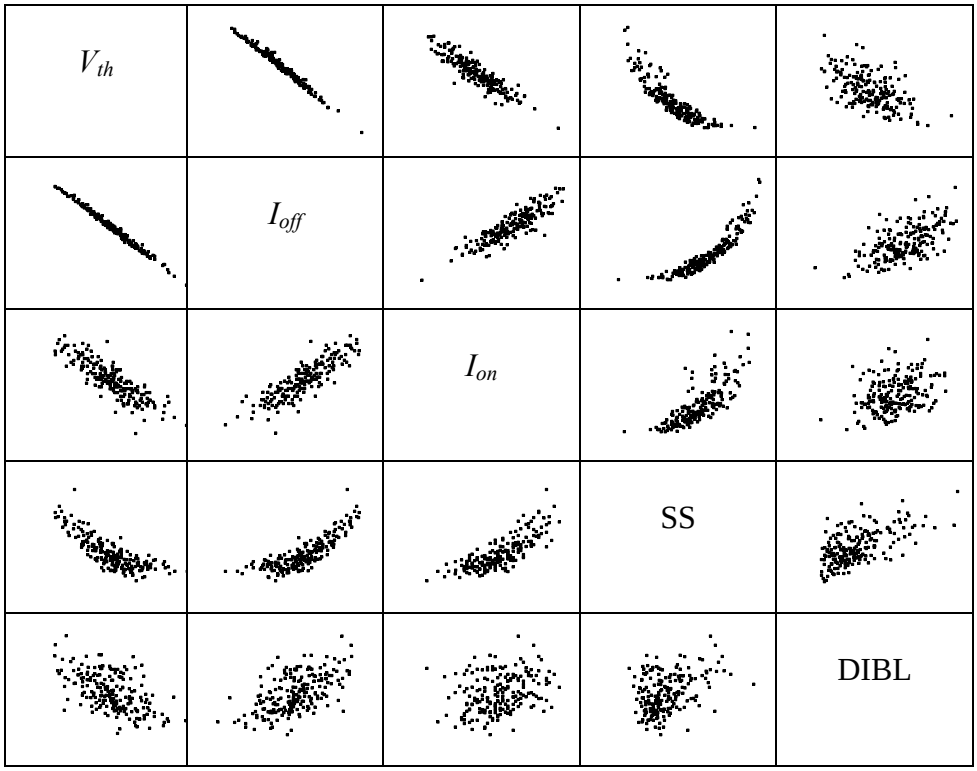


Figure 5.5 Histograms of sub-threshold slope and DIBL in 35 nm poly-gate MOSFETs.

The scatter plots of the statistical variability of the figures of merit for the 35 nm MOSFETs are presented in Table 5.1. Subject to statistical variability, the device parameters scatter over a certain range around the design point. As expected, there is a strong linear correlation between V_{th}, I_{off} and I_{on}, which are the three most important device parameters in determining the *I-V* characteristic of the individual transistors, all determined by the random dopant distribution near the source of the transistor. The I_{on}/I_{off} ratio apparently varies from device to device. A strong correlation exists between SS and V_{th}, I_{off} and I_{on}. The DIBL which is mainly determined by the random dopant distributions near the drain is less correlated to the rest of the parameters. Such scatter plots illustrate the influence of statistical variability on the figures of merit, and provides the benchmarking references for statistical compact models (SCM) [177].

Table 5.1 Scatter plots between figures of merit in 35 nm physical gate length MOSFETs, down-left: nMOSFET, up-right: pMOSFET.



5.3. Statistical variability of 25 nm poly-gate MOSFETs

Although metal gate stacks have been introduced by Intel in the 45 nm technology generation, the rest of the semiconductor players still use poly-gate technology in the 45 nm CMOS technology generation. Therefore, the evaluation of statistical variability in scaled poly-gate MOSFETs for the next generation technology nodes is still useful. As a result, the variability of a 25 nm physical gate length MOSFET is investigated in this section. The procedures are similar to that already followed for the 35 nm gate length MOSFET. The considered sources of variability are RDD, LER and PSG for the n-MOSFET, and RDD and LER for the p-MOSFET. An ensemble of 200 sample devices is simulated for each transistor type.

Figure 5.6 clearly shows the wide distribution of I_d-V_g characteristics of the 25 nm n-MOSFETs under the influence of statistical variability. The variation in leakage current has spanned over 4 orders of magnitude and could probably lead to malfunctioning of logic gates in large chips and to overall reduction in yield. Larger threshold voltage fluctuations compared to those found in the 35 nm device can be deduced from the semi-log scale I_d-V_g transfer characteristics. A similar spread can be observed in the characteristics of the 25 nm p-MOSFETs illustrated in Figure 5.7. This means that the performance of 25 nm

poly-silicon gate CMOS will be seriously impaired by increased statistical variability, giving a warning for next generation CMOS technology design. Improving short-channel effect and gate capacitance, and reducing the variability will be the driving force for switching to metal gate stacks in the following technology generations.

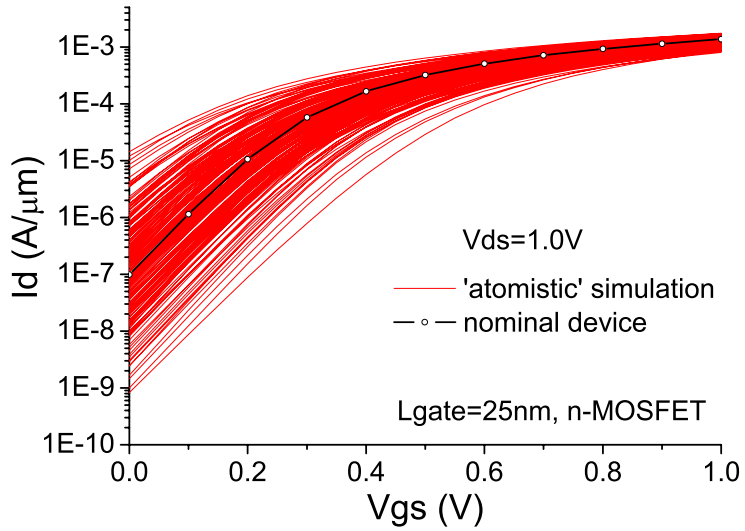


Figure 5.6 Id-Vg characteristics for 25 nm gate length poly-gate n-channel MOSFETs subject to RDD, LER and PSG statistical variability.

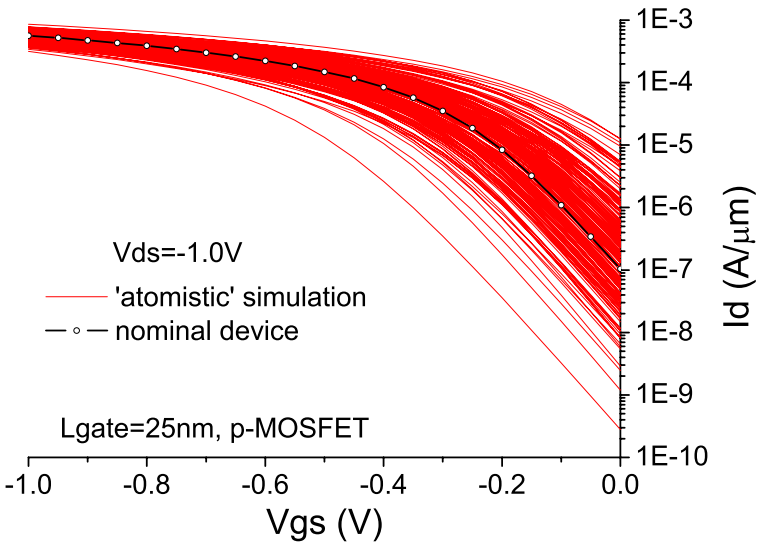


Figure 5.7 Id-Vg characteristics for 25 nm gate length poly-gate p-channel MOSFETs subject to RDD, LER statistical variability.

Similar to the analysis of the 35 nm MOSFET a set of important device parameters are extracted in order to understand the impact of statistical variability on the scaled 25 nm poly-gate MOSFET. Figure 5.8 shows histograms of V_{th}, log(I_{off}) and I_{on} distributions.

The threshold voltages are extracted using current criteria of 1.06×10^{-5}A/μm and 8.4×10^{-6}A/μm for the n-MOSFET and p-MOSFET respectively. The threshold voltages distributions in Figure 5.8 are for a drain voltage 1.0V. The V_{th} fluctuations are larger than in the 35 nm MOSFET. The standard deviation (σV_{th}) is 80.2mV and 81.8mV and the average threshold voltage is 0.206V and -0.185V for n-MOSFETs and p-MOSFETs respectively. The increase in variability compared to the 35 nm transistors is due to the combined effect of reduced gate area, increased channel doping and the increased impact of the poly-silicon depletion.

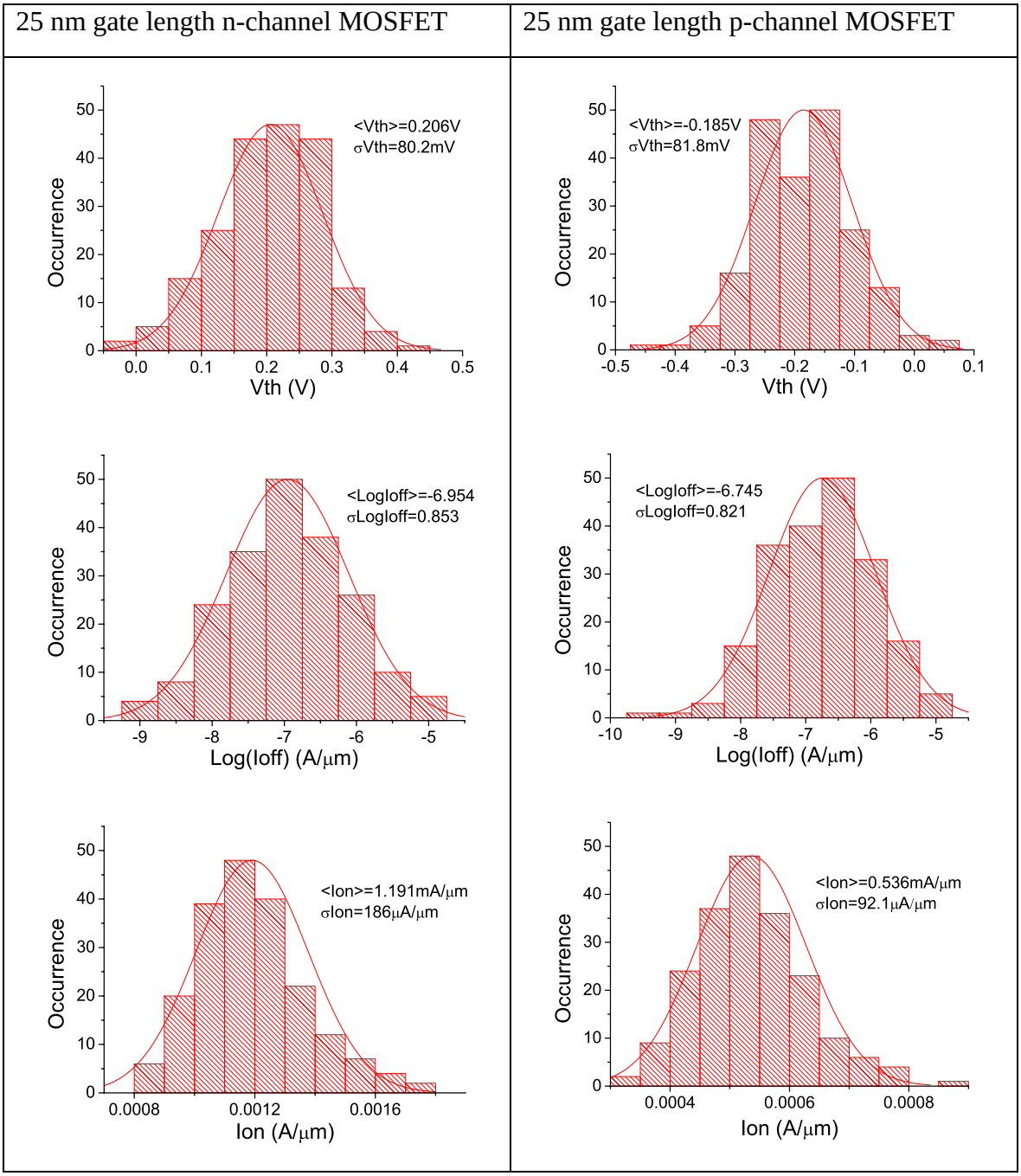


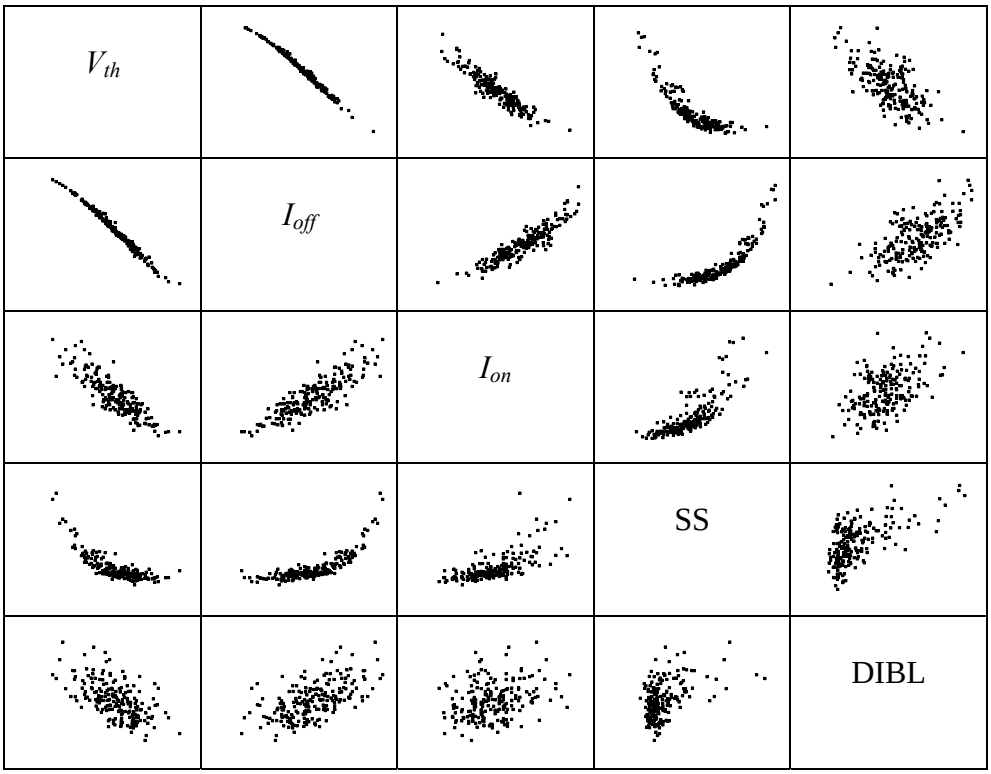
Figure 5.8 Histograms of intrinsic parameter fluctuation for 25 nm poly-gate MOSFETs.

The off-current variation in the 25 nm device is also bigger. The leakage current overshoots the 10^{-5} A/μm range, which is two orders of magnitude higher than the design target. The corresponding log(I_{off}) standard deviation is 0.853 and 0.821 for n-channel and p-channel MOSFETs respectively.

Finally, the standard deviations in the drive current are 188 μA/μm and 92 μ A/μm for n-channel and p-channel transistors, respectively, which are 15.8% and 17.2% of the average currents respectively. The distribution of the drive current magnitude for both n-channel and p-channel 25 nm gate length MOSFETs is wider compared to 35 nm gate length counterparts.

The scatter plots of the statistical variability in the figures-of-merit for 25 nm MOSFETs are shown in Table 5.2. Similarly to the 35 nm MOSFET results, V_{th}, I_{off} and I_{on} are strongly correlated among each other but less correlated individually to DIBL.

Table 5.2 Scatter plots between figures of merit in 25 nm physical gate length MOSFETs, down-left: nMOSFET, up-right: pMOSFET.



5.4. Statistical variability of 18 nm metal-gate MOSFETs

The introduction of metal gate stack has been driven by several performance benefits. As described in Chapter 4, the 18 nm gate length MOSFET has been designed utilizing a

combination of high-k and metal gate. This section presents statistical variability of the 18 nm n- and p-MOSFET induced by RDD and LER. The variability introduced by the metal gate granularity and the corresponding variability in the metal grain work functions [178] is not included in this study. The high drain voltage I_d-V_g characteristics of an ensemble of 200 microscopically different MOSFETs are shown in Figure 5.9 for n-MOSFETs and in Figure 5.10 for p-MOSFETs. The I_d-V_g characteristics of the nominal 18 nm devices with continuous doping profile are also shown in corresponding figures.

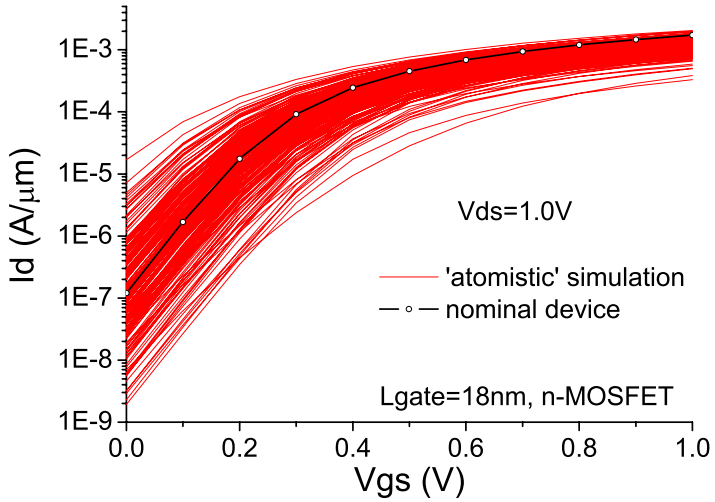


Figure 5.9 Id-Vg characteristics for 18 nm gate length high-k/metal gate n-channel MOSFETs subject to RDD, LER statistical variability.

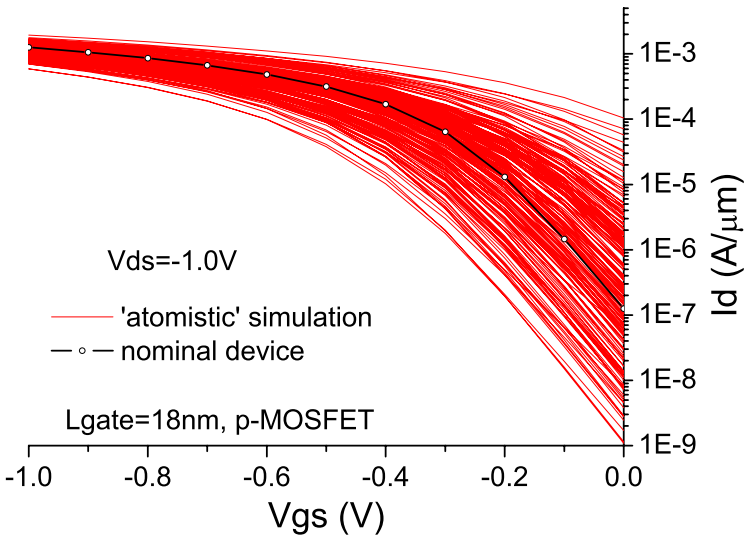


Figure 5.10 Id-Vg characteristics for 18 nm gate length high-k/metal gate p-channel MOSFETs subject to RDD, LER statistical variability.

It can be observed comparing Figure 5.6 and Figure 5.9, that spreading the I_d-V_g characteristics of the 18 nm n-MOSFETs in the sub-threshold region compared to the

25 nm generation has remained virtually the same. This is primarily due to the elimination of the poly-silicon depletion effect, leading to the improvement of gate capacitance and better electrostatic integrity. However, the comparison of Figure 5.7 and Figure 5.10 reveals that the 18 nm p-channel MOSFETs have larger sub-threshold variability even with the inclusion of a metal gate. Both types of 18 nm gate length MOSFETs have a larger drive current variation.

The statistical distribution of V_{th}, log(I_{off}) and I_{on} for the 18 nm high-k/metal gate MOSFETs are presented in Figure 5.11.

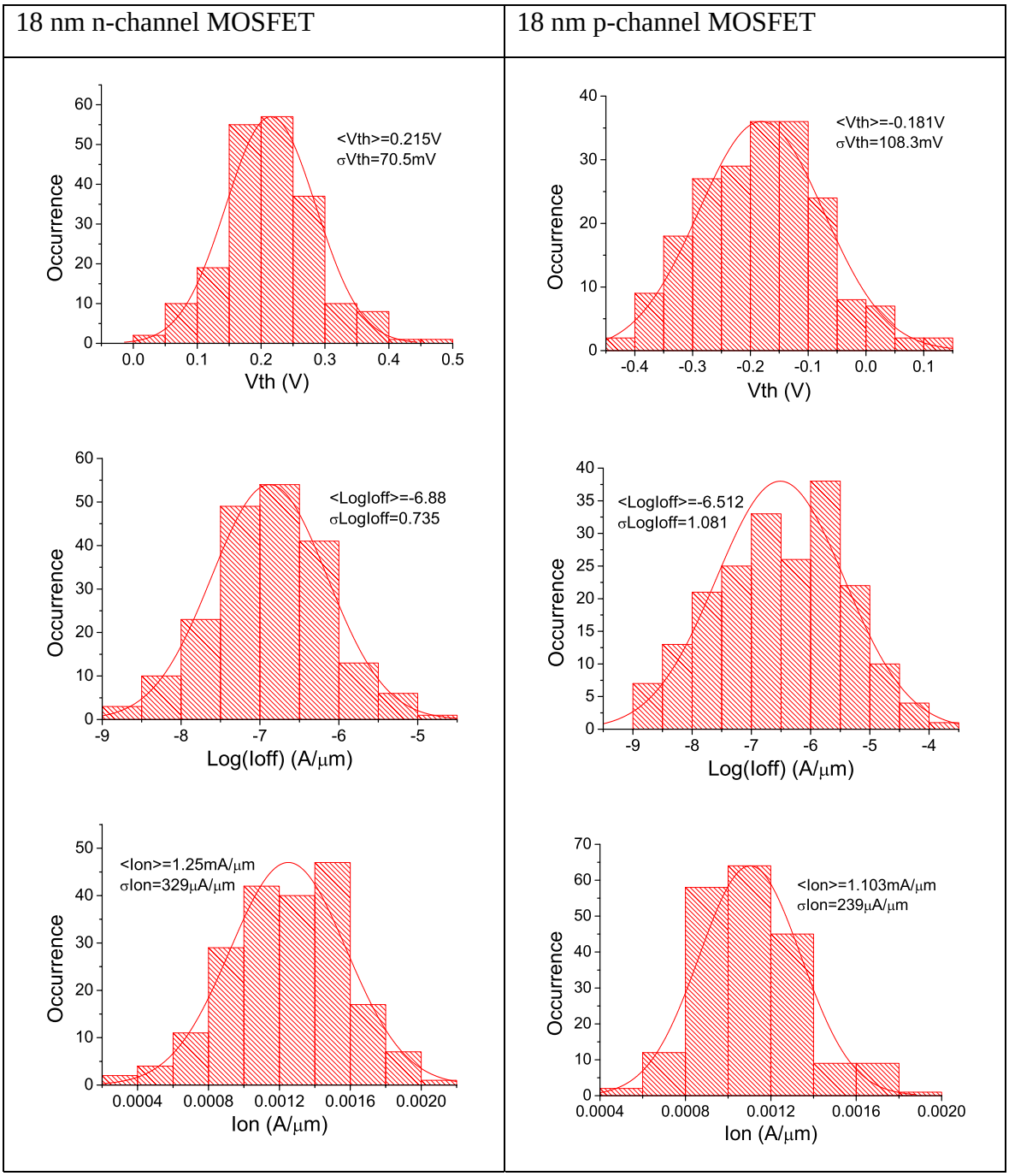


Figure 5.11 Device parameter variations for 18 nm high-k/metal gate MOSFETs.

The threshold voltage is estimated at drain currents of 1.76×10^{-5}A/μm and 1.31×10^{-5}A/μm for the n-MOSFETs and the p-MOSFETs respectively. The threshold voltage standard deviation (σV_{th}) of 18 nm high-k/metal gate n-MOSFETs is 70.5mV, smaller compared to that of 25 nm poly-gate n-MOSFETs. However for p-channel MOSFETs σV_{th} is 108.3mV, larger compared to that of the corresponding 25 nm MOSFETs. This mainly derives from the different SCE control of the n-channel and p-channel MOSFET. The strong SCE in the p-MOSFETs due to the deteriorating impact of SiGe on the S/D doping profiles significantly exacerbates the p-MOSFET statistical variability. As illustrated in Figure 5.12, the average value and the standard deviation of DIBL of the p-MOSFETs are much larger compared to that of the n-MOSFETs, indicating weaker gate control over channel and larger parameter variation.

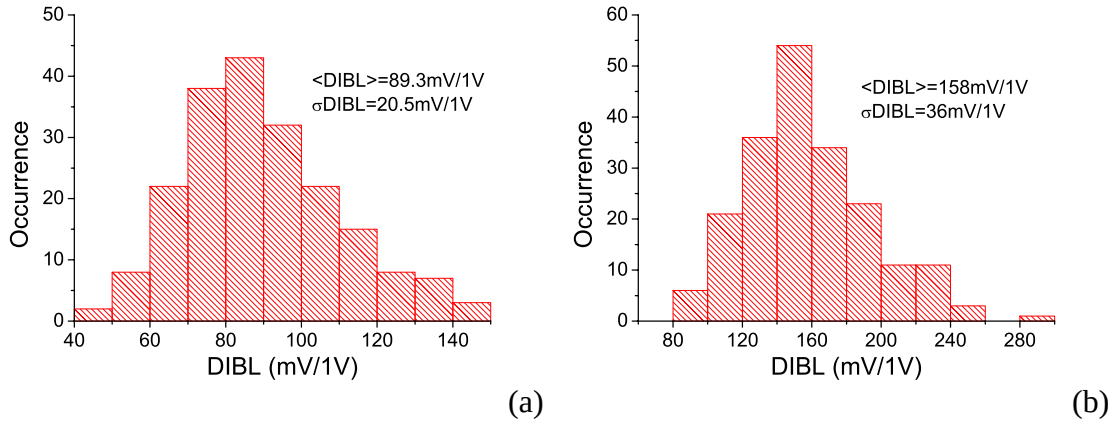
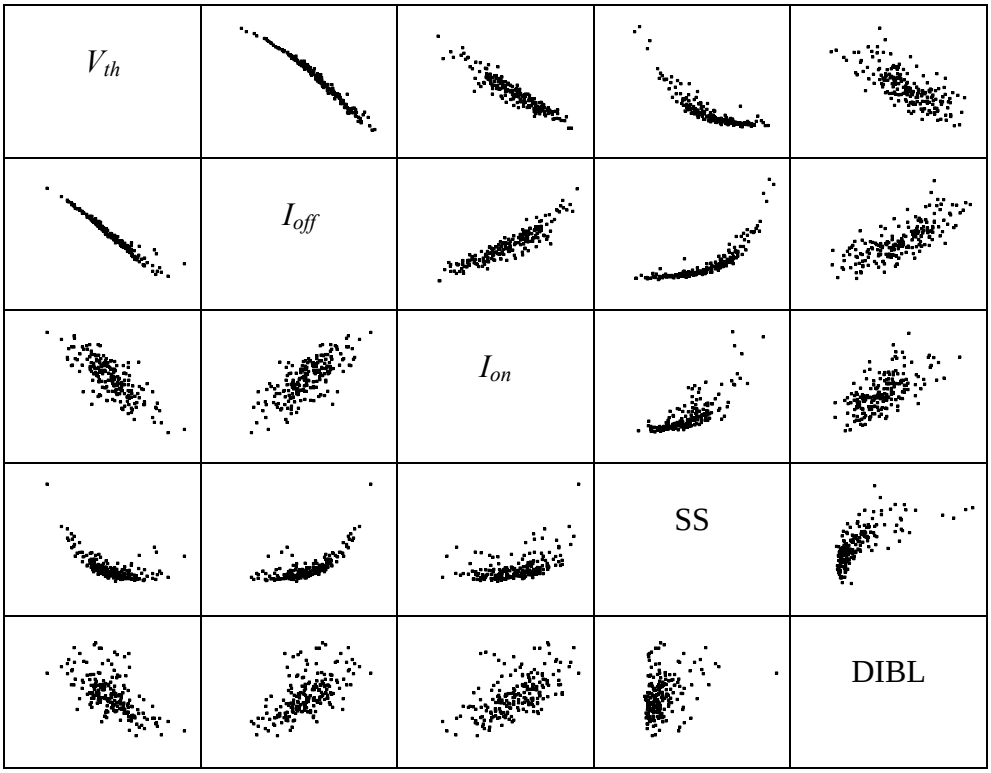


Figure 5.12 The drain induced barrier lowering (DIBL) as an index of SCE for 18 nm n-MOSFETs (a) and p-MOSFETs (b).

The off-current leakage variation is maintained at the same level as in the previous generation MOSFETs, but the drive current fluctuation becomes much larger. The drive current standard deviations become 21% and 26% of the corresponding n- and/p-MOSFETs average currents, which will cause significant issues due to timing variability.

The scatter plots of the 18 nm MOSFET's figures-of-merit under the influence of statistical variability are shown in Table 5.3. Similar correlations are observed in the 18 nm physical gate length MOSFETs when compared with the results for the 35 nm and 25 nm gate length devices.

Table 5.3 Scatter plots between figures of merit in 18 nm physical gate length MOSFETs, down-left: nMOSFET, up-right: pMOSFET.



5.5. Summary

This chapter presents a predictive simulation study of the statistical variability of scaled 35 nm, 25 nm and 18 nm gate length MOSFETs. The increasing trend of statistical variability of bulk MOSFETs is generally observed. It was illustrated also that careful design and technology boosters such as high-k/metal gate can moderate the increasing variability trend. The drastic increase of threshold voltage standard deviation from approximately 44mV to 80mV has been observed in the transition from 35 to 25 nm gate length in the presence of poly-Si-gate as shown in Figure 5.13.

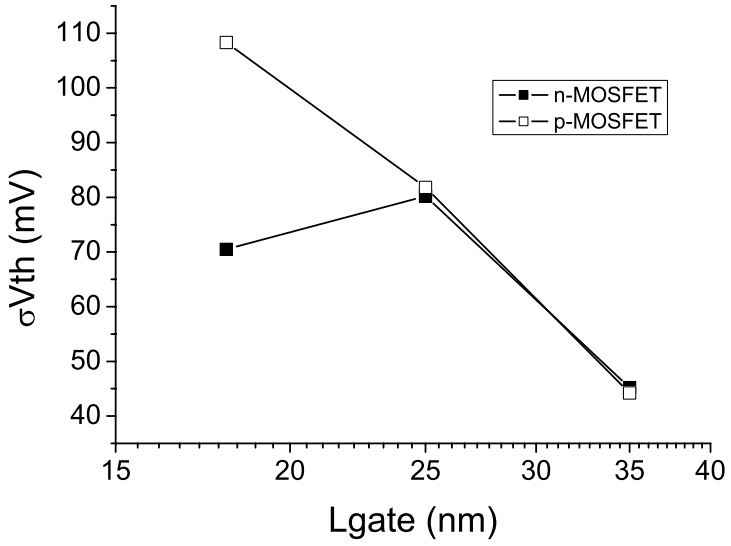


Figure 5.13 Threshold voltage standard deviation subject to statistical variability sources as a function of gate length.

The introduction of metal gate in 18 nm gate length MOSFETs reduce n-MOSFET σV_{th} to 70.5mV although the weak SEC control in p-MOSFETs renders its σV_{th} of 108mV. The scaling study of statistical variability carried out in this chapter affirms that the statistical variability will play an increasingly important role in future scaling and integration of bulk MOSFETs until the end of ITRS.

Chapter VI

6. Impact of strain and STI on variability

In this chapter we study the impact of realistic device structures on the variability of advanced MOSFETs. This includes the impact of strain or the impact of the shallow trench isolation on variability. Both effects have not been previously included in the variability simulations.

6.1. Strain enhanced LER variability

In last chapter, the major statistical variability sourced including RDD, LER and PSG are discussed in detail. Although RDD is still the dominant source of statistical device variability in bulk MOSFETs, LER can soon take over, if dramatic lithography improvements will not allow a drastic reduction in the current LER level. In modern devices, the process-induced stress engineering has been used to enhance device performance since the 90 nm technology generation. Deterministic studies and simulations of the impact of strain on devices subject to realistic layouts [179], and OPC lithography correction [180] have shown that geometry induced strain variations have strong impact on the transistor characteristics.

However, the details of how LER influences the channel strain distribution and how strain variability statistically influences device electrical characteristics introducing additional variability are still unclear. In addition, there has been no study of the effect of LER magnitudes on the variability in strained devices. This section presents 3D simulation methodology aimed to capture the effects of LER on statistical variability in strained devices, to quantitatively explore their magnitude. The simulation strategy and the device calibration will be covered. Channel strain and enhanced mobility variability due to strain variation will be investigated. Deterministic results are first presented and analysed. Statistical simulations have been done to investigate how strain variability gives rise to an increased variation in both drive current and leakage current, while indeed increasing the average on-current. The 3D strain variations introduced by LER have been investigated to illustrate the origin of the additional variation due to strain variability. Different

magnitudes of LER have a strong effect on the variability of MOSFET electrical characteristics both with and without the inclusion of strain [181][182].

6.1.1. Simulation methodology

Using the process and device TCAD simulation suite Sentaurus, process simulation is carried out to replicate the doping profiles of a real 35 nm physical gate length nMOSFET fabricated by Toshiba, whose 2D simulation details have already been provided in the first section of Chapter 4. The related three-dimensional simulation methodology is illustrated in Chapter 3. The channel width of the simulated device, shown in Figure 6.4, is 80 nm. We introduce statistically modelled LER at the polysilicon patterning stage as discussed in detail below. We also deposit a cap (contact etch stop) layer after source/drain formation in order to introduce tensile strain into the channel region. The process flow, from p-well implants to the final cap layer thermal processing, is depicted in Figure 6.1.

The process simulation is carried out in a 3D simulation domain and associated meshing, with the simulation of processes such as implantation and diffusion considered fully in 3D. Computationally more realistic than 2D [183] and pseudo 3D [155] approaches, a fully 3D simulation approach naturally allows the correct simulation of LER. It is necessary to simulate fully 3D devices to accurately model the effects of LER and its impact on strain. In addition, narrow devices are strongly affected by shallow trench isolation (STI), which makes 3D device structures produced using 3D process simulation vital for accuracy. LER is introduced by the photo-resist mask at the process simulation stage, and its traces are modelled based on real LER captured from different examples of 193nm lithography [155]. The spacer deposition following gate patterning almost mirrors the roughness of the gate edges, which guarantees effective transfer of strain from the cap layer.

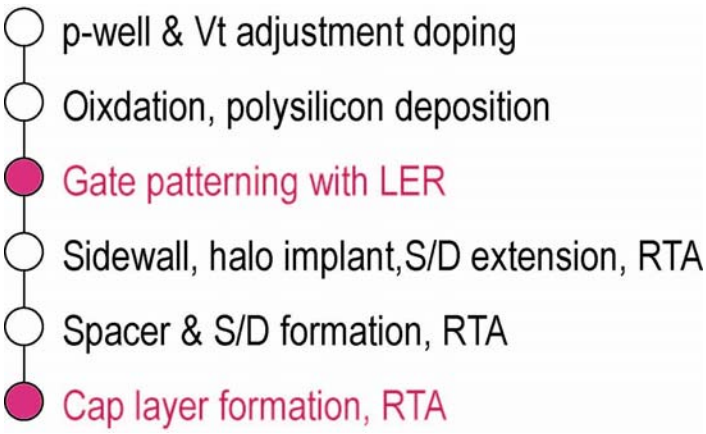


Figure 6.1 3D process flow of modelling LER. LER is introduced at gate patterning and the tensile stressor is a deposited cap layer.

In real devices, strain may be introduced by oxide growth, thermal mismatch and lattice spacing mismatch, but the majority of the strain in these devices is due to the cap stressor layer. This is therefore the only source of strain modelled in this analysis. Visco-elastic stress models are used to capture the strain in the various materials of the device. Silicon dioxide and silicon nitride are considered as visco-elastic materials, while silicon and poly-silicon considered as purely elastic materials.

6.1.2.LER and strain variability

6.1.2.1. Line edge roughness

In this work, LER is considered as a fundamental source of statistical variability which cannot be avoided. It is due to the underlying molecular granularity of the lithographic process and the corpuscular nature of the light. We use a Fourier synthesis-based approach developed in Glasgow to model device LER [155], parameterised by the rms magnitude of the LER, Δ and the correlation length of the variation, Λ. Typical values of LER are $3\Delta \approx 5$ nm and from 10 nm to 50 nm for Λ, depending on the physical nature of the lithography. The power spectrum of the associated Gaussian autocorrelation function is used to generate sample LER.

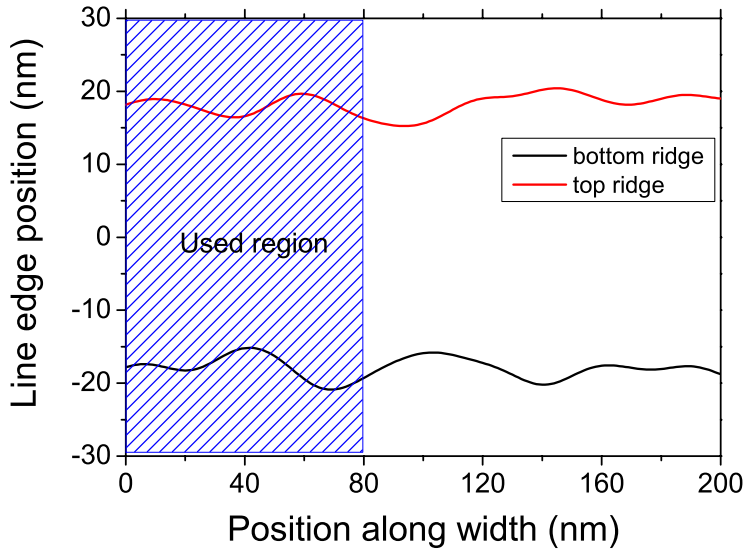


Figure 6.2 LER sample generated by inverse Fourier transform of a Gaussian autocorrelation function. Gate width is along x-axis.

Figure 6.2 shows an example of LER with $\Delta = 2$ nm and $\Lambda = 20$ nm. In our simulations, this magnitude and shape of LER is introduced to pattern the polysilicon gate. This introduces LER into the implanted doping profiles, a roughness only partially smoothed by the implantation spread and rapid thermal annealing later in the process flow. Results of

gate LER on the doping post-anneal are shown in Figure 6.3. LER clearly causes a fluctuation in the *p-n* junction definition, locally altering the effective gate length, and therefore the electrical properties of the transistor. Due to the gate acting as a mask to source/drain extension ion implantation, the gate LER is definitely duplicated to some degree in the roughness of the doping profile, causing the *p-n* junction line to fluctuate. This situation is especially evident just below the gate. However, it has to be noticed that even here, thermal annealing smoothes in part the junction roughness, and suppresses the highest frequencies.

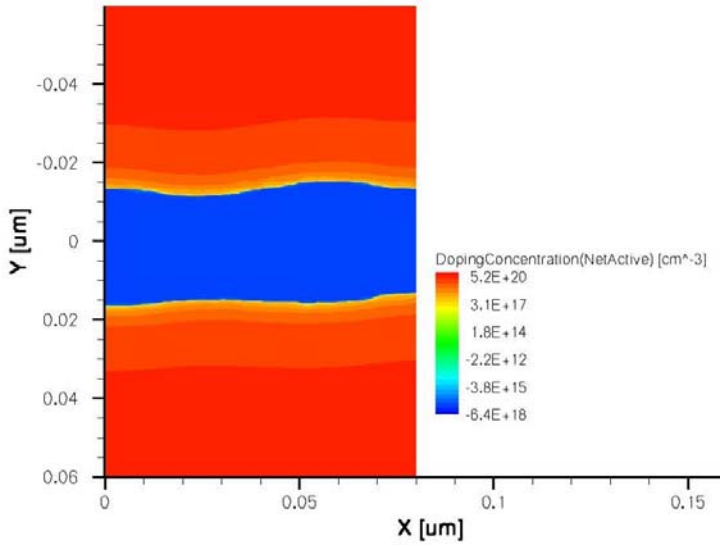


Figure 6.3 Channel doping profile 1 nm below the gate SiO_2/Si interface. Junction line subject to LER is smoothed by the RTA step, but still exhibits fluctuation.

6.1.2.2. LER induced strain variability

The initial intrinsic stress of the cap layer is set at 1.8 GPa, with shear stresses set to zero. The induced tensile strain is transferred to the whole device during an RTA process (Figure 6.4). This leads to a uniaxial tensile strain along the channel with a compressive strain normal to the channel plane. As an observation of stress components in the channel 1 nm below the oxide/silicon interface: σ_{yy} is on average 1.267×10^{8} Pa, σ_{xx} is -5×10^{6} Pa, σ_{zz} is -1.445×10^{8} Pa while shear stress components are comparatively small: less than 10^{6} Pa. A rough calculation [123] based on above stress levels shows the *yy* value of the mobility enhancement tensor to be 1.21, which means that strained devices should exhibit an ~21% drain current increase at low drain voltage.

A illustrated in Figure 6.4, the strain is no longer constant across the channel. This strain variability leads to mobility variations along the channel width (via the stress dependent mobility enhancement model) and enhances *I/V* curves variability from device to device.

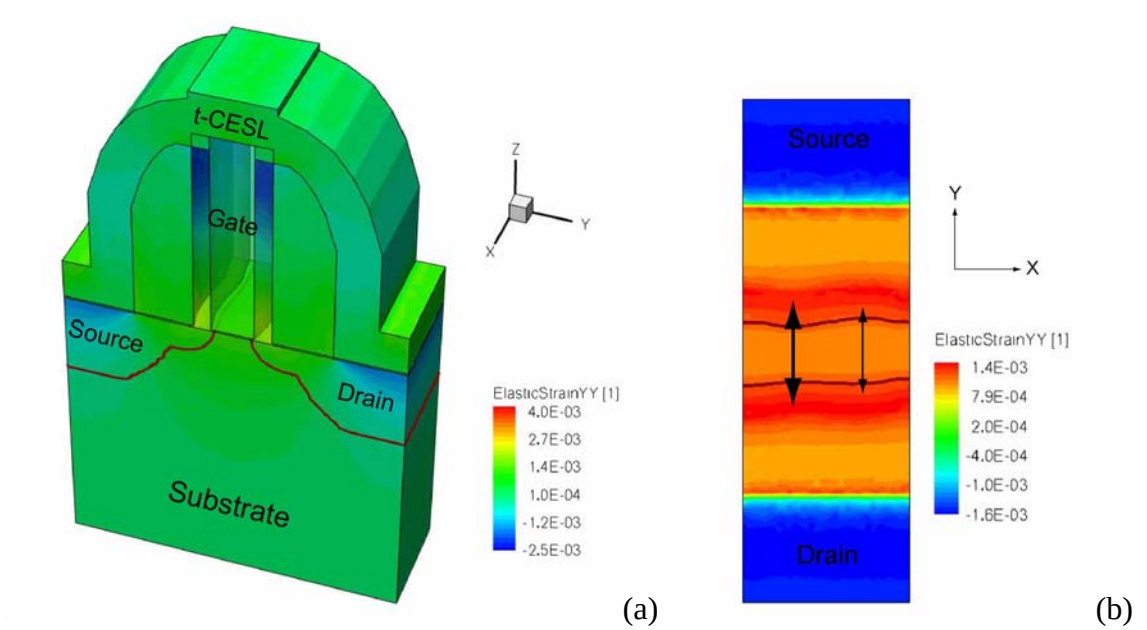


Figure 6.4 (a) Left graph is the distribution of channel direction normal strain. High tensile strain in the channel is transferred from the nitride cap layer above S/D; (b) Right graph is 2D cross section view of channel direction strain 1nm below the oxide/Si interface. Stronger tensile strain is induced in local channel shortening.

Due to gate LER, delivery of the strain to the channel also reflects the rough boundary of gate line, giving rise to an irregularity of strain across the width of the channel direction. This local strain variability further induces local variability in the carrier velocity/mobility enhancement, and hence overall current variability, in addition to that due to local channel length variability in unstrained devices. As a sample, the top trace in Figure 6.5 shows that the local shortening of the channel has higher local channel-direction normal strain, and the bottom figure shows that the average electron velocity in this strained device has increased by about 21.5% compared with a corresponding unstrained device (extracted when V_{gs} = V_{ds} = V_{dd}), with local fluctuations which follow the induced strain. LER in an unstrained device can induce local current variations, but local strain fluctuations induce additional mobility variability with a highly consistent fluctuation trend, overlaid on an average mobility increase due to the average tensile strain. The mobility of strained devices in the channel shortening regions of Figure 6.4 and Figure 6.5 increases more due to higher strain, and this in turn strengthens the 'hotspot' effect, introducing additional variability.

Notice that the strain variation across the channel is smaller than might otherwise be expected, because the gate boundary (subject to LER) is remote from the highest stress areas by the spacer thickness, and because substrate tensile stress provides a measure of stress compensation in the channel. Decreasing the spacer thickness between gate and source or drain from 52nm to 32nm can significantly enhance the tensile strain in the channel, and it also enhances the strain variation (Figure 6.6). For the 52nm spacer, the

mean value of strain component *yy* in Figure 6.5 is 1.0883×10^{-3}, and *Max-Min*=3.868×10^{-5}, therefore the fluctuation ratio *(Max-Min)/Average* is of the order of 3.6%. However for a thin 32 nm spacer (Figure 6.6), the average is 1.5156×10^{-3}, and *Max-Min*=6.2×10^{-5}, so the fluctuation ratio is approximately 4.1%. This indeed illustrates enhancement of the impact of LER on strain variability when the distance between gate and source or drain gets smaller.

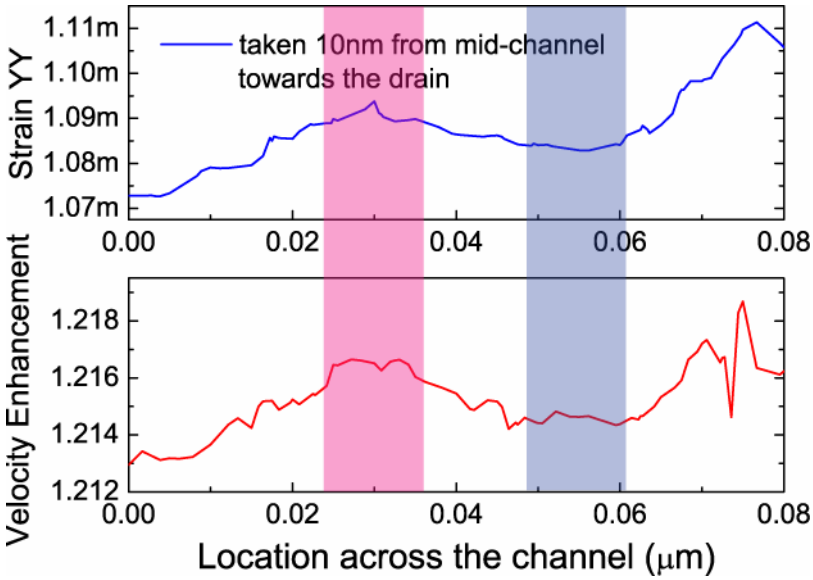


Figure 6.5 The upper trace is the 1D elastic strain yy value across the device width at 10nm away from the middle of the channel towards drain, 1nm under the gate dielectric. The bottom trace is the corresponding electron velocity enhancement due to the strain, indicating the strain induced mobility variability in the nMOSFET.

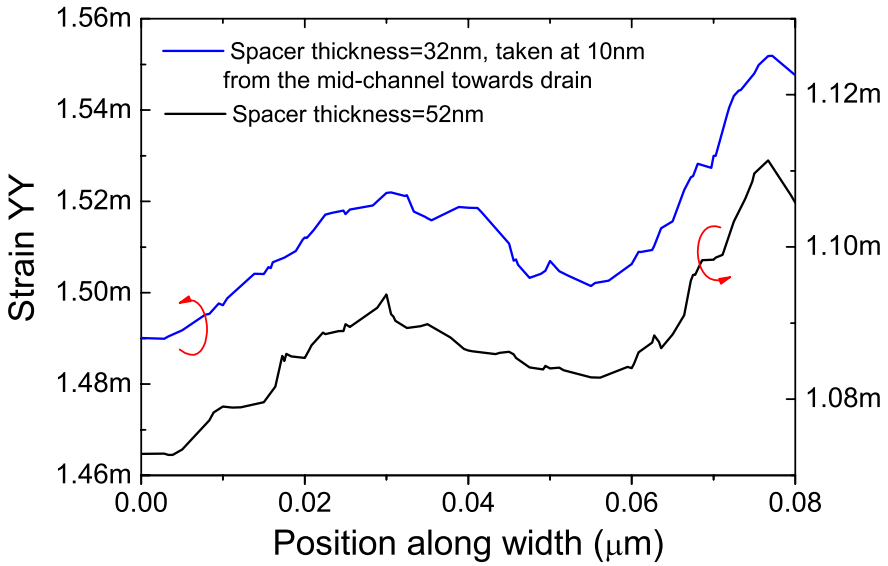


Figure 6.6 A thinner spacer guarantees increased strain, but also increases the strain variability due to LER.

6.1.3. Strain enhanced electrical variability

6.1.3.1. Deterministic performance enhancement of strain

The original intrinsic cap layer stress and cap layer thickness determine the overall strain, while gate length and spacer thickness are important factors affecting channel strain distribution. 2D simulations have been performed to explore how these factors determine the transistor drive current. The magnitude of the cap stress directly determines on-current enhancement; the more stress originated by the cap layer, the more drive current from the device. A linear relationship between drive current and intrinsic cap layer stress is shown in Figure 6.7. The cap layer thickness is also a direct factor influencing on-current. Thicker cap layers can deliver more strain and more performance enhancement. Obviously unlimited intrinsic stress is impossible, and only an effective cap layer with finite thickness can be deposited due to transistor size.

However, device scaling provides additional advantage in stress engineering. Right graph of Figure 6.7 shows that drive current will increase as gate length gets smaller. This wins back some benefits from the scaling limitations of small devices.

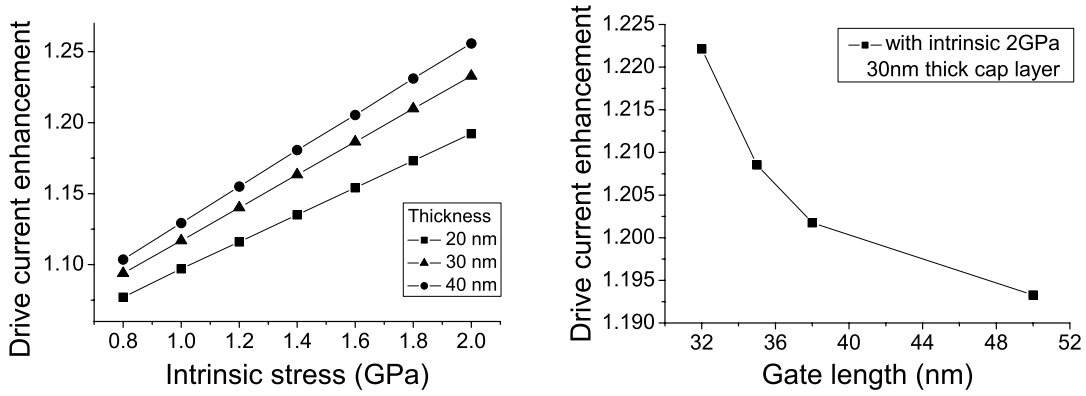


Figure 6.7 The left graph shows on-current enhancement dependence on intrinsic stress and tensile cap thickness. The right graph shows the relationship of drive current to gate length, with fixed spacer size.

6.1.3.2. Strain enhanced statistical variability

Intentional stress aimed at improving device performance can significantly increase mobility and therefore drive current. However, mobility changes due to LER and varying strain may also enhance statistical variability caused by LER. Based on our calibrated device, 50 nominally identical devices patterned with different LER have been simulated (Figure 6.8). The LER has typical parameters of root mean square (rms) $\Delta = 2.0$ nm and correlation length 20 nm.

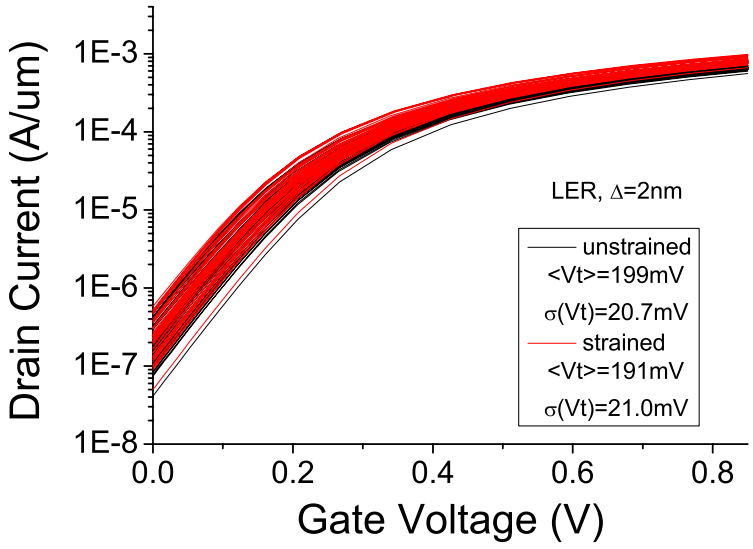


Figure 6.8 50 simulated high drain Id-Vg curves of nominal identical devices with/without stress, influenced by LER Δ=2 nm.

On-current distribution with and without the effects of strain is shown in Figure 6.9. Strained devices (simulated using the basic piezoresistance model) deliver more drive current in each device compared to the unstrained counterpart. LER leads to variation of drive current, no matter if the device is strained or not. A statistical analysis for LER Δ=2 nm shows that devices without strain have a standard deviation 58 µA/µm, with a mean value of 681 µA/µm. Strained devices have a standard deviation 69 µA/µm with a higher mean current of 813 µA/µm (Figure 6.9 (a)). Both the mean and standard deviation of the strained devices increase by 19.4%. It seems that the mobility enhancement indeed increases current, and makes both the mean and standard deviation proportionally bigger. Figure 6.10 emphasizes this point by showing the linear correlation between individual drive currents of unstrained and strained devices.

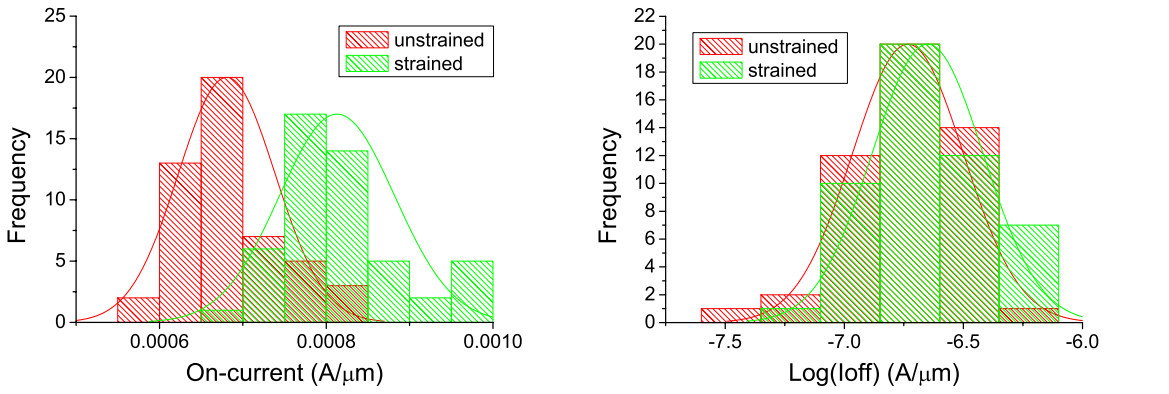


Figure 6.9 The left graph (a) shows statistics of on-current variation. Strained devices have larger current on average, but also bigger variation. The right graph (b) shows off-current distribution of devices with/without strain.

As illustrated in the logarithmic distribution plot of leakage current, Figure 6.9, the off-currents of the unstrained and strained devices can be approximated by a log-normal distribution. Interestingly, although the average leakage current log(I_{off}) is enhanced in the strained devices, from -6.73 to -6.65, the standard deviation in log scale is almost identical around 0.23.

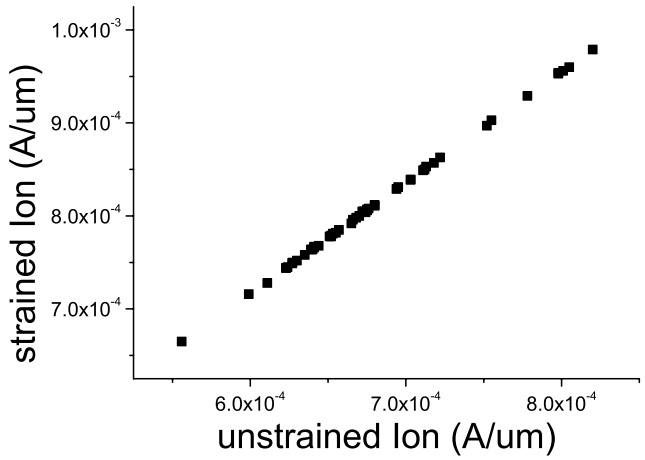


Figure 6.10 The correlation of strained device on-currents to unstrained device on-currents.

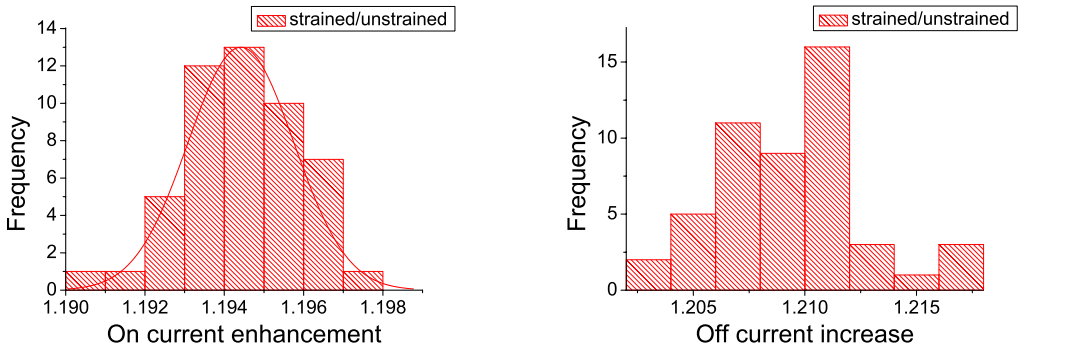


Figure 6.11 The left graph shows the statistics of on-current enhancement. The right graph shows the statistical distribution of off-current due to strain, which is wider compared with the drive current enhancement distribution. Strained devices show additional variation due to local fluctuations in mobility enhancement.

Figure 6.11 illustrates the distribution of the on- and off- current *enhancements* due to strain. The on-current enhancement, reflects the fact that the mobility for every device does not increase by the same factor. The specific electrical characteristics of each device are determined by the device's specific local channel strain variability, a function of the LER in the device, and the specific local strain produced by this LER. The enhancement has an approximatley normal distribution The average on-current enhancement is 19.44% with a standard deviation 0.13%. As shown in the previous subsection, the effect of gate spacers on delivery of strain to the channel is one reason for this small variation. This variability in

the overall on-current due to LER induced strain variations, produces a variability on top of the localised mobility variation. The variability of the off-current increase has a larger standard deviation than that of the on-current (shown in Figure 6.11 (b)). The off-current enhancement has an average increase 20.94% with standard deviation 0.31%.

Simulations for devices with LER rms Δ = 1 nm have also been carried out. Results comparing the magnitudes of LER are shown in Figure 6.12. The LER magnitude has a strong effect on both drive current and leakage current variability whether devices are strained or not. Larger LER increases the variation of I_{on} and I_{off}. In addition, an increase in LER mildly increases the average I_{on} and significantly increases the average I_{off}. The strained devices have as expected an average enhancement, but also suffer an increase in average leakage current. Independent of LER values, the strained devices have higher I_{on} variation and almost identical I_{off} variation compared with unstrained devices.

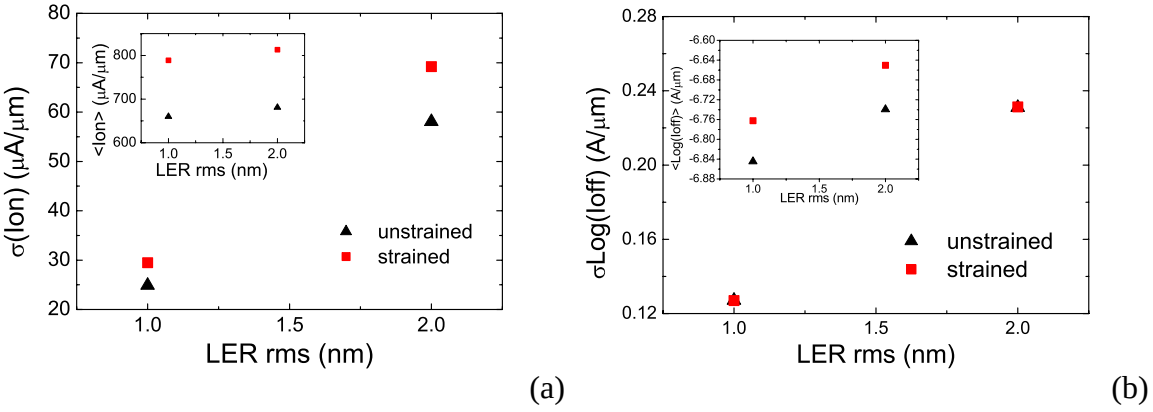


Figure 6.12 Statistical results of drive current (a) and off current (b) for LER Δ=1nm and 2nm.

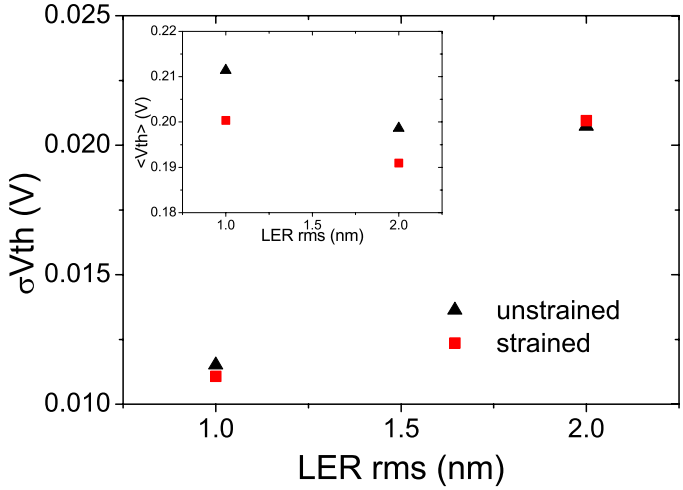


Figure 6.13 Statistical results of saturation threshold voltage for LER Δ=1nm and 2nm.

LER also causes threshold voltage fluctuations, and as the magnitude of the LER increased, the threshold voltage fluctuations increase. However there is little difference in the V_{th} fluctuations between two sets of strained and unstrained devices at a fixed LER (as illustrated in Figure 6.13). V_{th} has a standard deviation of 11 mV at Δ = 1 nm, but it increases to 21mV at Δ = 2 nm. From the insert figure, it is also clear that the LER magnitude also affects the average threshold voltage. Bigger LER results in a smaller average threshold voltage in both strained and unstrained cases. Strain induced increase in the sub-threshold current results in an overall threshold voltage lowering.

6.2. STI effects in decananometre MOSFETs

As already discussed, increasing transistor variability has become a critical issue to future CMOS scaling and integration. It has already dramatically affected SRAM and a variety of design measures have had to be implemented, including redundancy, differential biasing and localised sensing strategies to achieve gigabyte SRAM functionality in the presence of increasing variability [184]. The main reason for the acute SRAM sensitivity to variability is the minimum width of the SRAM transistors. This makes them extremely sensitive to systematic variations introduced by a variety of narrow-width effects and shallow trench isolation non-uniformities [185][186]. At the same time, the SRAM cell transistors are acutely susceptible to statistical variability introduced by the discreteness of charge and granularity of matter since in most of the cases, including the impact of random discrete dopants (RDD), the statistical variability increases with inverse proportionality to the active gate area of the transistors [32].

In this section, using simulations with both commercial TCAD tools and the Glasgow statistical variability simulator, we study several aspects of the systematic and statistical variability in contemporary narrow width MOSFETs, primarily from the point of view of SRAM applications. The template transistor design used in this study is presented. It first focuses on systematic variability associated with the narrow-width effect enhanced or suppressed by different designs of the STI structure. The statistical variability introduced by random discrete dopants, which is the main source of statistical variability in bulk MOSFETs, in the presence of STI is studied. Reliability issues related to STI in narrow-channel MOSFET are discussed and their role in statistical variability is elaborated in detail.

6.2.1.STI structure in narrow-channel MOSFETs

Here we use the redesigned 35 nm gate length MOSFET as the template transistor, whose design has been outlined in Chapter 4. However, the full scale 3D process is used, resulting in a device shown in Figure 6.14. The STI process is tuned to achieve reduced channel-width sensitivity. A combination of silicon etching, thermal growth, oxide deposition and CMP is used to achieve STI with minimum recess and rounded silicon corners to limit electric field crowding in the STI corners of the channel. The subsequent process steps include well doping, poly-gate patterning, source/drain extensions, halo doping, spacer formation, and source/drain doping. The poly-silicon gate is deposited and patterned over the STI. Device simulations are carried out employing the drift-diffusion approach with density gradient corrections. The stress effect of the STI is not simulated in order to isolate the study from compensation effects due to compressive stress, which results in n-MOSFET threshold voltage increase [187].

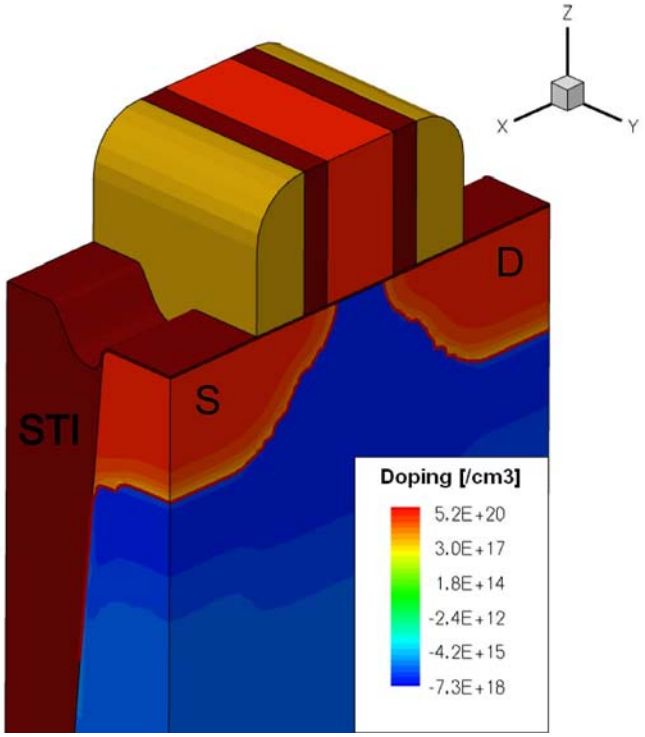


Figure 6.14 Process simulation of the 35 nm gate length n-MOSFET with edge STI is showing the structure of 35 nm channel width and net active doping.

The same device is also simulated using the Glasgow ‘atomistic’ simulator, which produces similar simulation results, after careful calibrations, to TCAD simulations in the absence of STI. The focus of the simulations is the statistical variability induced by random discrete dopants in the presence of STI. The isolation spacing is generally 70-100 nm in 45 nm technology [160] and STI spacing here is set to 70 nm (simulating one

half), in contrast to 100 nm in the TCAD simulations, to reduce the computational cost of the statistical simulations. This results in a small difference between the TCAD and 'atomistic' simulation results for a continuously-doped transistor in the presence of STI.

6.2.1.1. Inverse narrow channel effect

In narrow channel MOSFETs, the electrostatic behaviour near the STI and gate end determines the threshold voltage dependence on the channel width. The STI design has a strong impact on the narrow-channel effect, therefore different STI geometries are explored here using TCAD simulations. Four generic cases of STI isolation, illustrated in Figure 6.15, are simulated and their impact on the width dependence of the threshold voltage is investigated. The width dependence of the I_D-V_G characteristics of *STI 1* is shown in Figure 6.16. This is the worst structure in terms of narrow-width behaviour. At a channel width of 70 nm the leakage current is almost one order of magnitude larger than that in a control device without STI, and the saturation threshold voltage $V_{th,sat}$ is reduced by more than 100 mV in comparison to the control one. With the increase in the channel width both sub-threshold leakage and saturation current are reduced but stay above the corresponding values of the control structure.

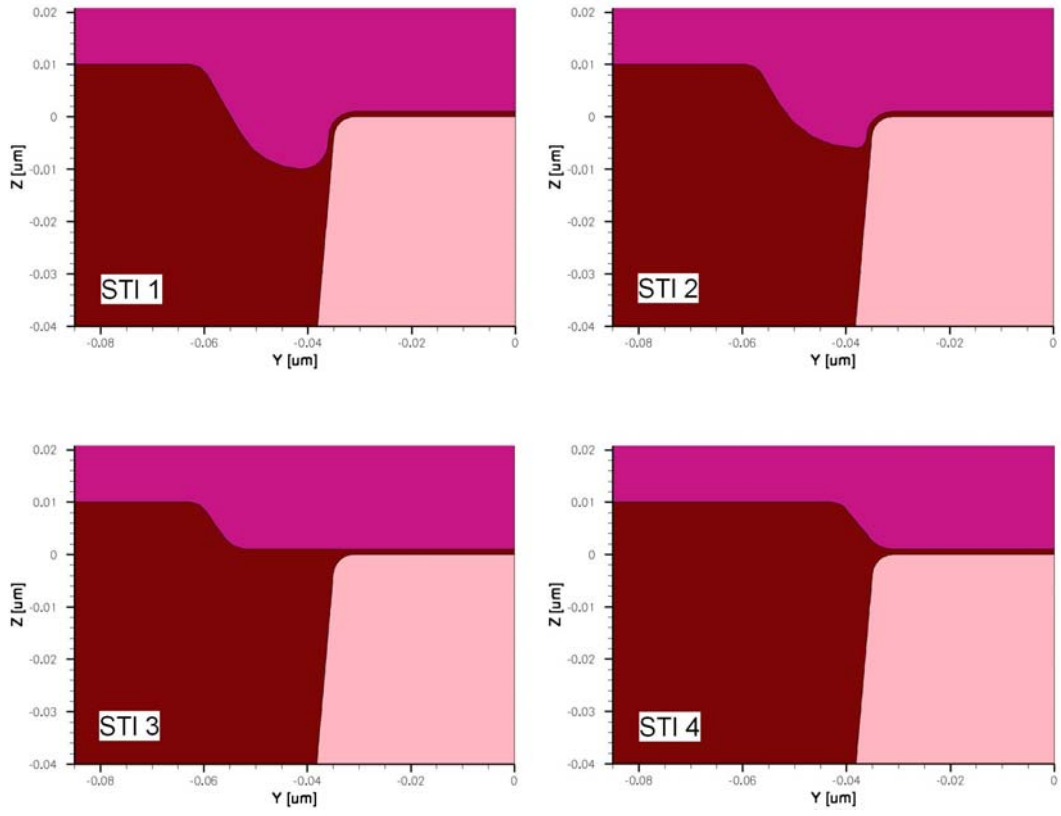


Figure 6.15 Various STI geometries are compared and studied to explore the narrow-channel effect.

Figure 6.17 compares the width dependence of the threshold voltage for those four STI structures. The rising STI oxide in *STI 4* results in a considerable decrease in the threshold voltage width-sensitivity compared to all other cases. This, in the case of SRAM design, will result in less variation in the threshold voltage due to problems associated with CD control, OPC and defocus. Our simulation results match well the experimental measurements of threshold voltage decrease in minimum-width transistors, which is of the order of 100 mV [188][189]. The significant width dependence of the threshold voltage observed in all STI cases means that threshold voltage CD sensitivity cannot be completely avoided in minimum-width design of SRAM applications [190]. Further optimisation of the STI design requires more realistic considerations regarding current technology limitations [191], which goes beyond the objectives of this work.

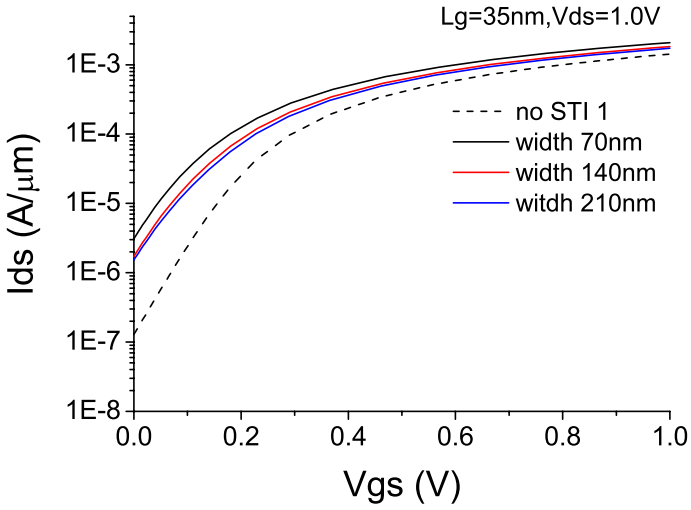


Figure 6.16 Width dependence of I_D-V_G characteristics of an n-MOSFET in the presence of structure *STI 1*, compared with a control device without STI.

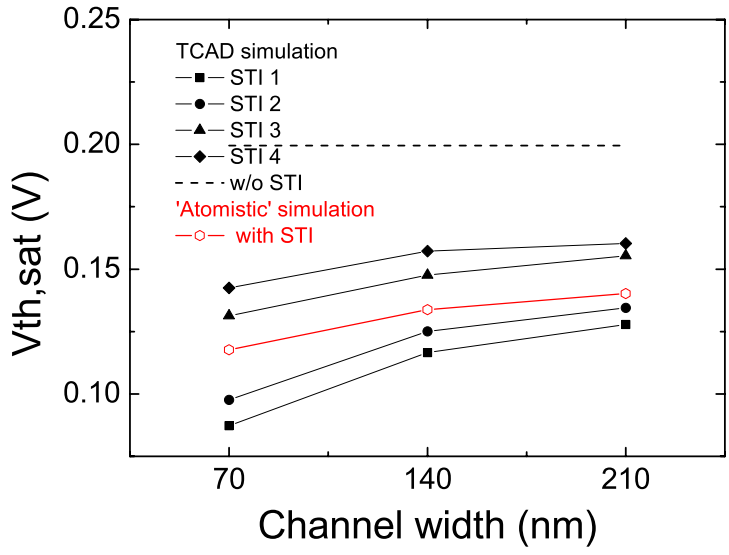


Figure 6.17 Width dependence of saturation threshold voltage for different STI architectures.

The threshold voltage decrease in the presence of STI results in the well known inverse-narrow-width effect [19]. The origin of this phenomenon is illustrated in Figure 6.18. The gate-STI over-ride results in electric-field crowding and early inversion in the STI corner. The potential distribution in left graph of Figure 6.18 illustrates the strong electric-field crowding from the poly-silicon into the wrap-round silicon corner indicating the excess fringing capacitance formed there. This results in an early inversion and increase of the local electron density in the corner, illustrated in the right graph. The corresponding local reduction in the threshold voltage is manifested in a high, localised current density near the STI edge. With the increase in the channel width, the relative contribution of this high current-density region decreases, resulting in a gradual increase in the threshold voltage.

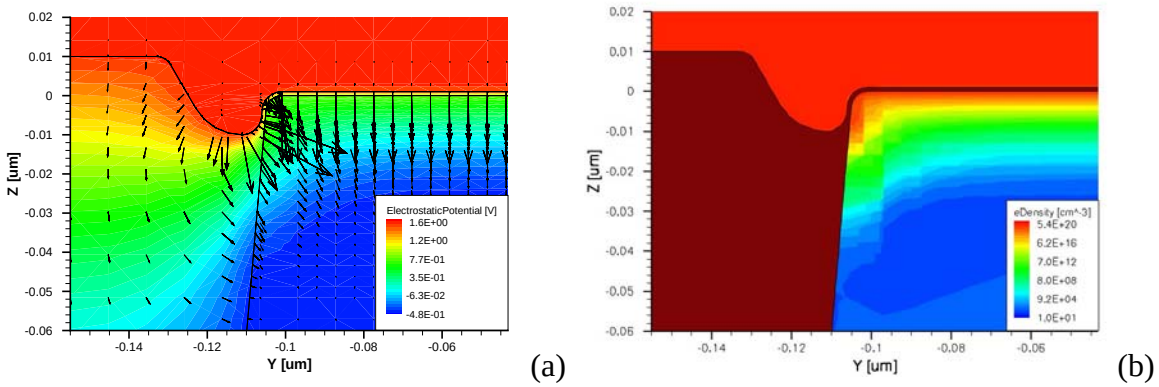


Figure 6.18 Mid-channel cross sections normal to the channel direction for (a) electrostatic potential with in-plane electric field, and (b) electron density.

6.2.2.RDD variability in the presence of STI

The Glasgow 3D 'atomistic' simulator is used to study the impact of the STI on RDD induced variability. As a starting point, simulations of continuously-doped transistors are carried out and compared with the TCAD simulation results in Figure 6.17. Similarity to the TCAD simulated channel width dependence is obtained, although a simplified rectangular STI structure is used in the 'atomistic' simulator.

Figure 6.19 illustrates the electron density distribution in one transistor from the statistical samples used to study the RDD induced variability. Both the extensions and channel are 'atomistically' doped. In Figure 6.20, which illustrates the potential distribution in the transistor, the acceptors induce potential spikes (barriers) for electrons in the channel and potential wells for the holes in the substrate while the donors create potential wells for the electrons in the extensions.

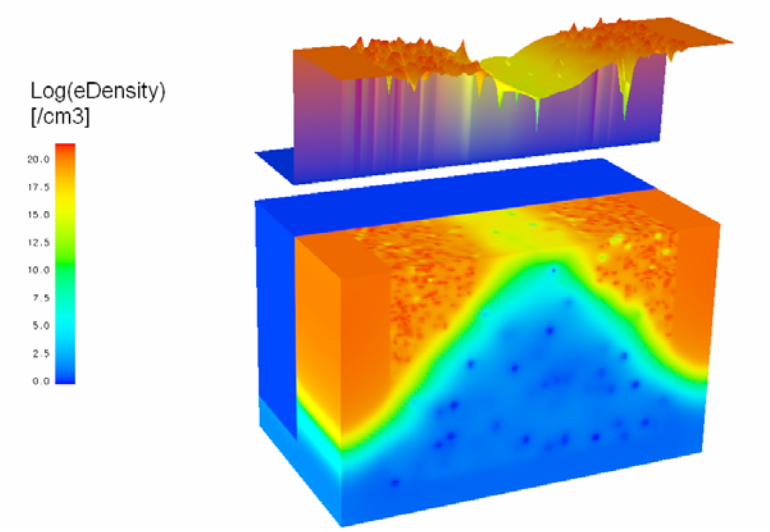


Figure 6.19 Electron concentration subject to random dopants, biased at zero, in the presence of the STI structure. The slice for the surface plot is taken at 0nm depth, namely silicon surface.

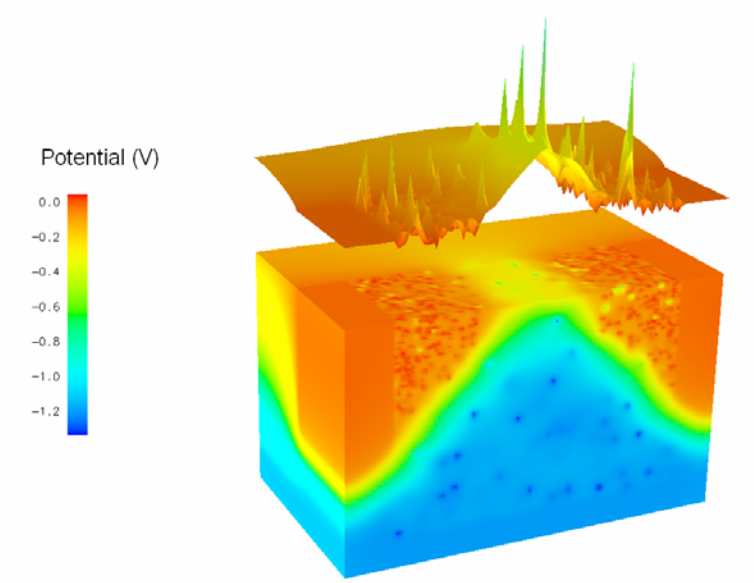


Figure 6.20 Electrostatic potential subject to random dopants with STI, biased at zero. The slice is surface potential referred to source contact.

6.2.2.1. Statistical impact of STI on V_{th} variability

Statistical simulations subject to RDD were carried out with and without STI for comparison. The corresponding distributions of the threshold voltages for the simulated samples of 200 transistors are illustrated in Figure 6.21. The standard deviations of the threshold voltage are 41.0 mV and 41.2 mV for devices without and with STI respectively. Due to the presence of narrow-width effects the average threshold voltage of the STI sample is reduced by 82.1 mV.

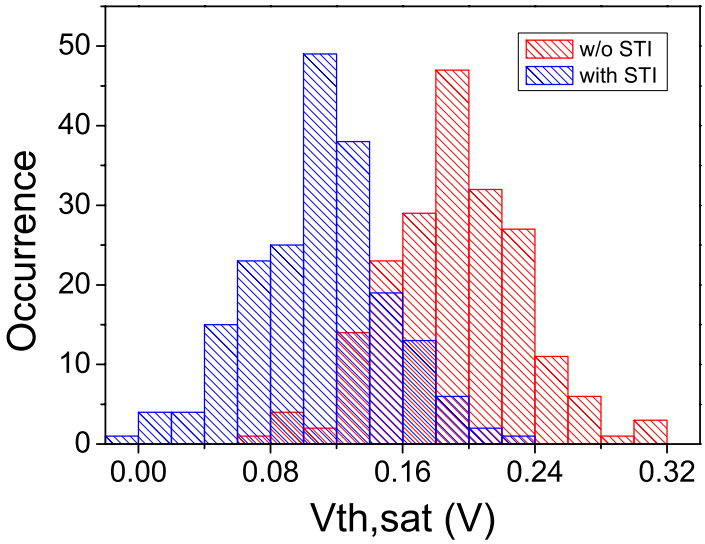


Figure 6.21 Saturation threshold voltage distribution under RDD induced variability for devices of channel width 70nm with/without STI.

The almost identical magnitude of threshold voltage variation for these two sets of devices derives from the fact that the partitioning of depletion charge due to additional fringing gate capacitance does not, in fact, change the localised impact of the random discrete dopants. This is also confirmed by the strong correlation of two sets of threshold voltages illustrated in Figure 6.22. The slope of the linear fit is surprisingly close to unity, which means almost identical standard deviations of two strongly correlated random variables.

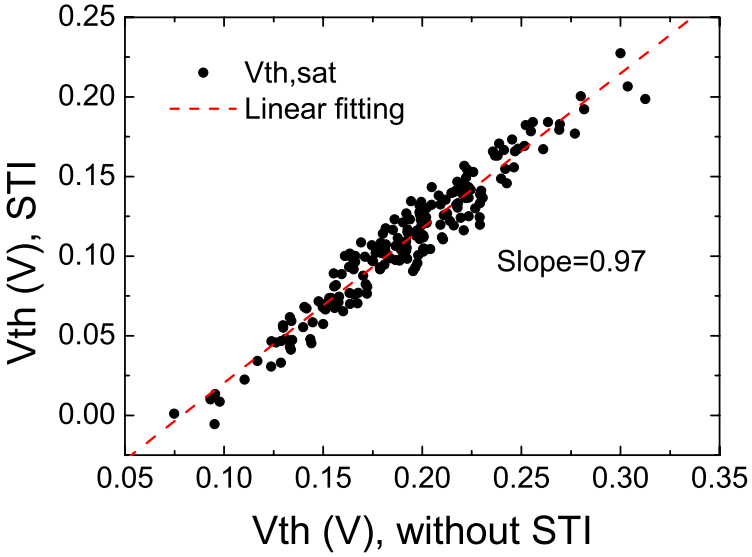


Figure 6.22 The correlation of threshold voltages between two sets of devices with/without STI.

6.3. Impact of STI on statistical variability and reliability of decananometre MOSFETs

The dominant source of statistical variability in contemporary bulk MOSFETs are the random discrete dopants (RDD) in the channel and source/drain regions. The statistical variability is further exacerbated by interface trapped charges (ITC) due to negative/positive bias temperature instability (NBTI/PBTI) [192] or hot carrier injection (HCI) [193]. The narrow channel SRAM-type transistors that are most susceptible to statistical variability and reliability are also strongly influenced by the shallow trench isolation (STI), which typically promotes enhanced current density near the STI edges [16] and could locally alter the junction shape [194]. The impact of RDD and ITC in isolation, or in combination with other sources of statistical variability, has been studied extensively using 3D numerical simulations [195]. However, to the best of our knowledge the impact of STI has not yet been considered in the simulation of RDD and ITC variability. Here, for the first time, we report a generic simulation study that highlights the impact of STI on the RDD and ITC-related threshold voltage variability.

6.3.1. Simulation methodology

The test-bed device is already described in last section. The simulations are carried out with the Glasgow 'atomistic' device simulator, with density gradient quantum corrections for electrons and holes used to avoid artificial carrier trapping in the sharply resolved Coulomb potential wells associated with discrete dopants and trapped charges[196][33]. A simplified STI structure is used in the simulations in order to highlight the generic impact on the variability, avoiding details of company-specific STI processing. It is well known that the presence of STI may affect the doping distribution in adjacent regions [194] and the shape of the *p-n* junction through local defect generation [197], strain [198][199], sidewall implantation [200], and implantation straggle [201]. To capture this effect, three generic cases of straight, inward and outward curved junctions are simulated, as illustrated in Figure 6.23. The STI is introduced only on one side of the channel, and Neumann boundary conditions are used at the other side (labelled as *with STI* in subsequent figures). The corresponding electron density and the potential distribution obtained from continuous doping simulations are illustrated in the same figure. For comparison simulations were carried out on devices with identical junction shape but with no STI, and Neumann boundary conditions on both sides of the channel (labelled as *w/o STI* in subsequent figures). For simulations with RDD and ITC, ensembles of 200 microscopically different

devices are simulated in each case. The threshold voltage is defined in the subthreshold region as the gate voltage that results in a drain current of 1.54×10^{-5}A/μm at drain voltage 1.0 V. The random discrete dopants are generated from the continuous doping distribution using standard methodology described in [149]. Discrete trapped charges, uncorrelated to the underlying discrete dopant distribution, are introduced randomly at the interface according to the trapped charge sheet density [202].

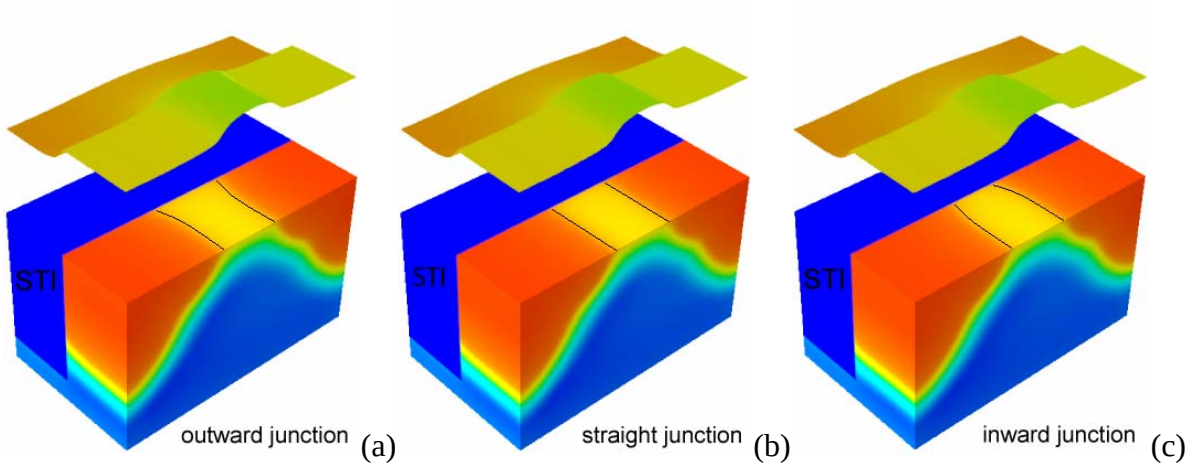


Figure 6.23 Simulation domains of shallow trench isolated 35×35 nm^2 channel area nMOSFETs for three channel junction shapes near STI: outward (a), straight (b), and inward (c) junctions, showing top surface potential (above) and electron density within the device (below) with biasing at V_{gs} =0.5V and V_{ds}=0.05V.

6.3.2. Results and discussions

The standard deviation of the threshold voltage, σV_{th}, for the three simulated junction shapes is compared in Figure 6.24 (a) in the presence and absence of STI. Results for RDD-only and for the combination of RDD and ITC are presented. We first discuss the results in the RDD-only case. The outward curved junction reduces the variability compared to the straight junction case, while the inward curved junction increases the variability. This is consistent with the expected inverse proportionality of $(\sigma V_{th})^2$ to the effective gate area. The localized high current density near the STI edges, and the corresponding inverse-narrow-width effect, is responsible for the lowering of the average threshold in the presence of STI, clearly illustrated in Figure 6.24 (b). In addition, threshold voltage roll-off leads to an average decrease in V_{th} when going from the outward to the inward junction shape. The increase of σV_{th} is stronger in the inward curved junction but the effect is somewhat reversed in the outward curved junction. However the differences in the straight and outward junction cases are within the statistical error of the σV_{th} estimate from the simulated sample of 200 devices. The picture seems to be much clearer and more definitive in the case of combined RDD and ITC simulations when a

relatively high trapped charge density of 1×10^{12} cm^{-2} is used in the simulations. Such trapped charge density results in 7-10 mV increase in σV_{th} depending on the junction shape. The presence of STI results in a higher sensitivity of V_{th} variability to the junction shape close to STI. The σV_{th} increases 3mV from the outward case to the inward junction case, compared with the increase of 1mV from the outward to inward junction cases without STI. This highlights the increased sensitivity of the STI-adjacent regions to RDD and ITC variability due to localized increase in the current.

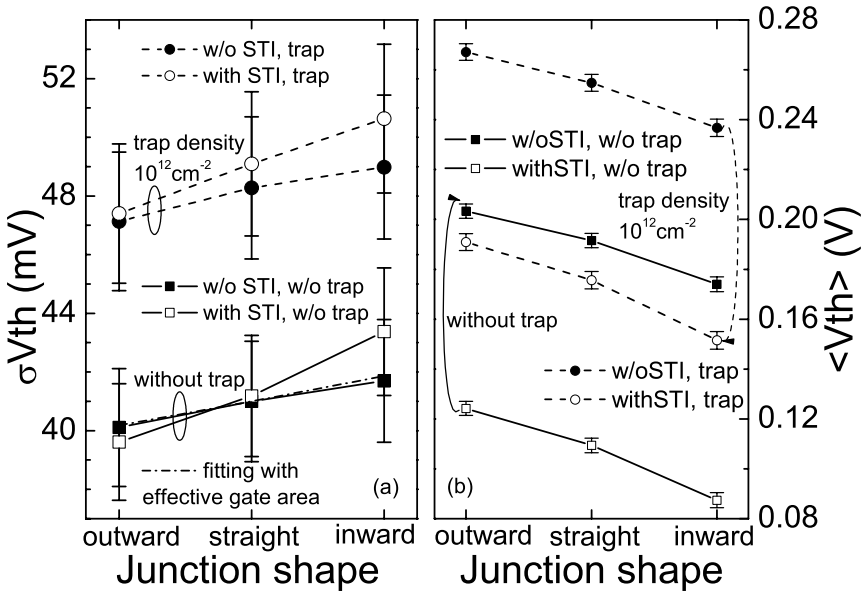


Figure 6.24 STI effect on statistical variability and reliability of threshold voltage (a) standard deviation and (b) mean for typical STI-adjacent junction cases.

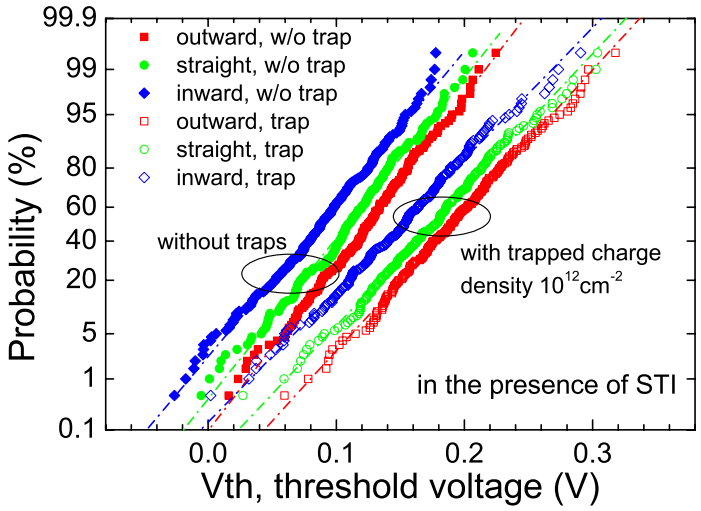


Figure 6.25 Threshold voltage distribution in the STI devices subject to different junction shapes close to the STI edge, and to degradation.

Figure 6.25 illustrates the V_{th} distribution with different junction shapes, with and without trapped charge. It is clear that STI devices with traps have higher average V_{th} and broader distributions.

Using 3D statistical simulations we have clearly demonstrated that the presence of STI results in an increased sensitivity of RDD and ITC variability to the junction shapes close to STI. This is linked to increased current density near the STI edge, which is also responsible for the inverse-narrow-width effect. The effect is even exacerbated by possible inward curvature of the p-n junctions near the STI edge. Careful control of the inverse-narrow-width effect and the junction shape could potentially reduce the STI impact on the statistical variability.

In summary, the variability in realistic modern device structures is studied for the first time in this chapter. At first, we explored the LER induced strain variability due to the stress engineering in modern devices. Strain variability enhances the current variability although threshold voltage fluctuations are little changed compared to unstrained devices. Secondly, the inverse narrow-width effect of the STI structure in narrow-channel MOSFETs is analyzed and this effect shifts the threshold voltage down. Junction shape has impact on V_{th} standard deviation, and STI structure affects statistical variability although depending on the junction shape. Meanwhile, STI exacerbates the statistical fluctuations of threshold voltage due to random traps at Si/Oxide interface.

From Chapter 4 and Chapter 5, coming to the end of Chapter 6, we have already achieved a comprehensive investigation of the technology projection of future CMOS generations, including the optimal process integrations, electrical characteristics, statistical variability, and the specialized variability subject to the realistic structures, all in aspects of electrostatic characteristics. The next chapter will study device performance under the small signal impulses and transient input conditions found in realistic circuits.

Chapter VII

7. Simulation of dynamic aspects of CMOS

In contrast to the quasi-static steady state (d.c.) conditions dealt with above, dynamic conditions exist for interconnected MOSFETs in circuits subject to analogue or digital inputs. Thus, a comprehensive analysis of MOSFET characteristics under a.c. stimulus is necessary to fully understand the dynamic behaviour of CMOS circuits. In this chapter analysis approaches are first reviewed, before results for the a.c. analysis of newly developed MOSFETs are presented and discussed.

7.1. Small signal a.c. analysis

To perform a.c. analysis on a device, perturbations (a 'small signal') are imposed upon it in a quasi-static state, and the resulting electrical response is analyzed, obtaining conduction and capacitance characteristics between the device terminals. Among the standard approaches to small signal analysis, sinusoidal perturbations are commonly used, giving simplicity of analysis as a function of frequency, and accuracy [203]. Small signal a.c. analysis results for particular 35 and 25 nm devices are discussed below.

7.1.1. Numerical approaches of small-signal a.c. analysis

The small signal analysis techniques are described in this subsection [203]. Given a device having N terminals subject to a.c. voltages $\tilde{V}_j$ with currents $\tilde{I}_i$ measured, the a.c. behaviour of the device can be expressed by an $N \times N$ admittance matrix Y determined by

$$\tilde{I} = Y\tilde{V} \tag{7.1}$$

where Y consists of a real-valued conductance matrix and purely imaginary capacitance matrix of the form $Y = G + j\omega C$. Each entry of the admittance matrix is determined by $Y_{ij} = \left.\frac{\tilde{I}_i}{\tilde{V}_j}\right|_{\tilde{V}_k = 0, k \neq j}$, therefore obtaining the capacitance and conductance of each entry.

Fourier decomposing the results of a transient excitation is one way to determine the a.c. behaviour of the device. A small step change in voltage, $\Delta V_j\, u(t)$, is imposed on contact j at time $t = 0$. The admittance matrix entry Y_{ij} is the frequency-domain ratio of the a.c. current at contact i to the a.c. voltage at contact j after Fourier transformation. The deduced low-frequency admittance has the following shape [203][204]

$$\begin{aligned} G_{ij}(\omega \to 0) &= \frac{I_i(\infty) - I_i(0)}{\Delta V_j} \\ C_{ij}(\omega \to 0) &= \frac{1}{\Delta V_j}\int_0^\infty [i_i(t) - I_i(0)]dt \end{aligned} \tag{7.2}$$

where I_i and i_i are the d.c. current and total current at terminal i. Although this Fourier decomposition method is a viable means to determine device a.c. behaviour, its major disadvantage is that it requires intensive computation to obtain useful simulation results, as a large number of closely spaced time steps are needed to reduce error.

In contrast, the Incremental Charge Partitioning approach has a far smaller burden of computation. A small d.c. voltage change is applied at the device contacts, and the incremental induced charges are partitioned to each terminal. The quasi-static capacitance is thus calculated by $C_{ij} = \Delta Q_i / \Delta V_j$. This method is especially useful when calculating gate capacitances as the gate current $i_i = \oint_{gate} \varepsilon_{ox} \frac{\partial \vec{E}}{\partial t} \cdot \vec{n} dA$ is, by nature, a displacement current. Substituting the displacement current into equation (7.2) results in the time derivative disappearing, and only the charge difference between the initial and final states are needed. However, this approach is heuristic in its treatment of how to assign / partition the charges, and this assignment is influenced by bias conditions and devices [205]. The method therefore lacks general utility and is of limited accuracy.

Sinusoidal steady-state analysis (S^3A) is a more accurate way to determine the a.c. behaviour of semiconductor devices. The governing equations are discretized at node i as formulated in Chapter 3,

$$\begin{aligned} F_{\phi i}(\phi, n, p) &= 0 \\ F_{ni}(\phi, n, p) &= \dot{G}_{ni}(n) \\ F_{pi}(\phi, n, p) &= \dot{G}_{pi}(p) \end{aligned} \tag{7.3}$$

where $\dot{G}$ are the time derivatives of generation terms. An 'infinitely' small sinusoidal signal is imposed on the steady state with the form $\xi = \xi_0 + \tilde{\xi} e^{j\omega t}$ where ξ represents ϕ, n

and p and ξ_0 are their steady state values. The control functions F are expanded around the steady state and to give $F(\phi_0, n_0, p_0) + \frac{\partial F}{\partial \phi}\tilde{\phi}e^{j\omega t} + \frac{\partial F}{\partial n}\tilde{n}e^{j\omega t} + \frac{\partial F}{\partial p}\tilde{p}e^{j\omega t}$ in which the steady-state term is zero and the right side of equation (7.3) is of the form $\dot{G} = \frac{\partial G(n_0 + \tilde{n}e^{j\omega t})}{\partial n}\frac{\partial (n_0 + \tilde{n}e^{j\omega t})}{\partial t} = \frac{\partial G}{\partial n} j\omega \tilde{n} e^{j\omega t}$ (this, for example, describes the electron density). Therefore the a.c. system at node i becomes [203][206]

$$\sum_j \begin{pmatrix} \frac{\partial F_{\phi i}}{\partial \phi_j} & \frac{\partial F_{\phi i}}{\partial n_j} & \frac{\partial F_{\phi i}}{\partial p_j} \\ \frac{\partial F_{ni}}{\partial \phi_j} & \frac{\partial F_{ni}}{\partial n_j} - j\omega\frac{\partial G_{ni}}{\partial n_j} & \frac{\partial F_{ni}}{\partial p_j} \\ \frac{\partial F_{pi}}{\partial \phi_j} & \frac{\partial F_{pi}}{\partial n_j} & \frac{\partial F_{pi}}{\partial p_j} - j\omega\frac{\partial G_{pi}}{\partial p_j} \end{pmatrix}_{dc} \begin{pmatrix} \tilde{\phi}_j \\ \tilde{n}_j \\ \tilde{p}_j \end{pmatrix} = 0 . \qquad (7.4)$$

In a.c. the Neumann boundary conditions are directly taken over from the d.c. simulation; Dirichlet boundary conditions just leave $\tilde{\phi}$ to excite the a.c. system whilst $\tilde{n} = \tilde{p} = 0$. Assembling nodes together the system may be written as

$$\left(J + jD\right)\tilde{X} = B \qquad (7.5)$$

where J is the Jacobian of the d.c. system; D is the contribution from generation terms; and $\tilde{X}$ is the solution vector. B is the real-valued boundary vector related to the a.c. excitation voltages. Re-writing $\tilde{X} = \tilde{X}_R + j\tilde{X}_I$, the a.c. system is transformed into

$$\begin{pmatrix} J & -D \\ D & J \end{pmatrix}\begin{pmatrix} \tilde{X}_R \\ \tilde{X}_I \end{pmatrix} = \begin{pmatrix} B \\ 0 \end{pmatrix}. \qquad (7.6)$$

After the a.c. system is solved, the a.c. current response, displacement current $\vec{\tilde{J}}_D$ and carrier currents $\vec{\tilde{J}}_n$ and $\vec{\tilde{J}}_p$, are calculated. Note that the factor $e^{j\omega t}$ does not appear in the a.c. system.

$$\begin{aligned} \vec{\tilde{J}}_D &= \frac{\partial \vec{D}}{\partial t} = -j\omega\varepsilon\nabla\tilde{\phi} \\ \vec{\tilde{J}}_n &= \left.\frac{\partial \vec{J}_n}{\partial \phi}\right|_{dc}\tilde{\phi} + \left.\frac{\partial \vec{J}_n}{\partial n}\right|_{dc}\tilde{n} + \left.\frac{\partial \vec{J}_n}{\partial p}\right|_{dc}\tilde{p} \\ \vec{\tilde{J}}_p &= \left.\frac{\partial \vec{J}_p}{\partial \phi}\right|_{dc}\tilde{\phi} + \left.\frac{\partial \vec{J}_p}{\partial n}\right|_{dc}\tilde{n} + \left.\frac{\partial \vec{J}_p}{\partial p}\right|_{dc}\tilde{p} \end{aligned} \qquad (7.7)$$

The small signal a.c. analysis admittance matrix is finally calculated according to equation (7.1).

7.1.2. Small signal response of 35 nm gate length nMOSFETs

In this subsection, the redesigned 35 nm physical gate length n-channel MOSFET is utilized as the main test bed device to demonstrate deca-nanometer MOSFET small signal response. The capacitance related physics is presented and the frequency response is obtained.

7.1.2.1. Capacitances in 35 nm gate length n-MOSFETs

The S^3A method was applied to deca-nanometer MOSFETs in order to observe their dynamic behaviour patterns. The nature of the terminal coupling capacitances is first examined in detail. Then the device physics leading to these coupling capacitances is explained and discussed

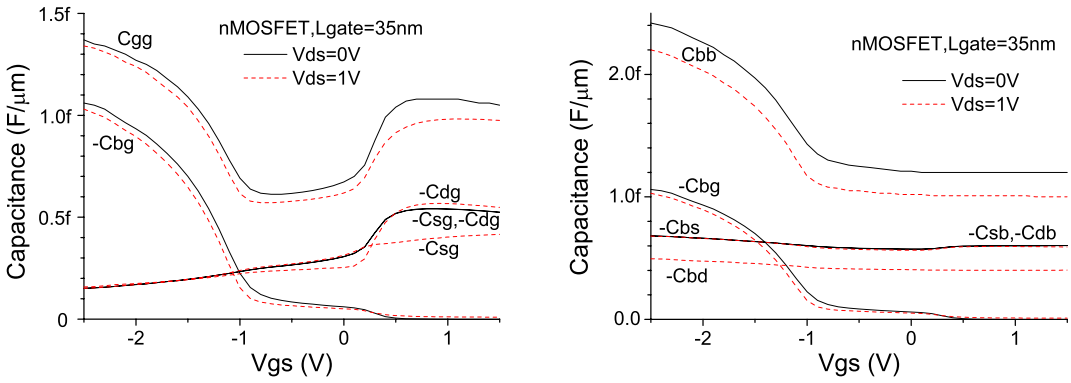


Figure 7.1 Capacitance against gate voltage characteristics are given for gate terminal (left) and bulk terminal (right) in a 35 nm nMOSFET. Here Cij is the coupling capacitance between electrodes i and j. Cgg is the total gate capacitance, and Cbb is the total capacitance related with bulk contact.

MOSFET gate capacitances are essential in determining device performance. Any small perturbation imposed on the steady-state gate voltage will cause current variations in the source, drain and bulk terminals, governed by the respective capacitances C_{sg}, C_{dg} and C_{bg}. Note that by charge conservation / current continuity, the sum of a column in the admittance matrix equals zero. For example, when the gate voltage is perturbed, the change in gate charge is equal to the sum of the change in charges at other terminals, and $\sum_{i=g,s,d,b} C_{ig} = 0$.

Figure 7.1 graphs the terminal capacitances of the 35 nm nMOSFET as a function of device bias. The simulations are carried out with a 1 MHz sinusoidal signal. When the drain bias is set to zero, the MOSFET is symmetric with respect to source and drain, and the total gate capacitance C_{gg} at inversion is 1.08 fF/μm (V_{gs} = 1 V). When the drain voltage is raised to 1.0 V, the inversion gate capacitance is reduced to 0.98 fF/μm (V_{gs} = 1 V). The inversion capacitance is expected to decrease with stronger gate field due to the poly-silicon depletion effect.

The detailed contributions to the gate capacitance can also be analyzed. In the left graph of Figure 7.1, C_{bg} dominates the gate capacitance in accumulation, whilst C_{sg}+C_{dg} is most important in inversion. Interestingly, in the depletion regime of this 35 nm n-MOSFET, C_{sg}+C_{dg} outweighs C_{bg}, raising the total gate capacitance above that expected in depletion.

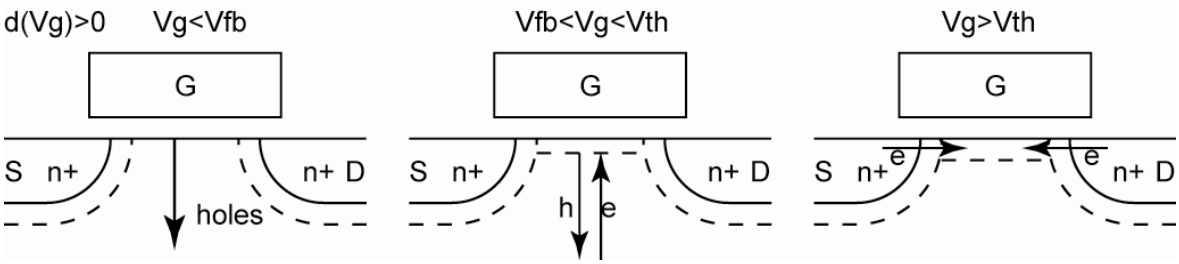


Figure 7.2 Physical view of carrier motion when forming gate capacitance in nMOSFETs, with zero biases of source, drain and bulk, namely source and drain in symmetry.

The physical reasons for these changes in gate capacitance can be understood with the help of the diagrams of Figure 7.2, which separate the three regimes of accumulation, depletion and inversion. In accumulation ($V_{gs} < V_{fb}$), holes accumulate in the channel close to the oxide/silicon interface. Under changes in the gate oxide electric field they will flow through the substrate towards the bulk contact rather than towards the source/drain because of the isolating *p*-*n* junctions. In the language of incremental charge partitioning, ΔQ_b (where current is outwards at bulk contact) is almost equal to ΔQ_g (where current is inwards at gate contact), therefore C_{bg} is the major contribution to the total gate capacitance.

When the device moves into depletion ($V_{fb} < V_{gs} < V_{th}$), a change in gate voltage will cause that charges are modulated at the depletion edge – at high frequency operation the generation rate in the depletion region is inadequate to supply the necessary charge directly, and charges must be obtained from the depletion edge. For a given gate voltage change the vertical field is lowered, as the source of mobile charge is further away, and C_{bg} becomes

much smaller. This is typical MOS capacitor behaviour. The charges from source/drain hardly approach channel still due to junction barriers.

However, it should be noted that overlap capacitance is playing an important role in modern devices [207]. When the gate voltage is greater than threshold, the n-plus source/drain extensions, which the gate overlaps, are in accumulation, and changes in gate voltage will cause source and drain currents through them, contributing to C_{sg} and C_{dg}. Even for gate voltages less than threshold there is a considerable effect and, in addition, the lateral face of the gate has fields linking to the source and drain. These effects are often labelled the source/drain overlap $C_{s/do}$, the inner fringe capacitance C_{if}, and the outer fringe capacitance C_{of} respectively (see Figure 7.3). Their values depend on geometrical device dimensions [208]. These overlap components always exist, and already contribute to C_{sg} and C_{dg} in the accumulation and depletion regimes. In the left graph of Figure 7.2, they constitute the non-zero portion of C_{sg} and C_{dg} when in accumulation and depletion, even although they are often assumed to be zero without overlap capacitance components [209]. In this template 35 nm nMOSFET, the overlap components make up a considerable portion of gate capacitance in accumulation and depletion, and around 50% of the inversion gate capacitance. They are important in accurately calculating the overall device characteristics, and its transient behaviour.

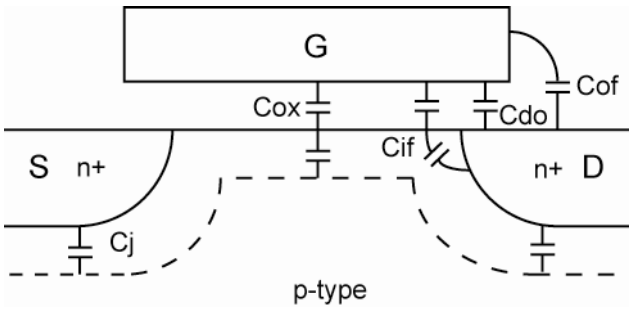


Figure 7.3 Schematic view of geometrical distribution of capacitances in nMOSFETs

When the gate d.c. voltage increases above threshold voltage ($V_{gs} > V_{th}$), the device moves into inversion, the channel linking the source and drain is well formed, and the substrate depletion boundary is stable at its maximum depth. In this situation, a small change in gate potential will bring plenty of electrons from the source/drain, and the bulk is effectively decoupled. C_{bg} is zero and the total gate capacitance is due to C_{sg} and C_{dg}. When a non-zero drain bias is introduced, inducing a lateral field in the channel, the depletion region close to drain is enlarged and the portion of inversion capacitance partitioned to the source decreases. This accounts for the decrease in C_{sg} and the increase in C_{dg} as V_{ds} is raised to 1 V.

Consider the capacitances connecting to bulk, in the right hand graph of Figure 7.1. The bulk contact is linked to the source, drain and gate by the bulk-to-source junction capacitance, bulk-to-drain capacitance, and oxide capacitance, with the total capacitance easily obtained by combining these. C_{bs} and C_{bd} are relatively constant, as these junction geometries change weakly as a function of V_g. They correspond, approximately, to the source and drain junction capacitances which are directly referenced in the common compact models used in SPICE circuit analysis. When V_d is biased (for instance to $V_d = 1$ V in Figure 7.1) the drain junction becomes widened and its capacitance reduces compared with that of the source junction as illustrated. Over the device operating range, C_{bg} trends are exactly as discussed in the gate capacitance discussion above.

7.1.2.2. Cut-off frequency

Generally the source and bulk contacts of MOSFETs in digital circuits are connected to d.c. voltage sources (usually the ground or supply voltage). As a consequence, a MOSFET may be simplified as a two-port a.c. network in circuits, as illustrated in (A) of Figure 7.4, with input port G-S(B) and output port D-S(B). A *hybrid-π* small signal compact model can be constructed based on this simple and general two-port network.

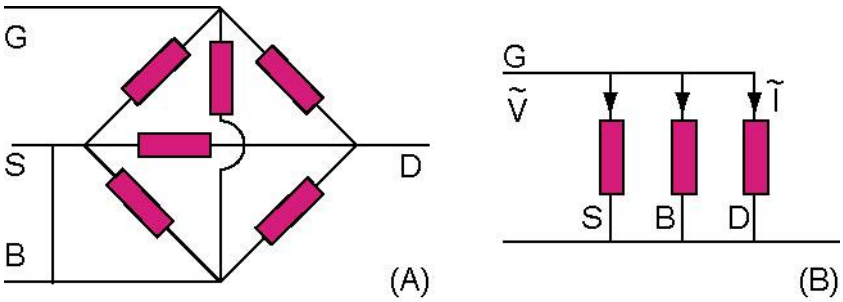


Figure 7.4 Two-port ac network of MOSFETs (A) with one port of G-S(B) and other port of D-S(B) and current-gain H_{21} parameter calculation in two-port MOSFETs (B).

To represent the characteristics of a 2-port a.c. circuit network subject to external excitation, various parameter matrices are possible, such as: the admittance Y-matrix previously discussed, the hybrid (H) matrix, or scattering (S) parameters. In the hybrid 2-port network, we have

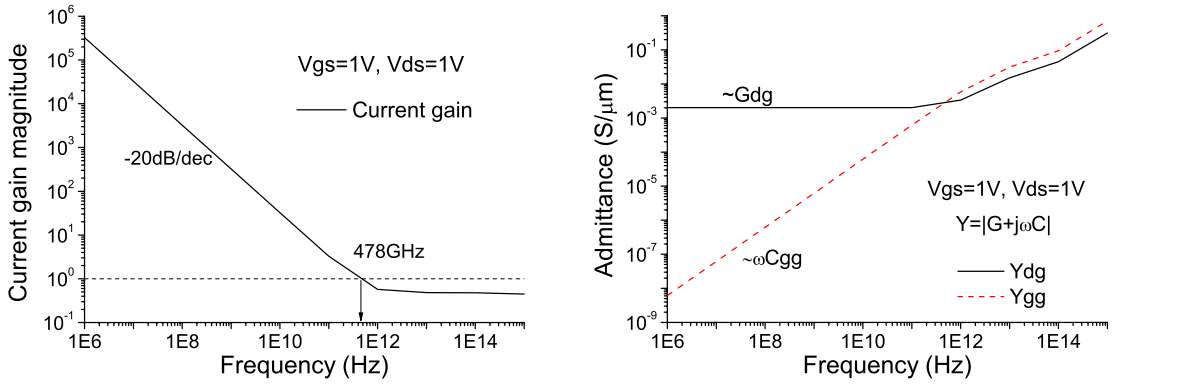
$$\begin{pmatrix} \tilde{v}_1 \\ \tilde{i}_2 \end{pmatrix} = \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix} \begin{pmatrix} \tilde{i}_1 \\ \tilde{v}_2 \end{pmatrix} \tag{7.8}$$

where port 1 is assigned to the gate-source and port 2 is assigned to the drain-source. $H_{21} = \left.\frac{\tilde{i}_2}{\tilde{i}_1}\right|_{\tilde{v}_2=0}$ is defined as the *current gain*, the ratio of output current to input current when

output port is shorted. This case is illustrated in (B) of Figure 7.4. Since in this MOSFET the input is the a.c. gate current and the output is the a.c. drain current, the current gain possesses the following form [203]

$$|A_I(\omega)| = \frac{|Y_{dg}|}{|Y_{gg}|} = \frac{|G_{dg} + j\omega C_{dg}|}{|G_{gg} + j\omega C_{gg}|} \tag{7.9}$$

A MOSFET's current gain declines with increasing frequency. When the current gain reaches unity, that frequency is defined as the *cut-off frequency,* f_T. Figure 7.5 graphs the small signal current gain vs. frequency response for the 35 nm gate length nMOSFET using equation (7.9). S^3A simulation predicts a cut-off frequency of 478 GHz. The typical operating frequency is usually smaller than f_T /10, or less than 47.8 GHz. Note that the current gain declines with frequency at a rate of –20dB/dec.



(a) (b)

Figure 7.5 Current gain magnitude versus frequency (a) and admittances of d-g and g-g versus frequency in 35 nm nMOSFET (b).

Specific observation of the admittance parameters Y_{dg} and Y_{gg} allows a clear understanding of why the current gain declines. In the right graph of Figure 7.5, for frequencies less than f_T, the magnitude of Y_{dg} is a constant, equalling the transconductance G_{dg} (commonly referred to as g_m), and the conductance G_{gg} is negligible compared with ωC_{gg}. Therefore, a well-known approximation for the cut-off frequency is

$$f_T = \frac{G_{dg}}{2\pi C_{gg}}. \tag{7.10}$$

The f_T value obtained from this approximation is f_T = 324 GHz, with an error of within 1.5% of the more accurate value when compared on a logarithmic scale

7.1.3. Split C-V analysis of 25 nm gate length pMOSFETs

As mentioned in last subsection, overlap components always exist in the capacitance of the gate with respect to the source and drain. These components are important in determining device performance of long-channel MOSFETs, and they are even more important for the case with current gate lengths in the deca-nanometer regime and with the diffusion of the SDE under the gate of the order of several nanometres. As a consequence there is a need to distinguish these extrinsic gate capacitance components from the intrinsic gate capacitance

7.1.3.1. Split C-V approach

Since parallel-plate capacitance is proportional to plate area, the total gate capacitance will increase with increasing gate length, even whilst the absolute values of the overlap components remain almost constant. Thus, by measurement at various gate lengths, it is possible to distinguish between overlap / gate-to-source / gate-to-drain capacitance and the intrinsic gate capacitance [210]. The total gate capacitance can be separated, approximately, into two parts:

$$C_{gg} = C_{in} + C_{ex} = C_{unit} L_{met} + C_{ex} \tag{7.11}$$

On varying the gate length, the intrinsic part can be calculated from

$$\Delta C_{gg} = C_{unit} \Delta L_{met} \tag{7.12}$$

where C_{unit} is the capacitance per unit gate length and L_{met} is the metallurgical gate length. This analysis is based on the assumption that consistent lateral channel doping profiles and symmetrical source/drain bias exist for variable gate length MOSFETs. As illustrated in Figure 7.6, long-channel MOSFETs almost meet the doping condition. However, in short-channel devices, the non-uniform halo-doping profiles diffused into the channel close to the source and drain overlap each other, increasing the channel doping concentration compared with long-channel MOSFETs. Although this effect will cause some error in extracting the depletion capacitance, the error will be negligible in the extraction of accumulation and inversion capacitances. The analysis provides a *first-order* extraction of intrinsic gate capacitance.

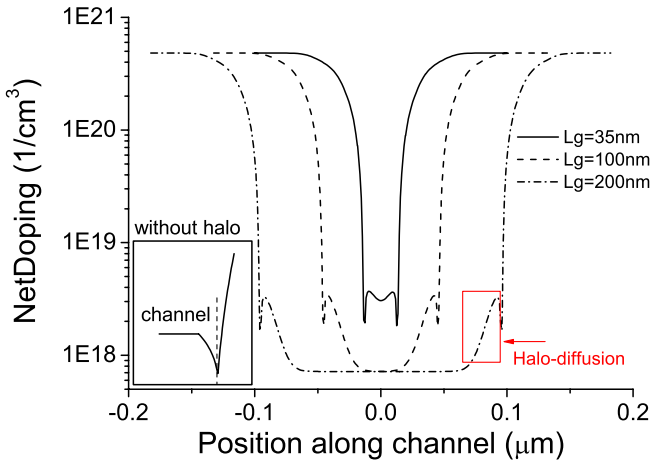


Figure 7.6 Lateral net active doping profiles for variable gate length n-MOSFETs.

7.1.3.2. Application to 25 nm high-k/metal replacement-gate pMOSFETs

Chapter 4 described a 25 nm high-k/metal replacement gate pMOSFET and the use of TCAD simulation tools in predicting the impact of additional device stress gain on its electrical characteristics (simulation results which agree well with experiment). The electrostatic improvement is evaluated by obtaining the intrinsic gate capacitance using the split *C-V* method [211].

Firstly, it should be noted that the gate capacitance is significantly increased, as the metal gate eliminates the poly-silicon depletion effect.

The results, presented in Chapter 4, show that the main improvement in high-*k*/metal gate device performance is associated with the reduction of the electrical EOT and the associated increase in inversion charge. The left graph of Figure 7.7 compares the total gate capacitance vs. gate voltage characteristics of the poly-silicon and the metal gate transistors at zero drain bias, with the capacitance normalized to a corresponding gate oxide parallel plate capacitor which has the same EOT and area. The comparison shows a 46% increase of the gate capacitance in the high-*k*/metal gate case which matches relatively well an observed 43% drain current improvement at low drain voltage in the absence of strain.

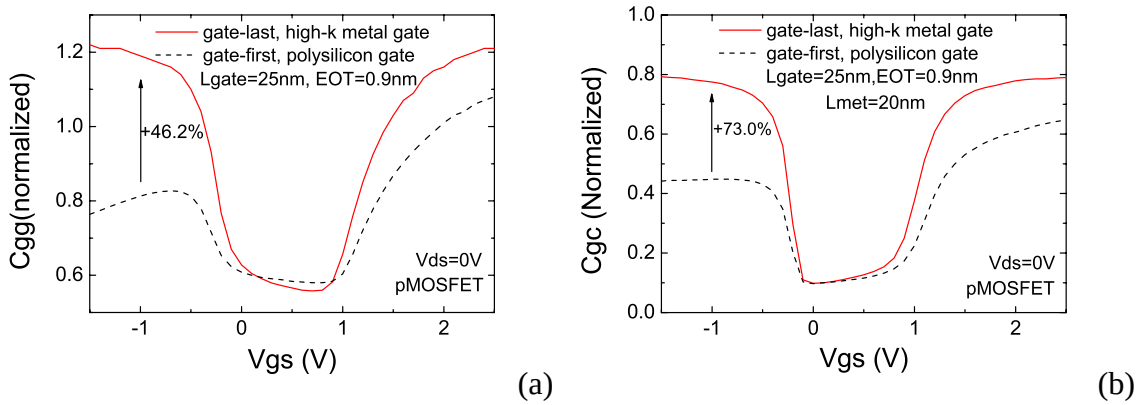


Figure 7.7 (a) The total capacitance spreads of 25 nm poly-gate and metal-gate pMOSFETs compared at zero drain bias. A 46% increase of inversion capacitance is achieved for the high-*k*/metal gate device. (b) Intrinsic gate capacitances for poly gate and metal gate pMOSFETs compared at zero drain bias. The intrinsic gate capacitance increases by about 73% in inversion for the metal gate case.

The right graph of Figure 7.7 compares the corresponding normalised extracted intrinsic gate capacitance C_{gc}-V characteristics. C_{gc}, which also includes partial capacitances related to the source and drain, is not C_{gb}. The significatnt improvement due to the reduction of electrical EOT in metal gate transistors and corresponding increase in the intrinsic gate capacitance are not fully translated into current improvements while the intrinsic gate capacitance increases 73% but drain current (at low drain voltage) just increases 43%. At low drain voltage this is due to the increased vertical component of the electric field corresponding to the larger inversion layer charge supported by the same gate voltage overdrive, which reduces channel mobility due to an increase in surface roughness scattering.

Comparing the intrinsic gate capacitance to total gate capacitance in the inversion regime, the extrinsic part accounts for nearly half of total gate capacitance for poly-gate devices. However, with the introduction of a metal gate, and the corresponding increase in intrinsic gate capacitance, the extrinsic part is reduced to about one-third of total gate capacitance. Thus, the metal gate significantly improves device a.c. performance

7.2. Transient simulations of 35 nm CMOS inverters

The transient behaviour of basic circuits under an a.c. bias signal is not only the direct measure of CMOS device performance in the circuit environment, but also a useful calibration-validating tool for MOSFET compact models. This section will focus on physical simulations of an inverter example, in which the redesigned 35 nm CMOS transistors presented in Chapter 4 are used as a test set, in order to obtain performance metrics and transient behaviour.

7.2.1.Mixed mode simulation

7.2.1.1. Transient simulation

When contact boundary conditions vary with time, the device is therefore subject to time-dependent simulation. In transient simulation, a set of physical equations used to describe semiconductor electrical behaviour have the following form [212]

$$\frac{d}{dt}q(z(t)) + f(t, z(t)) = 0\,. \tag{7.13}$$

where $z = (u, v, w)^T$, u, v and w are the normalized potential and the electron and hole quasi-Fermi levels respectively in terms of intrinsic *Debye* length $\sqrt{\varepsilon_{si} kT / q^2 n_i}$ [213][214]; $q(z) = (0, e^{u-v}, e^{w-u})^T$, and f comes from the boundary conditions. The d.c. state is at first solved, and then the equations solved iteratively increasing time step.

A simple Backward Euler (BE) discretisation scheme may be used to approximate the derivative between time step t_n and $t_n + h_n$,

$$q(t_n + h_n) + h_n f(t_n + h_n) = q(t_n)\,. \tag{7.14}$$

An improved discretisation scheme is the trapezoidal-rule/backward-differentiation formula (TRBDF). It makes an intermediate step $t_n + \gamma\, h_n$ (γ is usually chosen as $2-\sqrt{2}$) between t_n and $t_n + h_n$ through the trapezoidal rule

$$2q(t_n + \gamma h_n) + \gamma h_n f(t_n + \gamma h_n) = 2q(t_n) - \gamma h_n f(t_n)\,. \tag{7.15}$$

The second-order backward differentiation formula (BDF2) is applied to the time increment from $t_n + \gamma\, h_n$ to $t_n + h_n$

$$(2-\gamma)q(t_n + h_n) + (1-\gamma)h_n f(t_n + h_n) = (1/\gamma)[q(t_n + \gamma h_n) - (1-\gamma)^2 q(t_n)]\,. \tag{7.16}$$

7.2.1.2. Mixed-mode configuration for 35 nm gate length inverters

The transient behaviour of basic circuits is simulated using a mixed-mode technique involving device and circuit simulation. Semiconductor devices such as MOSFETs are physically simulated subject to time-dependent bias conditions, with other supporting components such as voltage sources, capacitors and resistors described by compact models. Mixed mode simulation has to take Kirchhoff's circuit laws (KCL and KVL) into account, besides each semiconductor device's governing equations. The disadvantage of simulating

circuits using mixed-mode is computational cost and complexity and such simulations are found to be unsuitable for large circuits and system, but because they combine the detailed physics of semiconductor devices into circuit models, they avoid possible inaccurate compact model extractions and the physical simplifications inherent in compact models and thus may be valuable in calibrating basic circuit simulations, compact models, and it provides insights into semiconductor device a.c. performance.

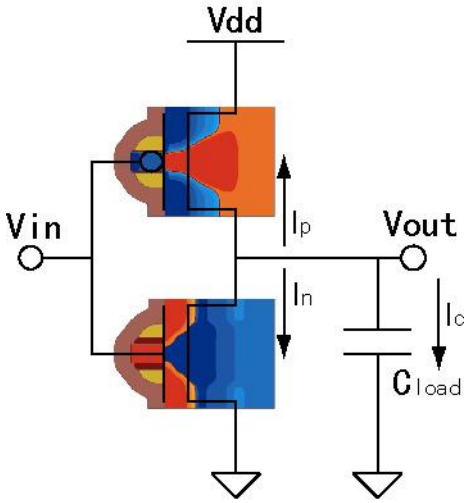


Figure 7.8 Schematic view of inverter configuration using mixed-mode simulation, where redesigned 35 nm gate length n-MOSFET and p-MOSFET in circuit environment are solved using numerical transient simulation.

A simple example of inverter simulation in the mixed mode simulation is explained below. Figure 7.8 illustrates the configuration of the inverter. n-MOSFET and p-MOSFET semiconductor devices are physically simulated, with the capacitor and voltage source simulated using compact models. Current and voltage equations are applied to the circuit. In this example, a 35 nm gate length nMOSFET is used, for which the source and bulk contacts are grounded. A counterpart 35 nm gate length pMOSFET has its source and bulk contacts connected to the supply voltage. The design details and electrical characteristics of these 35 nm CMOS transistors are elaborated in Chapter 4. Their gates are connected to an input voltage source V_{in}, and their drain contacts are linked to output capacitors (including wire-load capacitance and input capacitances of following inverters, here combined into a single load capacitance). The width ratio of 35nm physical gate length n/pMOSFET is 1μm : 2.3μm in order to match their drive currents.

7.2.2.a.c. performance of 35 nm MOSFETs in basic circuits

7.2.2.1. Inverter delay

Circuit speed and power dissipation are the two most critical figures of merit in VLSI design. Considerable efforts have been made in the literature to understand and formulate inverter delay and power dissipation. However, these efforts are typically based on compact simulations (for speed and viability of computation) and therefore lack accurate physical insights [215][216][217][218]; or introduce many, often non-physical, fitting parameters [219][220]. Here, physical simulation of inverters and inverter delay is analyzed according to physical observations.

Two factors influencing inverter delay are examined. The *inverter propagation delay* is defined as the time difference between the 50% transition points of the input and output signals. The load capacitance is an important factor in determining propagation delay. This load combines the interconnect capacitance with the input capacitances of the following circuit components. In addition, the input signal is affected by the drive characteristics of the preceding logic gate (whose driver is typically inverter-like) which, to first order, are parameterised an input signal slope or *slew*. In Figure 7.9, groups of experiments are carried out targeting these two factors. It is already evident that heavily capacitance-loaded inverters require a longer time to accomplish a logic change. At the onset of the input state change, an output voltage overshoot is noticed. With a fast input transition, this overshoot is high (typically up to 150% of the supply voltage), reducing as the input ramp decreases.

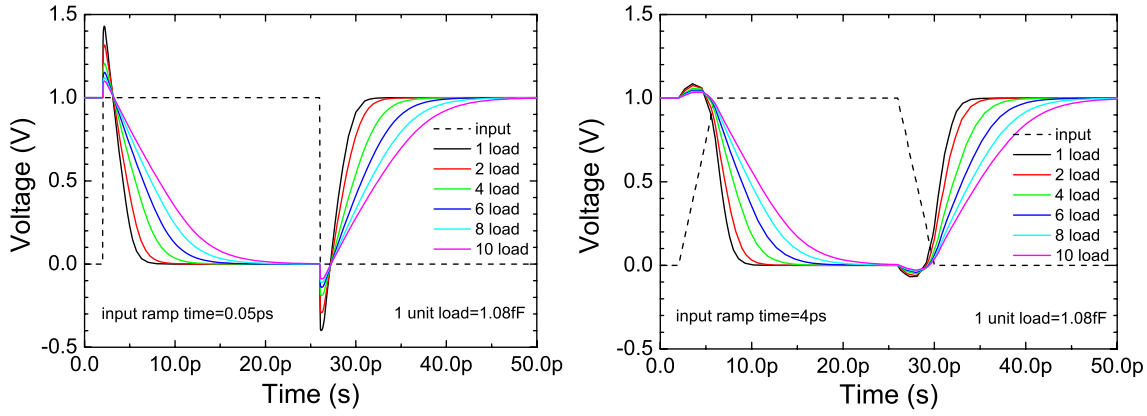


Figure 7.9 35 nm gate length CMOS inverter transfer characteristics with different load capacitance and different input ramp time.

A figure of merit for inverter delay in ITRS and textbooks has been described analytically as $\tau = C_{load} V_{dd} / 2 I_{dsat}$ [221]. This is a good approximation when inverter capacitances and inverter delay are both large. However, as devices shrink, parasitic components such as the

various extrinsic capacitances inside nano-CMOS devices will exhibit a larger relative their influence and should be taken into account. The MOSFET internal capacitances lead to intrinsic delay. Shown in Figure 7.10, propagation delay is linearly proportional to external load capacitance, but it shows a positive intercept of the delay axis when load capacitance is zero, which is called *intrinsic delay*. This portion of the delay is due to internal device capacitance. The increasing slope with load capacitance is defined as the *switching resistance* [221]. The simulated 35 nm gate length CMOS has a switching resistance of 420 Ω. Another point should be not neglected. With a slowly ramping input, the delay is bigger and the corresponding intrinsic delay is larger. The right hand graph of Figure 7.10 shows the effect of differing the input ramp. With a slowly transitioning ramp time (large ramp), the delay is linearly increasing. However the change of delay is small for rapid ramp times. The results here are in good agreement with Monte Carlo simulations [219].

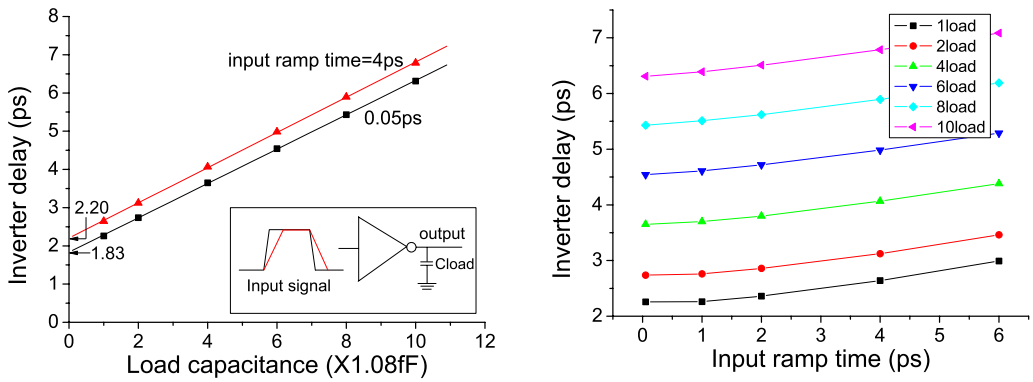


Figure 7.10 35 nm CMOS Inverter propagation delay as a function of capacitances and input ramp time, with Vdd=1.0V.

7.2.2.2. Underlying current relationship

When a MOSFET is turned on there is a conducting channel between source and drain for ohmic region, but the resistance will be large when the gate bias is below the threshold voltage. In addition, gate-to-drain capacitances and drain-to-bulk capacitances are conducting paths for displacement currents. Figure 7.11 schematically shows the possible conduction paths connected to the inverter output node.

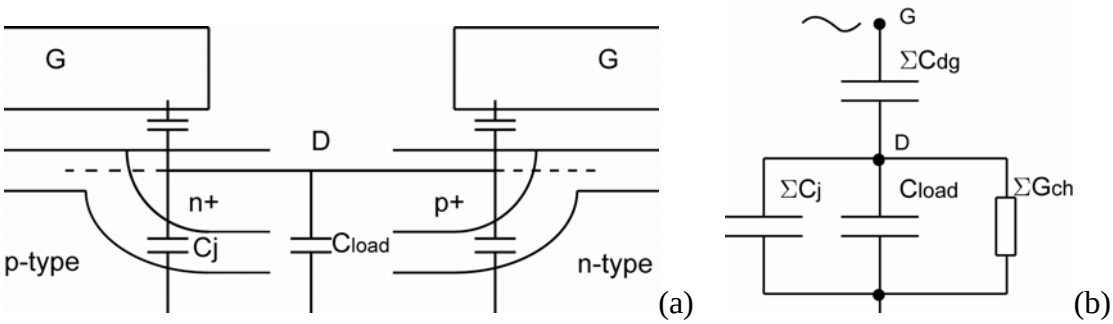


Figure 7.11 (a) Conduction components connected with drain contacts of 35 nm CMOS inverters. (b) The equivalent circuit during input initial ramp time.

When the input voltage is within certain voltage ranges, $V_{in} < V_{thn}$, $V_{in} > V_{dd} + V_{thp}$ (when $V_{out} < V_{dd}$), or $V_{in} > V_{out} + V_{thp}$ (when $V_{out} > V_{dd}$) , the channel is switched-off and the channel conductance can be ignored. However, the channel conductance may also be neglected compared to capacitances no matter what input voltage is if input has an extremely fast transition. The output overshoot happens due to the RC response to the input ramp excitation. Therefore, the overshoot value can derived according to the equivalent circuit indicated in (b) of Figure 7.11.

$$\Delta V_{overshoot} = \frac{|\, j\omega \sum C_{dg} \,|}{|\, j\omega(\sum C_{dg} + \sum C_j + C_{load}) + \sum G_{channel} \,|} \times \Delta V_{in} \qquad (7.17)$$

where C_{dg}, and C_j are the drain-to-gate and bulk-to-drain coupling capacitances. Figure 7.12 presents an estimate of overshoot voltage simply by omitting channel conductance and selecting constant capacitances of C_{dg} and C_{bd} (at V_{in}=0V and V_{out}=1V). When the load capacitance is large, the overshoot becomes smaller because of the relative increase of load capacitance to millar capacitance C_{dg}. In addition, the accuracy of the calculated orvershoot value from analytic model can improve when the inverter is loaded by a large capacitance.

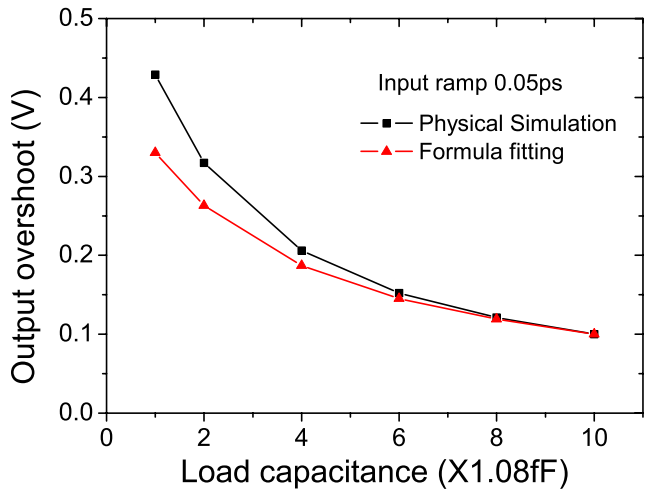


Figure 7.12 Output overshoot of 35 nm CMOS inverters with fast transition of input, Vdd=1V.

Understanding the current relationships during inverter switching time are helpful in understanding overall inverter a.c. behaviour. The extracted gate and drain currents are shown in Figure 7.13 during an input rise period. Currents are defined as positive if they are device-inwards. The gate currents are equal to the gate oxide displacement current, $C_{gg}(V_{in},V_{out})\cdot\frac{dV_{in}}{dt}$, therefore a 2ps ramp develops approximately twice the gate current of a 4ps ramp. Additionally the gate current in the pMOSFET is greater than that of the nMOSFET due to its larger gate area. N-channel formation is witnessed by a negative drain current when the input voltage is increasing but still below the nMOS threshold voltage. When above threshold voltage, the drain current is saturated due to high output voltage only if $V_{out}>V_{in}-V_{thn}$. As for drain current of pMOSFET, it receives a portion of displacement current by the coupling capacitance C_{dg} during switching time, short-circuit current through p-channel ($V_{out}<V_{dd}$, $V_{in}<V_{dd}+V_{thp}$) and also current from the drain diffusion capacitance.

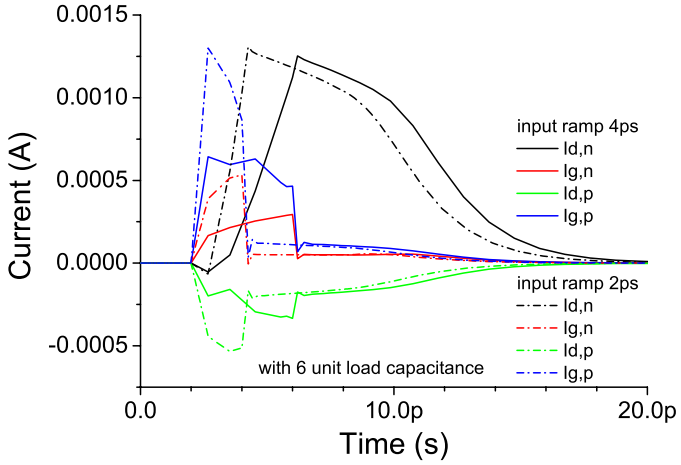


Figure 7.13 Gate currents and drain currents of 35 nm gate length nMOS and pMOS around input rise-up switching for different input ramp time.

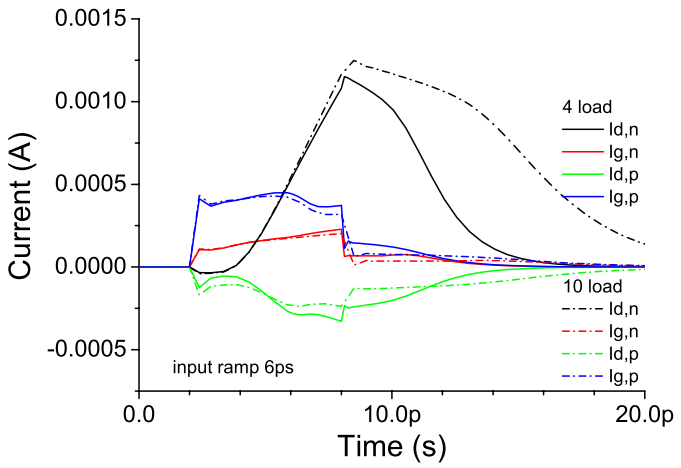


Figure 7.14 Gate currents and drain currents of 35 nm gate length nMOS and pMOS around input rise-up switching for different load capacitances.

With different loading capacitances, a similar analysis applies. The gate current is independent of load capacitance to first order, but drain currents are affected as shown in Figure 7.14. In particular, the nMOS drain current stays higher for longer due to the greater time it takes to discharge the larger load capacitance. A portion of the pMOS drain current is influenced by the output voltage and thus experiences changes when differently loaded. When the load capacitance does not receive a net positive current from the previous logic stage, the output voltage decreases, and the time point to approach overshoot peak value can by determined by equalling two drain currents during the switch time (mainly, the n-MOSFET drain saturation current [215] and the p-MOSFET drain current formed by gate displacement current), using the following analytical expression

$$\frac{\beta}{2}\left(V_{in}-V_{thn}\right)^2 = C_{dg}\frac{dV_{in}}{dt} \tag{7.18}$$

where β is current gain, here in estimation extracted as the gate voltage second derivative of drain current at $V_{gs}= V_{thn}$ and $V_{ds} = V_{dd}$ of I_d-V_g characteristics of the nMOSFET. In this simulation, the input transition rate is constant, making the treatment simple. If the rise time from zero to supply voltage takes t_T, then the transition rate is equal to V_{dd} / t_T . The required time for the output to reach the overshoot peak therefore is

$$t_1 = \sqrt{\frac{2C_{dg}t_T}{\beta V_{dd}}} + \frac{V_{thn}}{V_{dd}}t_T \,. \tag{7.19}$$

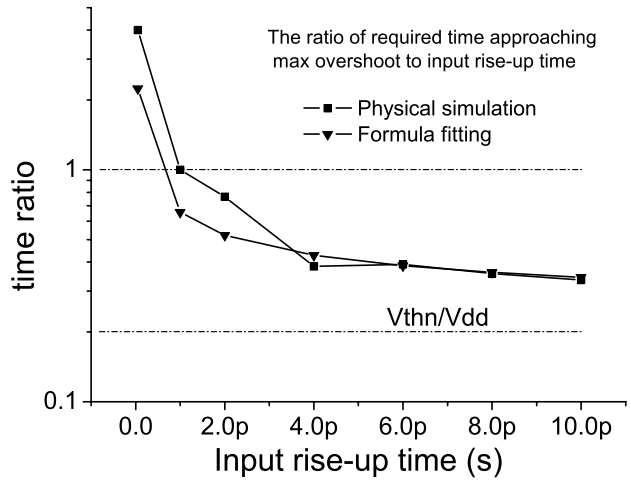


Figure 7.15 Analytical estimate of required time reaching maximum overshoot in 35 nm gate length CMOS inverters.

The accuracy of this analytical estimate is shown in Figure 7.15. Aside from circuits subjected to extremely fast ramp inputs, the overshoot peaks are located within the switching times, and ratio of required time to input ramp time falls with the increase of

ramp time, tending towards V_{th}/V_{dd}. This estimate essentially matches the data and follows the same trend.

As observed above, inverter delay increases with increasing input large ramp time. Correctly modelling propagation delays due to extremely fast input ramps will provide the data needed for good approximations of propagation delay over a wide range of input transition times. For a step response to input rise, V_{in} is at the supply voltage, and V_{out} falls from its overshoot value to $V_{dd}/2$ over the propagation delay time. The nMOS drain current discharges the capacitances connected to drain contacts and it may undergo a linear region at the end of switching if $V_d < (V_g - V_{th})/m$, of which m is body-effect coefficient. Short-circuit current via the p-channel device is switched off because of the gate bias of the pMOS being below threshold voltage. Therefore the inverter propagation delay for inputs with fast ramp time is approximated as

$$\tau_0 = \frac{\left(2V_{overshoot} + V_{dd}\right)\left(C_{dd,n} + C_{dd,p} + C_{load}\right)}{2I_{dsat}} \tag{7.20}$$

where $V_{overshoot}$ is obtained from equation (7.17), and $C_{dd,n}$ and $C_{dd,p}$ are the total capacitances associated with the drain contacts of the nMOS and pMOS devices respectively (mainly a combination of the drain diffusion junction capacitances and gate-to-drain capacitances)

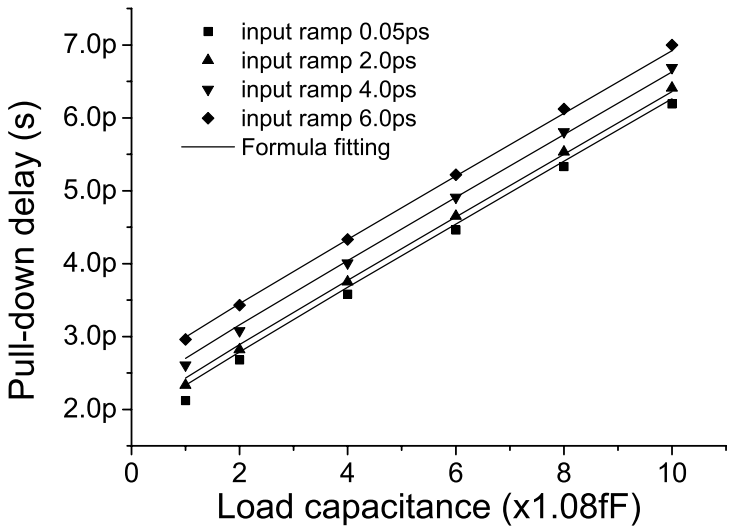


Figure 7.16 Physical fitting of 35 nm gate length CMOS inverter delay.

For slow transition inputs, the propagation delay increases with ramp time, as observed above in the physical simulations. This is modelled as [220]

$$\tau = \tau_0 + \alpha(t_T - t_{T0})u(t_T - t_{T0}) \tag{7.21}$$

where t_T is the input transition time and $u(\cdot)$ is a step function. t_{T0} is a cutoff value of ramp time for constant delay, here determined to be 1.35ps from simulations. α is the increasing rate of inverter delay with input ramp time, with $\alpha = 0.135$ obtained from above simulations. In Figure 7.16, the estimate of pull-down delay for rise-up input is expressed for a wide range of load capacitances and input ramp time. The errors, in the majority of cases, are within 5%, and the accuracy of the formula improves with heavy load capacitances and slowly ramping inputs. This formula can serve as an analytical platform for a better understanding of a.c. CMOS device behaviour.

7.3. Scaling of the a.c. performance of MOSFETs

In the previous two sections, a small signal analysis and transient simulations of particular MOSFETs were presented. It is also important to consider, and predict if possible, the scaling behaviour of these figures-of-merit describing device a.c. performance, in order to achieve a complete and practical scaling projection of bulk MOSFETs.

7.3.1. Small signal analysis of scaled MOSFETs

Figure 7.1 illustrates the total gate capacitances under inversion and accumulation conditions for the scaled MOSFETs presented in Chapter 4. As expected, the inversion gate capacitance is less than the accumulation gate capacitance in poly-gate MOSFETs due to an additional, series, poly-depletion capacitance in inversion, but there is practically no difference in the gate oxide capacitance under inversion and accumulation for metal-gate MOSFETs. The abrupt drop of both inversion and accumulation capacitances between 35 nm and 25 nm physical gate length devices reflects the combined effect of strong poly-depletion and slow gate oxide thickness scaling. The generally decreasing trend in gate capacitances for scaled MOSFETs indicates that vertical oxide scaling is slower than gate length scaling, as expected.

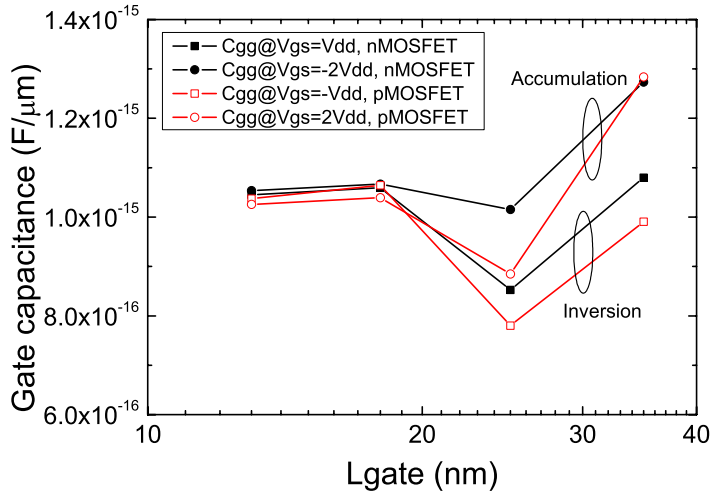


Figure 7.17 Scaling of total gate capacitance.

The cut-off frequency increases with MOSFET scaling. As pointed out in section 5 of Chapter 4, the saturation transconductance (per unit channel width) increases with scaling. In addition to the slightly decreasing of total gate capacitance noted above (Figure 7.17), this indicates that cut-off frequency increases with scaling according to the formula (7.10). It is demonstrated in Figure 7.18 although there is no obvious conclusion that it comply with linearity. This increase is mainly contributed from staturation transconductance improvement.

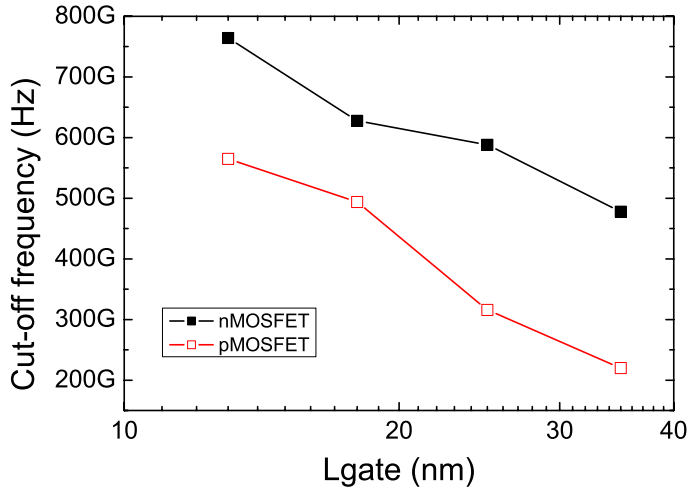


Figure 7.18 Cut-off frequencies of scaled MOSFETs.

The intrinsic and extrinsic gate capacitances are presented for scaled MOSFETs in Figure 7.19. As expected, a decreasing intrinsic portion and an increasing extrinsic portion, relative to total gate capacitance, are observed in poly-gate MOSFETs because the gate overlap length of SDE is on the order of several nanometres, relatively unchanged from generation to generation. The metal gate stack significantly boosts the intrinsic gate capacitance by the elimination of poly-depletion series capacitance, but does not increase

extrinsic gate capacitance. Therefore the drastic decreasing trend of relative intrinsic capacitance portion is somewhat slowed in metal-gate MOSFETs.

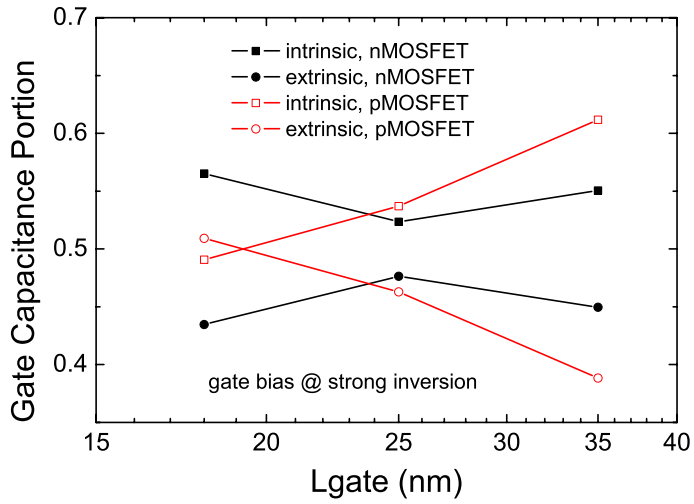


Figure 7.19 Intrinsic/extrinsic inversion gate capacitances of scaled MOSFETs.

7.3.2. The inverter delay projection

Scaled CMOS inverters are configured and simulated using the mixed-mode simulation method, and the inverter delays are obtained and illustrated in Figure 7.20. It is clear that the propagation delay time for unloaded inverters decreases with scaling. The capacitance (per unit channel width) slightly decreases as already shown in Figure 7.17. In addition, the drive current (per unit channel width) increases with scaling. Therefore the delay of the unloaded inverter improves. The abrupt decrease of inverter delay from poly-gate MOSFETs to metal-gate MOSFETs is caused by the drive current increase by the introduction of metal gate stack for the 18 nm and 13 nm gate length MOSFETs.

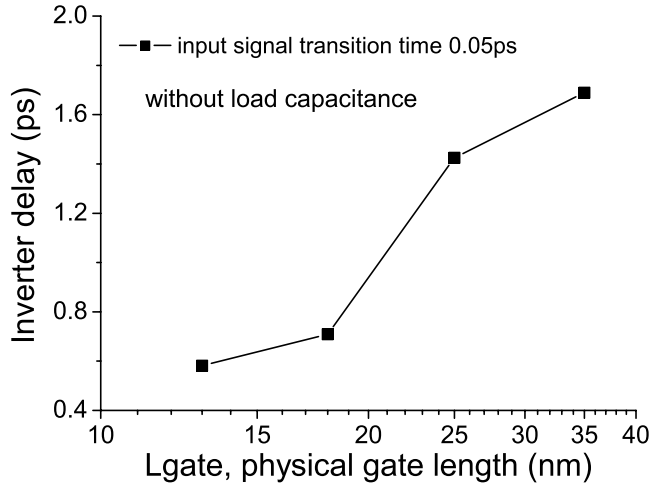


Figure 7.20 Intrinsic inverter delays of scaled MOSFETs.

In summary, this chapter provides a study of small signal analysis and transient analysis on scaled MOSFETs. It explores device ac performance and figures of merit, initially on specific devices. A detailed physical understanding of internal device capacitances is presented and the cut-off frequency is obtained for 35 nm gate length n-MOSFETs. The split C-V method is applied to analyze intrinsic/extrinsic gate capacitance. A detailed analysis of inverter behaviour is presented. Finally, the scaling behaviour of device and circuit a.c. performance is given. The gate capacitance deteriorates with scaling. The cut-off frequency increases with scaling. The inverter propagation delay becomes smaller with scaling, improving circuit speed.

Chapter VIII

8. Conclusion

Here it summarizes the conclusions of the thesis, which has presented the scaling of bulk MOSFETs and the results of a comprehensive simulation study of their corresponding statistical variability.

8.1. Summary

A comprehensive review of the scaling, challenges and technology boosters employed in bulk MOSFETs was presented in Chapter 2. The scaling trends of MOSFET physical dimensions, applied voltages and doping profiles were analyzed, and the rules of constant-field scaling and generalized scaling described. The new features of devices introduced in recent technology generations and incorporated into the ITRS projections of future bulk MOSFETs, were highlighted. The scaling of gate lengths, EOT and supply voltage is slowing down, and performance is being maintained through 'equivalent scaling', employing strain and high-k/metal gate stacks. Scaling challenges associated with the lithography, power dissipation, gate tunnelling leakage, poly-silicon depletion effects, short-channel effects, NBTI and statistical variability were discussed. This chapter provided an understanding of the challenges facing bulk MOSFET scaling and was a foundation for the scaling study presented in the thesis.

The simulation tools and methodology used to design and analyze scaled MOSFETs were described in Chapter 3. The techniques used to simulate important fabrication steps were first reviewed, including ion implantation, thermal annealing, film formation, and stress engineering, focusing on the related physical models and numerical techniques. The 3D simulation strategy used to study strain variability and STI was explained. The process simulator provides the realistic doping profiles and stress/strain distributions needed for accurate device simulation. The competent use of process simulation is critical to the analysis of realistic bulk MOSFET scaling. For the purpose of device simulation, the drift-diffusion approach with density gradient quantum corrections was used. The numerical methods applied to discretize and solve the Poisson's equation and current continuity

equations were outlined. The adopted mobility models, including impurity scattering, interface scattering, and high-field degradation effects, were presented. Special attention was paid to the modelling of local stress, which is key to the strain induced variability. Finally, it reviewed the techniques used to simulate statistical variability, which is at the heart of this thesis.

Based on the guiding principles described in Chapter 2 and utilizing the simulation tools described in Chapter 3, the successful calibration and the realistic scaling of bulk MOSFETs has been performed in Chapter 4. The calibration was based on detailed process information including channel indium and SDE arsenic doping profiles from physical Toshiba 35 nm gate length nMOSFETs. Accurate mobility tuning and precise matching of the measured electrical characteristics were achieved. In order to study more contemporary MOSFETs, the process was updated using information from 45 nm Intel and TSMC technologies, including matching the key dimensions and adopting stress engineering. This process update was simulated at device level and achieved the key performance indicators of the 45 nm Intel and TSMC bulk MOSFETs. The successful MOSFET scaling followed the generalized scaling principles, but with realistic physical constraints in gate oxide thickness, threshold voltage and supply voltage scaling. It strictly follows the projections for the gate length, EOT and supply voltage in the ITRS 2008 update edition. The simulated scaled MOSFETs achieve the performance prescribed by the ITRS. Still, there is a noticeable deterioration of the SCE in pMOSFETs with scaling in presence of eSiGe source/drain. High-k/metal gates were introduced into the process simulation, and after careful analysis and optimisation, it was found that high-k/metal gates improve device electrostatic integrity, reduce SCE, and increase device drive currents.

The statistical variability of the contemporary and scaled MOSFETs was systematically investigated for the first time using the Glasgow ‘atomistic’ simulator. The general trend of increasing threshold voltage standard deviation (σV_{th}) from approximately 44 ~ 45 mV to 80 ~ 82 mV was observed in the transition from 35 nm to 25 nm poly-gate MOSFETs. Meanwhile the corresponding off-current variation increases from ~3.4 times to approximately 7 times the average off-current. The introduction of a metal-gate in the 18 nm nMOSFETs reduces σV_{th} to 70.5 mV, but the poor SCE increases the σV_{th} of 18 nm pMOSFETs to 108 mV. Although DD simulation underestimates the on-current variation, the simulated standard deviation of on-current continuously increases from 8.5% ~ 8.2% of average on-current in the 35 nm MOSFETs to 15.8% ~ 17.2% in the 25 nm MOSFETs and to 21% ~ 26% in the 18 nm MOSFETs.

In Chapter 6 the impact of strain and STI on device variability was studied for the first time. It was demonstrated that due to LER induced channel length variations, the stressors induce different strain across the channel width. Stronger strain is induced by stressed CESL layer in channel shortening, which leads to localized increase in current density, enhancing the current variations. The stressor enhances the drive current standard deviation by a factor identical to that of the average current enhancement. Also, for the first time, the impact of the STI on the random dopant and trapped charge induced statistical variability was investigated. The STI enhances the current density near the isolation edge, resulting in an inverse-narrow-width effect. It enhances the impact of the edge part on the random dopant induced threshold voltage statistical variability. This effect is mediated by the junction shape near the STI effect. Similarly it enhances the statistical reliability on threshold voltage due to random traps at the interface of Si/Oxide.

In the last main chapter, the dynamic performance of the scaled MOSFETs was studied using small signal analysis and inverter transient simulations. The gate capacitances of the 35 nm nMOSFET were carefully examined. The results show that the extrinsic gate capacitance is occupying nonnegligible portion of the total gate capacitance. The split C-V analysis shows the use of metal-gate in the 25 nm pMOSFETs can significantly increase the active gate capacitance by removing the poly-depletion effect. Mixed-mode transient simulation was used to simulate the intrinsic and extrinsic delay in 35 nm gate length CMOS inverters. It was found that the output overshoot magnitude and peak time are directly related to the drain-to-gate capacitance. The inverter delay was studied in detail. It was found that the delay time is determined by discharging/charging the capacitive components connected to drain contacts of the n- and p-MOSFETs including not only load capacitances but also drain junction capacitances and miller capacitances. Finally, the scaling evolution of the dynamic performance of the scaled MOSFETs was presented. With MOSFET scaling, gate capacitance (per width) generally decreases due to slower scaling of gate dielectric thickness; the cut-off frequency increases owing to transconductance (per width) increases. The intrinsic portion of total gate capacitance decreases with scaling but metal-gate improves the intrinsic one. The inverter propagation delay reduces with scaling mainly due to drive current improvement.

8.2. Outlook

This study focused on the scaling of bulk MOSFETs, and presented a scenario for their performance evolution subject to technology challenges, electrical characteristic deterioration and statistical variability and reliability. Although significant efforts have

been invested in achieving best performance, the structural deficiencies of bulk MOSFETs have become evident. The short-channel effect is the first major problem, and high-doping related statistical variability is the second one. The fully-depleted SOI MOSFET is a promising alternative to contemporary bulk MOSFETs. It resolves both of these problems at the same time, achieving better electrostatic integrity and less dopant-related parameter variation. Therefore, the scaling and the variability study of SOI MOSFETs is an important subject of further research.

However, no new device is expected to replace bulk MOSFETs in the near future. Decreasing dimensions will start to influence device performance due to STI. Doping concentrations can be affected by STI through diffusion and stress, further influencing the localized nonuniformity in device electrical characteristics. Therefore, further investigation of STI effects in relation to statistical variability will benefit an understanding of impact of bulk MOSFETs on SRAM scaling, yield and design strategies.

References

[1] Gordon E. Moore, “Cramming more components onto integrated circuits,” *Electronics*, pp.114-117, April 19, 1965.

[2] R.H. Dennard, F.H. Gaensslen, H.-N. Yu, V.L. Rideout, E. Bassous, and A.R. LeBlance, “Design of ion-implanted MOSFET’s with very small physical dimensions,” *IEEE J. Solid-State Circuits*, Vol. SC-9, No.5, October 1974.

[3] K. Mistry, C. Allen, C. Auth, B. Beattie, D. Bergstrom, M. Bost, M. Brazier, M. Buehler, A. Cappellani, R. Chau, C.-H. Choi, G. Ding, K. Fischer, T. Ghani, R. Grover, W. Han, D. Hanken, M. Hattendorf, J. He, J. Hicks, R. Huessner, D. Ingerly, P. Jain, R. James, L. Jong, S. Joshi, C. Kenyon, K. Kuhn, K. Lee, H. Liu, J. Maiz, B. McIntyre, P. Moon, J. Neirynck, S. Pae, C. Parker, D. Parsons, C. Prasad, L. Pipes, M. Prince, P. Ranade, T. Reynolds, J. Sandford, L. Shifren, J. Sebastian, J. Seiple, D.Simon, S. Sivakumar, P. Smith, C. Thoms, T. Troeger, P. Vandervoorn, S. Williams, K. Zawadzki, “A 45 nm logic technology with high-k+metal gate transistors, strained silicon, 9 Cu interconnect layers, 193nm dry patterning, and 100% Pb-free packaging,” in *IEDM Tech. Dig.*, pp.247-250, 2007.

[4] Shekhar Borkar, “Obeying Moore’s law beyond 0.18 micron,” in *Proc. IEEE ASIC/SOC Conf.*, pp.26-31, 2000.

[5] Mark Bohr, “The new era of scaling in an SoC World,” in *ISSCC Dig. Tech. Papers*, pp.23-28, 2009.

[6] Yuan Taur, and Tak H. Ning, *Fundamentals of Modern VLSI Devices*, Cambridge University Press, 1998, Page 152.

[7] L. K. Wang, Y. Taur, D. Moy, R. H. Dennard, K. Chiong, F. Hohn, P.J. Coane, “0.5 micron gate CMOS technology using e-beam/optical mix lithography,” in *Symp. VLSI Tech. Dig.*, pp.13-14., 1986.

[8] W.-H. Chang, B. Davari, M.R. Wordeman, Y. Taur, C.C.-H. Hsu, and M. D. Rodriguez, “A high-performance 0.25-μm CMOS technology: I—Design and characterization,” *IEEE Trans. Electron Devices*, Vol. 39, No. 4, pp.959-966, April 1992.

[9] Y. Taur, S. Wind, Y.J. Mii, Y. Lii, D. Moy, K.A. Jenkins, C.L. Chen, P.J. Coane, D. Klaus, J. Bucchignano, M. Rosenfield, M.G.R. Thomson, and M. Polcari, “High performance 0.1 μm CMOS devices with 1.5 V power supply,” in *IEDM Tech. Dig.*, pp.127-130, 1993.

[10] P. Bai, C. Auth, S. Balakrishnan, M. Bost, R. Brain, V. Chikarmane, R. Heussner, M. Hussein, J. Hwang, D. Ingerly, R. James, J. Jeong, C. Kenyon, E. Lee, S-H. Lee, N.

Lindert, M. Liu, Z. Ma, T. Marieb, A. Murthy, R. Nagisetty, S. Natarajan, J. Neirynck, A. Ott, C. Parker, J. Sebastian, R. Shaheed, S. Sivakumar, J. Steigerwald, S. Tyagi, C. Weber, B. Woolery, A. Yeoh, K. Zhang, and M. Bohr, "A 65 nm logic technology featuring 35 nm gate lengths, enhanced channel strain, 8 Cu interconnect layers, low-k ILD and 0.57 μm^2 SRAM cell," in *IEDM Tech. Dig.*, pp.657-660, 2004.

[11] S. Inaba, K. Okana, S. Matsuda, M. Fujiwara, A. Hokazono, K. Adachi, K. Ohuchi, H. Suto, H. Fukui, T. Shimizu, S. Mori, H. Oguma, A. Murakoshi, T. Itani, T. Iinuma, T. Kudo, H. Shibata, S. Taniguchi, M. Takayanagi, A. Azuma, H. Oyamatsu, K. Suguro, Y. Katsumata, Y. Toyoshima, and H. Ishiuchi, "High performance 35 nm gate length CMOS with NO oxynitride gate dielectric and Ni salicide," *IEEE Trans. Electron Devices*, Vol. 49, No. 12, pp.2263-2270, December 2002.

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

[13] T. H. Ning, P.W. Cook, R.H. Dennard, C.M. Osburn, S.E. Schuster, and H.N.Yu, "1 μm MOSFET VLSI technology: Part IV—hot-electron design constraints," *IEEE J. Solid-State Circuits*, Vol. SC-14, pp.268-275, April 1979.

[14] S. Ogura, P.J. Tsang, W.W. Walker, D.L. Critchlow, and J.F. Shepard, "Design and characteristics of the lightly doped drain-source (LDD) insulated gate field-effect transistor," *IEEE J. Solid-State Circuits*, Vol. SC-15, No. 4, August 1980.

[15] G.J. Hu, C. Chang, and Y.-T. Chia, "Gate-voltage-dependent effective channel length and series resistance of LDD MOSFET's," *IEEE Trans. Electron Devices*, Vol. ED-34, No. 12, pp.2469-2475, December 1987.

[16] N. Shigyo, T. Hiraoka, "A review of narrow-channel effects for STI MOSFET's: a difference between surface- and buried-channel cases," *Solid-State Electronics*, Vol.43, pp.2061-2066, 1999.

[17] A. Bryant, W. Hänsch, and T. Mii, "Characteristics of CMOS device isolation for the ULSI age," in *IEDM Tech. Dig.*, pp.671-674, 1994.

[18] A. Chatterjee, J. Esquivel, S. Nag, I. Ali, D. Rogers, K. Taylor, K. Joyner, M. Mason, D. Mercer, A. Amerasekera, T. Houston, and I.-C. Chen, "A shallow trench isolation study for 0.25/0.18 μm CMOS technologies and beyond," in *Symp. VLSI Tech. Dig.*, pp.156-157, 1996.

[19] Lex A. Akers, "The inverse-narrow-width effect," *IEEE Electron Device Letters*, Vol. EDL-7, No. 7, pp.419-421, July 1986.

[20] G. Fuse, H. Ogawa, K. Tateiwa, I. Nakao, S. Odanaka, M. Fukumoto, H. Iwasaki, and T. Ohzone, "A practical trench isolation technology with a novel planarization process," in *IEDM Tech. Dig.*, pp.732-735, 1987.

[21] K. Shibahara, Y. Fujimoto, M. Hamada, S. Iwao, K. Tokashiki, T. Kunio; "Trench isolation with ∇(NABLA)-shaped buried oxide for 256mega-bit DRAMs," in *IEDM Tech. Dig.*, pp.275-278, 1992.

[22] B. Hoeneisen and C. Mead, "Fundamental limitations in microelectronics—I. MOS technology," *Solid-State Electronics*, Vol. 15, No. 7, pp. 819-829, July 1972.

[23] R. Swanson and J. Meindl, "Ion-implanted complementary MOS transistors in low-voltage circuits," *IEEE J. Solid-State Circuits,* Vol. SC-7, No. 2, pp.146-153, April 1972.

[24] Dale L. Critchlow, "MOSFET scaling—the driver of VLSI technology," in *Proc. of the IEEE*, pp.659-667, Vol. 87, No. 4, April 1999.

[25] Yuan Taur, and Tak H. Ning, *Fundamentals of Modern VLSI Devices*, Cambridge University Press, 1998, Chapter 4.

[26] G. Baccarani, M.R. Wordeman, and R.H. Dennard, "Generalized scaling theory and its application to a 1/4 micrometer MOSFET design," *IEEE Trans. Electron Devices*, Vol. ED-31, No. 4, pp.452-462, April 1984.

[27] F.M. Schellenberg, L. Capodieci, and B. Socha, "Adoption of OPC and the impact on design and layout," in *Proc. of DAC*, pp.89-92, 2001.

[28] Wolfgang M. Arden, "The international technology roadmap for semiconductors—perspectives and challenges for the next 15 years," *Current Opinion in Solid State and Material Science*, Vol. 6, pp.371-377, 2002.

[29] R. Pack, V. Axelrad, A. Shibkov, V. Boksha, J. Huckabay, R. Salik, W. Staud, R. Wang, W. Grobman, "Physical & timing verification of subwavelength-scale designs – Part I: Lithography impact on MOSFETs," in *Proc. of SPIE*, Vol. 5402, 2003.

[30] S.-H. Lo, D.A. Buchanan, Y. Taur, "Modeling and characterization of quantization, polysilicon depletion, and direct tunnelling effects in MOSFETs with ultrathin oxides," *IBM J. Res. Develop.*, Vol. 43, No. 3, May 1999.

[31] W.-C. Lee, T.-J. King, C. Hu, "Evidence of hole direct tunnelling through ultrathin gate oxide using P+ poly-SiGe gate," *IEEE Electron Device Letters*, Vol. 20, No. 6, pp.268-270, June 1999.

[32] A. Asenov, "Random dopant induced threshold voltage lowering and fluctuations in sub-0.1 μm MOSFET's: A 3-D 'atomistic' simulation study," *IEEE Trans. Electron Devices*, Vol. 45, No. 12, pp.2505-2513, December 1998.

[33] G. Roy, A.R. Brown, F. Adamu-Lema, S. Roy, and A. Asenov, “Simulation study of individual and combined sources of intrinsic parameter fluctuations in conventional nano-MOSFETs,” *IEEE Trans. Electron Devices*, Vol. 53, No. 12, pp.3063-3070, December 2006.

[34] A.R. Brown, G. Roy, and A. Asenov, “Poly-Si-gate-related variability in decananometer MOSFETs with conventional architecture,” *IEEE Trans. Electron Devices*, Vol. 54, No. 11, pp.3056-3063, November 2007.

[35] B. Cheng, S. Roy, G. Roy, F. Adamu-Lema, A. Asenov, “Impact of intrinsic parameter fluctuations in decanano MOSFETs on yield and functionality of SRAM cells,” *Solid-State Electronics*, Vol. 49, pp.740-746, 2005.

[36] S. Tyagi, C. Auth, P. Bai, G. Curello, H. Deshpande, S. Gannavaram, O. Golonzka, R. Heussner, R. James, C. Kenyon, S-H. Lee, N. Lindert, M. Liu, R. Nagisetty, S. Nataranjan, C. Parker, J. Sebastian, B. Sell, S. Sivakumar, A. St Amour, K. Tone, “An advanced low power, high performance, strained channel 65 nm technology,” in *IEDM Tech. Dig.*, pp.245-247, 2005.

[37] T. Ghani, M. Armstrong, C. Auth, M. Bost, P. Charvat, G. Glass, T. Hoffmann, K. Johnson, C. Kenyon, J. Klaus, B. McIntyre, K. Mistry, A. Murthy, J. Sandford, M. Silberstein, S. Sivakumar, P. Smith, K. Zawadzki, S. Thompson and M. Bohr, “A 90nm high volume manufacturing logic technology featuring novel 45 nm gate length strained silicon CMOS transistors,” in *IEDM Tech. Dig.*, pp978-980, 2003.

[38] C. Auth, A. Cappellani, J.-S. Chun, A. Dalis, A. Davis, T. Ghani, G. Glass, T. Glassman, M. Harper, M. Hattendorf, P. Hentges, S. Jaloviar, S. Joshi, J. Klaus, K. Kuhn, D. Lavric, M. Lu, H. Mariappan, K. Mistry, B. Norris, N. Rahhal-orabi, P. Ranade, J. Sandford, L. Shifren, V. Souw, K. Tone, F. Tambwe, A. Thompson, D. Towner, T. Troeger, P. Vandervoorn, C. Wallace, J. Wiedemer, C. Wiegand, “45 nm high-k+metal gate strain-enhanced transistors,” in *Symp. VLSI Tech. Dig.*, pp.128-129, 2008.

[39] International Technology Roadmap for Semiconductors, http://www.itrs.net/, access in 2009.

[40] M. Rothschild, T.M. Bloomstein, T.H. Fedynyshyn, R.R. Kunz, V. Liberman, M. Switkes, N.N. Efremow Jr., S.T. Palmacci, J. H.C. Sedlacek, D.E.Hardy, and A. Grenville, “Recent trends in optical lithography,” *Lincoln Laboratory Journal*, Vol. 14, No. 2, pp.221-236, 2003.

[41] Franklin M. Schellenberg, “A history of resolution enhancement technology,” *Optical Review*, Vol. 12, No. 2, pp.83-89, 2005.

[42] Masato Shibuya, “Resolution enhancement techniques for optical lithography and optical imaging theory,” *Optical Review*, Vol. 4, No. 1B, pp.151-160, 1997.

[43] Edward J. Nowak, “Ultimate CMOS ULSI performance,” in *IEDM Tech. Dig.*, pp.115-118, 1993.

[44] B. Davari, R.H. Dennard, and G.G. Shahidi, “CMOS scaling for high performance and low power—the next ten years,” in *Proc. of the IEEE*, Vol. 83, No. 4, April 1995.

[45] Y. Taur, D.A. Buchanan, W. Chen, D.J. Frank, K.E. Ismail, S.-H. Lo, G. A. Sai-Halasz, R.G. Viswanathan, H.-J.C. Wann, S.J. Wind, and H.-S. Wong, “CMOS scaling into the nanometer regime,” in *Proc. of the IEEE*, Vol. 85, No. 4, April 1997.

[46] R.H. Dennard, J. Cai, A. Kumar, “A perspective on today’s scaling challenges and possible future directions,” *Solid-State Electronics*, pp.518-525, 2007.

[47] C.-Y. Lu, J.M. Sung, H.C.Kirsch, S.J. Hillenius, T.E. Smith, and L. Manchanda, “Anomalous C-V characteristics of implantated poly MOS structure in n+/p+ dual-gate CMOS technology,” *IEEE Electron Device Letters*, Vol. 10, No. 5, pp.192-194, May 1989.

[48] K.S. Krisch, J.D. Bude, and L. Manchanda, “Gate capacitance attenuation in MOS devices with thin gate dielectrics,” *IEEE Electron Device Letters*, Vol. 17, No. 11, pp.521-523, November 1996.

[49] G. Baccarani, and M.R. Wordeman, “Transconductance degradation in thin-oxide MOSFET’s,” *IEEE Trans. Electron Devices*, Vol. ED-30, No. 10, pp.1295-1304, October 1983.

[50] R. Rios, N.D. Arora, and C.-L. Huang, “An analytical polysilicon depletion effect model for MOSFET’s,” *IEEE Electron Device Letters*, Vol. 15, No. 4, pp.129-131, April 1994.

[51] N.D. Arora, R.Rios, and C.-L. Huang, “Modeling the polysilicon depletion effect and its impact on submicrometer CMOS circuit performance,” *IEEE Trans. Electron Devices*, Vol. 42, No. 5, pp.935-943, May 1995.

[52] Frank Stern and W.E. Howard, “Properties of semiconductor surface inversion layers in the electric quantum limit,” *Phys. Rev.*, Vol. 163, No. 3, pp.816-835, 1967.

[53] Yasuyuki Ohkura, “Quantum effects in Si n-MOS inversion layer at high substrate concentration,” *Solid-State Electronics*, Vo. 33, No. 12, pp.1581-1585, 1990.

[54] Brain K. Ip, and John R. Brews, “Quantum effect upon drain current in a biased MOSFET,” *IEEE Trans. Electron Devices*, Vol.45, No. 10, pp.2213-2221, October 1998.

[55] A. Pirovano, A.L. Lacaita, and A.S. Spinelli, “Two-dimensional quantum effects in nanoscale MOSFETs,” *IEEE Trans. Electron Devices*, Vol. 49, No. 1, pp.25-31, January 2002.

[56] S.-H. Lo, D.A. Buchanan, Y. Taur, “Modeling and characterization of quantization, polysilicon depletion, and direct tunnelling effects in MOSFETs with ultrathin oxides,” *IBM J. Res. & Develop.*, Vol. 43, No. 3, pp.327-337, May 1999.

[57] W. Haensch, E.J. Nowak, R.H. Dennard, P.M. Solomon, A. Bryant, O.H. Dokumaci, A. Kumar, X. Wang, J.B. Johnson, M.V. Fischetti, “Silicon CMOS devices beyond scaling,” *IBM J. Res. & Dev.*, Vol. 50, No. 4/5, pp.339-361A, July/September 2006.

[58] Yuan Taur, and Tak H. Ning, *Fundamentals of Modern VLSI Devices*, Cambridge University Press, 1998, Page 95.

[59] J. Maserjian, “Tunnelling in thin MOS structures,” *J. Vac. Sci. Tech.*, Vol. 11, No. 6, pp.996, 1974.

[60] C. Chang, M-S Liang, C. Hu and R.W. Brodersen, “Carrier tunnelling related phenomena in thin oxide MOSFET’s,” in *IEDM Tech. Dig.*, pp.194-197, 1983.

[61] S. Nagano, M. Tsukiji, K. Ando, E. Hasegawa, and A. Ishitani, “Mechanism of leakage current through the nanoscale SiO_2 layer,” *J. Appl. Phys.*, Vol. 75, No. 7, pp.3530-3535, 1994.

[62] S.-H. Lo, D.A. Buchanan, Y. Taur, and W. Wang, “Quantum-mechanical modelling of electron tunnelling current from the inversion layer of ultra-thin-oxide nMOSFET’s,” *IEEE Electron Device Letters*, Vol. 18, No. 5, pp.209-211, May 1997.

[63] D.A. Buchanan, “Scaling the gate dielectric: materials, integration, and reliability,” *IBM J. Res. & Develop.*, Vol. 43, No. 3, pp.245-264, 1999.

[64] E.P. Gusev, V. Narayanan, M.M. Frank, “Advanced high-κ dielectric stacks with polySi and metal gates: recent progress and current challenges,” *IBM J. Res. & Dev.*, Vol. 50, No. 4/5, pp.387-410, July/September 2006.

[65] A. Toriumi, K. Kita, K. Tomida, Y. Zhao, J. Widiez, T. Nabatame, H. Ota and M. Hirose, “Materials science-based device performance engineering for metal gate high-k CMOS,” in *IEDM Tech. Dig.*, pp.53-56, 2007.

[66] G.D. Wilk, R.M. Wallace, J.M. Anthony, “High-k gate dielectrics: current status and materials properties considerations,” J. Appl. Phys., Vol. 89, No. 10, pp.5243-5275, May 2001.

[67] L.D. Yau, “A simple theory to predict the threshold voltage of short-channel IGFET’s,” *Solid-State Electronics*, Vol. 17, pp.1059-1063, 1974.

[68] T.-H. Kim, J. Keane, H. Eorn, and C.H. Kim, “Utilizing reverse short-channel effect for optimal subthreshold circuit design,” *IEEE Trans. Very Large Scale Integration (VLSI) Systems*, Vol. 15, No. 7, pp.821-829, July 2007.

[69] Shakir A. Abbas, and Robert C. Dockerty, “N-channel IGFET design limitations due to hot electron trapping,” in *IEDM Tech. Dig.*, pp.35-38, 1975.

[70] C. Hu, S.C. Tam, F.-C. Hsu, P.-K. Ko, T.-Y. Chan, and K.W. Terrill, “Hot-electron-induced MOSFET degradation—model, monitor, and improvement,” *IEEE Trans. Electron Devices*, Vol. ED-32, No. 2, pp.375-385, February 1985.

[71] S. Baba, A. Kita, J. Ueda, “Mechanism of hot carrier induced degradation in MOSFET’s,” in *IEDM Tech. Dig.*, pp.734-737, 1986.

[72] S. Ogawa, M. Shimaya, and N. Shiono, “Interface-trap generation at ultrathin SiO_2 (4-6 nm)-Si interfaces during negative-bias temperature aging,” *J. Appl. Phys.*, Vol. 77, No. 3, pp.1137-1148, February 1995.

[73] William Shockley, “Problems related to *p-n* junctions in silicon,” *Solid-State Electronics*, Vol. 2, No. 1, pp.35-67, 1961.

[74] A. Phillips Jr., R.R. O’Brien, R.C. Joy, “IGFET hot electron emission model,” in *IEDM Tech. Dig.*, pp.39-42, 1975.

[75] T.H. Ning, C.M. Osburn, and H.N. Yu, “Emission probability of hot electrons from silicon into silicon dioxide,” *J. Appl. Phys.*, Vol. 48, No. 1, pp.286-293, 1977.

[76] S. Tam, P.-K. Ko, C. Hu, and R.S. Muller, “Correlation between substrates and gate currents in MOSFET’s,” *IEEE Trans. Electron Devices*, Vol. ED-29, No. 11, pp.-1740-1744, November 1982.

[77] K.R. Hofmann, C. Werner, W. Weber, and G. Dorda, “Hot-electron and hole-emission effects in short n-channel MOSFET’s” *IEEE Trans. Electron Devices*, Vol. ED-32, No. 3, pp.691-699, March 1985.

[78] A. Goelzberger, A.D. Lopez, and R.J. Strain, “On the formation of surface states during stress aging of thermal Si-SiO_2 interfeces,” *J. Electrochemical Society*, Vol. 120, No. 1, pp.90-96, January 1973.

[79] D.K. Schroder, J.A. Babcock, “Negative bias temperature instability: road to cross in deep submicron silicon semiconductor manufacturing,” *J. Appl. Phys.*, Vol. 94, No. 1, pp.1-18, July 2003.

[80] S. Chakravarthi, A.T. Krishnan, V. Reddy, C.F. Machala and S. Krishnan, “A comprehensive framework for predictive modelling of negative bias temperature instability,” in *Int. Reliability Physics Symp.*, pp.273-282, 2004.

[81] N. Kimizuka, K. Yamaguchi, K. Imai, T. Iizuka, C.T. Liu, R.C. Keller and T. Horiuchi, "NBTI enhancement by nitrogen incorporation into ultrathin gate oxide for 0.10-μm gate CMOS generation," in *Symp. VLSI Tech. Dig.*, pp.92-93, 2000.

[82] G. Chen, K.Y. Chuah, M.F. Li, D.SH Chan, C.H. Ang, J.Z. Zheng, Y. Jin and D.L. Kwong, "Dynamic NBTI of PMOS transistors and its impact on device lifetime," in *Int. Reliability Physics Symp.*, pp.196-202, 2003.

[83] M.A. Alam, "A critical examination of the mechanics of dynamic NBTI for PMOSFETs," in *IEDM Tech. Dig.*, pp.345-348, 2003.

[84] M. Nishigohri, K. Ishimaru, M. Takuhashi, Y. Unno, Y. Okayama, F. Matsuoka, and M. Kinugawa, "Anomalous hot-carrier induced degradation in very narrow channel nMOSFETs with STI structure," in *IEDM Tech. Dig.*, pp.881-884, 1996.

[85] J.F. Chen, K. Ishimaru, and C. Hu, "Enhanced hot-carrier induced degradation in shallow trench isolated narrow channel PMOSFET's," *IEEE Electron Device Letters*, Vol. 19, No. 9, pp.332-334, September 1998.

[86] H.-S. Wong, and Y. Taur, "Three-dimensional 'atomistic' simulation of discrete random dopant distribution effects in sub-0.1μm MOSFET's," in *IEDM Tech. Dig.*, pp.705-708, 1993.

[87] T. Mizuno, J. Okamura, and A. Toriumi, "Experimental study of threshold voltage fluctuation due to statistical variation of channel dopant number in MOSFET's," *IEEE Trans. Electron Devices*, Vol. 41, No. 11, pp.2216-2221, November 1994.

[88] K. Takeuchi, T. Tatsumi, A. Furukawa, "Channel engineering for the reduction of random-dopant-placement-induced threshold voltage fluctuation," in *IEDM Tech. Dig.*, pp.841-844, 1997.

[89] K.R. Lakshmikumar, R.A. Hadaway, and M.A. Copeland, "Characterization and modeling of mismatch in MOS transistors for precision analogue design," *IEEE J. Solid-State Circuits*, Vol. SC-21, No. 6, pp.1057-1066, June 1986.

[90] W. Hinsberg, F. Houle, M. Sanchez, J. Hoffnagle, G. Wallraff, D. Mdeiros, G. Gallatin, and J. Cobb, "Extendibility of chemically applified resists: another brick wall," in *Proc. of SPIE,* Vol. 5039, 2003.

[91] S. Hasegawa, Y. Kitamura, K. Takahata, H. Okamoto, T. Hirai, K. Miyashita, T. Ishida, H. Aizawa, S. Aota, A. Azuma, T. Fukushima, H. Harakawa, E. Hasegawa, M. Inohara, S. Inumiya, T. Ishizuka, T. Iwamoto, N. Kariya, K. Kojima, T. Komukai, N. Matsunaga, S. Mimotogi, S. Muramatsu, K. Nagatomo, S. Nagahara, Y. Nakahara, K. Nakajima, K. Nakatsuka, M. Nishigoori, A. Nomachi, R. Ogawa, N. Okada, S. Okamoto, K. Okano, T. Oki, H. Onoda, T. Sasaki, M. Satake, T. Suzuki, Y. Suzuki, M. Tagami, K. Takeda, M. Tanaka, K. Taniguchi, M. Tominaga, G.

Tsutsui, K. Utsumi, S. Watanabe, T. Watanabe, Y. Yoshimizu, T. Kitano, H. Naruse, Y. Goto, T. Nakayama, N. Nakamura and F. Matsuoka, “A cost-conscious 32nm CMOS platform technology with advanced single exposure lithography and gate-first metal gate/high-k process,” in *IEDM Tech. Dig.*, 2008.

[92] A. Cathignol, K. Rochereau, G. Ghibaudo, “Impact of a single grain boundary in the polycrystalline silicon gate on sub 100nm bulk MOSFET characteristics – implication on matching properties,” in *Proc. ULIS*, pp.145-148, 2006.

[93] W.E. Taylor, N.H. Odell, and H.Y. Fan, “Grain boundary barriers in Germanium,” *Phys. Rev.*, Vol. 88, No. 4, pp.867-875, November 15, 1952.

[94] T.I. Kamins, “Hall mobility in chemically deposited polycrystalline silicon,” *J. Appl. Phys.*, Vol. 42, No. 11, pp.4357-4365, October 1971.

[95] John Y.W. Seto, “The electrical properties of polycrystalline silicon films,” J. Appl. Phys., Vol. 46, No. 12, pp.5247-5254, December 1975.

[96] Charles S. Smith, “Piezoresistance effect in germanium and silicon,” *Phys. Rev.*, Vol. 94, No. 1, pp.42-49, April 1, 1954.

[97] A. Steegen, A. Lauwers, M. de Potter, G. Badenes, R. Rooyachers, K. Maex, “Silicide and shallow trench isolation line width dependent stress induced junction leakage,” in *Symp. VLSI Tech. Dig.*, pp.180-181, 2000.

[98] R.A. Bianchi, G. Bouche, O. Roux-dit-Buisson, “Accurate modelling of trench isolation induced mechanical stress effects on MOSFET electrical performance,” in *IEDM Tech. Dig.*, pp.117-120, 2002.

[99] J. Welser, J.L. Hoyt, S. Takagi, and J.F. Gibbons, “Strain dependence of the performance enhancement in strained-Si *n*-MOSFETs,” in *IEDM Tech. Dig.*, pp.373-376, 1994.

[100] K. Rim, J. Welser, J.L. Hoyt, and J.F. Gibbons, “Enhancement hole mobilities in surface-channel strained-Si *p*-MOSFETs,” in *IEDM Tech. Dig.*, pp.517-520, 1995.

[101] K. Rim, J. Chu, H. Chen, K.A. Jenkins, T. Kanarsky, K. Lee, A. Mocuta, H. Zhu, R. Roy, J. Newbury, J. Ott, K. Petrarca, P. Mooney, D. Lacey, S. Koester, K. Chan, D. Boyd, M. Ieong, and H.-S. Wong, “Characteristics and device design of sub-100 nm strained Si N- and PMOSFETs,” in *Symp. VLSI Tech. Dig.*, pp.98-99, 2002.

[102] S. Ito, H. Namba, K. Yamaguchi, T. Hirata, K. Ando, S. Koyama, S. Kuroki, N. Ikezawa, T. Suzuki, T. Saitoh, T. Horiuchi, “Mechanical stress effect of etch-stop Nitride and its impact on deep submicron transistor design,” in *IEDM Tech. Dig.*, pp.247-250, 2000.

[103] P.R. Chidambaram, B.A. Smith, L.H. Hall, H. Bu, S. Chakravarthi, Y. Kim, A.V. Samoilov, A.T. Kim, P.J. Jones, R.B. Irwin, M.J. Kim, A.L.P. Rotondaro, C.F.

Machala and D.T. Grider, “35% drive current improvement form recessed-SiGe drain extensions on 37 nm gate length PMOS,” in *Symp. VLSI Tech. Dig.*, pp.48-49, 2004.

[104] S.E. Thompson, M. Armstrong, C. Auth, S. Cea, R. Chau, G. Glass, T. Hoffman, J. Klaus, Z. Ma, B. Mcintyre, A. Murthy, B. Obradovic, L. Shifren, S. Sivakumar, S. Tyagi, T. Ghani, K. Mistry, M. Bohr, and Y. El-Mansy, “A logic nanotechnology featuring strained-silicon,” *IEEE Electron Device Letters*, Vol. 25, No. 4, pp.191-193, April 2004.

[105] Roosevelt People, “Physics and applications of Ge_xSi_{1-x}/Si strained-layer heterostructures,” *IEEE J. Quantum Electronics*, Vol. QE-22, No. 9, pp.1696-1710, September 1986.

[106] S.-I. Takagi, J.L. Hoyt, J.J. Welser, and J.F. Gibbons, “Comparative study of phonon-limited mobility of two-dimensional electrons in strained and unstrained Si metal-oxide-semiconductor field-effect transistors,” *J. Appl. Phys.*, Vol. 80, No. 3, pp.1567-1577, August 1996.

[107] Y. Sun, S.E. Thompson, and T. Nishida, “Physics of strain effects in semiconductors and metal-oxide-semiconductor field-effect transistors,” *J. Appl. Phys.*, Vol. 101, pp.104503, 2007.

[108] C.M. Osburn, I. Kim, S.K. Han, I. De, K.F. Yee, S. Gannavaram, S.J. Lee, C.-H. Lee, Z.J. Luo, W. Zhu, J.R. Hauser, D.-L. Kwong, G. Lucovsky, T.P. Ma, M.C. Öztürk, “Vertically scaled MOSFET gate stacks and junctions: how far are we likely to go?” *IBM J. Res. & Dev.*, Vol. 46, No. 2/3, March/May 2002.

[109] E.P. Gusev, E. Cartier, D.A. Buchanan, M. Gribelyuk, M. Copel, H. Okorn-Schmidt, C. D’Emic, “Ultrathin high-K metal oxides on silicon: processing, characterization and integration issues,” *Microelectronic Engineering*, 59, pp.341-349, 2001.

[110] E.P. Gusev, C. Cabral Jr., M. Copel, C. D’Emic, M. Gribelyuk, “Ultrthin HfO2 films grown on silicon by atomic layer deposition for advanced gate dielectrics applications,” *Microelectronic Engineering*, 69, pp.145-151, 2003.

[111] K. Sekine, S. Inumiya, M. Sato, A. Kaneko, K. Eguchi, Y. Tsunashima, “Nitrogen profile control by plasma nitridation technique for poly-Si gate HfSiON CMOSFET with excellent interface property and ultra-low leakage current,” in *IEDM Tech. Dig.*, pp.102-106, 2003.

[112] M.V. Fischetti, D. Neumayer, E. Cartier, “Effective electron mobility in Si inversion layers in metal-oxide-semiconductor systems with a high-k insulator: the role of the remote phonon scattering,” *J. Appl. Phys.*, Vol. 90, pp.4587-4608, 2001.

[113] R. Chau, S. Datta, M. Doczy, B. Doyle, J. Kavalieros and M. Metz, "High-κ/metal-gate stack and its MOSFET characteristics," *IEEE Electron Device Letters*, Vol. 25, No. 6, pp.408-410, June 2004.

[114] O. Weber, M. Casse, L. Thevenod, F. Ducroquet, T. Ernst, S. Deleonibus, "On the mobility in high-k/metal gate MOSFETs: evaluation of the high-k phonon scattering impact," *Solid-State Electronics*, Vol. 50, pp.626-631, 2006.

[115] Y.-C. Yeo, T.-J. King, and C. Hu, "Metal-dielectric band alignment and its implications for metal gate complementary metal-oxide-semiconductor technology," *J. Appl. Phys.*, Vol. 92, No. 12, pp.7266-7271, December 15, 2002.

[116] J. Westlinder, G. Sjöblom, J. Olsson, "Variable work function in MOS capacitors utilizing nitrogen-controlled TiN_x gate electrodes," *Microelectronic Engineering*, 75, pp.389-396, 2004.

[117] K. Choi, P. Lysaght, H. Alshareef, C. Huffman, H.-C. Wen, R. Harris, H. Luan, P.-Y. Hung, C. Sparks, M. Cruz, K. Matthews, P. Majhi, B.H. Lee, "Growth mechanism of TiN film on dielectric films and the effects on the work function," *Thin Solid Films*, 486, pp.141-144, 2005.

[118] L.R. Fonseca, "First-principles calculation of the TiN effective work function on SiO_2 and on HfO_2," *Phys. Rev. B*, Vol. 74, pp.195304, 2006.

[119] H.Y. Yu, C. Ren, Y.-C. Yeo, J.F. Kang, X.P. Wang, H.H.H. Ma, M.-F. Li, D.S.H. Chan, and D.-L. Kwong, "Fermi pinning-induced thermal instability of metal-gate work functions," *IEEE Electron Device Letters*, Vol. 25, No. 5, pp.337-339, May 2004.

[120] A. Yagishita, T. Saito, K. Nakajima, S. Inumiya, Y. Akasaka, Y. Ozawa, G. Minamihaba, H. Yano, K. Hieda, K. Suguro, T. Arikado, K. Okumura, "High performance metal gate MOSFETs fabricated by CMP for 0.1μm regime," in *IEDM Tech. Dig.*, pp.785-788, 1998.

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

[122] K. Maitra, M. M. Frank, V. Narayanan, V. Misra, E. A. Cartier, "Impact of metal gates on remote phonon scattering in titanium nitride/hafnium dioxide n-channel metal-oxide-semiconductor field effect transistors-low temperature electron mobility study," *J. Appl. Phys.*, Vol 102, pp.114507, 2007.

[123] TCAD Sentaurus, Synopsys, Version A-2007.12.

[124] Asen Asenov, “Simulation of statistical variability in nano MOSFETs,” in *Symp. VLSI Tech. Dig.*, pp.86-87, 2007.

[125] A. Asenov, S. Roy, A.R. Brown, G. Roy, C. Alexander, C. Riddet, C. Millar, B. Cheng, A. Martinez, N. Seoane, D. Reid, M.F. Bkhori, X. Wang, U. Kovac, “Advanced simulation of statistical variability and reliability in nano CMOS transistors,” in *IEDM Tech. Dig.*, pp.421, 2008.

[126] James W. Mayer, “Ion implantation in semiconductors,” in *IEDM Tech. Dig.*, pp.3-5, 1973.

[127] Simon M. Sze, *Semiconductor Devices: Physics and Technology*, 2nd Edition, John Wiley & Sons, 2002, Chapter 13.

[128] B. E. Deal and A. S. Grove “General relationship for the thermal oxidation of silicon,” *J. Appl. Phys.*, Vol. 36, No. 12, pp.3770-3778, December 1965.

[129] H.K. Gummel, “A self-consistent iterative scheme for one-dimensional steady state transistor calculations,” *IEEE Trans. Electron Devices*, Vol. 11, No. 10, pp.455-465, October 1964.

[130] Siegfried Selberherr, *Analysis and Simulation of Semiconductor Devices*, Springer-Verlag Wien New York, 1984, Chapter 7.

[131] B.V. Gokhale, “Numerical solutions for a one-dimensional silicon n-p-n transistor,” *IEEE Trans. Electron Devices*, Vol. ED-17, No. 8, pp.594-602, August 1970.

[132] M.G. Ancona, H.F. Tiersten, “Macroscopic physics of the silicon inversion layer,” *Phys. Rev. B*, Vol. 35, No. 15, pp.7959-7965, May 15, 1987-II.

[133] M.G. Ancona, G.J. Iafrate, “Quantum correction to the equation of state of an electron gas in a semiconductor,” *Phys. Rev. B*, Vol. 39, No. 13, pp.9536-9540, May 1, 1989.

[134] Frank Stern, “Calculated temperature dependence of mobility in silicon inversion layers,” *Phys. Rev. Lett.*, Vol. 44 No.22, pp.1469-1472, June 1980.

[135] G. Masetti, M. Severi, and S. Solmi, “Modeling of carrier mobility against carrier concentration in arsenic-, phosphorus-, and boron-doped silicon,” *IEEE Trans. Electron Devices*, Vol. ED-30, No. 7, pp.764-769, July 1983.

[136] N.D. Arora, J.R. Hauser, and D.J. Roulston, “Electron and hole mobilities in silicon as a function of concentration and temperature,” *IEEE Trans. Electron Devices*, Vol. ED-29, No. 2, pp.292-295, February 1982.

[137] C. Lombardi, S. Manzini, A. Saporito, and M. Vanzi, “A physically based mobility model for numerical simulation of nonplanar devices,” *IEEE Trans. Computer-Aided Design*, Vol. 7, No. 11, pp.1164-1171, November 1988.

[138] 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 Trans. Electron Devices*, Vol. 44, No. 9, pp.1529-1538, September 1997.

[139] D.M. Caughey, R.E. Thomas, "Carrier mobilities in silicon empirically related to doping and field," in *Proc. of the IEEE*, pp.2192-2193, December 1967.

[140] C. Canali G. Majni, R. Minder, and G. Ottaviani, "Electron and hole drift velocity measurements in silicon and their empirical relation to electric field and temperature," *IEEE Trans. Electron Devices*, Vol. ED-22, No. 11, pp.1045-1047, November 1975.

[141] J. Bardeen and W. Shockley, "Deformation potentials and mobilities in non-polar crystals," *Phys. Rev.*, Vol. 80, No. 1, pp.72-80, October 1, 1950.

[142] M. Lades, J. Frank, J. Funk, G. Wachutka, "Analysis of piezoresistive effects in silicon structures using multidimensional process and device simulation," in *Simulation of Semiconductor Devices and Processes*, Vol. 6, pp.22-25, September 1995.

[143] Y. Kanda, "A graphical representation of the piezoresistance coefficients in silicon," *IEEE Trans. Electron Devices*, Vol. ED-29, No. 1, pp.64-70, January 1982.

[144] S. E. Thompson, G. Sun, K. Wu, J. Lim and T. Nishida, "Key differences for process-induced uniaxial vs. substrate-induced biaxial stressed Si and Ge channel MOSFETs," in *IEDM Tech. Dig.*, pp.221-224, 2004.

[145] T. Guillaume, and M. Mouis, "Calculations of hole mass in [110]-uniaxially strained silicon for the stress-engineering of p-MOS transistors," *Solid-State Electronics*, Vol.50, pp.701-708, 2006.

[146] B. Obradovic, P. Matagne, L. Shifren, E. Wang, M. Stettler, J. He and M.D. Giles, "A physically-based analytic model for stress-induced hole mobility enhancement," *Journal of Comptional Electronics*, Vol. 3, No. 3-4, pp.161-164, 2004.

[147] I. Martin-Bragado, M. Jaraiz, P. Castrillo, R. Pinacho, J.E. Rubio and J. Barbolla, "A kinetic Monte Carlo annealing assessment of the dominant features form ion implant simulations," *Materials Science and Engineering B*, Vol. 114-115, pp.345-348, December 2004.

[148] N. Sagara, M. Kuwahar, M. Makino and J. Chao, "An adaptive mesh generation of surfaces defined by lie algebra and its visualization toward intelligent communication system," in *Proc. TENCON2004, IEEE Region 10 Conference*, Vol. B, pp.454-457, 2004. Or, the original Japanese paper by H. Sano, M. Makino, and J. Chao, "An adaptive high quality mesh generation for surface defined by linear lie

algebra," *Journal of the Japan Society for simulation technology*, Vol. 20, No. 3, pp.251-258, 2001.

[149] D.J. Frank, Y. Taur, M. Ieong, and H.-S.P. Wong, "Monte Carlo modelling of threshold variation due to dopant fluctuations," in *Symp. VLSI Circuits Dig.*, pp.171-172, 1999.

[150] N. Sano, K. Matsuzawa, M. Mukai and N. Nakayama, "Role of long-range and short-range Coulomb potentials in threshold characteristics under discrete dopants in sub-0.1 μm Si-MOSFETs," in *IEDM Tech. Dig.*, pp.275-278, 2000.

[151] T. Ezaki, T. Ikezawa, A. Notsu, K. Tanaka, and M. Hane, "3D MOSFET simulation considering long-range Coulomb potential effects for analyzing statistical dopant-induced fluctuations associated with atomistic process simulator," in *Proc. SISPAD*, pp.91-94, 2002.

[152] T. Yamaguchi, H. Namatsu, M. Nagase, K. Yamazaki, and K. Kurihara, "Nanometer-scale linewidth fluctuations caused by polymer aggregates in resist films," *Appl. Phys. Lett.*, Vol. 71, No. 16, pp.2388-2390, October 1997.

[153] H. Namatsu, M. Nagase, T. Yamaguchi, K. Yamazaki, and K. Kurihara, "Influence of edge roughness in resist patterns on etched patterns," *J. Vac. Sci. Technol. B*, Vol. 16, No. 6, pp.3315-3321, Nov/Dec 1998.

[154] S. Kaya, A.R. Brown, A. Asenov, D. Magot and T. Linton, "Analysis of statistical fluctuations due to line edge roughness in sub-0.1μm MOSFETs," in *Proc. SISPAD*, pp.78-81, 2001.

[155] A. Asenov, S. Kaya, and A.R. Brown, "Intrinsic parameter fluctuations in decananometer MOSFETs introduced by gate line edge roughness," I*EEE Trans. Electron Devices*, Vol. 50, No. 5, pp.1254-1260, May 2003.

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

[157] A. Asenov, A. Cathignol, B. Cheng, K.P. McKenna, A.R. Brown, A.L. Shluger, D. Chanemougame, K. Rochereau, and G. Ghibaudo, "Origin of the asymmetry in the magnitude of the statistical variability of n- and p-channel poly-Si gate bulk MOSFETs," *IEEE Electron Device Letters*, Vol. 29, No. 8, pp913-915, August 2008.

[158] F. Adamu-Lema, G. Roy, A.R. Brown, A. Asenov and S. Roy, "Intrinsic parameter fluctuations in conventional MOSETs at the scaling limit: a statistical study," *Journal of Computational Electronics*, Vol.3 pp.203-206, 2004.

[159] W.R. Thurber, R.L. Mattis, and Y,M. Liu, "Resistivity-dopant density relationship for phorsporous doped silicon," J. Electrochm. Soc., Vol. 127, pp.1807-1812, 1980.

[160] Z. Luo, N. Rovedo, S. Ong, B. Phoong, M. Eller, H. Utomo, C. Ryou, H. Wang, R. Stierstorfer, L. Clevenger, S. Kim, J. Toomey, D. Sciacca, J. Li, W. Wille, L. Zhao, L. Teo, T. Dyer, S. Fang, J. Yan, O. Kwon, O. Kwon, D. Park, J. Holt, J. Han, V. Chan, J. Yuan, T. Kebede, H. Lee, S. Kim, S. Lee, A. Vayshenker, Z. Yang, C. Tian, H. Ng, H. Shang, M. Hierlemann, J. Ku, J. Sudijono, M. Ieong, "High performance transistors featured in an aggressively scaled 45 nm bulk CMOS technology," in *Symp. VLSI Tech. Dig.*, pp.16-17, 2007.

[161] K.-L. Cheng, C.C. Wu, Y.P. Wang, D.W. Lin, C.M. Chu, Y.Y. Tarng, S.Y. Lu, S.J. Yang, M.H. Hsieh, C.M. Liu, S.P. Fu, J.H. Chen, C.T. Lin, W.Y. Lien, H.Y. Huang, P.W. Wang, H.H. Lin, D.Y. Lee, M.J. Huang, C.F. Nieh, L.T. Lin, C.C. Chen, W. Chang, Y.H. Chiu, M.Y. Wang, C.H. Yeh, F.C. Chen, C.M. Wu, Y.H. Chang, S.C. Wang, H.C. Hsieh, M.D. Lei, K. Goto, H.J. Tao, M. Cao, H.C. Tuan, C.H. Diaz, and Y.J. Mii, "A highly scaled, high performance 45 nm bulk logic CMOS technology with 0.242 μm^2 SRAM cell," in *IEDM Tech. Dig.*, pp.243-246, 2007.

[162] T. Miyashita, K. Ikeda, Y. S. Kim, T. Yamamoto, Y. Sambonsugi, H. Ochimizu, T. Sakoda, M. Okuno, H. Minakata, H. Ohta, Y. Hayami, K. Ookoshi, Y. Shimamune, M. Fukuda, A. Hatada, K. Okabe, T. Kubo, M. Tajima, T. Yamamoto, E. Motoh, T. Owada, M. Nakamura, H. Kudo, T. Sawada, J. Nagayama, A. Satoh, T. Mori, A. Hasegawa, H. Kurata, K. Sukegawa, A. Tsukune, S. Yamaguchi, K. Ikeda, M. Kase, T. Futatsugi, S. Satoh, and T. Sugii, "High-performance and low-power bulk logic platform uilizing FET specific multiple-stressors with highly enhanced strain and full-porous low-*k* interconnects for 45-nm CMOS technology," in *IEDM Tech. Dig.*, pp.251-254, 2007.

[163] Joseph M. Steigerwald, "Chemical mechanical polish: the enabling technology," in *IEDM Tech. Dig.*, pp.--, 2008.

[164] J. Wang, Y. Tateshita, S. Yamakawa, K. Nagano, T. Hirano, Y. Kikuchi, Y. Miyanami, S. Yamaguchi, K. Tai, R. Yamamoto, J. Wang, Y. Tateshita, S. Yamakawa, K. Nagano, T. Hirano, Y. Kikuchi, Y. Miyanami, S. Yamaguchi, K. Tai, R. Yamamoto, S. Kadomura and N. Nagashima, "Novel channel-stress enhancement technology with eSiGe S/D and recessed channel on damascene gate process," in *Symp. VLSI Tech. Dig*., pp.46-47, 2007.

[165] S. Yamakawa, J. Wang, Y. Tateshita, K. Nagano, M. Tsukamoto, H. Ohri, N. Nagashima and H. Ansai, "Analysis of novel stress enhancement effect based on damascene gate process with eSiGe S/D for pFETs," in *Proc. SISPAD*, pp.109-112, 2007.

[166] S. Yamakawa, S. Mayuzumi, Y. Tateshita, H. Wakabayashi, and H. Ansai, “Stress enhancement concept on replacement gate technology with top-cut stress liner for nFETs,” in Proc. ESSDERC, pp.174-177, 2008.

[167] A. Oishi, O. Fujii, T. Yokoyama, K. Ota, T. Sanuki, H. Inokuma, K. Eda, T. Idaka, H. Miyajima, S. Iwasa, H. Yamasaki, K. Oouchi, K. Matsuo, H. Nagano, T. Komoda, Y. Okayama, T. Matsumoto, K. Fukasaku, T. Shimizu, K. Miyano, T. Suzuki, K. Yahashi, A. Horiuchi, Y. Takegawa, K. Saki, S. Mori, K. Ohno, I. Mizushima, M. Saito, M. Iwai, S. Yamada, N. Nagashima and F. Matsuoka, “High performance CMOSFET technology for 45 nm generation and scalability of stress-induced mobility enhancement technique,” in *IEDM Tech. Dig.*, pp.229-232, 2005.

[168] K.K. Ng, C.S. Rafferty, H.-I. Cong; “Effective on-current of MOSFETs for large-signal speed consideration,” in *IEDM Tech. Dig.*, pp.693-696, 2001.

[169] M.H. Na, E.J. Nowak, W. Haensch, J. Cai, “The effective drive current in CMOS inverters,” in *IEDM Tech. Dig.*, pp.121-124, 2002.

[170] J. Deng and H.-S.P. Wong, “Metrics for performance benchmarking of nanoscale Si and Carbon nanotube FETs including device nonidealities,” *IEEE Trans. Electron Devices*, Vol. 53, No. 6, pp.1317-1322, June 2006.

[171] E. Yoshida, Y. Momiyama, M. Miyamoto, T. Saiki, M. Kojima, S. Satoh, and T. Sugii, “Performance boost using a new device design methodology based on characteristic current for low-power CMOS,” in *IEDM Tech. Dig.*, pp.195-198, 2006.

[172] Hidetoshi Onodera, “Toward variability-aware design,” in *Symp. VLSI Tech. Dig.*, pp.91-92, 2007.

[173] V. Moroz, G. Eneman, P. Verheyen, F. Nouri, L. Washington, L. Smith, M. Jurczak, D. Pramanik, and X. Xu, “The impact of layout on stress-enhanced transistor performance,” in *Proc. SISPAD*, pp.143-146. 2005.

[174] Kelin J. Kuhn, “Reducing variation in advanced logic technologies: approaches to process and design for manufacturability of nanoscale CMOS,” in *IEDM Tech. Dig.*, pp.471-474, 2007.

[175] A. Cathignol, B. Cheng, D. Chanemougame, A.R. Brown, K. Rochereau, G. Ghibaudo, and A. Asenov, “Quantitative evaluation of statistical variability sources in a 45-nm technological node LP N-MOSFET,” in *IEEE Electron Device Letters*, Vol. 29, No. 6, June 2008.

[176] C. Alexander, G. Roy, and A. Asenov, “Random-dopant-induced drain current variation in nano-MOSFETs: a three-dimensional self-consistent Monte Carlo simulation study using ‘*ab initio*’ ionized impurity scattering,” *IEEE Trans. Electron Devices*, Vol. 55, No. 11, pp.3251-3258, November 2008.

[177] B. Cheng, D. Dideban, N. Moezi, C. Millar, G. Roy, X. Wang, S. Roy and A. Asenov, "Statistical-variability compact-modeling strategies for BSIM4 and PSP," *IEEE Design & Test of Computers*, Vol. 27 No. 2, pp.26-35, March/April 2010.

[178] H. Dadgour, K. Endo, V. De, and K. Banerjee, "Modeling and analysis of grain-orientation effects in emerging metal-gate devices and implicaitoins for SRAM reliability," in *IEDM Tech. Dig.*, pp.705-708, 2008.

[179] V. Moroz, L. Smith, X.-W. Lin, D. Pramanik, and G. Rollins, "Stress-aware design methodology," in *Proc. of ISQED*, pp.807-812, 2006.

[180] L. Sponton, L. Bomholt, D. Pramanik W. Fichtner, "A full 3D TCAD simulation study of line-width roughness effects in 65 nm technology," in *Proc. SISPAD*, pp.377-380, 2006.

[181] X. Wang, B. Cheng, S. Roy and A. Asenov, "Simulation of strain enhanced variability in nMOSFETs," in *Proc. ULIS*, pp.89-92, 2008.

[182] X. Wang, S. Roy and A. Asenov, "Impact of strain on LER variability in bulk MOSFETs," in *Proc. ESSDERC*, pp.190-193, 2008.

[183] P. Oldiges, Q. Lin, K. Pertillo, M. Sanchez, M. Ieong, and M. Hargrove, "Modeling line edge roughness effects in sub 100 nm gate length devices," in *Proc. SISPAD*, pp.131-134, 2000.

[184] K. Sohn, Y.-H. Suh, Y.-J Son, D.-S. Yim, K.-Y. Kim, et al., "A 100nm double-stacked 500MHz 72Mb separated-I/O synchronous SRAM with automatic cell-bias scheme and adaptive block redundancy," in *ISSCC Dig. Tech. Papers*, pp.386-622, 2008.

[185] G. Scott, J. Lutze, M. Rubin, F. Nouri, and M. Manley, "NMOS drive current reduction caused by transistor layout and trench isolation induced stress," in *IEDM Tech. Dig.*, pp.827-830, 1999.

[186] N. Wils, H.P. Tuinhout, and M. Meijer, "Characterization of STI edge effects on CMOS variability," *IEEE Trans. on Semiconductor Manufacturing*, Vol. 22, No. 1, pp.59-65, February 2009.

[187] C. Pacha, M. Bach, K. v. Arnim, R. Brederlow, D. Schimitt-Landsiedel, P. Seegebrecht, J. Berthold, and R. Thewes, "Impact of STI-induced stress, inverse narrow width effect, and statistical V_{TH} variations on leakage currents in 120 nm CMOS," in *Proc. ESSDERC*, pp.397-400, 2004.

[188] P. VanDerVoorn, D. Gan, and J. P. Krusius, "CMOS shallow-trench-isolation to 50-nm channel widths," *IEEE Trans. on Electron Devices*, Vol. 47, No.6, pp.1175-1182, June 2000.

[189] E. Augendre, R. Rooyackers, D. Shamiryan, C. Ravit, M. Jurczak and G. Badenes, "Controlling STI-related parasitic conduction in 90 nm CMOS and below," in *Proc. ESSDERC*, pp.507-510, 2002.

[190] M.J.M. Pelgrom, A.C.J. Duinmaijer, and A.P.G. Welbers, "Matching properties of MOS transistors," *IEEE J. Solid-State Circuits*, Vol. 24, No. 5, pp.1433-1440, October 1989.

[191] M. Nandakumar, A. Chatterjee, S. Sridhar, K. Joyner, M. Rodder, and I.-C. Chen, "Shallow trench isolation for advanced ULSI CMOS technologies," in *IEDM Tech. Dig.*, pp.133-136, 1998.

[192] M. Agostinelli, S. Lau, S. Pae, P. Marzolf, H. Muthali, S. Jacobs, "PMOS NBTI-induced circuit mismatch in advanced technologies," *Microelectronics Reliability*, Vol.46, pp.63-68, 2006.

[193] C. H. Tu, S. Y. Chen, A. E. Chuang, H. S. Huang, Z. W. Jhou, C. J. Chang, S. Chou, and J. Ko, "Transistor variability after CHC and NBTI stress in 90nm pMOSFET technology," *Electronics Letters*, Vol.45 No.16, 2009.

[194] H. Fukutome, Y. Momiyama, Y. Tagawa, T. Kubo, T. Aoyama, H. Arimoto, and Y. Nara, "Direct measurement of effects of shallow-trench isolation on carrier profiles in sub-50 nm N-MOSFETs," in *Symp. VLSI Tech. Dig.*, pp.140-141, 2005.

[195] B. Cheng, S. Roy, A. R. Brown, C. Millar and A. Asenov, "Evaluation of statistical variability in 32 and 22 nm technology generation LSTP MOSFETs," *Solid-State Electronics*, Vol.53, pp.767-772, 2009.

[196] A. Asenov, G. Slavcheva, A. R. Brown, J. H. Davies, and S. Saini, "Increase in the random dopant induced threshold fluctuations and lowering in sub-100 nm MOSFETs due to quantum effects: a 3-D density-gradient simulation study," *IEEE Trans. Electron Devices*, Vol.48 No.4, April 2001.

[197] M.L. Polignano, I. Mica, V. Bontempo, F. Cazzaniga, M. Mariani, A. Mauri, G. Pavia, F. Sammiceli, G. Spoldi, "The evolution of the ion implantation damage in device processing," *J Mater Sci: Mater Electron*, Vol.19, pp.182-188, 2008.

[198] M. Miyamoto, H. Ohta, Y. Kumagai, Y. Sonobe, K. Ishibashi, and Y. Tainaka, "Impact of reducing STI-induced stress on layout dependence of MOSFET characteristics," *IEEE Trans. Electron Devices*, Vol.51 No.3, March 2004.

[199] T.-H. Lee, Y.-K. Fang, Y.-T. Chiang, H.Y. Chiu, M.-S. Chen and O. Cheng, "Effect of STI stress on leakage and Vccmin of a sub-65 nm node low-power SRAM," *J. Phys. D: Appl. Phys.*, Vol. 41, p.195101, 2008.

[200] K. Ohe, S. Odanaka, K. Moriyama, T. Hori, and G. Fuse, “Narrow-width effects of shallow trench-isolated CMOS with n+-polysilicon gate,” *IEEE Trans. Electron Devices*, Vol.36 No.6, pp.1110-1116, June 1989.

[201] T. B. Hook, J. Brown, P. Cottrell, E. Adler, D. Hoyniak, J. Johnson, and R. Mann, “Lateral ion implant straggle and mask proximity effect,” *IEEE Trans. Electron Devices*, Vol.50 No.9, pp.1946-1951, Sept. 2003.

[202] M. F. Bukhori, S. Roy, and A. Asenov, “Statistical aspects of reliability in bulk MOSFETs with multiple defect states and random discrete dopants,” *Microelectronics Reliability*, Vol.48, pp.1549-1552, 2008.

[203] Steven E. Laux, “Techniques for small-signal analysis of semiconductor devices,” *IEEE Trans. Electron Devices*, Vol.ED-32, No.10, pp.2028-2037, October 1985.

[204] Martin Reiser, “A two-dimensional numerical FET model for DC, AC, and large-signal analysis,” *IEEE Trans. Electron Devices*, Vol.ED-20, No.1, pp.35-45, January 1973.

[205] D.E. Ward, R.W. Dutton, “A charge-oriented model for MOS transistor capacitances,” *IEEE J. Solid-State Circuit*s, Vol. SC-13, No. 5, pp.703-708, October 1978.

[206] M.A.Green, and J. Shewchun, “Application of the small-signal transmission line equivalent circuit model to the a.c., d.c. and transient analysis of semiconductor devices,” *Solid-State Electronics*, Vol. 17, pp.941-949, 1974.

[207] P. Klein, K. Hoffmann, B. Lemaitre, “Description of the bias dependent overlap capacitance at LDD MOSFETs for circuit applications,” in *IEDM Tech. Dig.*, pp.493-496, 1993.

[208] R. Shrivastava, and K. Fitzpatrick, “A simple model for the overlap capacitance of a VLSI MOS device,” *IEEE Trans. Electron Devices*, Vol. ED-29, No.12, pp.1870-1875, December 1982.

[209] F. Pregaldiny, C. Lallement, D. Mathiot, “A simple efficient model of parasitic capacitances of deep-submicron LDD MOSFETs,” *Solid-State Electronics*, Vol. 46, pp.2191-2198, 2002.

[210] J.-C. Guo, C.C.-H. Hsu, P.-S. Lin, and S.S. Chung, “An accurate ‘decoupled C-V’ method for characterizing channel and overlap capacitances of miniaturized MOSFET,” in *VLSITSA*, pp.256-260, 1993.

[211] X. Wang, S. Roy, and A. Asenov, “Impact of strain on the performance of high-k/metal replacement gate MOSFETs,” in *Proc. ULIS*, pp.289-292, 2009.

[212] R.E. Bank, W.M. Coughran, Jr., W. Fichtner, E.H. Grosse, D.J. Rose, and R.K. Smith, "Transient simulation of silicon devices and circuits," *IEEE Trans. Electron devices*, Vol. ED-32, No. 10, pp.1992-2007, October 1985.

[213] R. E. Bank, D. J. Rose, and W. Fichtner, "Numerical methods for semiconductor device simulation," *IEEE Trans. Electron Devices*, ED-30, No.9, pp.1031-1041, September 1983.

[214] W. Fichtner, D. J. Rose, and R. E. Bank, "Semiconductor device simulation," *IEEE Trans. Electron Devices*, ED-30, No.9, pp.1018-1030, September 1983.

[215] Harry J.M. Veendrick, "Short-circuit dissipation of static CMOS circuitry and its impact on the design of buffer circuits," *IEEE J. Solid-State Electronics*, Vol. SC-19, No. 4, pp.468-473, August 1984.

[216] N. Hedenstierna and K.O. Jeppson, "CMOS circuit speed and buffer optimization," *IEEE Trans. Computer-Aided Design*, Vol. CAD-6, No. 2, pp.270-281, March 1987.

[217] T. Sakurai, and A.R. Newton, "A simple MOSFET model for circuit analysis," *IEEE Trans. Electron Devices*, Vol. 38, No. 4, pp.887-894, April 1991.

[218] A. Hirata, H. Onodera, and K. Tamaru, "Estimation of propagation delay considering short-circuit current for static CMOS gates," *IEEE Trans. Circuits and System—I: Fundamental Theory and Applications*, Vol. 45, No. 11, pp.1194-1198, November 1998.

[219] M.-E. Arbey, S. Galdin, P. Dollfus, P. Hesto, "Predictive expression of propagation delay in short channel CMOS/SOI inverter using Monte Carlo simulation," in *Proc. ESSDERC*, pp.500-503 1997.

[220] S. Galdin, M.-E. Arbey, P. Dollfus, P. Hesto, "Accurate analytical delay expression for short channel CMOS SOI inverter using Monte Carlo simulation," *Solid-State Electronics*, Vol. 43, pp.1869-1877, 1999.

[221] Yuan Taur, and Tak H. Ning, *Fundamentals of Modern VLSI Devices*, Cambridge University Press, 1998, Chapter 5.