Predictive Modeling for Extremely Scaled CMOS and Post Silicon Devices

by

Chi-Chao Wang

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved March 2011 by the
Graduate Supervisory Committee

Yu Cao, Chair
Chaitali Chakrabarti
Lawrence Clark
Dieter Schroder

ARIZONA STATE UNIVERSITY

May 2011

ABSTRACT

To extend the lifetime of complementary metal–oxide–semiconductors (CMOS), emerging process techniques are being proposed to conquer the manufacturing difficulties. New structures and materials are proposed with superior electrical properties to traditional CMOS, such as strain technology and feedback field-effect transistor (FB-FET). To continue the design success and make an impact on leading products, advanced circuit design exploration must begin concurrently with early silicon development. Therefore, an accurate and scalable model is desired to correctly capture those effects and flexible to extend to alternative process choices.

For example, strain technology has been successfully integrated into CMOS fabrication to improve transistor performance but the stress is non-uniformly distributed in the channel, leading to systematic performance variations. In this dissertation, a new layout-dependent stress model is proposed as a function of layout, temperature, and other device parameters. Furthermore, a method of layout decomposition is developed to partition the layout into a set of simple patterns for model extraction. These solutions significantly reduce the complexity in stress modeling and simulation.

On the other hand, semiconductor devices with self-feedback mechanisms are emerging as promising alternatives to CMOS. Fe-FET was proposed to improve the switching by integrating a ferroelectric material as gate insulator in a MOSFET structure. Under particular circumstances, ferroelectric capacitance is effectively negative, due to the negative slope of its polarization-electrical field

curve. This property makes the ferroelectric layer a voltage amplifier to boost surface potential, achieving fast transition. A new threshold voltage model for Fe-FET is developed, and is further revealed that the impact of random dopant fluctuation (RDF) can be suppressed.

Furthermore, through silicon via (TSV), a key technology that enables the 3D integration of chips, is studied. TSV structure is usually a cylindrical metal-oxide-semiconductors (MOS) capacitor. A piecewise capacitance model is proposed for 3D interconnect simulation. Due to the mismatch in coefficients of thermal expansion (CTE) among materials, thermal stress is observed in TSV process and impacts neighboring devices. The stress impact is investigated to support the interaction between silicon process and IC design at the early stage.

TABLE OF CONTENTS

Chapter Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

CHAPTER 1

INTRODUCTION

For the past decades, transistor performance has been consistently improved by scaling down the device feature size by following Moore's law stating that the number of transistors in an integrated circuit doubles every two years. The 2009 International Technology Roadmap for Semiconductors [1], predicts that the half pitch of high-performance logic technology will shrink to 11.9 nm and the equivalent oxide thickness (EOT) will scale down to 0.59 nm by 2020. When the device dimension is further scaled, the property of carrier transportation will change and enter the region of ballistic transportation. Moreover, the oxide thickness is scaled to 1nm, about twice the silicon lattice, implying the scaling is approaching the physical limits to the atomistic level. With the relentless scaling of CMOS technology, many secondary effects become more pronounced, such as drain induced barrier lowering (DIBL) effect, process variations, gate tunneling leakage, mobility degradation, and increasing power consumption. These secondary effects lead to dramatic challenges to robust circuit design and system integration.

To extend the lifetime of CMOS, emerging process techniques are proposed to conquer the manufacturing difficulties. Strain technology is being introduced to elevate carrier mobility to improve device performance by applying mechanical stress to the transistors [2]. On the other hand, novel devices with new structures and materials are proposed with superior electrical properties to traditional CMOS, such as CNT-FET [3] and FB-FET [4], for the nanometer era.

To continue the design success and make an impact on leading products, advanced circuit design exploration must begin concurrently with early silicon development. Therefore, an accurate and scalable model is desired to correctly capture those effects and is flexible to extend to alternative process choices.

With scaling of oxide thickness, threshold voltage decreases and results in an increasing off-state leakage. DIBL effect is another major challenge during scaling of CMOS technology. When drain voltage is applied to a short-channel device, the energy barrier height between source and drain is lowered, leading to the decrease of threshold voltage and increasing off-state leakage. To maintain the scaling criteria of leakage power consumption, V_{th} needs to be kept in a certain value; therefore more channel dopants are introduced to maintain the threshold voltage. However, increased doping concentration degrades carrier mobility and

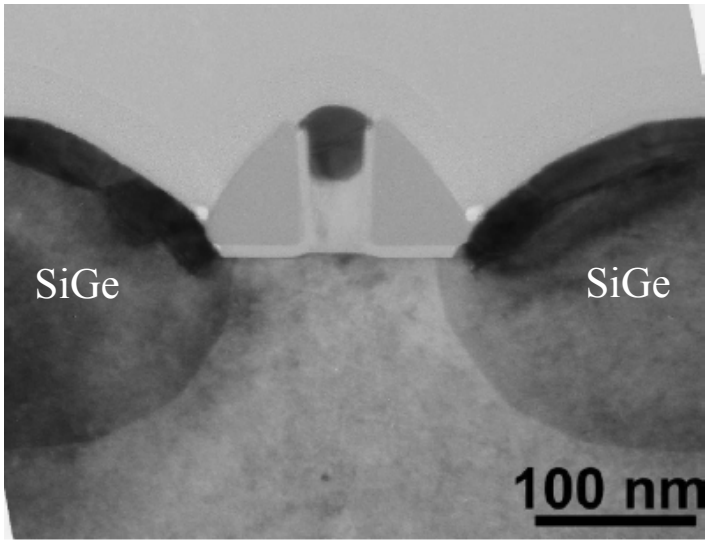


Figure 1.1 Illustration of eSiGe technology [75]

reduces the driving current. Strain technology, which employs mechanical stress to alter band structure of silicon and reduces carrier effective mass and scattering rate, is introduced to elevate carrier mobility for continual scaling, as shown in Fig. 1.1. Therefore, carriers can move faster and device performance is improved. Strain technology has been successfully integrated into CMOS fabrication to enhance carrier transport properties since 90nm node. However, since the stress is non-uniformly distributed in the channel, the enhancement in carrier mobility, velocity, and threshold voltage shift strongly depend on circuit layout patterns, leading to extra process variations. A compact stress model that physically captures this behavior is essential to bridge the process technology with design optimization.

The other scaling issue is management of power consumption. As the scaling is approaching the physical limit down to nanometer regime, the fundamental limits on minimum operational voltage and switching energy of transistors is restricted by the subthreshold slope [5]. In a MOSFET, the gate voltage must be changed at least by 2.3kT/q to change subthreshold current by ten times. At room temperature, 60mV is the minimum value required to modulate subthreshold current by an order of magnitude. To overcome this limit, new devices with self-feedback mechanism such as Ferro-electrical FET (Fe-FET) [6] is proposed to achieve the steep subthreshold slope, lowering the switching energy and operational voltage. A compact model is of significance to capture the behavior of steep subthreshold slope for early stage design exploration.

Meanwhile with scaling, device performance becomes more sensitive to manufacturing process and device parameters, contributing to the increasing process variations. The variation mainly results from the manufacturing imperfection and profoundly impacts all aspects of circuit performance, posing a grand challenge to future robust IC design, and become one of bottlenecks for further scaling [6]. Although in tradition variability issues are mostly controlled with improvements in the manufacturing process, the industry starts to accept the fact that some of the negative effects can be better mitigated during the design process. In this work, the variation mechanism in Fe-FET is investigated and shed a possibility that a device with internal feedback could suppress intrinsic variations.

On the other hand, through-silicon-via (TSV) is a promising technology that enables the 3D integration of chips with diverse functionalities, providing improved packing density, better noise immunity and faster speed due to reduced interconnect wire length. To evaluate the impact of TSV in circuit performance, a piecewise CV model is proposed for efficient simulation. Furthermore, a stress-aware design approach is utilized to optimize the required TSV area and the mobility variation influenced by the thermal stress during TSV process.

This dissertation is organized as following:

In chapter 2, compact modeling of mobility, velocity, and threshold voltage under eSiGe, DSL, and STI stresses is proposed for the first time. It physically captures the dependence on primary layout parameters, temperature, and other device characteristics. Methods of layout decomposition are proposed

for efficient stress extraction. The entire layout is decomposed into both vertical and lateral directions, reducing the analysis complexity while maintaining the accuracy.

In chapter 3, under particular circumstance, ferroelectric capacitance is effectively negative, due to the negative slope of its polarization-electrical field (P-E) curve. This property makes the ferroelectric layer a voltage amplifier to boost surface potential, achieving fast transition. A new threshold voltage model is developed to capture the subthreshold IV characteristics of Fe-FET. It is further revealed that the impact of random dopant fluctuation (RDF) on leakage variability can be significantly suppressed in Fe-FET.

In chapter 4, a piecewise capacitance model is proposed for efficient simulation to evaluate the impacts of TSV parameters on circuit performance. TSV interconnect capacitance varies with the applied voltage and is sensitive to the radius of TSV, thickness of barrier layer, and the doping concentration in the substrate. Besides TSV capacitance, TSV resistance also strongly depends on radius of conducting material. The dependences on TSV geometry parameters are discussed. TSV induced thermal stress is modeled and can be used to optimized the required TSV area.

The last chapter summarizes the major research results of this dissertation.

CHAPTER 2

MODELING OF LAYOUT-DEPENDENT STRESS EFFECT

Strain technology has been successfully integrated into CMOS fabrication to improve carrier transport properties since 90nm node. However, since the stress is non-uniformly distributed in the channel, the enhancement in carrier mobility, velocity, and threshold voltage shift strongly depend on circuit layout, leading to systematic performance variations among transistors. A compact stress model that physically captures this behavior is essential to bridge the process technology with design optimization. In this chapter, starting from the first principle, a new layout-dependent stress model is proposed as a function of layout, temperature, and other device parameters. Furthermore, a method of layout decomposition is developed to partition the layout into a set of simple patterns for efficient model extraction. These solutions significantly reduce the complexity in stress modeling and simulation. They are comprehensively validated by TCAD simulation and Si data, including the state-of-the-art strain technologies and the STI stress effect.

2.1 INTRODUCTION

For the past decades, the miniature of device feature size has driven the improvement in transistor performance [1]. However, many secondary effects become pronounced during scaling. For example, threshold voltage (V_{th}) decreases with scaling the channel length because of DIBL effect. To maintain the scaling criteria of leakage power consumption, V_{th} needs to be kept in a certain value. Therefore, more channel dopants are introduced to maintain the threshold voltage. However, increased doping concentration degrades carrier mobility and

reduces the driving current. Strain technology, which employs mechanical stress to alter band structure of silicon and reduces carrier effective mass and scattering rate, is introduced to elevate carrier mobility for continual scaling. In general, such a technology can be classified into biaxial stress and uniaxial stress, both of which contribute significant mobility enhancement. Based on the lattice mismatch between Si and SiGe, the biaxial stress is exerted by depositing a pseudomorphic Si layer on a relaxed SiGe substrate [8]. On the other hand, the uniaxial stress is applied to one direction, usually to the direction of channel, and has been adopted as standard process since 90nm node because of lower integration complexity and smaller threshold voltage shift [9]. The major techniques to introduce uniaxial stress include embedded SiGe technology (eSiGe), dual stress liner (DSL), stress memorization technique (SMT), and the parasitic stress from shallow trench isolation (STI). eSiGe technology embedded SiGe in the source and drain area to introduce compressive stress for PMOS [9]. DSL introduces the stress by depositing a highly stressed silicon nitride layer, tensile stress for NMOS region and compressive stress for PMOS region, over the entire wafer to elevate carrier mobility [9]. In SMT, the stress in the channel is transferred from the stressed deposited dielectric and is memorized during the re-crystallization of the active area and poly-gate when thermal annealing is activated [10]. STI stress results from the difference in thermal expansion coefficients between SiO_2 and Si. It is an intrinsic stress source and not intentionally built up for enhancing device performance [11][12].

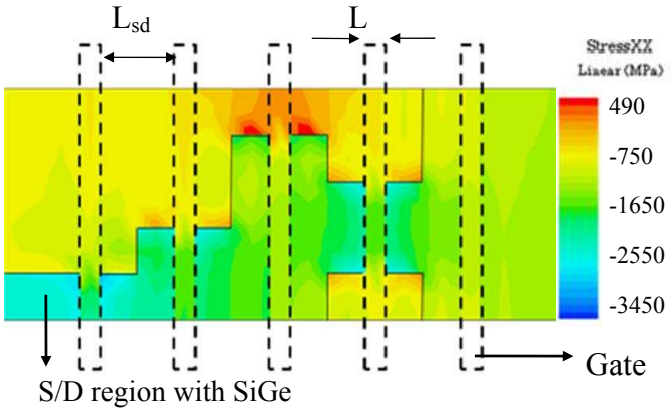
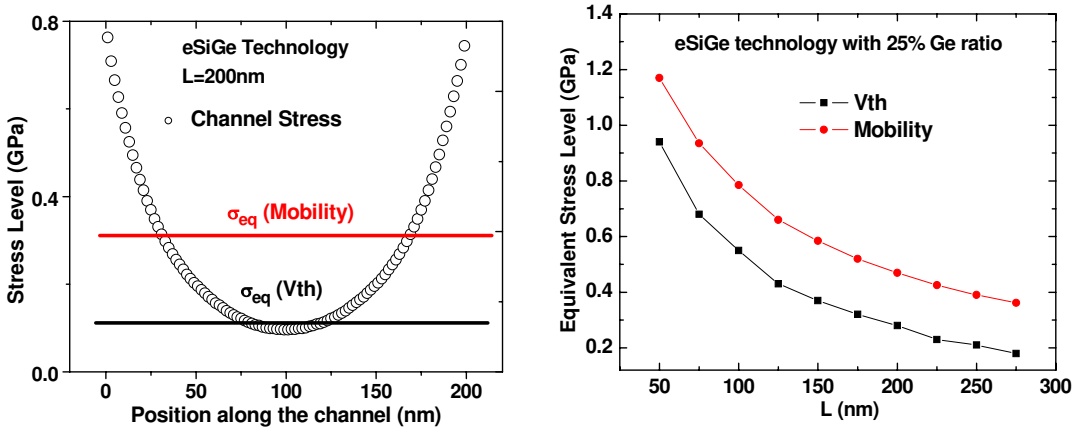


Figure. 2.1. Top view of stress contours in a five-finger layout pattern with SiGe embedded in source/drain area.

In the most strain technologies, the amount of performance enhancement depends on both the applied stress magnitude and circuit layout parameters, such as gate length, source/drain size, and the distance from gate edge to STI [13][14] because of the nature of mechanical stress in silicon: non-uniform distribution. Figure 2.1 illustrates the TCAD simulation [15] of stress distribution in a 45nm standard cell under restrictive design rules, where SiGe with 25% Ge composition is embedded in the S/D area. The stress level is widely different across the cell, depending on transistor size, layout pitch, etc. Such non-uniformity results in pronounced variations among transistors as well as circuit performance and further increases the complexity in modeling and simulation. To capture such a systematic effect, traditional efforts resort to TCAD simulation [16] to extract the stress level from the entire layout and analyze performance enhancement. This approach usually requires expensive computation, especially when chip complexity per unit area keeps increasing along with technology scaling. Therefore, it is necessary to develop a more effective modeling approach that is

able to extract the stress effect for each device and embed it into standard model parameters for circuit simulation. On the other hand, the stress dependence provides designers another alternative to optimize the circuit performance. For example, higher gate density per unit area can be achieved by exploiting the stress effect [13]. The insertion and placement of active area are optimized by STI stress to improve the circuit performance [17]. A new stress-aware layout design is proposed to reduce leakage power [18]. Therefore, an accurate model that physically captures the stress effect is required to bridge the process technology and design communities.

The layout-dependent stress effect is first observed and reported from STI stress [11]. Models regarding layout-dependent STI effect are proposed on the basis of the experimental observation that the changes of drive current and threshold voltage follow the trend of the length of oxide definition area (LOD) [12]. The work in [19] indicates that the trend of STI stress effect is attributed to



(a) (b)

Figure 2.2 (a) The stress distribution in the channel, (b) the equivalent stress levels for strain-induced shifts of V_{th} and mobility.

the non-uniform stress distribution in the channel, but it does not quantitatively explain how this distribution impacts the electrical properties. Instead, an empirical model is proposed. Moreover, STI width effect is also investigated and modeled to enhance circuit performance. On the other hand, the layout dependent stress effects are also observed in the state-of-the-art strain technologies [13][14]. In Fig. 2.2 (a), TCAD simulation using eSiGe technology shows that the stress profile in the channel, with higher stress level at the edges and lower stress in the center of the channel [21]. The stress profile is sensitive to the primary layout parameters such as channel length and source/drain diffusion length. To capture this layout dependence, Dunga et al. [23] proposes a modeling approach to finding an equivalent stress level in the channel accounting for the mobility enhancement with an assumption that the mobility enhancement is proportional to the applied stress. However, this approach also results in empirical fitting and not adequate for strain-induced V_{th} shift because the impacts of stress on different electrical properties are dominated by different stress regions. For example, the strain-induced threshold voltage shift is mainly dominated by the bottom stress level in the channel (more details in section 2.2.5), while the entire channel stresses are required to be taken into consideration for the enhancement of mobility. In Fig. 2.2(b), TCAD simulation [15][24] shows the obvious difference of the equivalent stresses between the shifts in threshold voltage and mobility for different devices with various channel lengths.

In this work, a new general modeling approach [76][77] is proposed to capture the layout-dependent stress effect. This model should physically capture

the impact of circuit layout on transistor performance such that model scalability is guaranteed for future technology generations. Although there are many layout parameters that affect the stress effect, recent trend in layout regularity suggests that it is feasible to decompose the entire layout into a pre-defined set of patterns. Therefore, the modeling efforts can be localized and only focus on pre-defined patterns for efficient extraction.

The important contributions of this chapter include:

- Compact modeling of mobility, velocity, and threshold voltage under eSiGe, DSL, and STI stresses. It physically captures the dependence on primary layout parameters, temperature, and other device characteristics.
- Layout decomposition methods for efficient stress extraction. The entire layout is decomposed in both vertical and lateral directions, reducing the analysis complexity while maintaining the accuracy;
- Circuit performance benchmark at the 45nm node. The interaction between layout and circuit performance is accurately predicted by the new stress models.

These solutions are developed and demonstrated in a 45nm design using restrictive design rules. They are comprehensively evaluated with TCAD simulations and published Si data. Based on them, the impact of the stress effect is well assessed and scalable into future technology generations.

The chapter is organized as following. Section 2.2 presents the model derivation for the dependence on layout and other parameters. In section 2.3, the proposed model is validated by TCAD simulation and Si data. Section 2.4

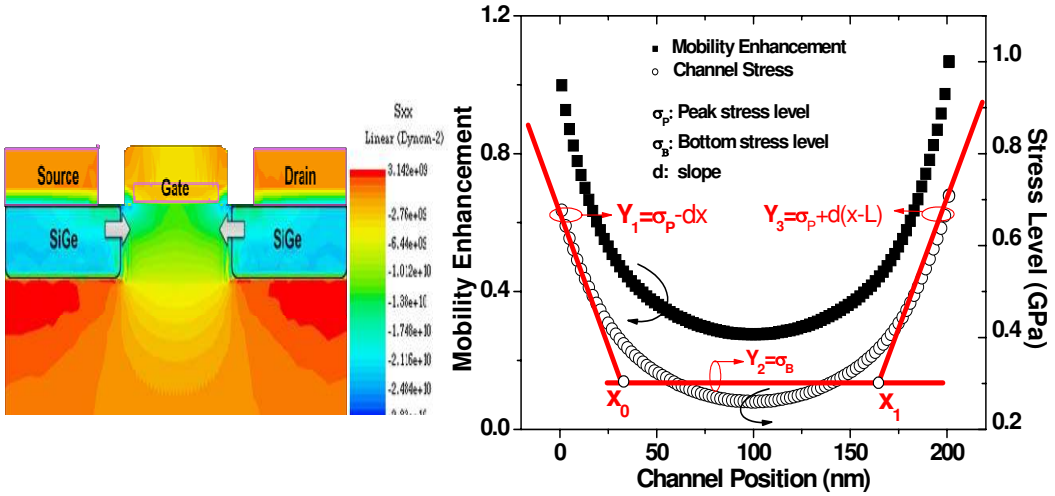
describes the method of layout decomposition, which systematically partitions the layout into simple patterns to analyze the stress impact on transistor performance. Section 2.5 concludes this chapter.

2.2 COMPACT STRESS MODELS

Strain technology has been successfully implemented since 90nm node to improve transistor performance. However, compact modeling for the stress effect is still at the infant stage to analyze and predict the circuit performance. In this section, based on the mechanical properties of the stress and its impacts on electrical properties, an analytical model is proposed to capture the layout and temperature dependence in this section.

2.2.1. Bathtub Stress Distribution

As investigated in [21], the magnitude of mechanical stress in a substrate with cubic symmetry, such as silicon, decays sharply from the edge of stress



(a) The stress induced by SiGe S/D

(b) The piecewise linear approximation of the stress distribution in the channel.

Figure 2.3 Modeling approach of the mechanical stress.

source and becomes less dependent on the distance when the location is far from the origin of the applied stress. For the state-of-the-art uniaxial strain technologies such as eSiGe and DSL, the stress in the channel mainly come from the source and drain area. Figure 2.3(a) shows the cross section of stress distribution for eSiGe technology, indicating that the stress in the channel results from the S/D stressors and non-uniformly distributed. Figure 2.3(b) shows the corresponding stress profile in the channel. Some models regarding the non-uniform stress distribution are proposed [12][25], but most of them are still too complicated for mobility analyses. Instead, in this work, without losing the generality we propose a piecewise linear model to capture the non-uniform stress distribution, as shown Y_1, Y_2, and Y_3 in Fig. 2.3(b) with expressions as Eqs. (2.1)-(2.3),

$$Y_1 = \sigma_P - dx \tag{2.1}$$

$$Y_2 = \sigma_B \tag{2.2}$$

$$Y_3 = \sigma_P + d(x - L) \tag{2.3}$$

where σ_P and σ_B denote the peak and bottom stress levels in the channel, respectively. d is the slope. Y_1 and Y_3 intercept with Y_2 at points of x_0 and x_1, respectively. Thus, x_0 and x_1 are expressed by Eqs. (2.4) and (2.5).

$$x_0 = \frac{\sigma_P - \sigma_B}{d} \tag{2.4}$$

$$x_1 = L - \frac{\sigma_P - \sigma_B}{d} \tag{2.5}$$

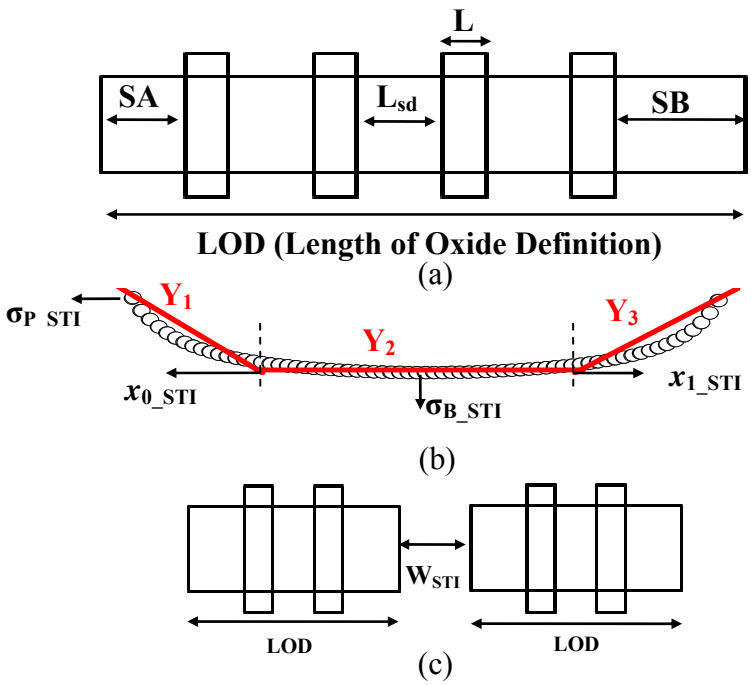


Figure. 2.4. Modeling of STI stress distribution: (a) layout pattern of MOSFET, (b) piecewise linear model for the lateral STI stress, (c) Definition of STI width (W_{STI})

In addition, STI stress is one of the important stress sources and originates from the edges of the oxide definition region. Figure 2.4(a) shows a layout pattern of a MOSFET. SA and SB represent the distances of the leftmost and rightmost fingers to STI boundary. LOD is the length of oxide definition. Figure 2.4(b) illustrates the corresponding STI stress distribution along the region of oxide definition. STI stress distribution is similar to eSiGe, and can be captured by Eqs. (2.1)-(2.5). To clearly express STI stress, the channel length (L) in Eqs (2.1)-(2.5) is replaced by LOD. Meanwhile, the peak and bottom stress levels are attached with subscripts of STI (σ_{P_STI} and σ_{B_STI}), making them distinguishable from eSiGe stress. Moreover, the piecewise linear model is capable of capturing the stress profile along the width direction because the similar bathtub stress distribution is observed in width direction [41]. These equations capture the

essential behavior of the stress distribution, as well as their important dependence on device size. Furthermore, they are simple enough to support further model derivation without losing sufficient model scalability.

2.2.2. Layout Dependence

Since the stress is non-uniformly distributed in the channel, the enhancement in carrier mobility, velocity, and threshold voltage shift strongly depend on circuit layout, leading to systematic performance variations among transistors. In this sub-section, the stresses from both lateral direction and width direction are discussed and modeled for further model derivation.

2.2.2.1. Stresses from lateral direction

Based on the piecewise linear model, the stress distribution in the channel is determined by peak and bottom stress levels as well as the decreasing rate of stress from the channel edge. Peak and bottom stresses are highly correlated to layout parameters and the manufacturing process [13][14]. For example, as channel length shrinks, the overall channel stress grows up because the channel is closer to the origin of the stressor in the source and drain regions, as shown in Fig. 2.5(a). Therefore, peak and bottom stress levels are accordingly elevated. Moreover, when S/D diffusion length (L_{sd}) decreases, σ_P and σ_B decline since the amount of stressor material in the source and drain regions is lower [14], as shown in Fig. 2.5(b). Note that when S/D diffusion length reaches a critical length, the stress level becomes saturated [14]. To combine these issues, σ_P is modeled as functions of channel length (L) and S/D diffusion length (L_{sd}) as Eq. (2.6),

$$\sigma_P = \left(1 + \frac{m}{L} + \frac{m}{L + L_{sd}} + \frac{m}{2L + L_{sd}}\right) \cdot \frac{L_{sd}}{A + L_{sd}} \cdot \sigma_m \tag{2.6}$$

where σ_m is the saturation stress level. A and *m* are fitting parameters accounting for the dependence on L_{sd} and the stress decreasing rate over distance from neighboring transistors, respectively. Each term in the parenthesis represents the contribution by a diffusion region, depending on their separation distance to the channel. The first two terms in Eq. (2.6) account for the contribution from the source/drain area of the target transistor, as shown the middle finger in Fig. 2.5(b)), while the rest two terms represent the stress sources from the closest two neighboring transistors. Equation (2.6) assumes that all diffusion regions in the neighboring transistors have the same size L_{sd}. If they are different, L_{sd} should be replaced with the exact value. The other important stress parameter is σ_B. As channel length becomes shorter and approaches zero, σ_B grows up to the limit of peak stress level. This dependence can be modeled by Eq. (2.7) with a fitting parameter C:

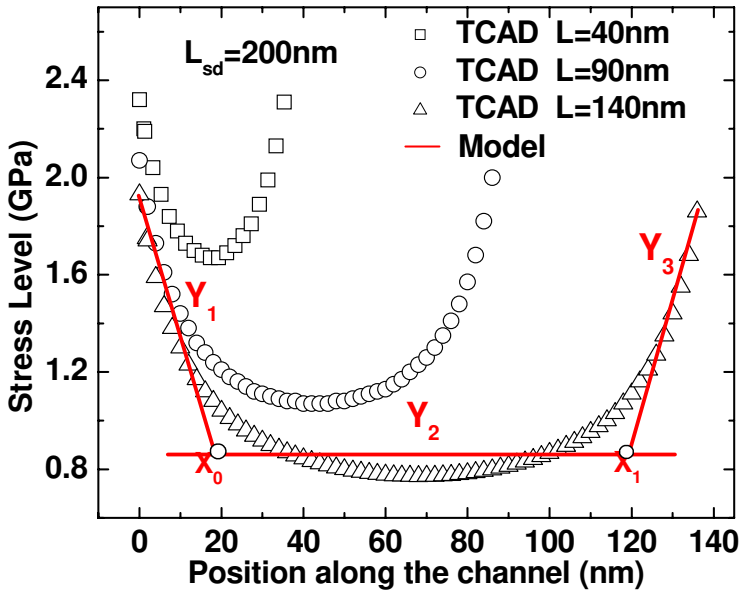
$$\sigma_B = \frac{C}{C + L} \cdot \sigma_P \tag{2.7}$$

Model coefficients are dependent on the fabrication technology. Figure 2.5(b) shows the excellent agreement with the TCAD simulation results.

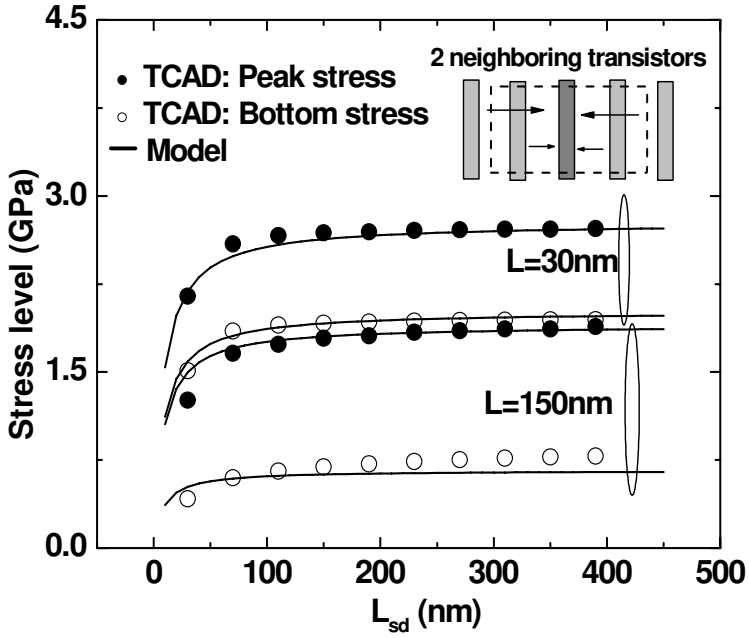
Similar to L_{sd} dependence for eSiGe/DSL, larger the STI width (W_{STI}), the distance between regions of oxide definition shown in Fig. 2.4(c), introduces more STI stress [16]. The peak stress level of STI can be modeled as Eq. (2.8),

$$\sigma_{P\_STI} = \left(1 + \frac{m}{LOD}\right) \cdot \frac{W_{STI}}{A_{STI} + W_{STI}} \cdot \sigma_{m\_STI} \tag{2.8}$$

where LOD is the length of oxide definition, W_{STI} denotes the STI width and σ_{m_STI} represent the STI saturation stress. A_{STI} is a fitting parameter to capture the



(a)



(b)

Figure. 2.5. Modeling of the layout dependence on (a) channel length (b) source/drain diffusion length.

dependence on STI width. Moreover, the bottom stress level of STI is related to σ_{P_STI}, and can be modeled as Eq. (2.9),

$$\sigma_{B\_STI} = \frac{C_{STI}}{C_{STI} + LOD} \cdot \sigma_{P\_STI} \tag{2.9}$$

where C_{STI} is a fitting parameter accounting for the LOD dependence. Note that the STI stress in the channel, which directly impacts the device performance, does not distribute as a bathtub curve. Instead, the stress distribution in the channel is part of the whole stress along the oxide definition area.

2.2.2.2. Stresses from width direction

The stress from the width direction also impacts the device performance. For example, STI stress from the width direction leads to hole mobility degradation. Similar to the stress in the lateral direction, the stress profile in width direction follows the bathtub curve as well, and the dependence on layout parameters can be modeled through the peak and bottom stress levels, as shown in Eqs. (2.10)-(2.11):

$$\sigma_{PW} = \left(1 + \frac{m}{W}\right) \cdot \frac{W_W}{A_W + W_W} \cdot \sigma_{mW} \tag{2.10}$$

where W denotes the channel width. A_W is a fitting parameter to capture width dependence, and σ_{mW} is the saturation stress. W_w is the width of stress source from vertical direction and determines the stress magnitude at the edge channel width. Moreover, the bottom stress is modeled in Eq. (2.11),

$$\sigma_{BW} = \frac{C_W}{C_W + W} \cdot \sigma_{PW} \tag{2.11}$$

where C_W is a fitting parameter accounting for the width dependence.

2.2.3. Equivalent Low-Field Mobility

In this sub-section, starting from the first principle, an equivalent mobility is derived to account for the non-uniform mobility enhancement in the channel. When the stress is applied, the band structure is altered and further changes the symmetrical ellipsoids of constant energy of silicon. This results in carrier redistribution and reduction of carrier scattering rate. At low applied stress, the mobility enhancement is modeled proportional to the applied stress by the piezoresistance coefficient [26][27]. However, piezoresistance coefficient is insufficient to predict at larger applied stress because the mobility enhancement becomes saturated [28]. Moreover, in scaled devices, due to the strong quantum confinement, the piezoresistance coefficients cannot be used to calculate the channel mobility [28]. On the other hand, the valley-occupancy model based on the strain-induced band splitting is proposed to account for the mobility enhancement [29], and can be expressed in a general form, as Eq. (2.12),

$$\frac{\mu}{\mu_0} = 1 + B \cdot \left[\exp(\frac{\Delta E}{kT}) - 1 \right] \tag{2.12}$$

where B is a function of electron longitudinal and transfer masses in the subvalley. For the hole mobility, B is corrected with the light and heavy hole masses. μ_0 is the unstrained low field mobility. k and T represent Boltzmann constant and the temperature. ΔE denotes the strain-induced energy splitting of conduction band or valence band and can be calculated based on the deformation potential theory [9][30] by Eq. (2.13),

$$\Delta E = P \cdot \sigma \tag{2.13}$$

where P is a function of deformation potential constants [9][30] and depends on the directions of the applied stress as well as the temperature [41]. Therefore, the temperature-dependent behavior can be modeled as Eq. (2.14),

$$P(T) = P_0 \cdot \left(\frac{T}{T_0} \right)^{\alpha} \tag{2.14}$$

where P_0 denotes its value at room temperature (T_0) and α is a fitting parameter.

Since the stress level varies along the channel, the enhancement in mobility is also non-uniform. Based on the principle of current continuity, the non-uniform mobility can be modeled as an equivalent mobility [31], μ_L, by using Eq. (2.15),

$$\frac{1}{\mu_L} = \frac{1}{L} \int_0^L \frac{1}{\mu(x)} dx \tag{2.15}$$

where $\mu(x)$ denotes the mobility distribution along the channel. Furthermore, if the channel stress is symmetrically distributed, only half of the channel is required to determine the impact of layout parameters. Otherwise, a linear superposition will be used to sum the contribution from both sides. Based on the linear piecewise models of Eqs.(2.1)-(2.5) and valley-occupancy mobility model in Eq. (2.12), the mobility enhancement factor is expressed as Eq. (2.16),

$$\frac{1}{M_L} = \frac{\mu_0}{\mu_L} = \frac{2kT}{dPL(B-1)} \cdot \left\{ \frac{-dPx_0}{kT} + \ln \left[\frac{1 + B\left(\exp\left(\frac{P\sigma_P}{kT} \right) - 1 \right)}{1 + B\left(\exp\left(\frac{P\sigma_B}{kT} \right) - 1 \right)} \right] \right\}$$

$$+ \frac{L - 2x_0}{L \cdot \left[1 - B + B \exp\left(\frac{P\sigma_B}{kT} \right) \right]} \tag{2.16}$$

where the first term denotes the contribution of the stress distributing from the channel edge to the bottom, which is modeled in Eq. (2.1). σ_P denotes the stress at the source end while σ_B represents the bottom stress in the center of channel. The second term contributes from the bottom stress as modeled in Eq. (2.2). On the other hand, for STI stress shown in Fig. 2.4, if the transistor is located in the region of Y_1, only the first term in Eq. (2.16), with the term $-dPx_0/kT$ replaced by $-dPL/kT$, is responsible for the mobility enhancement. For example, the stress effect on the leftmost finger shown in Fig. 2.4 only takes the first term in Eq. (2.16) into account. Moreover, if the transistor is placed within the region of Y_2, only the second term in Eq. (2.16) is required for stress calculation. Furthermore, if the transistor is located in the transition region between Y_1 and Y_2, then both terms in Eq. (2.16) are required. In this final equation for mobility enhancement, the physical dependence on L and temperature (T) is well preserved during the derivation, ensuring the model scalability. Moreover, the impact of stress variation due to different layout parameters such as L and L_{sd} is modeled through σ_P and σ_B in Eqs. (2.6)-(2.9).

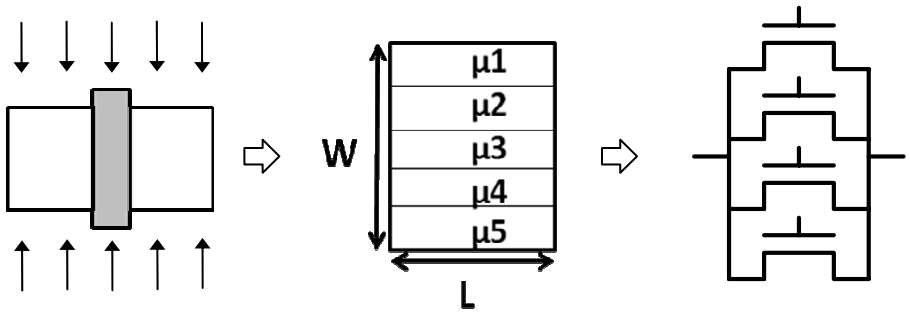


Figure 2.6. Modeling of width dependence on mobility.

On the other hand, the stress from the width direction also impacts the device performance. To capture the non-uniform mobility distribution along the channel width, the transistor is equivalently regarded as several sub-transistors with small slice width in parallel as shown in Fig. 2.6. Therefore, the equivalent mobility considering the stress from width direction is modeled as Eq. (2.17):

$$\mu_W = \frac{1}{W}\int_0^W \mu(y)dy \tag{2.17}$$

where $\mu(y)$ denotes the mobility distribution in vertical direction and W is the channel width. By integrating the proposed piecewise stress models and the valley-occupancy mobility model, the equivalent mobility enhancement factor is expressed as Eq. (2.18):

$$M_W = \frac{\mu_w}{\mu_0} = \frac{2BkT}{WPd}\cdot\left[\exp(\frac{P\sigma_{PW}}{kT}) - \exp(\frac{P\sigma_{PW} - Pdy_0}{kT})\right]$$

$$+\frac{W-2By_0}{W} + (\frac{BW-2By_0}{W})\left[\exp(\frac{P\sigma_{BW}}{kT}) - 1\right] \tag{2.18}$$

where σ_{PW} and σ_{BW} are the peak and bottom stress levels in the width direction and modeled by Eqs. (2.10)-(2.11). y_0 and d are the intercept point and the edge

stress decreasing rate in the vertical direction, with a similar expression to that in lateral direction. By considering the stress effect from lateral and vertical directions, the total mobility enhancement factor is expressed as Eq. (2.19).

$$M_{total} = 1 + B\left[\left(1 + \frac{M_L - 1}{B}\right) \cdot \left(1 + \frac{M_W - 1}{B}\right) - 1\right] \tag{2.19}$$

where M_L and M_W are the mobility enhancement factors for lateral and width directions, respectively.

2.2.4. Velocity in the Saturation Region

The effect of stress-induced mobility enhancement with layout dependence is modeled in section 2.2.3. However, carrier mobility is mainly responsible for the linear operation region, but not adequate to explain the carrier behavior of high electrical field in the saturation region. Instead, velocity is usually used to describe the high E-field behavior. Equation (2.20) is a simplified solution of the energy balance equation [33], which accounts for the velocity overshoot behavior in a short channel device. This simplified solution considers how mobility influences the high E-field behavior:

$$V_{sat} = V_{sat\,0} + 0.13\,\mu_{eff}\sqrt{\tau\mu_{eff}\,kT/q} \cdot \left(V_d / L^2\right) \tag{2.20}$$

where V_{sat0} is the saturation velocity in a long channel device. μ_{eff} is the effective mobility and is a linear function of low-field mobility [20]. τ is the relaxation time and L is the channel length. By Eq. (2.20), the layout dependence on velocity can be modeled through the mobility variation.

2.2.5. Strain Induced Threshold Voltage Shift

In addition to strain-induced mobility variation, strain-induced threshold voltage shift (ΔVth) is also observed in the strained devices. The change in threshold voltage is attributed to strain-induced variation of energy bandgap, electron affinity, and density of states (DOS) [34]. Note that the effect of density of states is ignored due to its insignificant impact. As the band diagram shown in Fig. 2.7, the strain-induced flat-band voltage shift is expressed as Eq. (2.21)

$$\Delta V_{FB} = \Delta E_G - \Delta E_C \tag{2.21}$$

where ΔE_C and ΔE_G denote the strain-induced change in conduction band and bandgap. Based on the deformation potential theory [9][30], the strain-induced change in flat-band voltage is proportional to the applied stress magnitude. Moreover, in principle, applying stress affects the intrinsic carrier density (n_i),

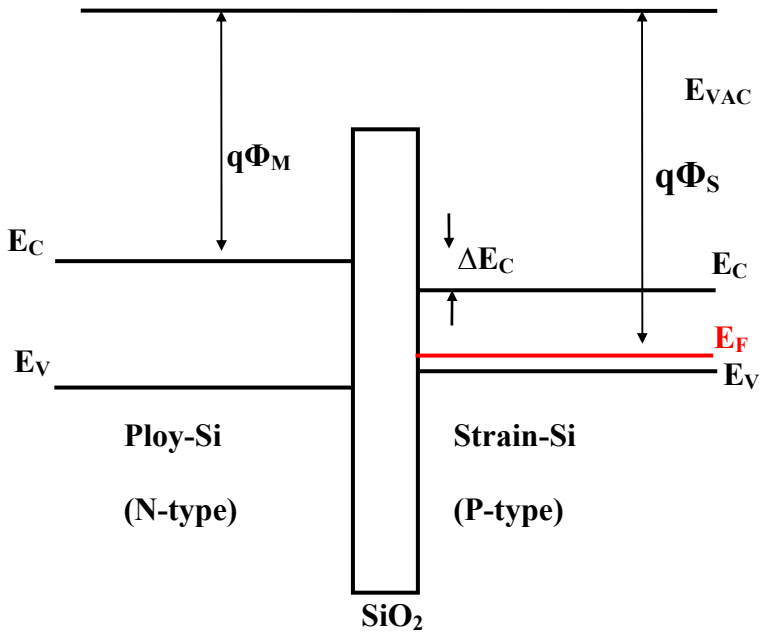


Figure 2.7. Band diagram of strained-Si NMOSFET

which is an exponential function of bandgap, further changing the barrier between source and channel area (qV_{bi}) and the bulk potential (Φ_B) [35]. The expression for V_{bi} and Φ_B are in Eqs. (2.22)-(2.23).

$$V_{bi} = \frac{kT}{q} \ln\left(\frac{N_S N_{ch}}{n_i^2}\right) \tag{2.22}$$

$$\phi_B = \frac{kT}{q} \ln\left(\frac{N_{ch}}{n_i}\right) \tag{2.23}$$

where N_S and N_{ch} denote the doping concentrations in source and channel regions. Note that the flat-band voltage and bulk potential are the major components of threshold voltage. Therefore, to the first order, the strain-induced threshold voltage shift equals the summation of the changes from flat-band voltage and bulk potential. The flat-band voltage shift and the bulk potential change are modeled linearly proportional to the applied mechanical stress by Eq.(2.24),

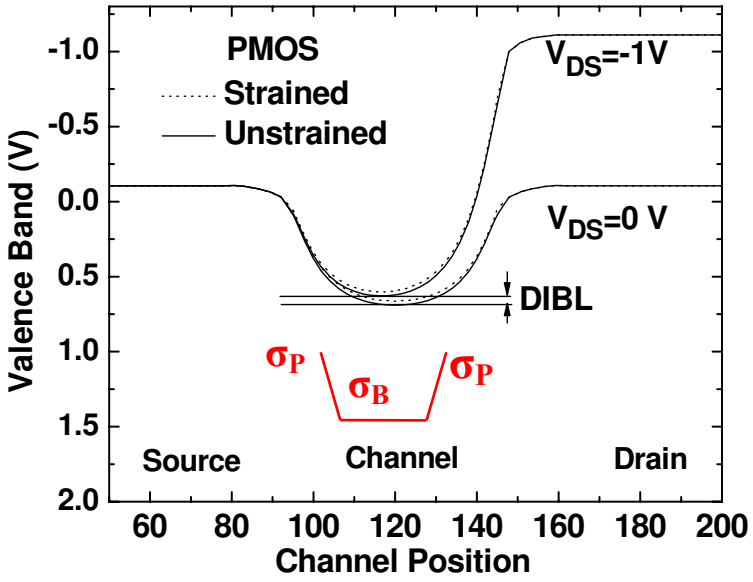


Figure 2.8. Simulated band diagram along the channel

$$\Delta V_{th} \approx VTH\_STR \cdot \sigma_B \tag{2.24}$$

where *VTH_STR* is a fitting parameter to capture the linear relationship between threshold voltage shift and the applied stress magnitude. Note that the bottom stress level is responsible for threshold voltage shift because the barrier peak between source and substrate (qV_{bi}) is controlled by σ_B, as shown the band diagram of PMOS in Fig. 2.8. The simulated valence bands of unstrained/strained PMOS indicate that energy barrier between channel and source is lowered and thus it becomes easier for holes to pass through the channel. This lowering valence band confirms the strain-induced threshold shift.

2.2.6. Other Secondary Effects

Drain induced barrier lowering (DIBL) effect is becoming pronounced with scaling the device feature size, and to the first order can be modeled as Eq. (2.25) [36],

$$\Delta V_{th}(DIBL) \approx \left[3(V_{bi} - 2\phi_B) + V_{ds}\right] e^{-L/l} \tag{2.25}$$

where V_{bi} and Φ_B have the same meaning as discussed in the previous section. V_{ds} is the applied drain bias. L denotes the channel length and *l* is the DIBL characteristic length as expressed in Eq. (2.26),

$$l = \sqrt{3T_{ox}X_{dep}} \tag{2.26}$$

where Tox denotes the oxide thickness and Xdep is the depletion depth with expression in Eq. (2.27),

$$X_{dep} = \sqrt{\frac{2\varepsilon_{si}(2\phi_B - V_{BS})}{qN_{ch}}} \tag{2.27}$$

V_{bi} and ϕ_B are major factors to DIBL effect. Although applying stress changes V_{bi} and ϕ_B, the impact of stress on DIBL effect is canceled. The term of (V_{bi} - $2\phi_B$) in Eq. (2.25) is independent of n_i, i.e., the applied stress. On the other hand, Φ_B is changed and influences the depletion depth (X_{dep}), further affecting DIBL characteristic length. Therefore, DIBL effect is modulated through Eq. (2.25). Note that DIBL characteristic length (l) is the major factor to account for stress effect and is proportional to $Eg^{1/4}$, making DIBL effect relatively insensitive to the applied mechanical stress. Due to the change of depletion depth caused by the applied stress, the capacitance between substrate and channel (C_{dep}) is subject to be changed by Eq. (2.28).

$$C_{dep} = \frac{\varepsilon_{si}}{X_{dep}} \tag{2.28}$$

Therefore, the the impact of stress on subthreshold swing is modified through Eq. (2.29).

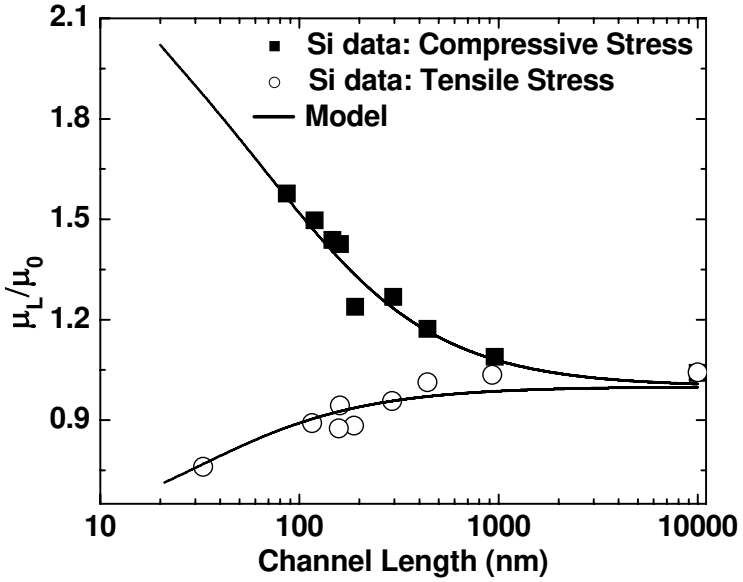
$$S.S. = 2.3\frac{kT}{q}(1 + \frac{C_{dep}}{C_{ox}}) \tag{2.29}$$

where C_{ox} represents the oxide capacitance. k and T represent Boltzmann constant and the temperature. Moreover, the threshold voltage shift induced by body bias is also modified and is modeled through the change of bulk potential (ϕ_B) in Eq. (2.30).

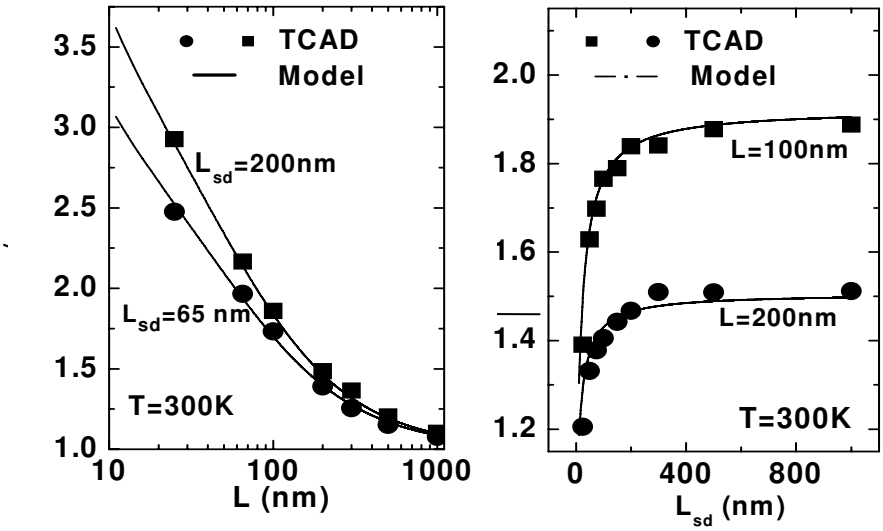
$$\Delta V_{th} = \frac{\sqrt{2\varepsilon_{si} q N_{ch} \cdot (2\phi_B - V_{bs})}}{C_{ox}} \tag{2.30}$$

2.3 MODEL VALIDATION

In this section, the newly developed models are validated by TCAD simulations as well as Si data. To capture the stress effect, the process simulator [15] is performed to include process induced mechanical stresses, which are further loaded into device simulator [24] to evaluate the impacts of stresses on



(a)



(b) (c)

Figure. 2.9. (a) DSL induced hole mobility variation with dependence on L; (b) eSiGe induced hole mobility changes with dependence on L and (c) L_{sd}.

device performance. During the simulation, the mobility is extracted in the linear region with gate bias at -1 volt and drain voltage at -0.1volt for PMOS.

In Fig. 2.9 (a), the experimental data shows DSL induced mobility enhancement has a strong dependence on channel length [40]. By applying compressive stress liner, the enhancement of hole mobility increases with scaling channel length. On the other hand, the tensile stress liner degrades hole mobility. Moreover, in Fig. 2.9 (b), TCAD simulation regarding eSiGe technology shows similar layout dependence to DSL technique. As channel length becomes shorter, the enhancement grows up, i.e., strain technology is more effective. On the other hand, in Fig. 2.9 (c) as L_{sd} becomes smaller, the enhancement becomes lower because of lower stressor material in the source and drain regions.

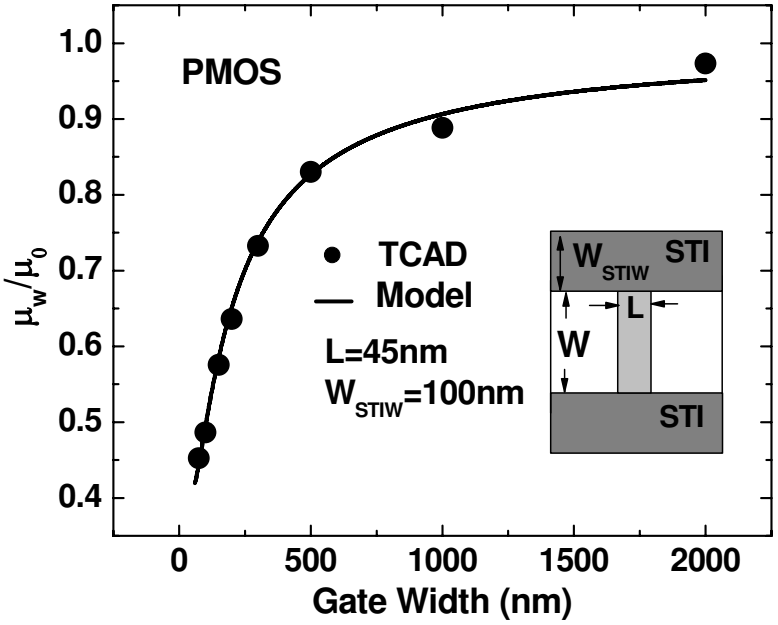


Figure. 2.10. STI stress induced hole mobility variation with dependence on gate width.

Figure 2.10 shows the width dependence on mobility. Given a constant vertical stress source, higher stress magnitude is observed in the device with smaller channel width. Higher compressive stress applied in the width direction leads to more hole mobility degradation. Excellent agreement between the new model and Si data and TCAD simulation are achieved.

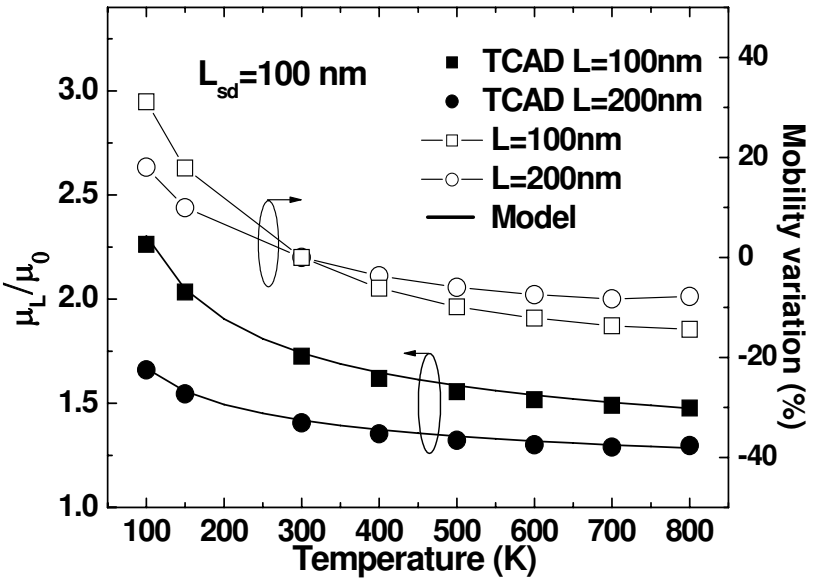


Figure. 2.11. Equivalent mobility with dependence on temperature

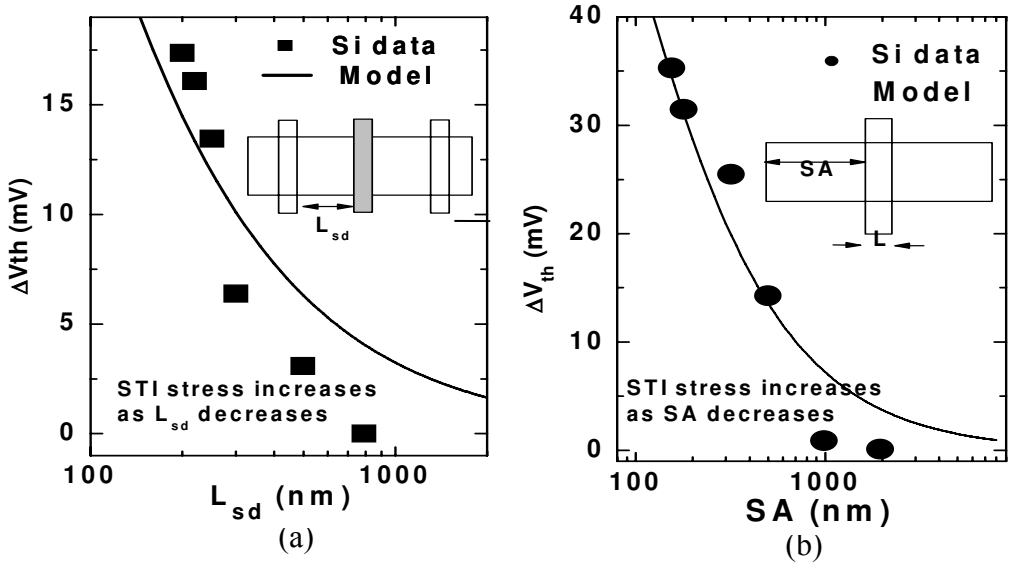


Figure. 2.12. STI induced V_{th} shift with dependence on (a) L_{sd} and (b) SA.

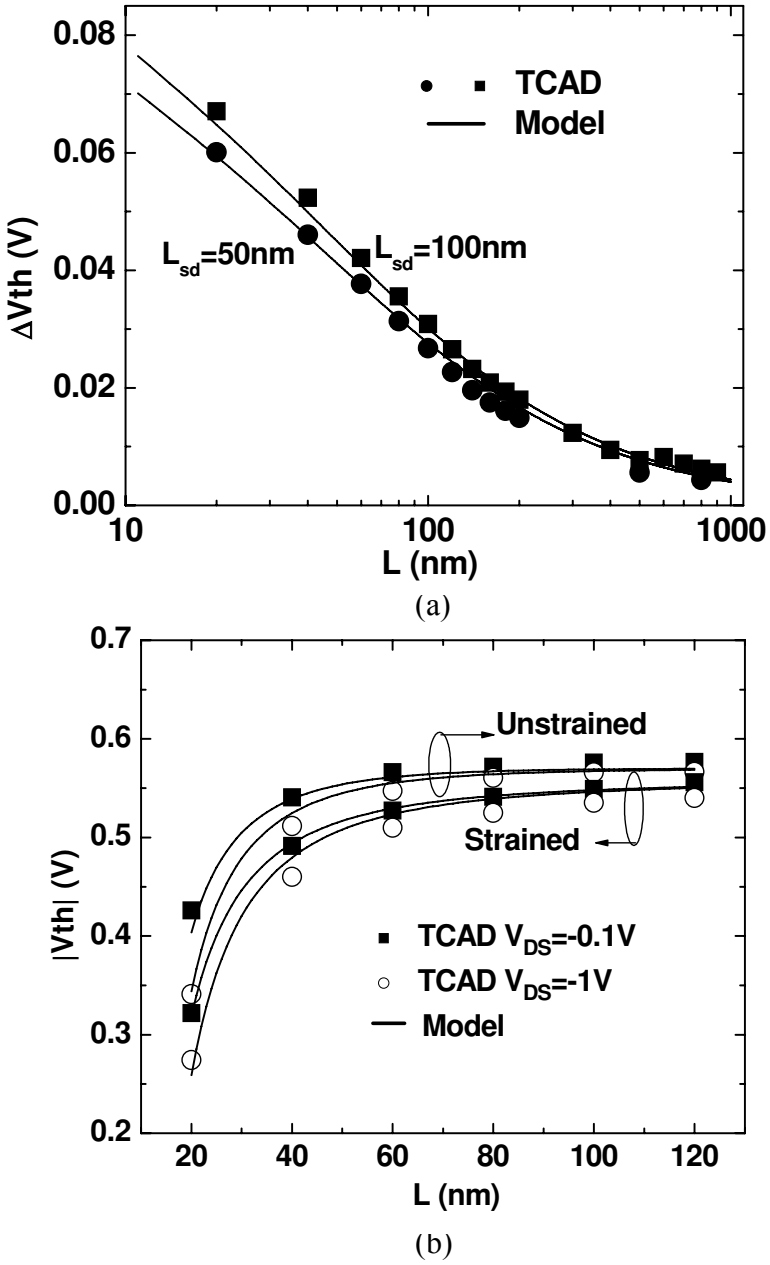


Figure. 2.13. eSiGe induced V_{th} shift with dependence on (a) L and (b) DIBL effect.

In addition to layout parameters, the stress effect depends on the temperature (Eqs. (2.12) and (2.14)). Figure 2.11 demonstrates the temperature dependence on strain enhanced mobility. To assess this effect, the device is operated at different temperatures from 100K to 800K. TCAD simulation shows

that at 100K, the mobility enhancement increases 31% more than that at room temperature for the device with both L and L_{sd} at 100nm. On the contrary, as temperature increases more than the room temperature, the mobility enhancement declines. The sensitivity of mobility enhancement to temperature is higher at lower temperatures. This behavior can be explained by Eq. (2.12), where the temperature term is in the exponential function, so that the change is more dramatic under low temperatures. In addition, simulation results illustrate that the device with longer channel length is less sensitive to the temperature variation.

In Fig. 2.12, the experimental data regarding STI stress show the strain-induced V_{th} shift is dependent on the layout parameters, L_{sd} and SA [39]. As L_{sd} and SA increase, the stress in the channel area decreases and results in lower V_{th} shift. In Fig. 2.13 (a), TCAD simulation based on eSiGe technology shows that the strain-induced V_{th} shift also correlates to the L and L_{sd}; the strain-induced ΔV_{th} increases as L decreases, and lower ΔV_{th} is observed for a smaller L_{sd} due to lower S/D stressors. Figure 2.13 (b) shows V_{th} decrease as L decreases due to the DIBL effect for strained and unstrained devices. Simulation results show the DIBL effect is very insensitive to the applied stresses. All these properties are

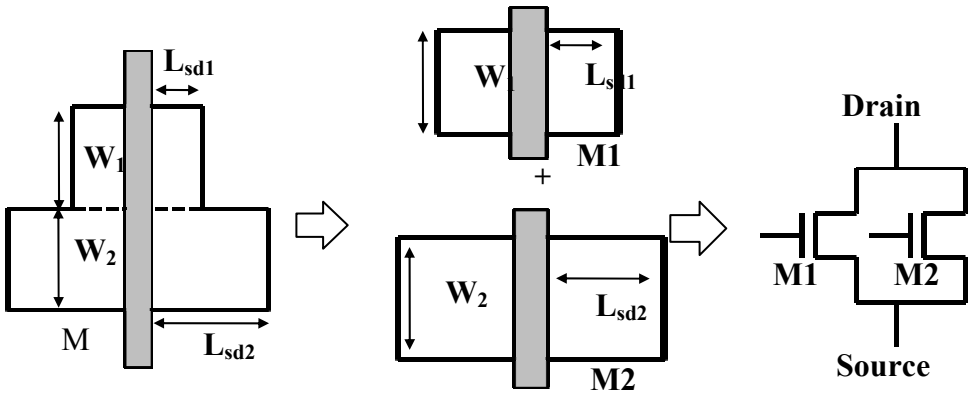


Figure 2.14. The decomposition along the width direction.

well modeled in the proposed analytical solutions.

2.4 DESIGN APPLICATION

Although applying mechanical stress improves the device performance, it introduces extra variation among transistors because of their different size and layout parameters. In this section, a decomposition method is proposed to partition a large layout into several sections, in order to simplify the extraction procedure without losing the accuracy. This study uses a 45nm technology that is customized from PTM [37]. Gate length (L) and source/drain diffusion length (L_{sd}) are assumed at 40nm and 130nm, respectively, following regular layout rules [38]. TCAD stress simulations are conducted on layout samples to determine the distribution in both vertical and lateral directions, and to evaluate the decomposition methods. For the process simulation, a typical eSiGe stress technology is used with 25% of Ge composition. After the process simulation, the device simulation is conducted to investigate how the stress process impacts the electrical characteristics of a transistor. The maximum on-current (I_{on}) and linear current (I_{lin}) are sampled at the gate bias at -1V and drain biases at -1 and -0.1 V, respectively, from a PMOS device.

2.4.1. Vertical Layout Decomposition

The first step is to decouple the layout along the vertical direction, as shown in Fig. 2.14. By separating a transistor with an irregular diffusion region in the width direction, it can be regarded as two transistors with different width in parallel. An equivalent circuit representation is shown in Fig. 2.14. All sub-transistors now have a regular diffusion region that defines the stress, i.e., L_{sd1} and

L_{sd2}, and each individual section can be described by the same stress model, as in Section 2.2.

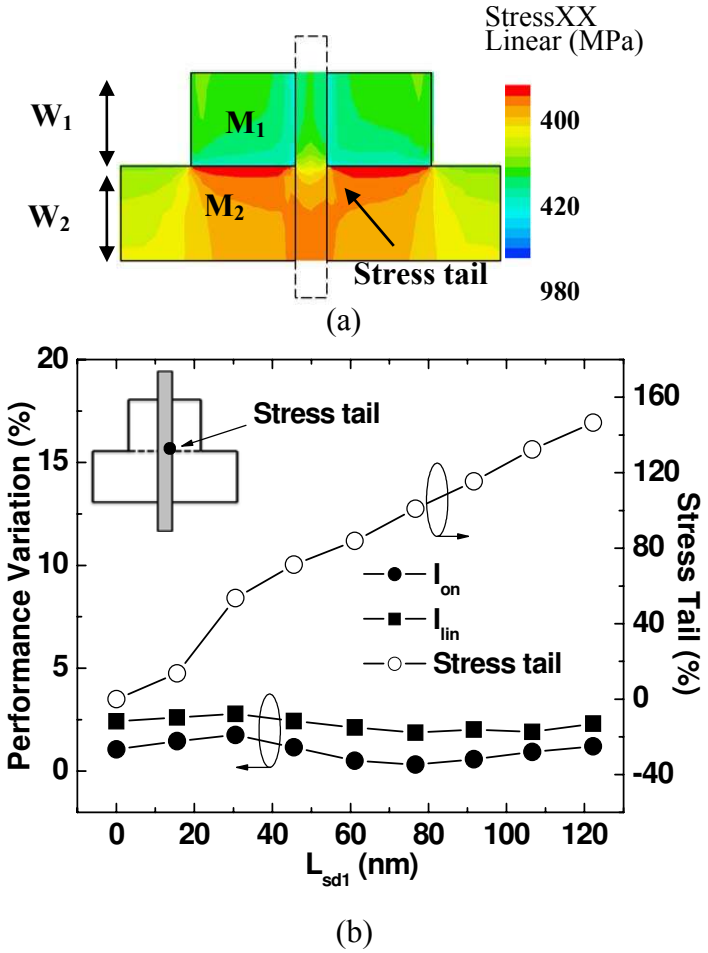


Figure. 2.15. (a) Top view of the stress distribution when stressor is only applied to S/D area of M1, (b) the stress tail only has a marginal impact on currents. (W_1=100nm, W_2=200nm)

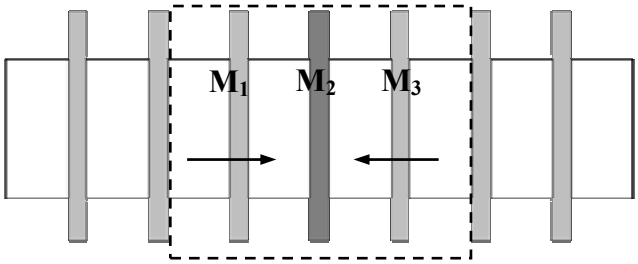
Such decomposition is feasible only if the performance enhancement due to the stress effect is localized and does not impact the neighbors along the width direction. For the distribution of the mechanical stress, this may not be true: Fig. 2.15 (a) illustrates the stress distribution induced by adding SiGe only in M_1; yet it shows a stress tail penetrating into M_2.

This effect becomes more significant when the diffusion size of M_1, L_{sd1}, increases, but degrades fast along width direction, as shown in Fig. 2.15(b). Fortunately, this tail only has a marginal influence on the IV characteristics of M_2, which are determined by the average stress level throughout the entire channel. As shown in Fig. 2.15 (b), the change of I_{on} and I_{lin} of M_2 is only 1-2% due to this tail, even when L_{sd1} is comparable to L_{sd2}. It is safe to neglect the influence of the stress tail from the vertical neighbors when evaluating the device performance and thus, this decomposition method is valid.

2.4.2. Lateral Layout Decomposition

To determine the stress influence from horizontal neighboring transistors, a seven-finger pattern, as shown in Fig. 2.16 (a), is used as the basic test layout. The middle transistor (M2) is served as the target transistor for model extraction. TCAD simulation in Fig. 2.16 (b) shows that more than 20% derivation of the peak stress is observed when no neighboring transistors are considered, leading to a large performance error. The error in IV prediction is reduced when more neighboring stresses are included. This indicates that different from the vertical decomposition, it is not accurate to only consider each individual gate in the lateral direction. Figure 16 (b) further illustrates that the error in I_{on} and I_{lin} decreases to 1-2% if the nearest two neighboring transistors are included into the stress calculation. Based on this observation, it is concluded that a pattern with three fingers is an appropriate set for stress modeling and extraction (i.e., M1 M2 and M3 in Fig. 2.16 (a)). This is the basic pattern for the lateral decomposition layout pattern.

From the above discussions, the irregular layout can be decomposed into several basic layout components in order to reduce the complexity in model



(a)

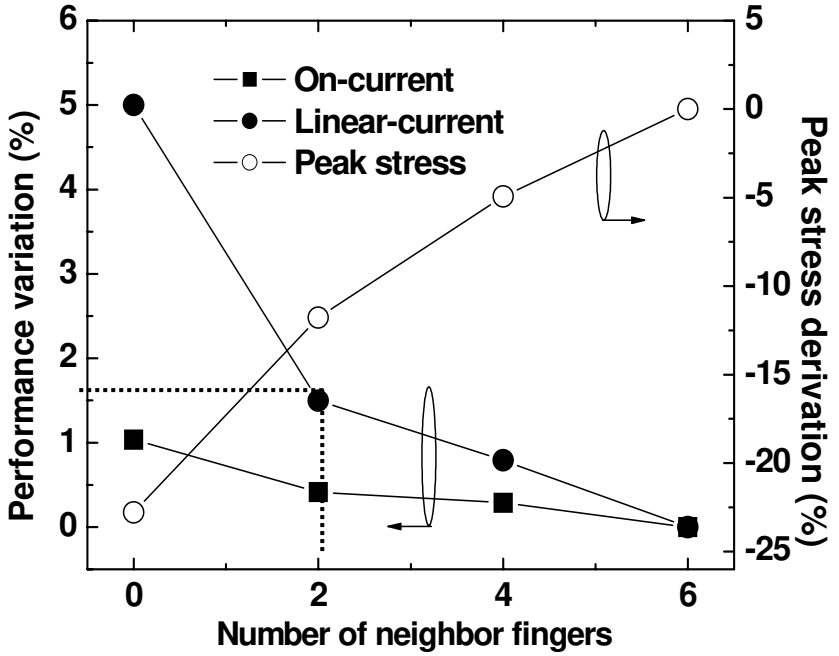


Fig. 2.16. Method for lateral decomposition: (a) The basic pattern for lateral decomposition (b) Performance variation with different numbers of neighboring gates

development. They are general enough to cover all layouts under restrictive design rules. After the decomposition, a general and physical model is applied to model the performance enhancement in each sub-transistor, as presented in the following section. These sub-transistors will be combined for further circuit simulation and analysis.

2.4.3. Circuit Performance Benchmark

A physical layout-dependent stress model is developed and is comprehensively verified by the Si data and TCAD simulation. In this section, we combine the decomposition approach together with the stress model, and demonstrate its capability in circuit simulation as well as performance analysis. As the first step, the stress model is calibrated with IV characteristics from TCAD simulations, in order to attain the accurate device behavior under irregular stress distribution. The enhancement in mobility and velocity as well as the V_{th} shift are

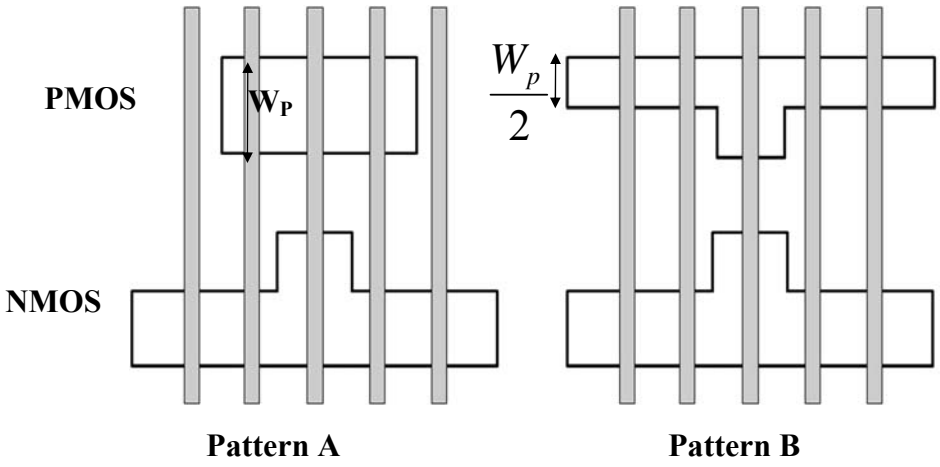


Figure. 2.17. Layout patterns of two NAND3 gates with the same size, with SiGe in source/drain only.

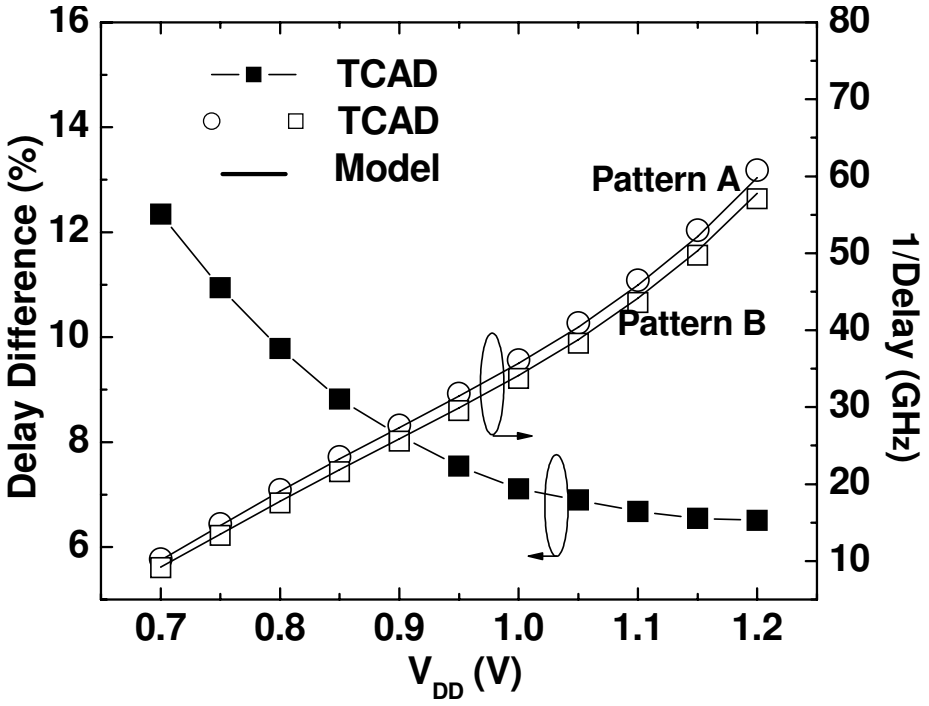


Figure 2.18. The systematic delay variation due to layout dependent stress effect.

accordingly embedded into standard device models. Given a circuit layout, it is decomposed in both directions; each sub-transistor is then translated by the stress model for further circuit simulation, depending on their size and layout parameters.

To benchmark the layout effect on circuit performance, two different layout patterns of NAND3 are generated as shown in Fig. 2.17. They are decomposed and modeled by the procedure described above. Figure 2.18 shows SPICE simulation results of high-to-low switching. The delay in Pattern B is longer than that in Pattern A because the pull-up network of Pattern B is stronger, benefiting from the larger and closer diffusion area, therefore leading to stronger

stress from neighboring gates. The performance difference increases at lower V_{DD} as the linear region is more sensitive to the stress effect. The decomposition method together with the layout dependent stress model demonstrates an excellent accuracy as compared to TCAD simulations.

2.5 Chapter Summary

With the scaling of device dimension, the strain-induced variation becomes more pronounced. Therefore, it is essential to develop compact models of the stress effect for circuit analysis and optimization. In this work, we propose the solution that significantly reduces the complexity in modeling and simulation: (i) compact models of stress-induced mobility, velocity, and threshold voltage, which are general enough to model the scalability to layout, temperature and other device parameters for different strain technologies (ii) the decomposition method to partition a large layout into simple patterns for easy model extraction and localize the modeling effort. As demonstrated in circuit benchmark, these solutions can be easily integrated into physical design tools to predict circuit performance under layout-dependent stress effect.

CHAPTER 3

MODELING OF FE-FET AND IMPLICATIONS ON VARIATION-INSENSITIVE DESIGN

Semiconductor devices with self-feedback mechanisms are considered as a promising alternative to traditional CMOS, in order to achieve faster operation and lower switching energy. Examples include IMOS and FBFET that are operated in a non-equilibrium condition to rapidly generate mobile carriers. More recently, Fe-FET was proposed to improve the switching by integrating a ferroelectric material as gate insulator in a MOSFET structure. Under particular circumstance, ferroelectric capacitance is effectively negative, due to the negative slope of its polarization-electrical field (P-E) curve. This property makes the ferroelectric layer a voltage amplifier to boost surface potential, achieving fast transition. In this chapter, a new threshold voltage model is developed to capture the feedback of negative capacitance and IV characteristics of Fe-FET. It is further revealed that the impact of random dopant fluctuation (RDF) on leakage variability can be significantly suppressed in Fe-FET, by tuning the thickness of the ferroelectric layer.

3.1. INTRODUCTION

It is of significant importance to have a device with a steep subthreshold slope to achieve low supply voltage and power dissipation. However, in a traditional MOSFET structure, the subthreshold swing (SS) is defined as Eq. (3.1):

$$SS = \frac{\partial V_g}{\partial(\log I_d)} = 2.3\frac{k_B T}{q}\cdot\left(1+\frac{C_S}{C_{ox}}\right) \tag{3.1}$$

where k_B is Boltzmann constant, T is the absolute temperature, and q is the unit charge. C_S and C_{ox} denote the depletion capacitance and oxide capacitance, respectively. Eq. (3.1) indicates that a fundamental limit of 2.3 k_BT/q of gate voltage change is required to modulate the subthreshold current by 10 times, which is also known as Boltzmann tyranny [43]; 60mV is needed at room temperature.

New device structures such as Impact Ionization MOS (IMOS) [44] and Feedback FET (FBFET) [45] are proposed to circumvent this limit. IMOS, a partially gated reverse biased p-i-n diode, is operated in the region of avalanche breakdown by modulating the electrical field to generate a vast amount of electron-hole pairs (EHPs), contributing to the fast subthreshold transition. On the other hand, in FBFET the charge is deliberately implemented and stored in the spacer regions to form the built-in potential barriers, creating an internal feedback loop to achieve the fast subthreshold switch.

Different from IMOS and FBFET, Fe-FET is proposed by integrating

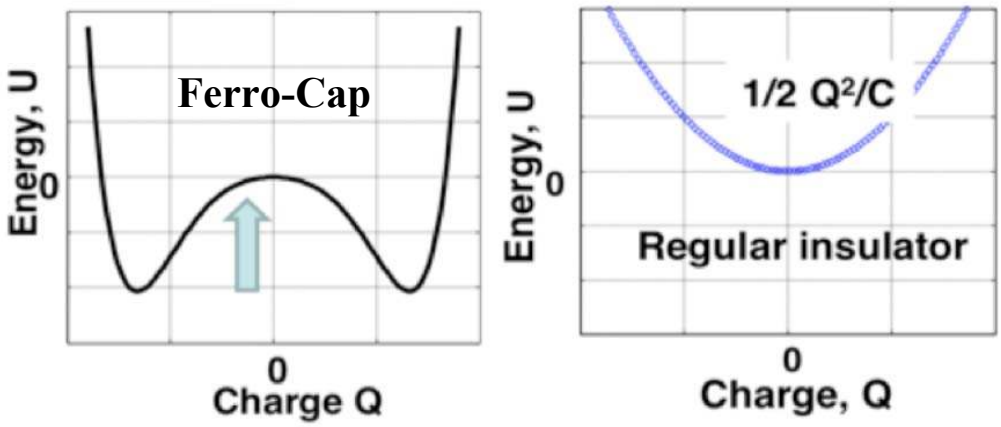


Figure 3.1. The energy landscape of a ferroelectric and regular capacitor. [46].

ferroelectric material in a traditional MOSFET structure to achieve the fast transition in subthreshold region [46]. Under particular circumstance, ferroelectric capacitance is effectively negative, due to the negative slope of its polarization-electrical field (P-E) curve [47]. Figure 3.1 shows the energy landscapes for a ferroelectric and regular capacitor, respectively. The curvature of energy landscape is negative in the ferroelectric insulator while a regular capacitor presents a positive one. The capacitance of the ferroelectric is negative. This property makes the ferroelectric layer an internal voltage amplifier to boost surface potential, leading to fast channel formation. The concept of negative capacitance as a voltage amplifier can be understood by two capacitances in series, shown in Fig. 3.2, where the charge stored in both capacitances are the same. When C_1 and C_2 are regular capacitances and the voltage drops in C_1 and C_2 follow the ratio of C_2/C_1 as expressed in Eqs. (3.2) and (3.3). This scheme is usually served as a voltage divider in the circuit.

$$V_1 - V_2 = \frac{C_2}{C_1 + C_2} \cdot V_1 \tag{3.2}$$

$$V_2 = \frac{C_1}{C_1 + C_2} \cdot V_1 \tag{3.3}$$

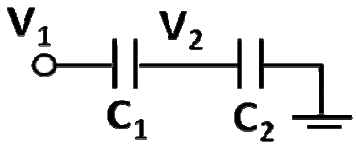


Figure 3.2. Series capacitance as a voltage divider.

V_2 is determined by the capacitances and the applied voltage, V_1. As V_1 increases, V_2 also increases because the overall charge is increased. The overall charge is expressed in Eq. (3.4).

$$Q_{total} = \frac{C_1 C_2}{C_1 + C_2} \cdot V_1 \tag{3.4}$$

Moreover, since C_1 and C_2 are in series, the charges stored in C_1 and C_2 are the same and can be expressed in Eqs. (3.5) and (3.6).

$$Q_1 = C_1 \cdot (V_1 - V_2) \tag{3.5}$$

$$Q_2 = C_2 \cdot V_2 \tag{3.6}$$

Now, let us consider the scheme with positive C_1 and negative C_2. As V_1 increases, the overall charge increases as long as the equivalent capacitance (C_1 and C_2 in series) remains positive and Q_1 equals Q_2. To maintain the charge conservation, V_2 is required to stay negative to ensure that Q_2 equal to Q_1. Note that the voltage drop in C_1, (V_1-V_2), becomes larger than the originally applied voltage, V_1, owing to the negative V_2. Therefore, more voltage is allocated to regular capacitance when in series with negative capacitance and the system remains stable. This behavior is explicitly different from scheme of regular capacitance in series and can be served as an internal voltage amplifier in a MOSFET device to boost the surface potential, implying the channel can be formed quickly. On the other hand, back to Eq. (3.1), if the C_{ox} is replaced by the negative ferroelectric capacitance, it is possible to break down the limit of Boltzmann tyranny and achieve the subthreshold swing lower than 60 mV/dec.

To investigate the device performance, traditional efforts resort to TCAD simulation to analyze the dependence of geometry and process on electrical characteristics. This approach usually requires expensive computation, especially when chip complexity per unit area keeps increasing along with technology scaling. To enable the early-stage exploration of Fe-FET on circuit design and its impact on performance variability, it is necessary to develop an effective model that is able to physically capture the steep subthreshold slope and embed it into standard model parameters for circuit simulation.

On the other hand, the aggressive scaling of CMOS technology inevitably leads to a drastic challenge in process variations, such as channel length variation [48], channel dopant fluctuations [49], and other layout-dependent proximity effects [50]. Among these variations, random dopant fluctuation (RDF) represents the intrinsic variation source in CMOS structure, posed by fundamental physical and manufacturing limits. RDF has become one of the most major barriers in the progress of large-scale integration scaling. In this work, a new threshold voltage model of Fe-FET is proposed to capture the steep subthrehsold behavior and is used to study the performance variation. This model is derived from the first principle and physically captures the subthreshold behavior of Fe-FET such that model scalability is guaranteed for future technology generations [78].

The important contributions of this chapter include:

i. A new threshold voltage model is developed to capture the feedback of negative capacitance and IV characteristics of Fe-FET;

ii. It is further revealed that the impact of random dopant fluctuation (RDF) on leakage variability can be significantly suppressed in Fe-FET, by tuning the thickness of the ferroelectric layer.

The chapter is organized as following. Section 3.2 presents the basics of Fe-FET and compact V_{th} model derivation to describe the subthreshold behavior. In section 3.3, the performance variability of Fe-FET is investigated by using the new threshold voltage model and TCAD simulation. Section IV concludes this chapter.

3.2. COMPACT MODELING OF FE-FET

Fe-FET is proposed to speed up the transition by stacking a ferroelectric material on top of silicon dioxide as an internal voltage amplifier to boost surface potential in a MOSFET structure, as shown in the schematic in Fig. 3.3 (a). In effect, this ferroelectric layer amplifies the gate voltage internally because of a negative gate capacitance; the inversion layer charge actually increases at lower gate voltages. Fig. 3.3 (b) illustrates the equivalent capacitance model for a Fe-FET device, where the ferroelectric capacitance (C_{FE}) is in series with the oxide capacitance (C_{OX}) and the substrate depletion capacitance (C_S). V_{int} denotes the internal voltage between C_{FE} and C_{OX}. Since Cox and C_{FE} are in series, the amount of charge stored in three capacitances is the same and is expressed in Eq. (3.7)

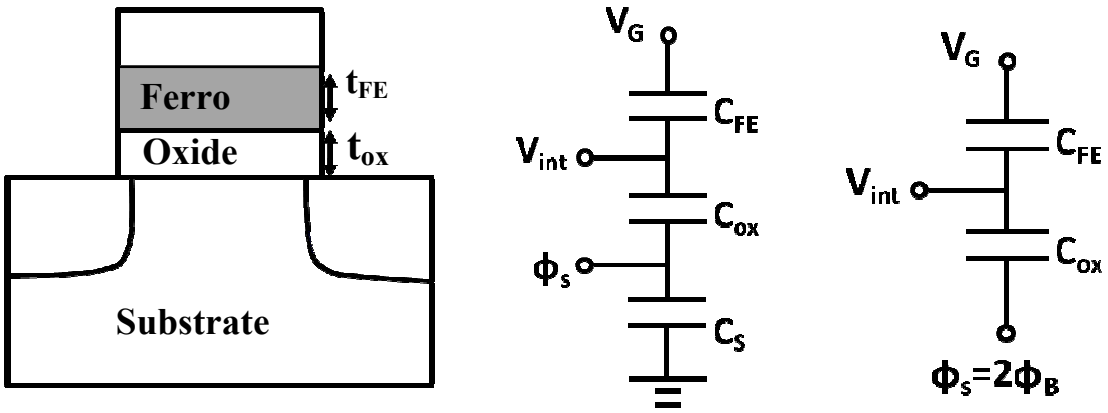


Figure 3.3. (a) Cross-section of Fe-FET. (b) The equivalent capacitance model in the sub-V_{th} region (c) The capacitance model in the super-V_{th} region.

$$Q = C_{eq} \cdot (V_g - \phi_s)$$

$$= C_{ox} \cdot (V_g - V_{\text{int}})$$

$$= C_{FE} \cdot (V_{\text{int}} - \phi_s)$$

$$= C_s \cdot \phi_s \tag{3.7}$$

where C_{eq} is the equivalent capacitance of C_{ox} and C_{FE} in series. If the equivalent capacitance is positive, the charge increases with applying gate voltage. The applied voltage is divided into three components in C_{FE}, C_{ox}, and C_s and the sum of voltage drops on each capacitance is equal to the outer gate voltage. However, since C_{FE} is negative, the voltage drop in ferro-capacitance ($V_{int} - \Phi_s$) stays negative, implying the voltage drops on the other two regular capacitances, C_{ox} and C_s, increase. Therefore, the voltage drop on substrate depletion capacitance, equivalent as surface potential in a MOSFET structure, is increasing, leading to the fast formation of channel.

The negative capacitance of the ferroelectric dielectric originates from the intrinsic feedback mechanism between the induced charge and the voltage drop on the capacitance; this feedback between charge and voltage drop is modeled by Eq. (3.8) [46].

$$V_G - V_{\text{int}} \approx \alpha_0 Q + \beta_0 Q^3 + \gamma_0 Q^5 + \rho_0 \frac{dQ}{dt} \tag{3.8}$$

where Q denotes the charge stored in the capacitance. The internal voltage (V_{int}), which accounts for the change of voltage drop in C_{FE}, is formulated in Eq. (3.9).

$$V_{\text{int}} = \phi_s + \frac{\sqrt{2\varepsilon_{si} q N_{ch} \phi_s}}{C_{ox}} \tag{3.9}$$

where the q is the unit charge and N_{ch} is the channel doping concentration. Note that the second term in Eq. (3.9) accounts for the voltage drop on C_{ox} and only stays valid in the subthreshold region. On the other hand, α_0 and β_0 are negative to account for the negative capacitance. γ_0 is positive to describe the behavior of normal capacitance. ρ_0 is the resistivity for the voltage drop. Those parameters (α_0, β_0, γ_0, and ρ_0) are proportional to the thickness of ferroelectric material (t_{FE}), and are modeled in Eqs. (3.10) - (3.13) [46],

$$\alpha_0 = 2\alpha t_{FE} \tag{3.10}$$

$$\beta_0 = 4\beta t_{FE} \tag{3.11}$$

$$\gamma_0 = 6\gamma t_{FE} \tag{3.12}$$

$$\rho_0 = \rho\ t_{FE} \tag{3.13}$$

where α, β, γ and ρ are material coefficients of the ferroelectric material. When the ferroelectric dielectric layer is integrated into the MOSFET structure, the feedback mechanism is enabled and starts to interact with other elements.

In general, the device performance of traditional MOSFET is modeled by electrostatic behavior and carrier transportation. In a device numerical solver, for example Sentaurus Device [30], both electrostatics and transport properties are self-consistently solved to analyze the device performance. This modeling approach is also known as Technology Computer-Aided Design (TCAD). Expensive computation is required to perform TCAD simulation. On the other hand, compact models such as BSIM [20] and PSP [51] are proposed to efficiently analyze the device performance with closed-form solutions, instead of self-consistent iterative solution. , As illustrated in the dash box in Fig. 3.4, the electrostatic property such as surface potential or threshold voltage is analytically calculated to estimate the carrier transportation for device performance prediction. Compact model is very effective in analyzing a VLSI chip, where more than one billion transistors are integrated.

When the ferroelectric dielectric layer is integrated into the MOSFET system, the feedback mechanism is enabled and interacting with other elements in the traditional MOS system. This can be modeled by inserting an additional feedback path in a traditional CMOS model, as shown in Fig. 3.4. The induced charge in the substrate will impact the electrostatic properties through the feedback loop. By combining Eqs. (3.8)-(3.13) and ignoring the high-order

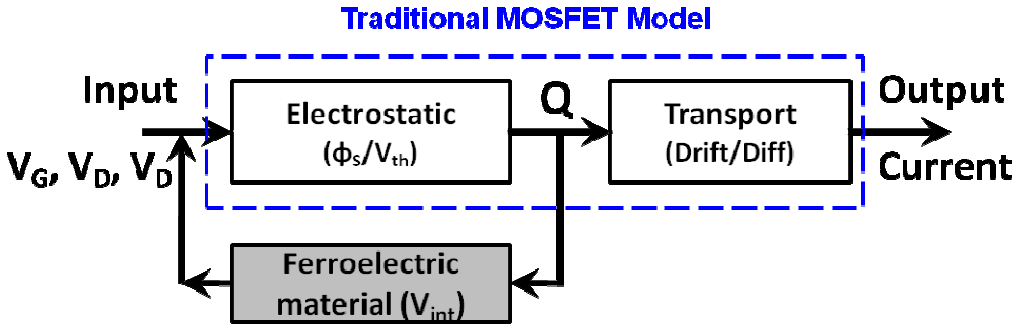


Figure 3.4. The integration of the ferroelectric material into traditional MOSFET model.

feedback effects, the surface potential (Φ_S) in the subthreshold region is derived as Eqs. (3.14) and (3.15) in a steady state condition.

$$\phi_s = \left(\frac{-A + \sqrt{A^2 + 4V_G}}{2} \right)^2 \tag{3.14}$$

$$A = \sqrt{2\varepsilon_{si} q N_{ch}} \cdot \left(\frac{1}{C_{ox}} + \alpha_0 \right) \tag{3.15}$$

Note that in Eq. (3.15) the feedback parameter α_o, a negative value, is introduced to modulate the effect of oxide capacitance. Moreover, to the first order, the threshold voltage (V_{th}) in MOS system is derived as a function of gate voltage and surface potential, as shown in Eq. (3.16) [52].

$$V_{th} = V_G + 2m\phi_B - m\phi_s \tag{3.16}$$

where m is the body effect coefficient and expressed as Eq. (3.17).

$$m = 1 + \frac{\sqrt{\varepsilon_{si} q N_{ch} / 4\phi_B}}{C_{ox}} \tag{3.17}$$

where Φ_B is the bulk potential and expressed as Eq. (3.18).

$$\phi_B = \frac{kT}{q}\ln\left(\frac{N_{ch}}{n_i}\right) \tag{3.18}$$

Eqs. (3.14)-(3.18) describe the closed-loop behavior of threshold voltage in Fe-FET, which is adaptively changed with various gate voltages. In addition to the dependence on channel doping concentration and oxide thickness, the thickness of ferroelectric dielectric is also one of the major factors of threshold voltage in Fe-FET according to Eq. (3.10). To achieve faster subthreshold switching, V_{th} should vary dynamically toward a smaller value when gate voltage is increasing. Therefore, the sum of $(1/C_{ox} + \alpha_0)$ needs to stay below zero, which ensures the negative derivative of V_{th} with respect to V_G (i.e., dV_{th}/dV_G), as shown in Eq. (3.19).

$$\frac{\partial V_{th}}{\partial V_G} = 1 - \left(\frac{-A + \sqrt{A^2 + 4V_G}}{\sqrt{A^2 + 4V_G}}\right) \tag{3.19}$$

The negative value of $(1/C_{ox} + \alpha_0)$ indicates that the series capacitance of C_{FE} and C_{ox} is negative. In this situation, the system could be stable in the subthreshold region as long as the total capacitance in series (Fig. 3.3(b)) is positive. However, when the channel is formed, i.e., the surface potential is fixed at $2\Phi_B$, the inversion layer shields C_s and thus, the device could become unstable if the series capacitance of C_{FE} and C_{OX} remains negative, as shown in Fig. 1(c). This implies Fe-FET could be a good alternative for the subthreshold circuits [53]. The subthreshold circuit is operated in the subthreshold region to achieve high energy efficiency as a priority rather than the traditional emphasis on speed [54][55].

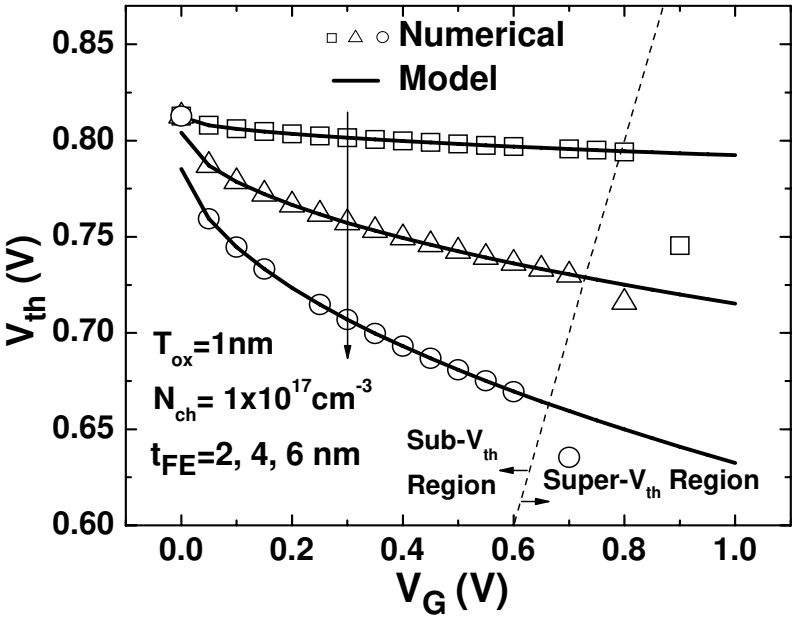


Figure 3.5. The dependence of V_{th} on gate voltage and t_{FE}.

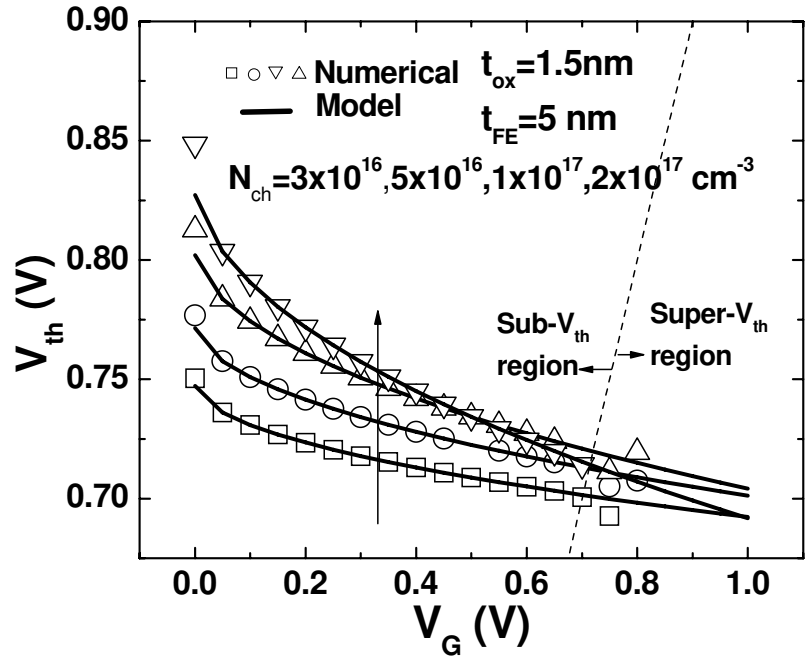


Figure 3.6. The dependence of V_{th} on gate voltage and N_{ch}.

The values of material coefficient are calibrated with experimental data in [56], where α is estimated at the order of -10^{10}m/F. Figure 3.5 shows the threshold

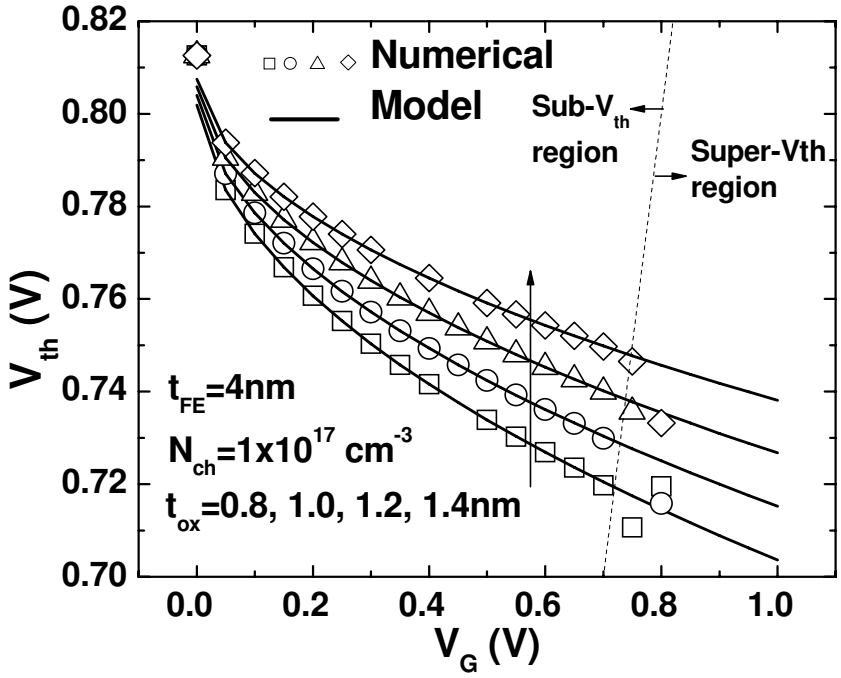


Figure 3.7. The dependence of V_{th} on gate voltage and t_{ox}.

voltage of Fe-FET is dependent on the applied gate voltage and decreases when higher gate voltage is applied. A stronger declining rate of V_{th} is observed with a thicker ferroelectric layer (t_{FE}) introduced. A thicker ferroelectric layer contributes to stronger effect of negative capacitance. Moreover, a thicker ferroelectric material makes the series capacitance of C_{ox} and C_{FE} less negative, leading to the smaller initial threshold voltage. Note that the V_{th} model is only applicable in the subthreshold region. Fig. 3.6 shows the dependence of V_{th} on gate voltage as well as channel doping concentration. The threshold voltage decreases with increasing V_G for various channel doping concentrations. As channel doping concentration is increasing, the initial threshold voltage (when V_G=0) increases, making the channel harder to be inverted. As N_{ch} increases, V_{th} decreases faster at larger V_G, presenting a more rapid switching behavior in the subthreshold region. This effect

is mainly reflected through the term of A in Eq. (3.15) and bulk potential in Eq. (3.18). The model fails to fit the numerical data well in the region of super threshold because the inversion charge is not considered in the model. The model only takes the depletion charge into consideration so that it stays valid only in the subthreshold region.

Fig. 3.7 shows the dependence of V_{th} on gate bias when oxide thickness is varying. As long as Eq. (3.19) is set negative, the threshold voltage decreases with the applied gate voltage. Moreover, with thinner oxide thickness, threshold voltage goes down more rapidly when biased at a higher gate voltage, further benefiting the faster transition. Finally, by integrating this V_{th} model with the standard MOSFET model, the characteristics of Fe-FET can be successfully reproduced. Fig. 3.8 shows the agreement between model-predicted IV and

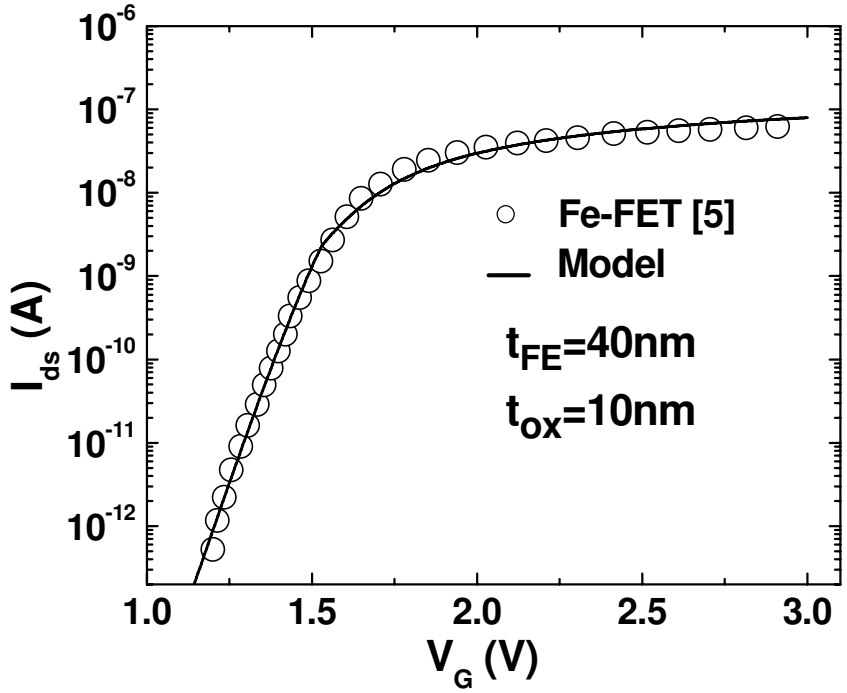


Figure 3.8. The matching between model prediction and measurement data.

published data [56], where the device is fabricated with 40nm ferroelectric dielectric and 10nm oxide thickness.

3.3 IMPLICATION OF PERFORMANCE VARIABILITY

The aggressive scaling of CMOS technology inevitably leads to a drastic challenge in process variations [79]. Based on the underlying mechanisms, variations in CMOS can be divided into intrinsic variations and manufacturing-induced variations [80]. The manufacturing induced variations arise from the imperfection during the fabrication process, and vary from foundries to foundries. Moreover, it exhibits the strong dependence on layout patterns, such as layout-dependent stress effect [12]. The process-induced variations could be improved by a better control of the process. On the other hand, intrinsic variations are limited by fundamental physics. They are inherent to CMOS structure, considered as one of the ultimate bottlenecks during the scaling of CMOS, such as random dopant fluctuations (RDF), and line edge roughness (LER) [48]. Among these variations, random dopant fluctuation is an intrinsic variation source in CMOS structure, posed by fundamental physical and manufacturing limits and has long been regarded as one of the ultimate major bottlenecks during large-scale integration scaling. RDF is caused by the random discrete placement of dopant atoms that follow a Poisson distribution in the channel region. As the device size scales down, the total number of channel dopants decrease, resulting in an elevated variation of dopant numbers, and significantly impacting threshold voltage.

Different from traditional MOSFET, there is an intrinsic feedback path in Fe-FET (Fig. 3.4), exploiting this feedback to modulate the subthreshold swing [46]. Furthermore, the loop of feedback implies that the sensitivity of device performance to certain parameters could be suppressed, if the feedback strength is appropriately tuned by Eq. (3.15) in section 3.2. The device performance would become more stable under microscopic variations. Based on TCAD and the new model, this section investigates this phenomenon, particularly on random dopant fluctuations (RDF) and explores the possibility to control the fundamental variation source in the MOSFET system with ferroelectric material integrated.

In a traditional MOSFET, the long-channel threshold voltage is expressed in Eq. (3.20)

$$V_{th} = V_{FB} + 2\phi_B + \frac{\sqrt{4\varepsilon_{si} q N_{ch} \phi_B}}{C_{ox}} \tag{3.20}$$

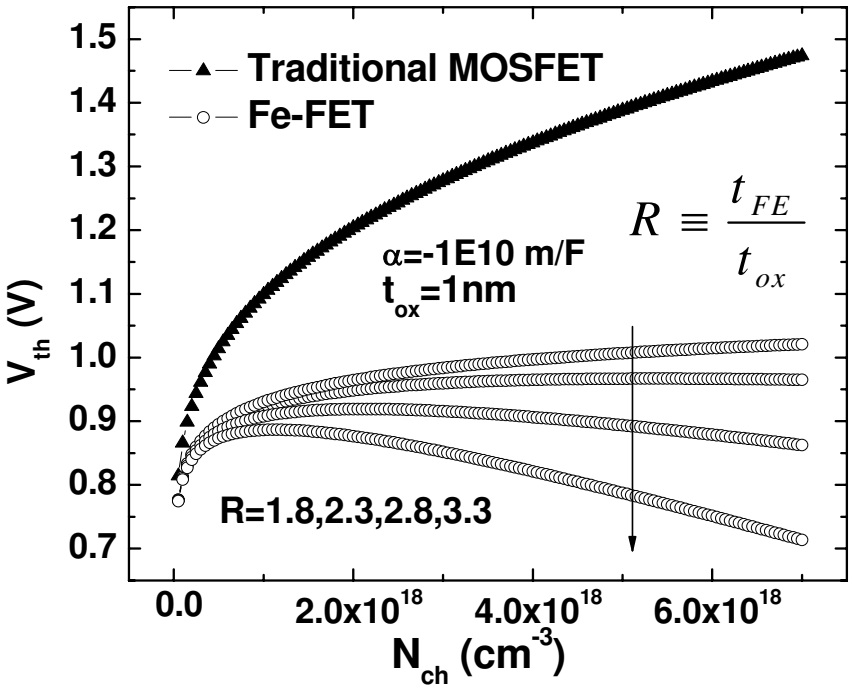


Figure 3.9. Dependence of V_{th} on N_{ch}. With the intrinsic feedback mechanism, V_{th} can be tuned to be insensitive to channel doping.

where V_{FB} and Φ_B denote flat-band voltage and bulk potential, respectively. The

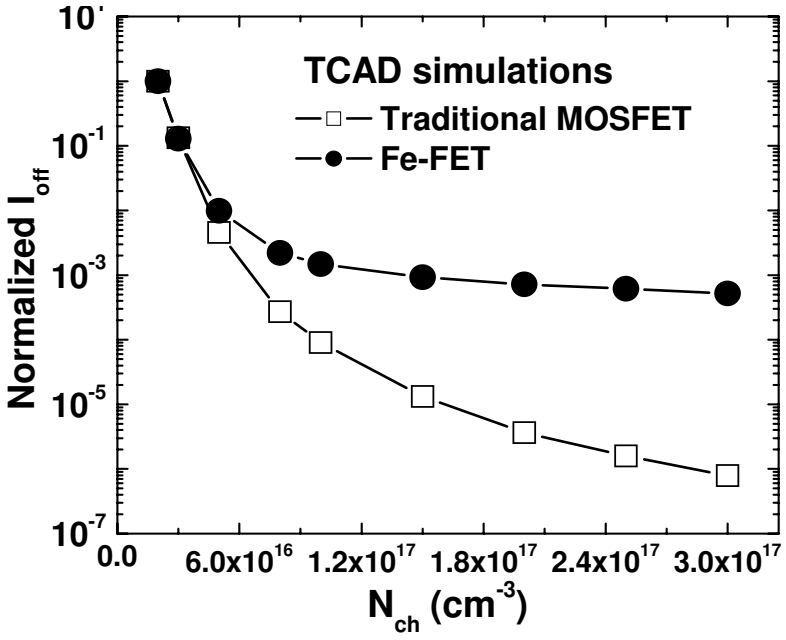


Figure 3.10. TCAD simulation of normalized off-state current for traditional MOSFET and Fe-FET.

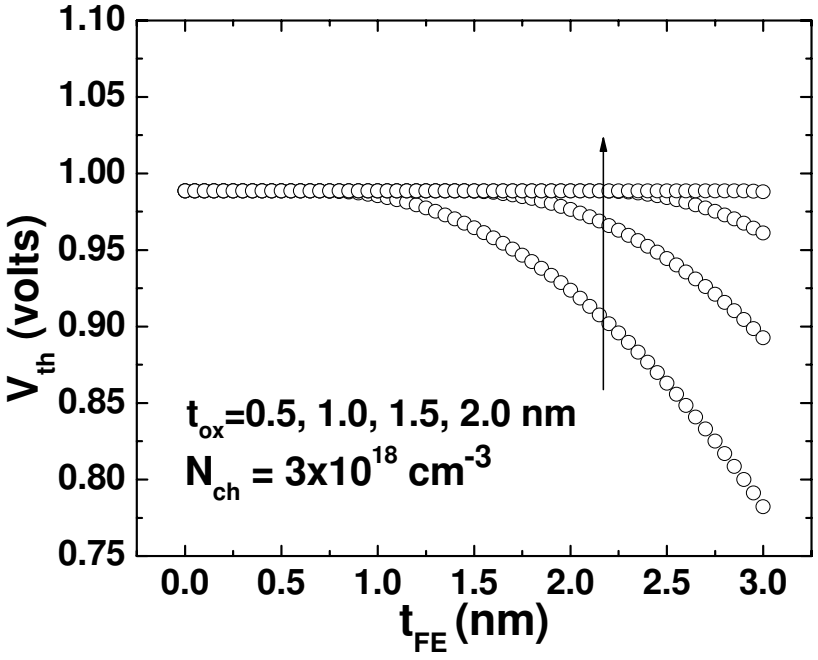


Figure 3.11. Dependence of V_{th} on t_{FE} and t_{ox} in Fe-FET. Higher t_{ox} suppresses t_{FE} -induced V_{th} variation.

impact of channel dopant fluctuations in a traditional MOSFET is mainly reflected through the last term in Eq. (3.20). Fig. 3.9 shows the sensitivity of the threshold voltages to the variation of channel dopant for traditional MOSFET and Fe-FET. The threshold voltage of Fe-FET is less sensitive to the dopant fluctuation than that in traditional MOSFET. Moreover, the sensitivity of V_{th} to N_{ch} in a Fe-FET can be modulated by the thickness of ferroelectric layer (t_{FE}). A larger t_{FE} leads to a stronger feedback path that compensates the RDF effect. When increasing the ferroelectric thickness, the threshold voltage starts to become much less sensitive to channel dopant and then declining with heavier dopant concentration introduced, which is different from that in a traditional MOSFET. This phenomenon demonstrates the great possibility for Fe-FET to suppress RDF effect by the proper design on thickness of ferroelectric material. TCAD simulation is further performed to validate this effect. Using TCAD tools, the dielectric constant of the gate insulator is adjusted to a negative value to simulate the effect of negative capacitance. In Fig. 3.10, TCAD simulation results compare the normalized off-state current between traditional MOSFET and Fe-FET, with various conditions of channel doping concentration. In the same range of dopant fluctuations, off-state current deviations in a Fe-FET and in a traditional MOSFET are 2.8 times and 114 times, respectively. The dependence on channel dopants becomes weaker with a negative capacitance integrated, implying that Fe-FET helps suppress RDF effect.

Figs. 3.11 and 3.12 further study the dependence of threshold voltage on ferroelectric layer thickness, oxide thickness, and channel dopant concentration.

Threshold voltage of Fe-FET changes with ferroelectric layer thickness and a thicker oxide thickness helps suppress t_{FE}-induced V_{th} variation as shown in Fig. 3.11. Fig. 3.12 illustrates that the dependence of threshold voltage on ferroelectric layer thickness for different doping concentrations. The lower doping concentration reduces the t_{FE}-induced variation as well as the nominal threshold voltage. Thus, lower N_{ch} and thicker t_{ox} improve device robustness; meanwhile, with a weaker feedback, the transition is degraded.

The reduction in V_{th} variability directly decreases the distribution of the off-state current (I_{off}) under RDF. Fig. 3.13 shows this effect for 45nm Fe-FET and traditional MOSFET, at the same on-state current (I_{on}). 1500 monte carlo

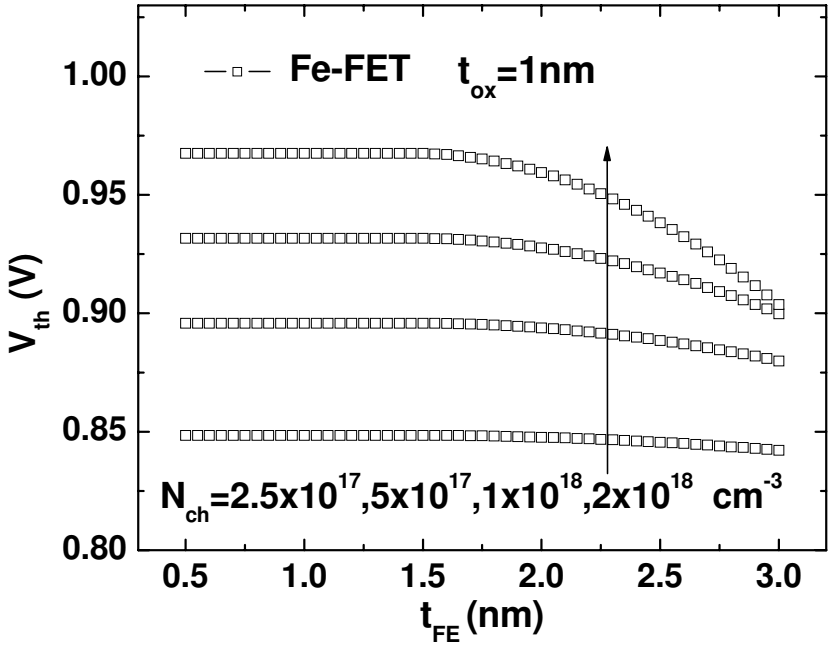


Figure 3.12. Dependence of V_{th} on t_{FE} and N_{ch}. Higher N_{ch} increases t_{FE} -induced V_{th} variation.

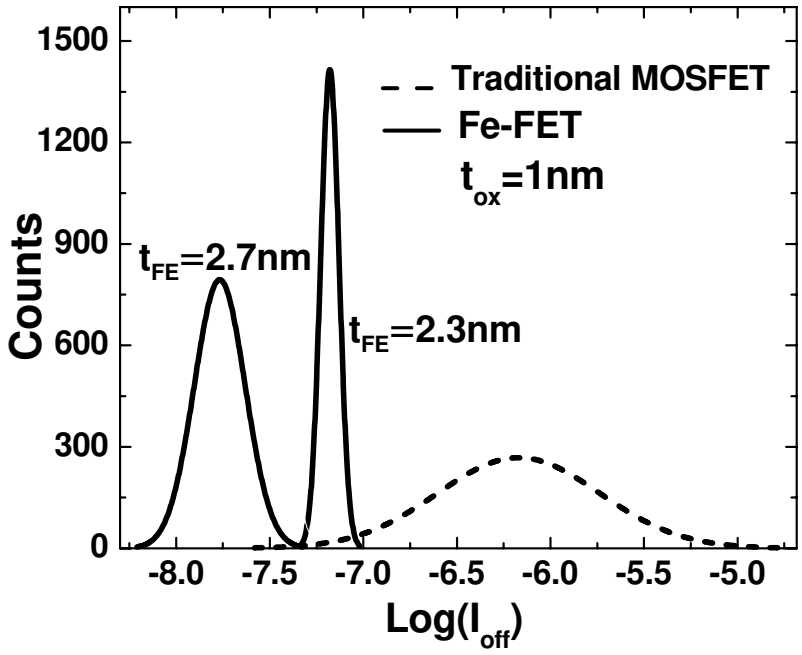


Figure 3.13. The off-state current distribution under RDF, with the same on-current.

simulations with 30% channel dopant fluctuation are performed to investigate the performance variability in a Fe-FET. As t_{FE} decreases to 2.3nm (the least dependence on N_{ch} in Fig. 3.9), the standard deviation (σ_{Ioff}) is minimized. In the log scale, the standard deviation of off-state current in a traditional MOSFET is 0.43, while in an optimized Fe-FET the standard deviation is 0.0492, showing a 8.8 times improvement in log-scale off-state current variation. When t_{FE} further increases, the subthreshold slope becomes steeper and thus, a lower I_{off} is achieved; however, a thicker t_{FE} leads to a larger I_{off} fluctuation. Fig. 3.14 illustrates that as t_{FE} increases, the mean value (μ_{Ioff}) decreases due to stronger positive feedback, while σ_{Ioff} increases. The ratio of variation to the mean value is increasing rapidly when thicker ferroelectric thickness is introduced. When the thickness is larger than 3.1nm, the off-state current variation becomes larger than

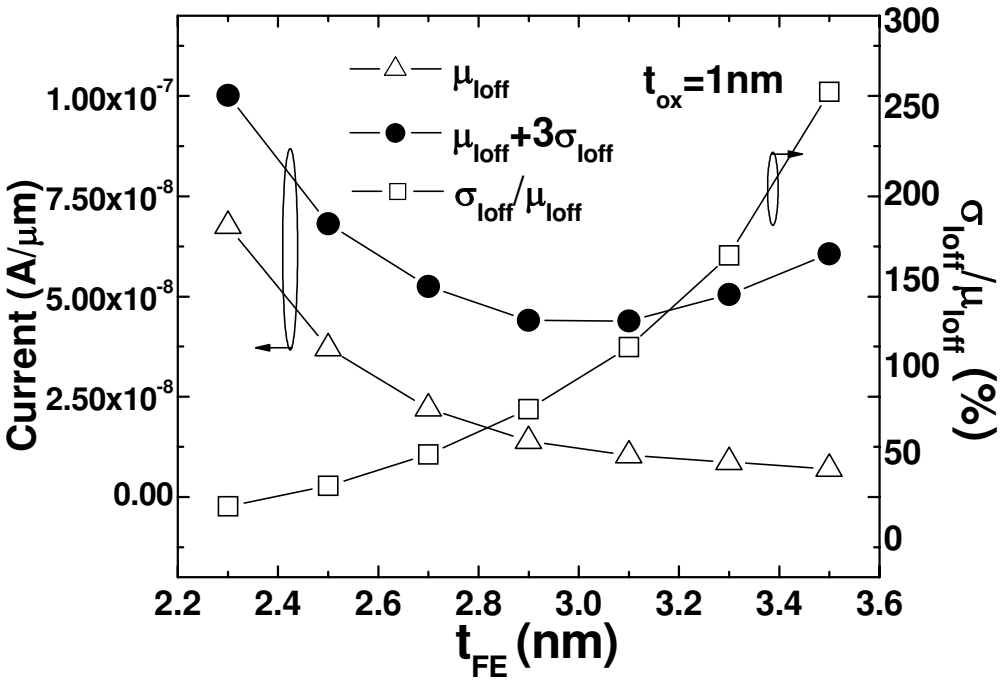


Figure 3.14. The mean value (μ_{Ioff}), standard derivation (σ_{Ioff}), and the worst case (μ_{Ioff}+ $3\sigma_{Ioff}$) of off-state current, at a fixed on-state current.

the nominal value, which eliminates the low nominal off-state current benefited from the stronger feedback. A balance between t_{ox} and t_{FE} is required to achieve the minimum worst case I_{off} (μ_{Ioff}+$3\sigma_{Ioff}$).

3.4 CHAPTER SUMMARY

By integrating ferroelectric material into the traditional MOSFET structure as an internal voltage amplifier, Fe-FET becomes one of the promising candidates for energy efficient application. In this study, a new compact model of Fe-FET is developed, by adding a feedback path for the ferroelectric layer to capture the feedback of negative capacitance and IV characteristics of Fe-FET. It is scalable with technology specifications. Based on the model, it demonstrates that a thicker ferroelectric material, higher channel doping concentration, and thinner oxide thickness improve the transition in the subthreshold region. On the other hand, it

is further revealed that the impact of random dopant fluctuation (RDF) on leakage variability can be significantly suppressed in Fe-FET, by tuning the thickness of ferroelectric layer. This implies the internal feedback in a device could possibly suppress the process variation. A balance between the feedback and feedthrough paths is required to stabilize device performance under the RDF effect, and further to optimize the circuit performance.

CHAPTER 4

MODELING OF THROUGH SILICON VIA CAPACITANCE AND THE STRESS EFFECT

Different from the previous chapters with focus on modeling device performance and variation management in active devices, the modeling effort of 3D integrated chips (ICs) [57][58] is emphasized in this chapter. From the perspective of performance, 3D ICs have short interconnects among each function block, leading to better RC delay. Through silicon via (TSV) is a promising technology to enable the 3D chip integration [59][60][61]. However, during the implementation of TSV, the interconnect capacitance is voltage dependent and sensitive to the diameter of TSV, thickness of barrier layer, and the substrate doping concentration. A model that physically captures the dependence of the voltage dependent TSV capacitance is essential to optimize the chip performance. Moreover, during the process, due to the mismatch of coefficient of thermal expansion (CTE) between copper and silicon, thermal stresses is observed at the interface between TSV and silicon substrate, impacting device performance of neighboring transistors [61]. A stress model is presented to model TSV induced stress effect to predict the influence of the stress and help designers optimize the circuit performance. The results help investigate the optimal process and material choice for high-speed TSV applications. Practical guidelines are extracted for the planning of TSV with active devices. These compact modeling results are well verified with TCAD simulations, supporting the interaction between silicon process and IC design at the early stage.

4.1. INTRODUCTION

With shrinkage of device dimension, the number of transistors in a chip doubles every 18 months according to Moore's law. However, more and more difficulties are encountered when device dimension are further scaling. For example, the oxide thickness of MOSFET in 45nm technology now is about one nanometer, roughly twice the silicon lattice, implying that the scaling is approaching the physical limits [1]. On the other hand, 3D integration is emerged as a new solution by connecting heterogeneous chips vertically and integrates more transistors. Through silicon via (TSV) is a promising and key technology to integrate chips with diverse functionalities by stacking chips vertically for implementation of 3D ICs with less space and better performance, as shown in Fig. 4.1. Moreover, TSV technology could significantly reduces the global interconnect wire length, and reduce the RC delay, which is especially beneficial

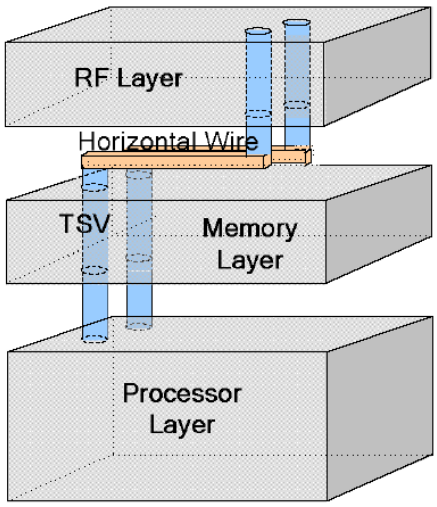


Figure 4.1. In 3D-IC, the heterogeneous chips are integrated vertically by through silicon vias (TSVs). [63]

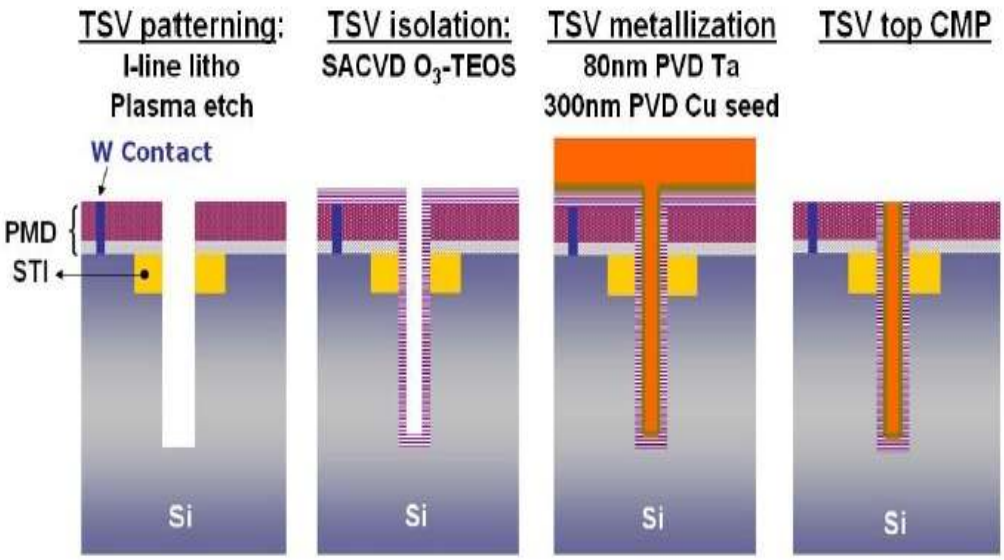


Figure 4.2. TSV process flow [65]

for high-speed application. One of the advantages in 3-D ICs, it serves as the most promising technology to implement "More-than-Moore" technologies, bringing heterogeneous materials and technologies on a single chip [63].

The basic structure of TSV is composed of two components: conducting metals in the via and the barrier layer around metal. Copper is one of the materials frequently used to server as interconnect between devices due to its better immunity of electron migration (EM) and lower resistivity [64]. The barrier layer is used to suppress the out-diffusion of copper. Fig. 4.2 shows the brief TSV process flow. First, lithography and etching technologies are used to transfer the mask pattern to create the high aspect ratio via in silicon. The barrier layer is deposited on the surface of via to avoid metal diffusion. Metal is deposited to fill via and planarized by CMP technology. Then etch the silicon substrate back to expose via and bonded to the landing wafer by using metal thermo-compression [65][66] .

Since TSV is developed based on metal-insulator-semiconductor (MIS) structure, the parasitic capacitance is different from the traditional interconnect capacitance. Instead, it has the similar behavior to that of a MOS capacitance and strongly depends on the applied voltage and sensitive to TSV physical dimensions, including the TSV metal radius and thickness of barrier layer. Moreover, the substrate doping concentration is one of the main factors contributing to the depletion capacitance in silicon substrate, making TSV capacitance sensitive to the applied voltage.

On the other hand, during TSV fabrication, thermal stress is observed at the interface between TSV and silicon due to the mismatch in coefficient of thermal expansion (CTE) between silicon substrate and metal, where copper is usually adopted as the conducting via [67][68]. The thermal stress is a penetrating field and decays over distance. Moreover, the stress magnitude is sensitive to TSV geometry structure; when the radius of TSV metal increases, more stress is introduced and becomes saturated. TSV stress has the similar property to the leading edge strain technology we mentioned earlier. This property makes it sensitive to the layout pattern. We'll further discuss in the following sections. Since TSV stress is not intentionally applied to impact the device performance, a zone around TSV is required to keep devices unaffected by TSV stress.

This chapter is organized as following. Section 4.2 introduces the basics of TSV capacitance and resistance. TCAD simulation is further performed to investigate the dependence of capacitance on TSV geometry parameters and doping concentrations. In section 4.3, the impact of thermal stress induced during

TSV process is analyzed and the mobility change is modeled. Section 4.4 concludes this chapter.

4.2. MODELING OF TSV RESISTANCE AND CAPACITANCE

TSV shortens the wire distance between chips and significantly reduced the line capacitance and resistance, i.e., the RC delay, making TSV a promising technology for high speed application. The prevalent TSV structure is composed of a cylindrical copper surrounded by an insulator in silicon substrate as shown in Fig. 4.2. This structure is essentially a metal-insulator-semiconductor (MIS) capacitance, rendering TSV the similar capacitance-voltage (CV) characteristics to a typical MOS capacitance. On the other hand, the unique architecture and process of TSV increases the complexity in modeling and analysis. In the following, we will study the advances of TSV modeling work and discuss the trade-off between performance and required area.

4.2.1 TSV Resistance

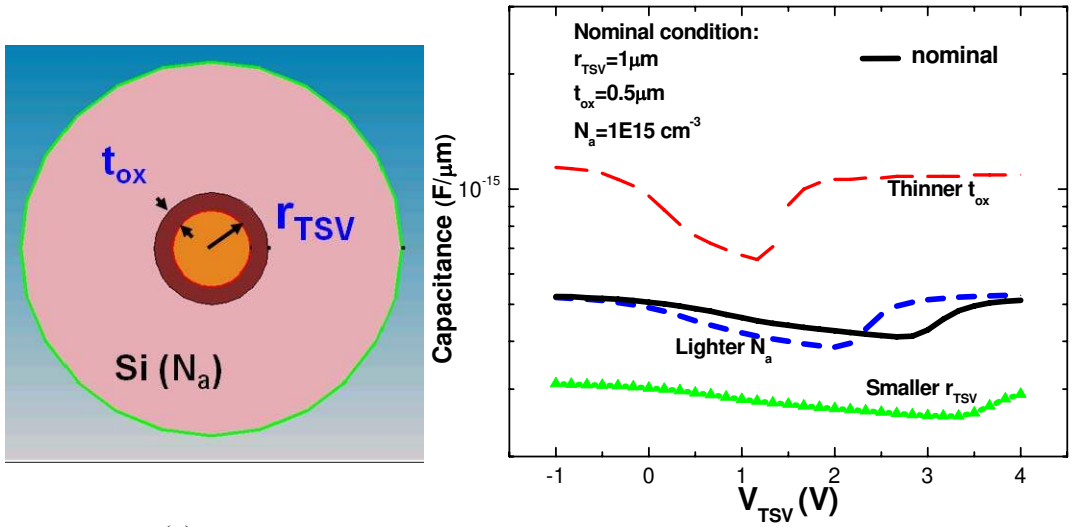
TSV resistance is an important factor to the interconnect delay. The resistance increases with the length of conducting metal and decreases as the cross-sectional area grow, which is expressed in Eq. (4-1),

$$R = l \cdot \frac{\rho}{A} \tag{4-1}$$

where l and A are the length and cross-sectional area of TSV respectively. ρ is the resistivity of the conducting metal. Since TSV is formed with a cylindrical shape through silicon substrate, the cross-sectional area is calculated in Eq. (4-2),

$$A = \pi \cdot r_{TSV}^2 \tag{4-2}$$

where r_{TSV} represents the radius of the conducting material. To reduce TSV resistance for high speed interconnection, the resistivity of conduction material needs to stay with a small value. Larger radius and shorter length of TSV conducting material are required to minimize the resistance. However, a larger TSV radius represents more area is required, indicating a trade-off between the performance and the cost. Moreover, the area is also related to the capacitance. In the following section, TSV capacitance is discussed to analyze RC delay.



(a) (b)

Figure 4.3. (a) TSV simulation structure with circular cross-section. (b) TSV capacitance for different geometry conditions and dopant concentrations.

4.2.2 TSV Capacitance

Since TSV consists of a conduction material surrounded by a barrier layer through silicon substrate, this unique structure makes the parasitic capacitance has the similar behavior to that of a MOS capacitance and strongly depends on the applied voltage and sensitive to TSV geometry parameters, including the TSV

radius and the thickness of barrier layer [69]. Compared with the traditional interconnect capacitance, TSV capacitance changes with the applied voltage and sensitive to the doping concentration in the substrate. To investigate the dependence, TCAD simulation is performed to analyze TSV capacitance.

TCAD simulation solves the Poisson equation with metal radius of 1 um, barrier-layer thickness of 0.5 um, and substrate doping concentration of ten to fifteen, as shown in Fig. 4.3 (a) [30]. Note that the materials of TSV metal and barrier layer in the simulation are copper and silicon dioxide, respectively. The capacitance changes with the applied voltage (V_{TSV}) and holds the similar CV characteristics to MOS capacitance. When the silicon in the substrate starts to deplete carriers, the overall capacitance is seen as the oxide capacitance and depletion capacitance in series, leading to the reduction of TSV capacitance. When the voltage keeps growing up, the depletion depth grows and results in a smaller TSV capacitance until the voltage reaches the threshold voltage, which represents the formation of sufficient minority carrier density around TSV. Therefore, the mobile carriers in the surface shield the depletion capacitance and leave the oxide capacitance while seen from the edge of copper.

To further investigate the dependence on the unique geometry parameters, TSV structures with a smaller copper radius and a thinner oxide layer are simulated, as shown in Fig. 4.3 (b). Compared with the nominal CV characteristics, the overall TSV capacitance increases with a thinner oxide thickness applied. It becomes easier for minority carrier around TSV to reach the sufficient amount to invert the substrate conduction polarity, also known as

threshold voltage in the MOS system, shielding the electrical field from the applied voltage. Moreover, Fig. 4.3 (b) shows the portion of depletion capacitance plays an increasing role on the series capacitance with a thinner oxide layer, where 40% drop from the oxide capacitance is observed. On the other hand, when a smaller radius of TSV is adopted, the overall capacitance decreases and the threshold voltage becomes larger, which poses an opposite impact with a thinner oxide layer. The effect of series capacitance becomes less significant. The smallest series capacitance shows 17% drop from the oxide capacitance, implying a less significant effect of depletion capacitance in the case of a smaller TSV radius.

In addition to the radius of copper and oxide thickness, the doping concentration in silicon impacts the depletion capacitance, which is reversely proportional to the depletion depth as modeled in Eq. (4-3)

$$C_{dep} = A\frac{\varepsilon_{si}}{d} \tag{4-3}$$

where d is the depletion depth and ε_{Si} is the dielectric constant of silicon. The depletion depth is reversely proportional to the square root of doping concentration, as Eq. (4-4).

$$d = \sqrt{\frac{2\varepsilon_{si}\phi_s}{qN_{ch}}} \tag{4-4}$$

where N_{ch} is the doping concentration and ϕ_s denotes the surface potential around TSV. As doping concentration increases, the depletion depth decreases, resulting in the growth of depletion capacitance. Note that the impact of doping

concentration on depletion depth is in the square root. Thus, the impact on the depletion capacitance is not significant. Fig. 4.3 (b) shows that when less doping concentration is introduced, TSV C-V characteristics shift to the left while the oxide capacitance remains the same. The threshold voltage is reduced and the depletion capacitance is not impacted significantly.

Fig. 4.4 shows the contour of electrical field distribution. No electrical field inside copper is observed and the E-filed is uniformly distributed across the oxide. In the silicon, the electrical field decays over distance from the TSV interface, contributing to the depletion capacitance. Fig. 4.5 shows TSV capacitance is composed of oxide capacitance (C_{ox}) and depletion capacitance (C_{dep}) in series, where C_{dep} varies with the applied voltage to copper.

To capture the dependence on TSV geometry and process conditions, traditional efforts resort to TCAD simulation to quantitatively analyze the change of TSV capacitance. This approach usually requires expensive computation, especially when chip complexity per unit area keeps increasing along with technology scaling. Therefore, developing an effective model that is able to predict the scaling trend and circuit performance in 3D ICs is essential.

Katti et al., [69] proposed a capacitance model by solving 1D Poisson equation in a cylindrical coordinate system, as given in Eq. (4-5),

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial \varphi}{\partial r}\right)=\frac{qN_a}{\varepsilon_{si}} \tag{4-5}$$

where q is the unit charge, N_a is the doping concentration in silicon substrate, and ε_{Si} is the permittivity of silicon. Since the applied voltage is distributed in the silicon oxide and the depletion region, the potential and the electrical field is zero at the boundary between depletion region and neutral silicon, with expressions in Eqs. (4-6) – (4-7)

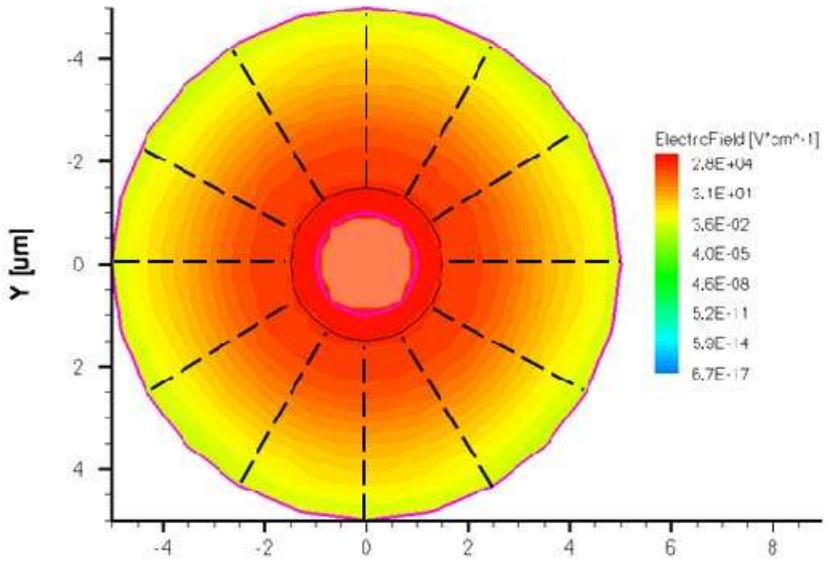


Figure 4.4. Contour of electrical field distribution in the silicon substrate

$$\varphi \mid r_{dep} = 0 \tag{4-6}$$

$$\frac{\partial \varphi}{\partial r} \mid_{r_{dep}} = 0 \tag{4-7}$$

where r_{dep} is the radius from the center to the boundary between depletion region and neutral silicon. The oxide capacitance and the depletion capacitance are calculated as in Eq. (4-8) and Eq. (4-9), respectively.

$$C_{OX} = \frac{2\pi \cdot \varepsilon_{ox}}{\ln\left(1 + \frac{t_{ox}}{r_{TSV}}\right)} \tag{4-8}$$

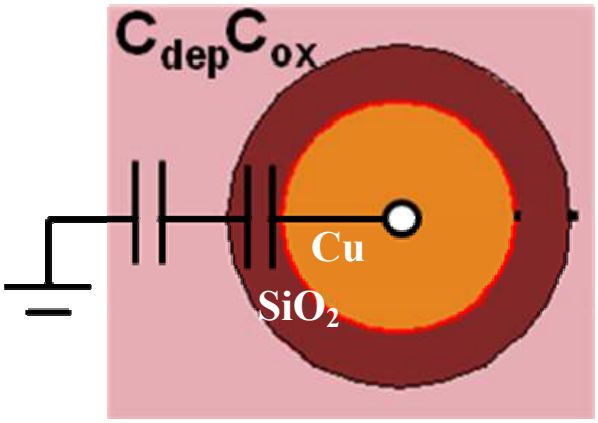


Figure 4.5. Schematic of series capacitance of depletion capacitance and oxide capacitance.

$$C_{dep} = \frac{2\pi \cdot \varepsilon_{si}}{\ln\left(\frac{r_{dep}}{r_{TSV} + t_{ox}}\right)} \tag{4-9}$$

where ε_{ox} and t_{ox} are the permittivity and thickness of silicon dioxide, respectively. r_{TSV} denotes the radius of copper. Note that the oxide capacitance is formulated as a function of metal radius and oxide thickness. Total TSV capacitance is composed of depletion capacitance and oxide capacitance in series, as in Eq. (4-10).

$$C_{total} = \frac{C_{dep} C_{OX}}{C_{dep} + C_{OX}} \tag{4-10}$$

Moreover, the threshold voltage is expressed in Eq. (4-11).

$$V_{th\_TSV} = V_{fb} + 2\phi_B + \frac{Q_B}{C_{OX}} \tag{4-11}$$

where V_{fb} and ϕ_B are flat-band voltage and bulk potential. Q_B is the depletion charge and can be calculated in Eq. (4-12).

$$Q_B = \pi q N_a \cdot \begin{pmatrix} (r_{TSV} + t_{ox} + d_{\max})^2 \\ -(r_{TSV} + t_{ox})^2 \end{pmatrix} \tag{4-12}$$

where d_{max} is the maximum depletion depth in silicon substrate and is calculated using surface potential equal to twice bulk potential.

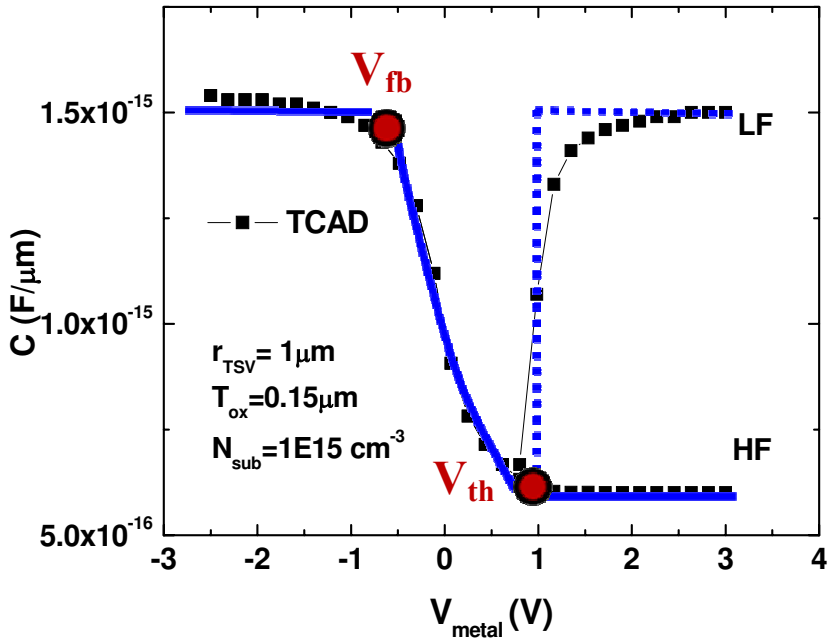


Figure 4.6. Compact capacitance model

With Eqs. (4-8) to (4-12), threshold voltage and flat-band voltage are defined. Moreover, the capacitances in accumulation and depletion regions are well calculated. Therefore, a piecewise capacitance model is proposed for an efficient circuit simulation. Figure 4.6 shows the TCAD simulation and the corresponding piecewise models for low-frequency and high-frequency operations, respectively. In low frequency, the depletion capacitance is shielded by the minority carrier after threshold voltage, so the total capacitance is seen as a single oxide capacitance. While in high frequency operation, the generation of minority carrier is not fast enough to catch the frequency of the applied voltage.

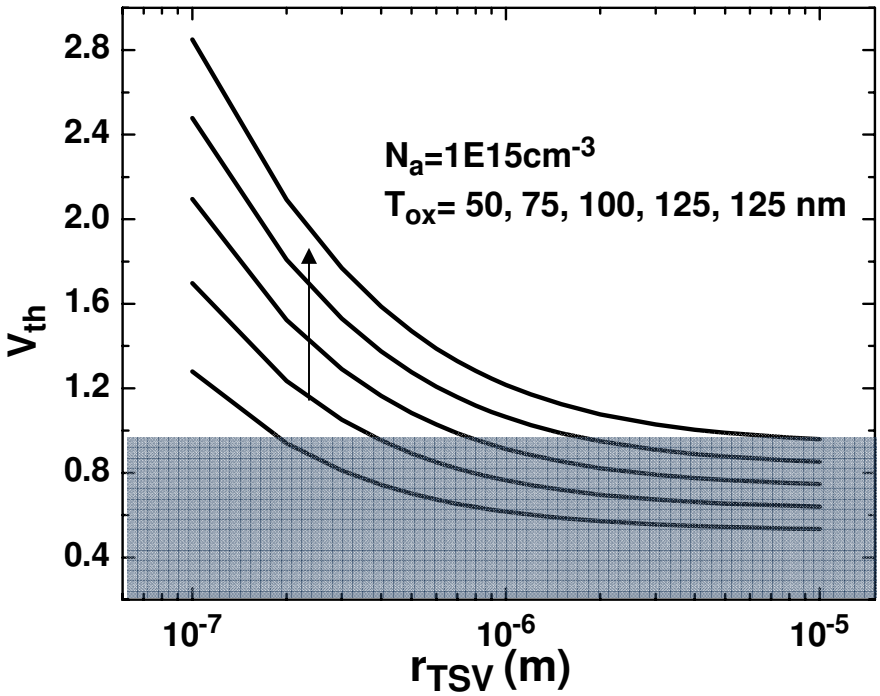


Figure 4.7. Threshold voltage varies with the radius of copper for different oxide thickness.

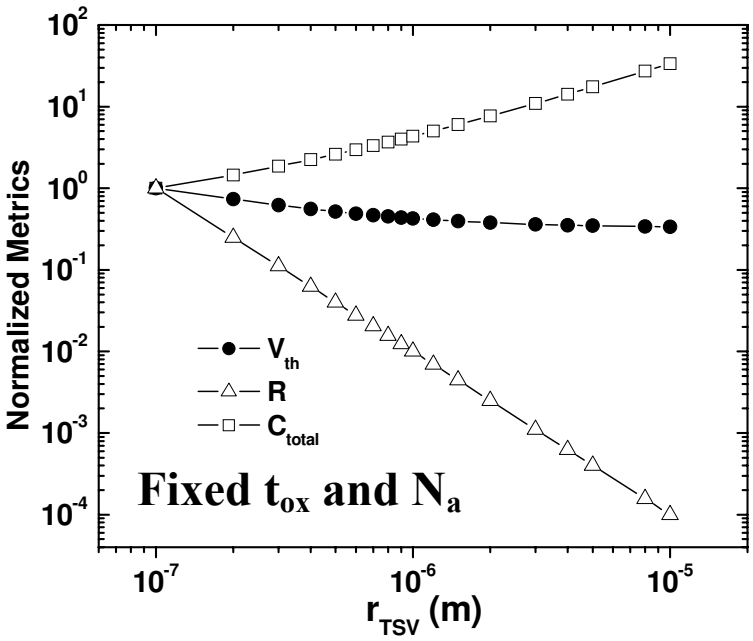
Therefore, the total capacitance is regarded as a series capacitance of oxide capacitance and depletion capacitance.

In 45nm technology node, the standard applied voltage (V_{dd}) is 1.0 volt. An optimized TSV structure is required to minimize the parasitic capacitance. For example, for a low-frequency operation, the applied voltage slightly below threshold voltage is preferred to have a low TSV capacitance. For a high-frequency operation, the applied voltage larger than threshold voltage is desired for high speed application. In Fig. 4.7, the threshold voltage changes with the radius of TSV metal. As the radius increases, the threshold voltage decreases; while as oxide thickness increases, the threshold voltage increases. For a high-frequency operation, the threshold voltage should be smaller than the applied voltage, as shown in the highlighted area in Fig. 4.7.

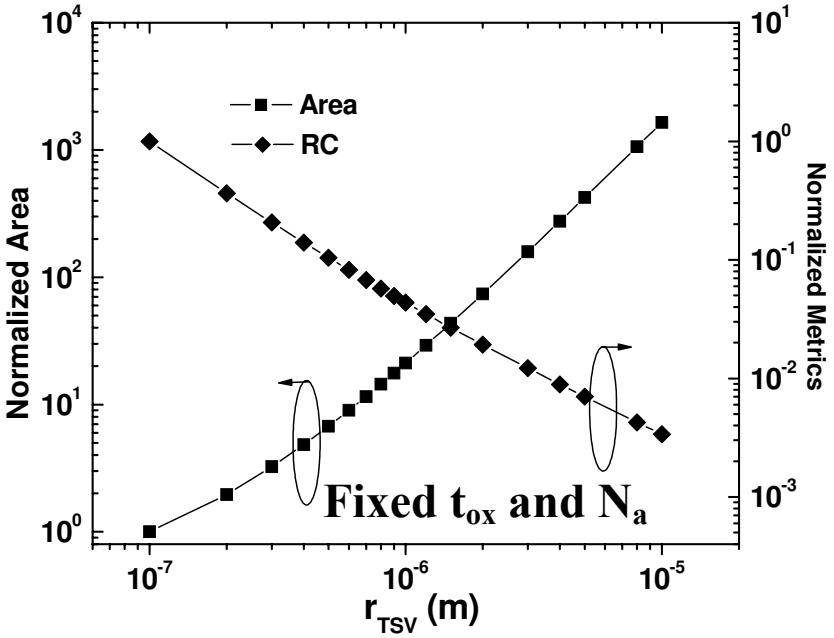
4.2.3 Trade-off between TSV Performance and Area

Low capacitance, low resistance and high density integration are the main desired features in a TSV structure. Low capacitance and low resistance implies the better interconnect performance, i.e. RC delay. High density integration indicates the small area required during TSV process. Both required area and interconnect performance heavily rely on TSV process and structures, including the design of metal radius (r_{TSV}), thickness of barrier layer (t_{ox}), and the doping concentration in silicon substrate (N_a). The materials for conducting metal and barrier layer of a TSV are copper and silicon dioxide, respectively. In this sub-section, TSV performance and the required area are studied with respect to various process parameters, including doping concentration, copper radius, and oxide thickness.

Figure 4-8 (a) shows the impact of copper radius on threshold voltage, resistance, and the smallest capacitance. Note that the smallest total capacitance (C_{total}) represents the series capacitance with surface potential equal to twice bulk potential. As the radius of copper (r_{TSV}) increases, the resistance decreases because of the larger copper cross-section. On the other hand, according to Eqs. (4-8) to (4-10), the capacitance grows up. Moreover, the threshold voltage drops because C_{ox} increases with the radius of copper. In Fig. 4-8 (b), RC delay and the area requirement are further calculated for various copper radiuses. Although the opposite changing trend of capacitance and resistance to different copper radius is observed in Fig. 4-8 (a), RC product decreases as larger copper radius is introduced, implying the impact of resistance reduction is stronger than the



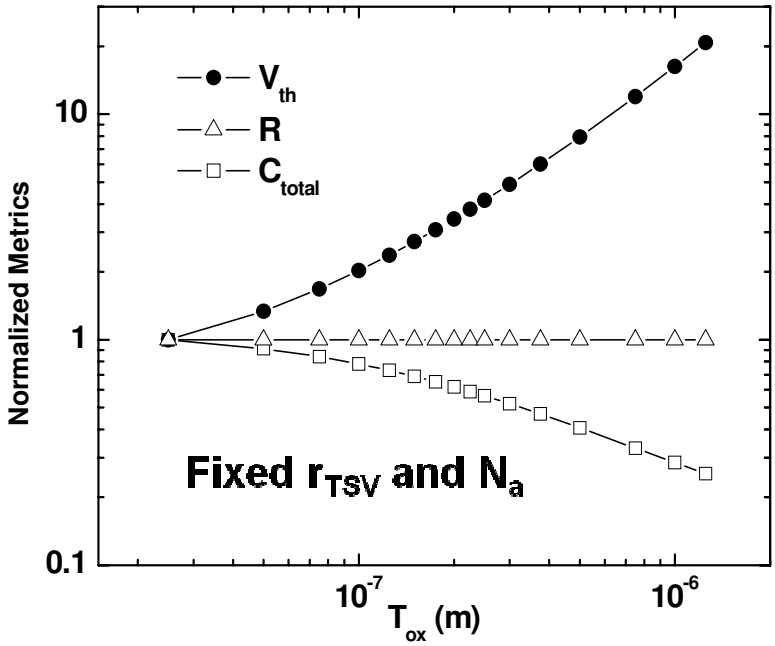
(a)



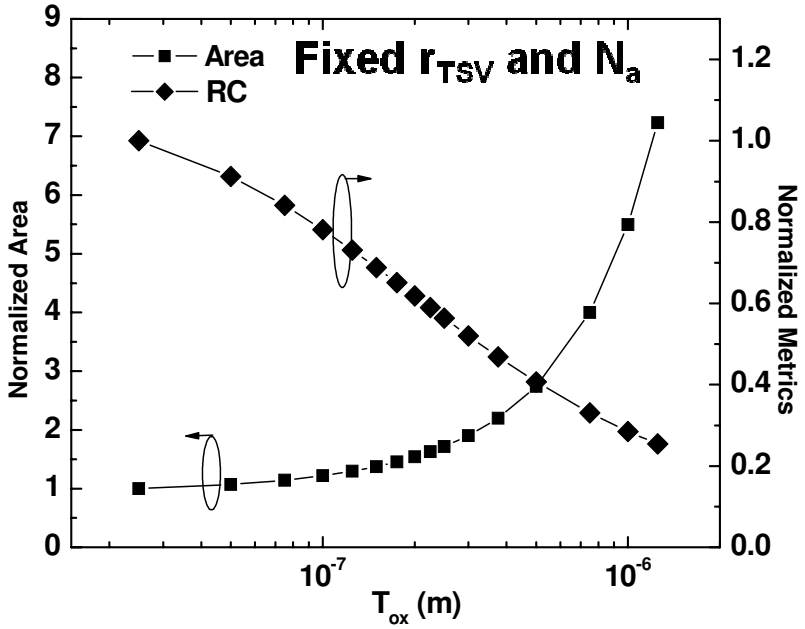
(b)

Figure 4.8. (a) Threshold voltage, resistance, and capacitance change with various metal radiuses. (b) trade-off between RC delay and TSV area.

increasing capacitance. On the other hand, with the growth of copper radius, a



(a)



(b)

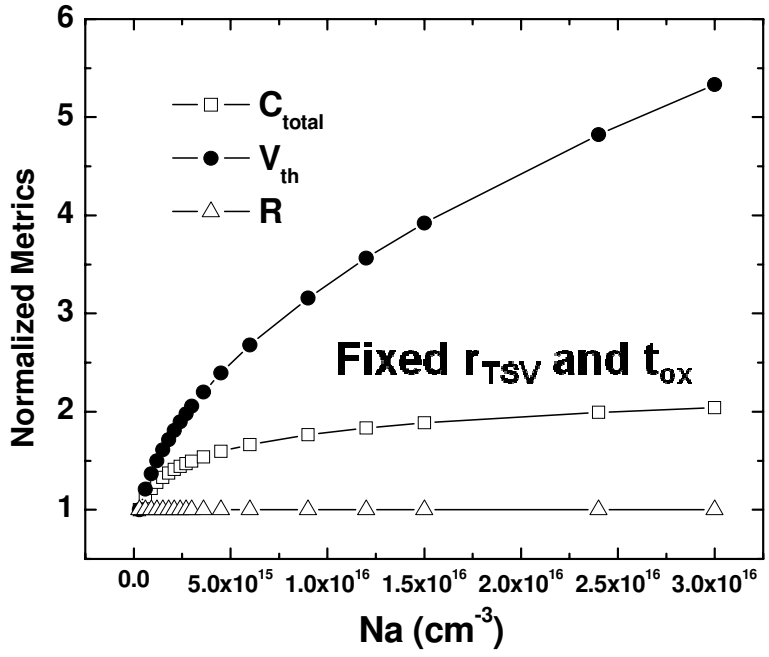
Figure 4.9. (a) Threshold voltage, resistance, and capacitance change with various T_{ox}. (b) trade-off between RC delay and TSV area.

larger TSV area is demanded, showing a trade-off between area and TSV performance. In addition to designing TSV with different copper radius, thickness

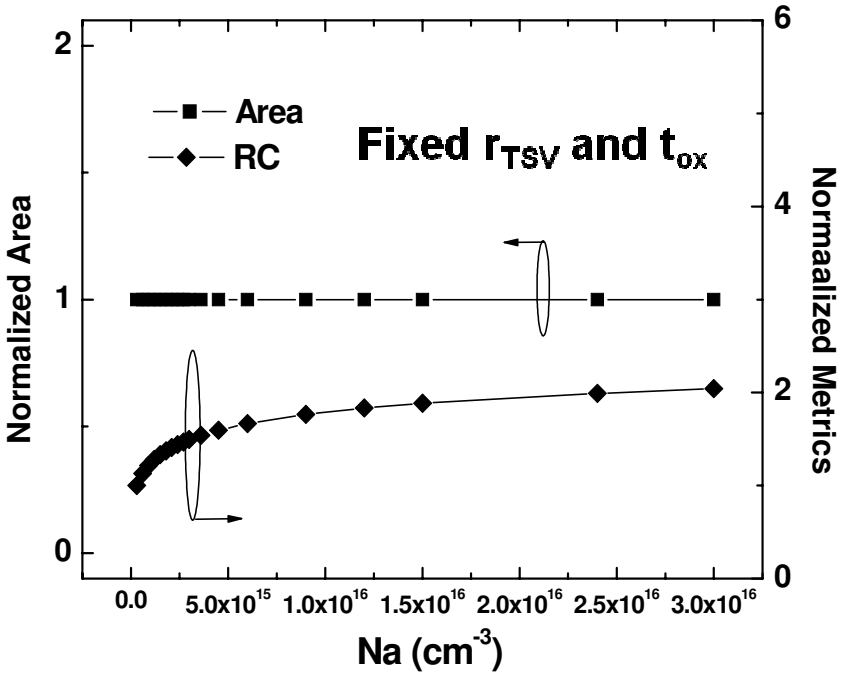
of barrier layer and the doping concentration are the other choices available for optimizing TSV structures and interconnect performance.

Figure 4.9 shows the normalized TSV characteristics with various thickness of barrier layer (t_{ox}) with a fixed metal radius and doping concentration. In Fig. 4.9 (a), a thicker barrier layer results in a smaller oxide capacitance. Therefore, the total capacitance declines. This behavior is similar to the traditional MOS capacitance. Due to the reduction of oxide capacitance, it becomes harder to invert the surface and therefore the threshold voltage increases with the growing barrier layer thickness. On the other hand, the resistance is independent of the barrier thickness. Fig. 4.9 (b) further discusses the dependence of the thickness of barrier layer on RC delay and TSV area. The area increases with the increasing barrier thickness, while the RC delay reduces. Similar to the impact of increasing metal radius, there is a trade-off between area and the RC delay when tuning the barrier thickness.

Figure 4.10 shows how TSV performance changes with various doping concentrations with fixed copper radius (r_{TSV}) and barrier thickness (t_{ox}). A heavier doping concentration leads to a larger depletion capacitance because of the shorter depletion depth. Therefore, the increasing series capacitance is observed in Fig. 4.10 (a). Changing doping concentration does not impact the resistance of the conducting metal. On the other hand, with the introduction of more dopants in the substrate, the bulk potential increases and harder to reach the criterion of surface inversion, leading to an increasing threshold voltage. This behavior is also similar to the traditional MOS capacitance. Fig. 4.10 (b) further



(a)



(b)

Figure 4.10. (a) Threshold voltage, resistance, and capacitance change with various N_a. (b) N_a does not impact TSV area

investigates how the doping concentration impacts on RC delay and TSV area. Different from the results in Fig. 4.8(b) and Fig. 4.9 (b), TSV area does not

change with various doping concentrations. Moreover, RC delay increases because of the increasing depletion capacitance. Therefore, a lighter doping concentration is desired for a high speed TSV application with the unchanged area.

4.3 IMPACT OF TSV STRESS ON NEIGHBOR DEVICES

In addition to the influence of TSV parameters on the parasitic capacitance and the required area, one major challenge during TSV process is the reliability due to the thermal stress [70][71]. Unlike the state-of-the-art strain technologies

Material	E (GPa)	ν	CTE (ppm/°C)
Copper	111.5	0.343	17.7
Low-k	9.5	0.3	20.0
Epoxy	6.1	0.35	33.0
Si_3N_4	190.0	0.27	3.2
SiO_2	71.7	0.16	0.51
Silicon	162.0	0.28	3.05
Tungsten	366.0	0.296	4.6

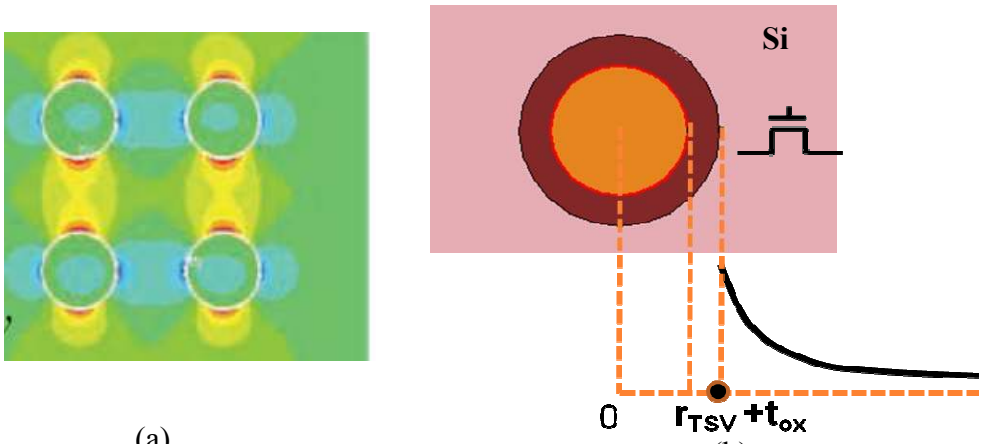
Table 4.1. material property for CTEs [62]

such as eSiGe or DSL, TSV thermal stress is not an intentional technique applied to improve device performance. Instead, the thermal stress results in additional process variation and reliability issue. The management and control of this parasitic stress is important for 3D IC development.

The thermal stress originates from the mismatch in coefficient of thermal expansion (CTE) between TSV fill material and silicon substrate. Tungsten, poly-silicon, and copper are being considered as TSV conducting metals. Due to the

low resistivity of copper, it becomes a major TSV fill material, leading to a low RC delay. However, owing to the CTE difference between copper and silicon, the thermal stress is observed when temperature ramp is involved during TSV process. The copper electroplating and annealing temperature is higher than the operating temperature, resulting in tensile stress in silicon substrate after cooling down to the room temperature [72]. Table 4.1 shows the material choices from the perspective of thermal expansion. It is necessary to assess the impact of thermal stress around the vias on the carrier mobility in silicon. It has been reported that the stress of 100MPa can result in the mobility variation by 7% [73]. The stress is a penetrating field and decays over distance. The profile of TSV induced thermal stress in silicon follows the similar distribution to the leading-edge strain technology we mentioned in chapter 2. This property makes it sensitive to the layout pattern. However, TSV stress is not confined in the channel region of a transistor. Instead, the stress profile is openly distributed and could impact transistors around TSV, as shown in Fig. 4.11. Moreover, when the radius of TSV increases, the stress increases and becomes saturated.

In the TSV process, TSV is fabricated as an array as in Fig. 4.11 (a), where the thermal stress is non-uniformly distributed. Fig.4.11 further shows that stress magnitude decays dramatically from TSV edge and will impact the neighbor transistors. In a TSV array, the stress profile between two TSVs behaves like a bathtub curve, similar to the leading-edge strain technologies such as eSiGe technology. Due to the bathtub stress distribution, the same modeling approach in chapter two is adopted; a linear piecewise stress model is proposed to capture



(a) (b)

Figure 4.11. (a) Stress distribution [74] (b) schematic of stress profile in silicon substrate

layout dependence of TSV stress effect. Shown in Fig. 4.12, the stress profile is well captured by the piecewise models, which are to be used to calculate the impact on mobility and other electrical characteristics without losing the fundamental physics. To account for layout dependence of TSV stress, the peak and bottom stresses are modeled as a function of layout parameters in Eqs (4-13) and (4-14)

$$\sigma_P = \left(1+\frac{m}{W_{TSV}}\right)\cdot\frac{r_{TSV}}{A_{TSV}+r_{TSV}}\cdot\sigma_{m\_TSV} \tag{4-13}$$

$$\sigma_B = \frac{C_{TSV}}{C_{TSV}+W_{TSV}}\cdot\sigma_P \tag{4-14}$$

where W_{TSV} and σ_{m_TSV} are the distance between two vias and the maximum stress introduced during TSV process, respectively. A_{TSV} is a fitting parameter to account for the dependence of TSV metal radius on the stress; when the radius of TSV increases, the stress increases and becomes saturated. C_{TSV} is responsible for the dependence of the distance between two TSVs on peak and bottom stresses.

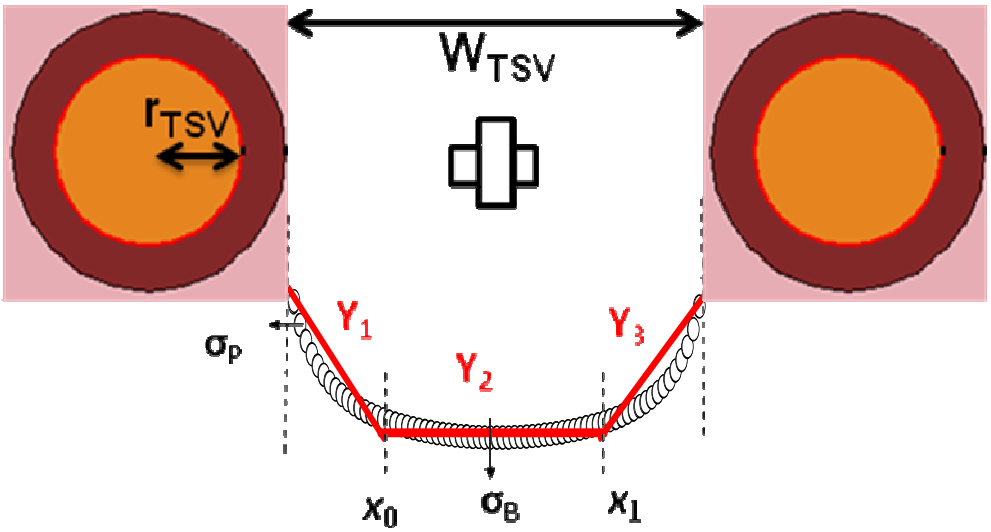


Figure 4.12. Linear piecewise stress model to capture the bathtub stress profile between two TSVs.

When the two TSVs are placed close, the bottom stress is approaching the peak stress.

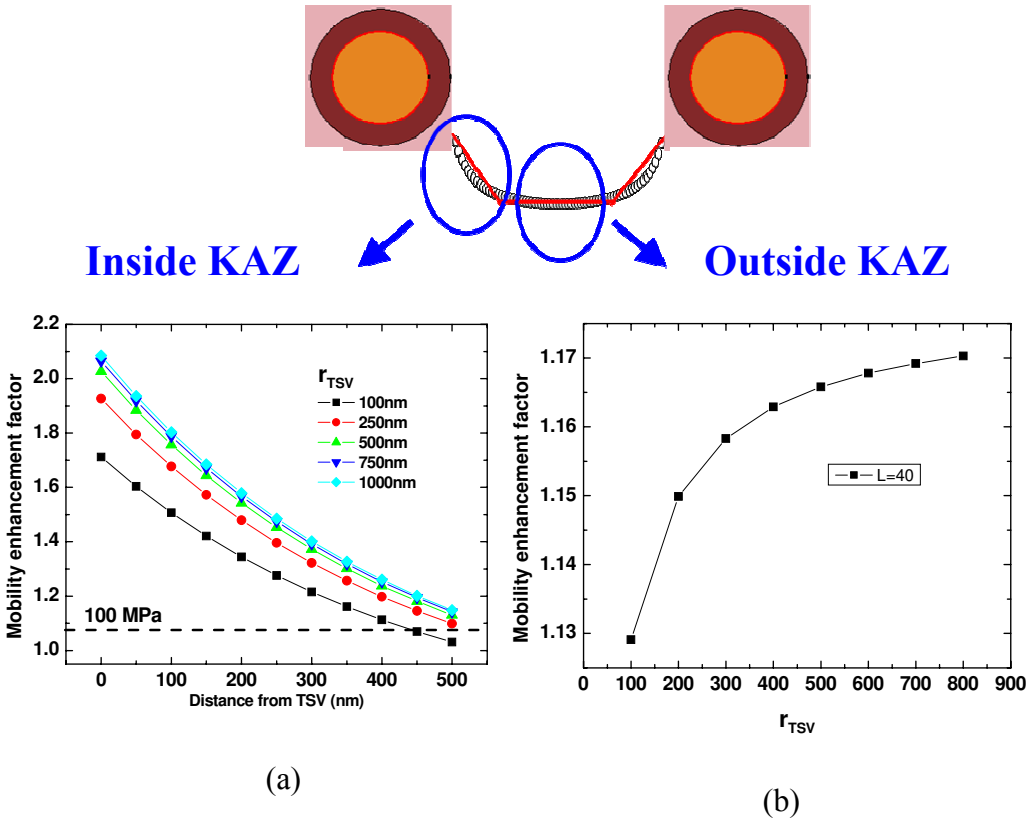
The impact of TSV thermal stress on carrier mobility is modeled by integrating the stress in a transistor in Eq. (4-15).

$$\frac{1}{\mu_L} = \frac{1}{L}\int_{x'}^{x'+L} \frac{1}{\mu(x)} dx \tag{4-15}$$

where x' is the distance from the edge of TSV to device and L is the channel length. $\mu(x)$ is the stress-induced mobility variation factor and is modeled in Eq. (4-16).

$$\frac{\mu}{\mu_0} = 1 + B \cdot \left[\exp\left(\frac{\Delta E}{kT}\right) - 1 \right] \tag{4-16}$$

where B are physical constant and different for holes and electrons. ΔE is the band splitting caused by the thermal stress. Detailed model derivation can be found in chapter two. On the other hand, since TSV stress is not intentionally



(a) (b)

Figure 4.13 characteristics for a device placed (a) inside the keep-away zone (b) outside the keep-away zone.

applied to impact the device performance, there's keep-away zone (KAZ) to keep devices unaffected by TSV stress. Inside keep-away zone, the stress is significant to impact the device performance, so there are no devices allowed in this zone, leading to more area cost during TSV process. Therefore, if the impact of thermal stress can be well predicted, KAZ can be utilized with stress-aware design for area efficiency. Shown as the schematic on top of Fig. 4.13, the keep-away zone can be defined in the region from peak stress to the bottom stress. Within keep-away zone, the device performance varies from location to location, while device performance is not sensitive to the stress effect outside the keep-away zone. Using the similar modeling approach, the impact of TSV-induced thermal stress is assessed. Fig. 4.13 (a) shows the mobility variation changes with the distance

from TSV edge within the keep-away zone. Mobility enhancement factor decays over the distance from the TSV edge; the farther the device is located the less stress effect and less mobility variation. Moreover, as the TSV radius increases, the stress effect becomes stronger and finally saturates. On the other hand, when the device is located outside the keep-away zone, the device performance is very stable and hardly unaffected by the stress. From the modeling perspective, the mobility variation is dominated by the bottom stress in Eq. (4-14). The stress out of the keep-away zone is assumed the same, leading to the identical impacts on the transistors. However, bottom stress is still affected by the radius of TSV conducting material. Fig. 4.13 (b) shows the mobility variation factor changes with TSV radius (r_{TSV}). As r_{TSV} increases, the bottom stress grows up towards saturation. This phenomenon is similar to that the stress is saturated in eSiGe technology when source/drain length increases, and the same modeling approach is applicable to TSV-induce thermal stress effect.

With the assistance of the stress model, the impact of TSV thermal stress on mobility variation is predicted. Fig. 4.14 shows the mobility variation with the distance from TSV edge. The mobility varies significantly inside keep-away zone, while stays stable out of keep-away zone. To keep devices unaffected from the thermal stress, the area of keep-away zone is required to keep the device performance stable, implying extra cost is needed during TSV process. The area of keep-away zone is calculated by the Eq. (4-17).

$$Area = \pi\left(r_{TSV}^2 + t_{ox}^2 + d_{KAZ}^2\right) \tag{4-17}$$

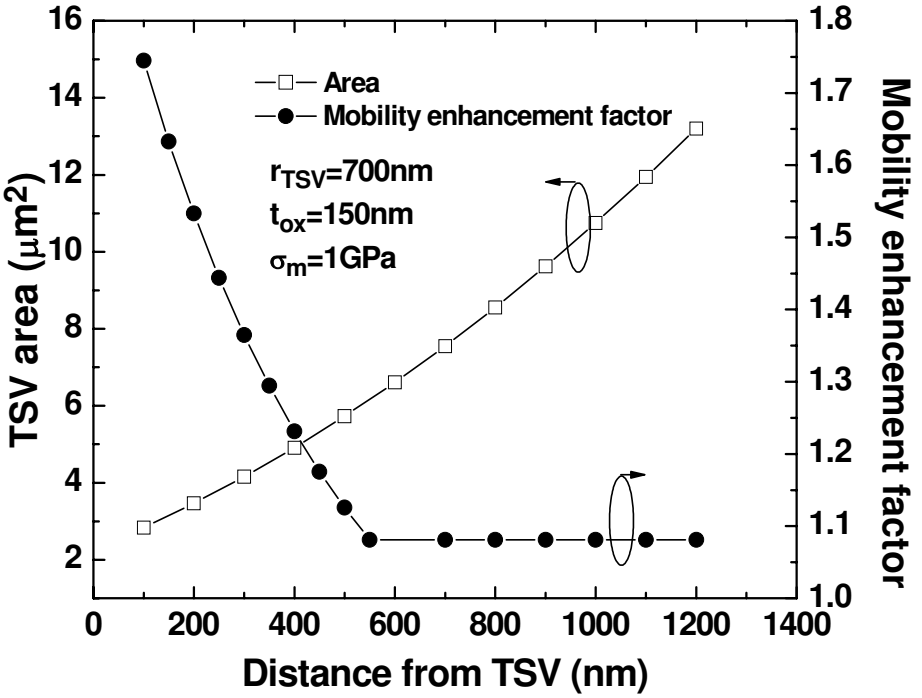


Figure 4.14. Trade-off between TSV area and the influence of mobility enhancement factor.

where d_{KAZ} is the distance from TSV edge. r_{TSV} and t_{ox} denote TSV metal radius and the thickness of barrier layer, respectively. More KAZ area reduces the impacts of thermal stress on devices, ensuring the stable process variation. However, the keep-away zone be utilized with stress-aware design approach if the mobility variation can be well modeled, illustrating the opportunities from joint device-design perspective.

4.4 Chapter Summary

During TSV implementation, the interconnect capacitance varies with the applied voltage and is sensitive to the radius of TSV, thickness of barrier layer, and the doping concentration in the substrate. TSV resistance also strongly depends on the radius of conducting material. To evaluate the impacts of TSV parameters on circuit performance, a piecewise capacitance model is proposed for

efficient simulation. Moreover, to minimize RC delay, a large TSV metal radius and a thicker barrier layer are required, indicating that a trade-off between the required TSV area and the interconnect performance.

On the other hand, thermal stress is observed due to the CTE mismatch between copper and silicon and impacts the neighboring transistors. The stress magnitude and distribution profile is dependent on TSV geometry parameters. For example, with increasing TSV radius, more stress is introduced and finally reaches saturation. Moreover, to minimize the impact of TSV stress, keep-away zone is adopted. However, keep-away zone consumes more silicon area and increases the cost of TSV fabrication. Therefore, an accurate model for mobility variation is required to predict the device performance and to be utilized with stress-aware design for area efficiency. Using the similar modeling approach in the eSiGe/DSL strain technology, the impact of TSV-induced thermal stress is well assessed.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Thesis Conclusion

With the rapid scaling of CMOS technology, manufacturing difficulties and emerging effects lead to design challenges, including the issues of power consumptions, mobility degradation and process variation. To conquer these challenges and continue the success on the leading products, circuit design must begin concurrently with advanced process development. Predictive modeling is essential for assisting early stage design exploration and bridges design communities with process technologies. The specific contributions of this dissertation include:

A new compact model of layout dependent stress effect for nanoscale CMOS technology and the method of layout decomposition for efficient model extraction are developed. Compact models of mobility, velocity, and threshold voltage under state-of-the-art strain technologies such as eSiGe and DSL stress techniques are presented. It physically captures the dependence on primary layout parameters, temperature, and other device characteristics. Moreover, the entire layout is decomposed in both vertical and lateral directions, reducing the analysis complexity while maintaining the accuracy. Circuit performance benchmark is verified at the 45nm node. The interaction between layout and circuit performance is accurately predicted by the new stress models. These solutions are developed and demonstrated in a 45nm design using restrictive design rules. They are comprehensively evaluated with TCAD simulations and published Si data. Based

on them, the impact of the stress effect is well assessed and scalable into future technology generations.

On the other hand, Fe-FET is considered as one of promising alternatives to traditional CMOS for further scaling. By integrating ferroelectric material into the traditional MOSFET structure as an internal voltage amplifier, Fe-FET becomes one of the promising candidates for energy efficient application. A new threshold voltage model for Fe-FET is proposed to capture the essence of steep subthreshold slope. It is scalable with technology specifications and indicates that a thicker ferroelectric material, higher channel doping, and thinner oxide thickness further improve the transition in the subthreshold region. Furthermore, it is further revealed that the impact of random dopant fluctuation (RDF) on leakage variability can be significantly suppressed in Fe-FET, by tuning the thickness of the ferroelectric layer. This implies the possibilities to suppress the intrinsic process variation by introducing the internal feedback loop in device level and meanwhile shed lights for device designers to construct a variation-insensitive device.

Moreover, through-silicon-via (TSV), a promising technology that enables 3D integration of chips with diverse functionalities, provides improved packing density, better noise immunity and faster speed due to reduced interconnect wire length. Most TSVs are filled with copper with a barrier layer inserted between metal and silicon substrate. This structure makes TSV a cylindrical voltage dependent MOS capacitor. A voltage-dependent TSV capacitance model is required to optimize 3D interconnect structure and to investigate the optimal

process and material choice for high-speed applications. A piecewise capacitance model is proposed for efficient simulation. Furthermore, due to the mismatch of coefficient of thermal expansion (CTE) between copper and silicon, thermal stress is observed, impacting the neighbor transistors and TSV capacitance. The thermal stress is essential during the TSV process. The impact of TSV stress is well assessed by the proposed stress model, supporting the interaction between silicon process and IC design at the early stage.

5.2 Future Work

The scaling of logic devices is slowing down in recent years, as silicon technology approaches physical and manufacturing limits. On the other side, research efforts on various types of memory devices are escalating, supporting increasingly higher data storage capability in a large scale system. These memory devices range from scaled SRAM, DRAM, and Flash memory that have been implemented in microprocessors today, to promising alternatives in the future, such as Magnetoresistive Random Access Memory (MRAM), Ferroelectric RAM (FeRAM), Phase Change Memory (PCM), and Resistive random-access memory (RRAM).

As compared to logic devices that are primarily field-effect transistors, memory devices are tremendously diverse in their physical mechanisms as well as performance characteristics. Some important features to evaluate a memory device include: read/write cycles, access time, data stability, and yield; additional concerns address the manufacturability, cost and technology compatibility. These common metrics significantly affect the design decision at circuit, architecture,

and system levels. However, current modeling efforts mainly focus on the unique physics of each technology and therefore, compact memory models are usually constrained to each individual type. Such a bottom-up approach creates the gap between memory technology developers and IC designers, especially when the types of promising memory devices rapidly increase in a heterogeneous system. To overcome this barrier, a top-down modeling methodology toward a universal memory model is required to enable convenient design benchmarking and optimization.

REFERENCE

[1] The International Technology Roadmap for Semiconductors (ITRS), 2009.

[2] E. E. Thompson, M. Armstrong, C. Auth, M. Alavi, M. Buehler, R. Chau, S. Cea, T. Chani, G. Glass, T. Hoffman, C.-H. Jan, C. Kenyon, J. Klaus, K. Kuhn, Z. Ma, B. Mcintyre, K. Mistry, A. Murthy, and B. Obradovic, R. Nagisetty, P. Nguyen, S. Sivakumar, R. Shaheed, L. Shifren, B. Tufts, S. Tyagi, M. Bohr, and Y. El-Mansy, "A 90-nm Logic Technology Featuring Strained-Silicon," *IEEE TED*, pp. 1790–1797, 2004,

[3] H. S. P. Wong, "Field effect transistors—From silicon MOSFETs to carbon nanotube FETs," in *Proc. MIEL,* vol. 1, 2002, pp. 103–107.

[4] A. Padilla, C. W. Yeung, C. Shin, C. Hu, and T.-J. King Liu, "Feedback FET: A novel transistor exhibiting steep switching behavior at low bias voltages," IEEE International Electron Devices Meeting Technical Digest, 2008.

[5] M. Horowitz, E. Alon, D. Patil, S. Naffziger, R. Kumar, K. Bernstein, "Scaling, Power, and the Future of CMOS," IEEE International Electron Devices Meeting Technical Digest, pp. 7-15 2005

[6] G.A. Salvatore, D. Bouvet, and A. M. Ionescu, "Demonstration of Subthreshold Swing Smaller than 60mV/decade in Fe-FET with P(VDF-TrFE)/SiO2 Gate Stack," in *IEDM*, pp. 167-170, 2008.

[7] M. J. M. Pelgrom, H. P. Tuinhout, and M. Vertregt, "Transistor matching in analog CMOS applications," in IEDM Tech. Dig., 1998, pp. 915–918.

[8] K. Rim, J.L. Hoyt, and J.F. Gibbons, "Fabrication and analysis of deep submicron strained-Si N-MOSFET's," TED, Vol. 47, no. 7, pp. 1406-1415, July 2000.

[9] J-S Lim, S.E. Thompson, J.G. Fossum, "Comparison of threshold-voltage shifts for uniaxial and biaxial tensile-stressed n-MOSFETs," *EDL*, vol. 25, no. 11, pp. 731-733, Nov. 2004.

[10] H. Nii, et al., "A 45nm High Performance Bulk Logic Platform Technology (CMOS6) using Ultra High NA(1.07) Immersion Lithography

with Hybrid Dual-Damascene Structure and Porous Low-k BEOL," in *IEDM*, pp. 685-688, 2006

[11] K. Ota, K. Sugihara, H. Sayama, T. Uchida, H. Oda, T.Eimori, H. Morimoto, and Y. Inoue, "Novel locally strained channel technique for high performance 55nm CMOS," in *IEDM*, pp.27-30, 2002.

[12] 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.*, 1999, pp. 827-830.

[13] R.A. Bianchi, G. Bouche, and O. Roux-dit-Buisson, "Accurate modeling of trench isolation induced mechanical stress effects on MOSFET electrical performance," *IEDM*, pp.117-120, 2002.

[14] H. Aikawa, E. Morifuji, T. Sanuki, T. Sawada, S. Kyoh, A. Sakata, M. Ohta, H.Yoshimura, T. Nakayama, M. Iwai, and F. Matsuoka, "Variability aware modeling and characterization in standard cell in 45 nm CMOS with stress enhancement technique," in *IEEE VLSI Symp,* pp.90-91, 2008.

[15] G. Eneman, P. Verheyen, R. Rooyackers, F. Nouri, L. Washington, R. Schreutelkamp, V. Moroz, L. Smith, A. D. Keersgieter, M. Jurczak, and K. D. Meyer, "Scalability of the Si1-xGex Source/Drain technology for the 45-nm technology node and beyond," *TED*, vol. 53, no. 7, pp.1647-1656, Jul. 2006.

[16] Taurus Tsuprem4, Manual, Oct. 2005. Version X-2005. 10.

[17] V. Moroz, G. Eneman, P. Verheyen, F. Nouri, L. Washington, L. Smith, M. Jurczakl, D. Pramanik, and X. Xu, "The impact of layout on stress-enhanced transistor performance," pp. 143-146, SISPAD 2005.

[18] A. B. Kahng, P. Sharma, and R.O. Topaloglu, "Exploiting STI stress for performance," pp. 83-90, ICCAD, 2007

[19] V. Joshi, B. Cline, D. Sylvester, D. Blaauw, K. Agarwal, "Leakage power reduction using stress-enhanced layouts," pp. 912-917, DAC 2008.

[20] K-W Su, Y-M Sheu, C-K Lin, S-J Yang, W-J Liang, X Xi, C-S Chiang, J-K Her, Y-T Chia, C H. Dim, and C. Hu, "A scaleable model for STI mechanical stress effect on layout dependence of MOS electrical characteristics," Custom Integrated Circuits Conference, pp. 245-248, 2003.

[21] BSIM4 Manual, Univ. California, Berkeley, CA, 2005.

[22] C. E. Murray, "Mechanics of edge effects in anisotropic thin film/substrate systems," *Journal of Applied Physics*, vol. 100, 103532, 2006.

[23] M. V. Dunga, C-H Lin, X Xi, D.D. Lu, A.M. Niknejad, and C. Hu, "Modeling advanced FET technology in a compact model," *TED*, vol. 53, No. 9, pp.1971-1978, Sept. 2006.

[24] Taurus Medici, Manual, June 2006. Version Y-2006.06.

[25] X-W Lin, "Modeling of Proximity Effects in Nanometer MOSFET's," *IEEE/ACM Workshop on Compact Variability Modeling* 2008.

[26] C. S. Smith, "Piezoresistance effect in Germanium and Silicon," Phys. Rev., Vol.94, no. 1, pp.42-49,1954.

[27] Y. Kanda, "A graphical representation of the piezoresistance coefficients in silicon," *TED*, vol. 29, No. 1, pp.64-70, Jan. 1982.

[28] E. Ungersboeck, S. Dhar, G. Karlowatz, V. Sverdlov, H. Kosina, and S. Selberherr, "The effect of general strain on the band structure and electron mobility of silicon," *TED*, vol. 54, No. 9, pp.2183-2190, Sept. 2007.

[29] J.L. Egley and D. Chidambarrao, "Strain Effects on Device Characteristics: Implementation in Drift-Diffusion Simulators," *Solid-State Electronics,* 36(12), pp. 1653-1664, 1993.

[30] Sentaurus Device, Manual, June 2005. Version Y-2006. 06.

[31] G.L. Bir and G. E. Pikus, "Symmetry and Strain-Induced Effects in Semiconductors," Wiley, New York, 1974.

[32] F. Payet, F. Payet, F. Boeuf, C. Ortolland and T. Skotnicki, "Nonuniform Mobility-Enhancement Techniques and Their Impact on Device Performance," *TED*, Vol. 55, no. 4, pp. 1050-1057, April 2008

[33] D. Sinitsky, "Physics of future very large-sclae integration (VLSI) MOSFETs," Ph.D. dissertation, Univ. California, Berkeley, CA, 1997.

[34] W. Zhang, J. G. Fossum, "On the threshold voltage of strained-Si–$Si_{1-x}Ge_x$ MOSFETs," *TED*, Vol. 52, no. 2, pp.263-268, Feb. 2005.

[35] S. M. Sze, Semiconductor Devices: Physics and Technology, 2nd ed., New York: John Wiley & Sons, 2002.

[36] Z.-H. Liu, C. Hu, J.-H. Huang, T.-Y. Chan, M.-C. Jeng, P. K. Ko, and Y.C. Cheng, "Threshold Voltage Model for Deep-Submicrometer MOSFET's," TED, Vol. 40, no. 1, pp. 86-95, Jan. 1993.

[37] W. Zhao, and Y. Cao, "New generation of predictive technology model for sub-45nm design exploration," *IEEE TED*, vol. 53, no. 11, pp. 2816-2823, Nov. 2006. (Available at http://www.eas.asu.edu/~ptm)

[38] A. R. Subramaniam, R. Singhal, C.-C. Wang, and Y. Cao, "Design Rule Optimization of Regular Layout for Leakage Reduction in Nanoscale Design," in *ASP-DAC*, pp. 474-479, 2008.

[39] H. Tsuno, K. Anzai, M. Matsumura, S. Minami, A. Honjo, H. Koike, Y. Hiura, A. Takeo, W. Fu, Y. Fukuzaki, M. Kanno, H. Ansai and N. Nagashima, "Advanced analysis and modeling of MOSFET characteristic fluctuation caused by layout variation," in *IEEE VLSI Symp,* pp.204-205, 2007.

[40] F. Andrieu, T. Ernst, F. Lime, F. Rochette, K. Romanjek, S. Barraud, C. Ravit, F. Boeuf, M. Jurczak, M. Casse, O. Weber, L. Brévard, G. Reimbold, G. Ghibaudo, S. Deleonibus, "Experimental and Comparative Investigation of Low and High Field Transport in Substrate- and Process-Induced Strained Nanoscaled MOSFETs," in *IEEE VLSI Symp,* pp.176-177, 2005.

[41] Marc J. Madou, "Fundamentals of microfabrication: the science of miniaturization," pp.198, CRC Press, 2nd edition, 2002.

[42] R. Li, L. Yu, H. Xin, Y. Dong, K. Tao, and C. Wang, "A Comprehensive Study of Reducing the STI Mechanical Stress Effect on Channel-Width-Dependent Idsat," Semiconductor Science and Technology, pp. 1292-1297, 2007.

[43] V. V. Zhirnov and R. K. Cavin, "Negative Capacitance to the Rescue?" Nanoelectronics, Vol.3, pp. 77-78, 2008.

[44] K. Gopalakrishnan and P. B. Griffin, and J. D. Plummer, "I-MOS: A Novel Semiconductor Device with a Subthreshold Slope Lower than kT/q," in *IEDM Tech. Dig.*, 2002, pp. 289–292.

[45] C. W. Yeung, A. Padilla, T-J King Liu, and C. Hu, "Programming Characteristics of Steep Turn-on/off Feddback FET (FBFET)," in *VLSI Symp.*, pp.176-177, 2009.

[46] S. Salahuddin, and S. Datta, "Use of Negative Capacitance to Provide Voltage Amplification for Low Power Nanoscale Devices," Nano Letters, Vol.8, No.2, pp. 405-410, 2008.

[47] S. Salahuddin, and S. Datta, "Can the Subthreshold Swing in a Classical FET Be Lowered below 60 mV/decade?" in *IEDM*, pp. 693.696, 2008.

[48] K. Bernstein, D. J. Frank, A. E. Gattiker, W. Haensch, B. L. Ji, S. R. Nassif, E. J. Nowak, and N. J. Rohrer, "High-performance CMOS variability in the 65-nm regime and beyond," IBM Journal of Research & Development, vol. 50, no. 4/5, pp. 433.449, Jul./Sep., 2006.

[49] K. Takeuchi, T. Fukai, A. T. Putra, A. Nishida, S. Kamohara, and T. Hiramoto "Understanding Random Threshold Voltage Fluctuation by Comparing Multiple Fabs and Technologies," in *IEDM*, pp. 467-470, 2007.

[50] R.A. Bianchi, et al., "Accurate modeling of trench isolation induced mechanical stress effects on MOSFET electrical performance," IEDM, pp.117-120, 2002.

[51] PSP 103.1 Manual, Arizona State University, Tempe, AZ, 2009.

[52] Y. Taur, and T. H. Ning, "Fundamentals of Modern VLSI Devices," Cambridge University Press, 1998.

[53] A. Wang, and A. Chandrakasan, "A 180mV FFT Processor Using Subthreshold Circuit Techniques," in *ISSCC*, 16.4, 2004.

[54] S. Hanson, M. Seok, D. Sylvester, and D. Blaauw, "Nanometer Device Scaling in Subthreshold Circuits," in *DAC*, pp. 700-705, 2007.

[55] B. H. Calhoun, A. P. Chandrakasa, "Ultra-dynamic voltage scaling (UDVS) using sub-threshold operation and local voltage dithering," IEEE Journal of Solid-State Circuits (JSSC), vol. 41, no. 1, pp.238-245, Jan. 2006.

[56] G.A. Salvatore, D. Bouvet, and A. M. Ionescu, "Demonstration of Subthreshold Swing Smaller than 60mV/decade in Fe-FET with P(VDF-TrFE)/SiO2 Gate Stack," in *IEDM*, pp. 167-170, 2008.

[57] P. Morrow, B. Black, M. J Kobrinsky, S. Muthukumar, D. Nelson, C.-M. Park, C. Webb, "Design and Fabrication of 3D Microprocessors," Material Research Society Symposium Proceedings, Vol. 970, pp. 91-102, 2007, Enabling Technologies for 3-D Integration.

[58] P. D. Moor, W. Ruythooren, P. Soussan, B. Swinnen, K. Baert, C. V. Hoof, and E. Beyne, "Recent Advances in 3D Integration at IMEC," Materials Research Society Symposium Proceedings, v 970, p 3-12, 2007, Enabling Technologies for 3-D Integration.

[59] P.S. Andry, C. Tsang, E. Sprogis, C. Patel, S. L. Wright and B. C. Webb, "A CMOS-compatible Process for Fabricating Electrical Through-vias in Silicon," Proceedings of the 56th Electronic Components and Technology Conference, 2006, pp. 831-837.

[60] B. Vandevelde, C. Okoro, M. Gonzalez, B. Swinnen and E. Beyne, "Thermo-mechanics of 3D-Wafer Level and 3D Stacked IC Packaging Technologies" 9th International Conference on Thermal, Mechanical and Multi-Physics Simulations and Experiments in Microelectronics and Micro-Systems, (EuroSimE), 2008, pp. 1-7.

[61] K. Takahashi and M. Sekiguchi, "Through Silicon Via and 3-D Wafer/Chip Stacking Technology," 2006 Symposium On VLSI Circuits, Digest of Technical Papers, 2006, pp. 89-92.

[62] A. P. Karmarkar, X. Xu and V. Moroz, "Performanace and Reliability Analysis of 3D-Integration Structures Employing Through Silicon Via (TSV)," in IEEE International Reliability Physics Symposium (IRPS), pp.682-687, 2009.

[63] C. Xu, H. Li, R Suaya and K. Banerjee, "Compact AC Modeling and Analysis of Cu, W, and CNT based Through-Silicon Vias (TSVs) in 3-D ICs" in IEDM Tech. Dig., pp. 521–524, 2009.

[64] C. -K. Hu, B. Luther, F. B. Kaufman, J. Hummel, C. Uzoh and D. J. Pearson, "Copper interconnection integration and reliability," Elsevier Thin Solid Films Vol. 262, Issues 1-2, pp. 84-92, 1995.

[65] B. Swinnen, W. Ruythooren, P. D. Moor, L. Bogaerts, L. Carbonell, K. D. Munck, B. Eyckens, S. Stoukatch, D. S. Tezcan, Z. Tőkei, J. Vaes, J. Van Aelst, and E. Beyne, "3D integration by Cu-Cu thermocompression bonding of extremely thinned bulk-Si die containing 10 μ m pitch through-Si vias," Technical Digest of the International Electron Device Meeting, pp. 371-374 2006.

[66] J. Van Olmen , A. Mercha, G. Katti1, C. Huyghebaert, J. V. Aelst, E. Seppala, Z. Chao, S. Armini, J. Vaes, R. C. Teixeira, M. V. Cauwenberghe, P. Verdonck. K. Verhemeldonck, A. Jourdain, W. Ruythooren, M. de Potter de ten Broeck, A. Opdebeeck, T. Chiarella, B. Parvais, I. Debusschere, T.Y. Hoffmann, B. De Wachter, W. Dehaene1, M. Stucchi, M. Rakowski, Ph. Soussan, R. Cartuyvels, E. Beyne, S. Biesemans, B. Swinnen, "3D Stacked IC Demonstration using a Through Silicon Via First Approach," Technical Digest of the International Electron Device Meeting, pp. 603-606, 2008.

[67] B. Wunderle, R. Mrossko, O. Wittler, E. Kaulfersch, P. Ramm, B. Michel, H. Reichl, "Thermo-Mechanical Reliability of 3D-Integrated Microstructures in Stacked Silicon", Material Research Society Symposium Proceedings, Vol. 970, 2007

[68] C.S. Selvanayagam, J.H. Lau, X. Zhang, S. Seah, K. Vaidyanathan, T.C. Chai, "Nonlinear thermal stress/strain analyses of copper filled TSV

(through silicon via) and their flip-chip microbumps," in Electronic Components and Technology Conference, pp. 1073-1081, 2008.

[69] G. Katti, M. Stucchi, K. D. Meyer, and W. Dehaene, "Electrical Modeling and Characterization of Through Silicon via for Three-Dimensional ICs," IEEE Transaction on Electron Devices, Vol. 57, No. 1, pp.256-262, 2010.

[70] K. H. Lu, X. Zhang, S.-K. Ryu, J. Im, R. Huang, and P. S. Ho, "Thermo-Mechanical Reliability of 3-D ICs containing Through Silicon Vias," in IEEE Electronic Components and Technology Conference, pp. 630-634, 2009.

[71] C. S. Selvanayagam, X. Zhang, R. Rajoo, D. Pinjala, "Modelling Stress in Silicon with TSVs and its Effect on Mobility," in IEEE Electronics Packaging Technology Conference, pp. 612-618, 2009.

[72] J.-S. Yang, K. Athikulwongse, Y.-J. Lee, S. K. Lim, and D. Z. Pan, "TSV Stress Aware Timing Analysis with Applications to 3D-IC Layout Optimization," in Proceedings of the Design Automation Conference, pp.803-806 2010.

[73] S.E. Thompson, G. Sun, Y. S. Choi, T. Nishida,"Uniaxial-Process-Induced Strained-Si: Extending the CMOS Roadmap", IEEE Transaction on Electron Device Vol. 53, pp. 1010-1020, 2006.

[74] T. Chidambaram, C. McDonough, R. Geer, and W. Wang, "TSV Stress Testing and Modeling for 3D IC Applications," in Proceedings of Physical and Failure Analysis of Integrated Circuits, 2009.

[75] L. Smith, V. Moroz, G. Eneman, P. Verheyen, F. Nouri, L. Washington, M. Jurczak, O. Penzin, D. Pramanik, and K. D. Meyer, "Exploring the Limits of Stress-Enhanced Hole Mobility, " IEEE Electron Device Letters, Vol. 26, pp.652-654, 2005.

[76] C.-C. Wang, W. Zhao, F. Liu, M. Chen, and Y. Cao, "Compact Modeling of Stress Effects in Scaled CMOS," in Proc. of IEEE Int'l Conference on Simulation of Semiconductor Processes and Devices (SISPAD), pp.131-134 , 2009.

[77] C.-C. Wang, W. Zhao, F. Liu, M. Chen, and Y. Cao, “Modeling of Layout-Dependent Stress Effect in CMOS Design," in Proc. IEEE/ACM International Conference on Computer Aided Design (ICCAD), pp. 513-520, 2009.

[78] C.-C. Wang, Y. Ye, and Y. Cao, “Compact modeling of Fe-FET with the implication on variation-insensitive design,” to appear in Proc. IEEE Int’l Conference on Simulation of Semiconductor Processes and Devices (SISPAD), 2010.

[79] Y. Cao, C.-C. Wang, Y. Ye, S. Gummalla, C. Chakrabarti, “Intrinsic Variability in Nano-CMOS Design and Beyond,” to be published at International Electron Devices Meeting (IEDM), 2010. [invited]

[80] Y. Ye, S. Gummalla, C.-C. Wang, C. Chakrabarti and Y. Cao, “Random Variability Modeling and Its Impact on Scaled CMOS Circuits,” to be published in Journal of Computational Electronics (JCE).