Transistor/Gate Level Reliability Modeling

LIU XU

School of Electrical & Electronic Engineering

A thesis submitted to the Nanyang Technological University

in particular fulfillment of the requirement for the degree of

Doctor of Philosophy

2019

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

[16/01/2019] [LIU XU]

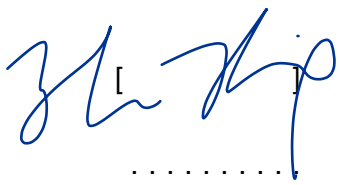
. . . . . .

[16/01/2019] [LIU XU]

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

[16/01/2019] [ ]



. . . .

[16/01/2019] [Zhou Xing]

Acknowledgement

First of all, I would like to give my sincere thanks to my supervisor, Dr. Zhou Xing, for opening the door of the research work. Over the past four years, he has been helping me so much, both academically and mentally. His patient guidance leads me in the correct direction. I acquire sufficient knowledge about the transistor level modeling, which lays the foundation of my research.

Equally importantly, my co-supervisor, Prof. Dr. Ing, Ulf Schlichtmann from Technical University of Munich (TUM) took great care of me during my stay in Munich. His advice was inspiring and innovative. With his help I was able to incorporate my device level knowledge into the gate level. I want to say thanks to him.

I give my great gratitude to my senior team members, Dr. Chiah Siau Ben and Dr. Zhou Hongtao from Nanyang Technological University (NTU). They did so much favor to me, especially for software application and bug fixing. My research work would have been delayed without them. My teammate Arjun Ajaykumar, who joined the team with me in the meantime, is a very good friend. We keep in track with each other all the way, sharing ideas and improving ourselves together. I want to say, thanks, my brother.

My colleagues in Munich also gave me adequate support from many aspects. I want to thank M.Sc. Alessandro Bernardini, not only for the suggestions and review of my work, but also for the delicate espresso and the fantastic hot spring; I owe my thanks to M.Sc. Maximilian Ulrich Neuner for the friendship and games we played together; My dear teammates, Alexandra Listl, and Shushanik Karapetyan accompanied me with ”Die Kicker” game almost

everyday, I really appreciate that; M.Sc. Andreas Hermann and Liu Chunfeng inspired with the enthusiasm for sports; M.Sc. Petra Maier is always optimistic with a strong sense of humour, and I really admire her; M.Sc. Zhang Li is keen on the research work, and she pushed me forward when I was lazy, I must say she's amazing; M.Sc. Hu Yong made my first few days' life in Munich more convenient, I would like to thank him for everything; my friend M.Sc. Chou Pang-yen is kind, and I appreciate it so much that we joined so many activities together. I would like to give a special thanks to Dr. Dominik Lorenz: we never met each other, but I've been reading his thesis, from where I enrich my knowledge of gate level modelling. In fact, this thesis lies on the frame work he has done.

My family members also play key roles during the past years. Their unconditional support is a most important thing in my life. Thanks, papa, mama, and my brother. Unfortunately my grandma, who brought me up, cannot witness that day when I become a Ph.D. Hopefully she could enjoy peace in heaven. My niece came into this world in last April- she must be an angel who helps mama out of the sorrow.

Abstract

The development of CMOS technology is a double-edged sword: for one thing, it provides faster, lower power-consuming, and smaller-size devices; for another, reliability issues such as Negative Bias Temperature Instability (NBTI) and Hot Carrier Injection (HCI) become severer, resulting in device/gate performance degradation. Compact and accurate modelling of these issues is required in aid of IC design. Reliability modelling could be done at the transistor level: the device parameter shift Δp(such as the threshold voltage ΔV_{th}) or performance (for example, the current I_{on}) shift are described as time-related (t_{stress}) functions. At the gate level, the delay degradation is obtained with the insertion of the transistor-level parameter/performance shift equations. The timing analysis at the gate level is done with methods such as the static timing analysis (STA) rather than the time-consuming SPICE simulation, but the latter could be adopted for characterisation owing to its accuracy. The reliability models at the transistor level and the gate level interact with each other: the gate level model takes Δp as an input, which must be derived from the transistor level model; the transistor level model determines Δp with the value of t_{stress}, which is a statistical parameter in the gate level model. The models achieve three goals: simplicity, time efficiency, and accuracy. The gate model called "Aging-Gate" has been built up to take both NBTI (ΔV_{th}) and HCI (ΔI_{on}) into account; algorithms for t_{stress} determination could be adopted for transistor level modelling. To simplify this model, HCI is also modelled with ΔV_{th} such that the gate model equations could be simplified. A surface-potential (ϕ_s)-based transistor model is also available, and t_{stress} is to be incorporated in for ΔV_{th} determination. The so called "recovery effect" in NBTI results in iterative methods that suffer from time-efficiency problems. The new model provides a compact algorithm that

achieves equivalent accuracy, which significantly improves the time efficiency. To ensure the efficiency, the original AgingGate model ignores the recovery effect by adopting a simple DC NBTI model, but this leads to overestimation of the delay degradation. The new model alleviates this since the recovery effect is considered, yet the time efficiency is maintained in that the new model is as fast as the DC model. Therefore, the accuracy is improved.

Contents

List of Figures

List of Tables

Symbols

Symbol	REMARK
N_{IT}	Interface states density
V_{th}	Threshold voltage
ϕ_s	Surface potential
t_{stress}	Time duration that a device is under stress
μ	Mobility
I_{on}	MOSFET on current
t	Time
ΔV_{th}	Threshold voltage shift
ΔI_{on}	On current shift
E_{ox}	Electric field strength in the gate oxide
I_{sat}	Saturation current of a MOSFET
g_m	Transconductance
V_{gs}	Gate-to-source voltage
$\Delta V_{th}(t)$	Threshold voltage shift after a time duration of t
V_g	Voltage applied to the gate terminal of a MOSFET
T	Temperature in Kelvin
k_B	Boltzmann constant
L_{eff}	Effective channel length
W	MOSFET channel width

Continued on next page

Table 1 – continued from previous page

Symbol	REMARK
ϕ_{it}	Critical energy to create interface states
λ_e	Channel mean free path of an electron
E_m	Lateral electric field strength in the velocity saturation region
V_{ds}	Drain-to-source voltage
V_{dsat}	Drain-to-source saturation voltage
t_{ox}	Gate oxide thickness
x_j	Drain junction depth
E_{sat}	Lateral electric field strength at velocity saturation
V_b	Voltage applied to the bulk terminal of a MOSFET
V_s	Voltage applied to the source terminal of a MOSFET
V_d	Voltage applied to the drain terminal of a MOSFET
V_{gb}	Gate-to-bulk voltage
V_{FB}	Flatband voltage
ϕ_{MS}	Work functions' difference between the poly gate and the bulk of a MOSFET
V_{ox}	Voltage drop across the gate oxide
v_{tm}	Thermal voltage
N_{ch}	Channel doping concentration
N_g	Poly gate doping concentration
n_i	Intrinsic electron concentration
q	Unit charge
C_{ox}	Gate oxide capacitance per area
ϵ_{ox}	Permittivity of S_iO_2
Q_g	Gate charge density (per area)
Q_{ox}	Effective oxide charge density (per area)
Q_{sc}	Bulk charge density (per area)
ϕ_p	Surface potential in the poly gate

Continued on next page

Table 1 – continued from previous page

Symbol	REMARK
N_A	Concentration of the acceptor
N_D	Concentration of the donor
ϕ_F	Fermi potential
V_{cb}	Channel-to-bulk voltage
ϕ	Potential representing the band bending
ρ	Charge density
E_{si}	Vertical electric field strength at the oxide-bulk interface
ϵ_{si}	Permittivity of S_i
γ	Body factor
$\phi_{s,acc}$	Surface potential in the accumulation region
$\phi_{s,dep}$	Surface potential in the depletion region
$\phi_{s,str}$	Surface potential in the strong inversion region
$\phi_s(y)$	Surface potential at the position in the channel marked "y"
Q_i	Inversion charge density (per area)
Q_b	Depletion charge density (per area)
I_{drift}	Drift current
I_{diff}	Diffusion current
$I_{dlinear}$	MOSFET current in the linear region
μ_{eff0}	Channel mobility
$V_{gt,s}$	Inversion charge density at the source side normalised to C_{ox}
$A_{b,s}$	Bulk charge factor at the source side
$\Delta\phi_s$	Surface potential difference between the drain and source sides
ν_{sat}	Saturation velocity
t_{deg}	Time period for device degradation
$I_d(t)$	Drain current at time t
$I_{sub}(t)$	Substrate current at time t

Continued on next page

Table 1 – continued from previous page

Symbol	REMARK
$I_g(t)$	Gate current at time t
Δp	Shift of parameter p
C_L	Output load capacitance
s_{in}	Input slope
s_{out}	Output slope
TD	Transition density
D_{aged}	Degraded gate delay
D_{fresh}	Gate delay before degradation
ΔD	Delay degradation
V_{dd}	Supple voltage of a circuit
V_{supply}	Supply voltage
k_F	NBTI forward dissociation rate constant
N_0	initial $S_i - H$ bond density
k_r	NBTI annealing rate constant
N_H	Concentration of the hydrogen species
δ	NBTI reaction interface thickness
$N_H^{(0)}$	Hydrogen species density at the $S_i - S_iO_2$ interface
D_H	Diffusion coefficient of the hydrogen species
$N_H^{(x)}$	Hydrogen species density at a distance of x from the interface
$N_{IT}(t)$	Interface states density at time t
t_0	NBTI recovery time
ξ	Possessing a value of 0.5 for double-sided diffusion
t_{stress}	Stress time
$t_{recovery}$	Recovery time
t_{total}	Total time including t_{stress} and $t_{recovery}$
SF	The stress factor

Continued on next page

Table 1 – continued from previous page

Symbol	REMARK
SP	The signal probability
f	Frequency
t_{sim}	Time required to accomplish a simulation
$\Delta V_{th,DC}(t_{total})$	DC case NBTI caused threshold voltage shift after t_{total}
$\Delta V_{th,AC}(t_{total})$	AC case NBTI caused threshold voltage shift after t_{total}
G	The set of all transistors of a digital gate
P	The set of parameters for reliability modelling
$\Delta V_{th,o}$	Threshold voltage shift without the consideration of recovery
$\Delta V_{th,w}$	Threshold voltage shift with recovery considered
P_{on}	The probability that the gate terminal of a PMOSFET is biased to low
P_i	PMOSFET of interest for NBTI modelling
SF_i	Stress factor of P_i
P_j	PMOSFET marked with j on the conducting path
$P_{on,k}$	The probability that a PMOSFET marked with ”k” is on
N_{hole}	inversion hole density (per area) of a PMOSFET
E_a	Activation energy of NBTI
ΔI_{on}	On current degradation
$I_{on,fresh}$	On current of a fresh MOSFET (without degradation)
$I_{on,aged}$	On current of an aged MOSFET
$E_{a,HCI}$	Activation energy of HCI
V_{deg}	Voltage source to model HCI influence on V_{th}
I_{deg}	Current source to model HCI impact on channel mobility
$N_{on,k}$	Probability that an NMOSFET marked ”k” is on
N_i	An NMOSFET marked ”i” for HCI modelling
SF_i	Probability that N_i suffers from HCI
μ_{fresh}	Mobility before HCI degradation

Continued on next page

Table 1 – continued from previous page

Symbol	REMARK
μ_{aged}	Mobility after HCI degradation
$V_{th,NBTI}$	NBTI induced threshold voltage shift
$V_{th,HCI}$	HCI induced threshold voltage shift
Δl	Length of the velocity saturation region
$V_{ds,sat,V_{th}}$	Saturation V_{ds} from the V_{th}-based model
V_{ds,sat,ϕ_s}	Saturation V_{ds} from the ϕ-based model
$I_{ds,0}$	Drain-source current for V_{th} determination
$V_{th,linear}$	Threshold voltage in the linear region
$I_{critical}$	Critical current value for V_{th} extraction

Chapter 1

Introduction

In circuit/logic designs, the mainstream CMOS (Complementary Metal Oxide Semiconductor) industry has been benefitting from transistor compact models (CMs) calibrated to a ”golden die” from a fab. Based on the ”nominal” device model, transistor/gate level simulations are conducted to evaluate the given performance figures-of-merit (FOM), which is the current paradigm of chip design/fabrication. As the industry practice for decades, this paradigm has been suffering from more and more reliability issues due to the continuous scaling of MOS devices. For one thing, device degradation is inevitable; for another, circuit performance targets must be met and a satisfying yield has to be maintained. This contradiction may be the bottleneck for future IC industry development. As a consequence, research and innovation on reliability modelling become urgent, aiming at guaranteeing the performance and yield of the coming generation chips. In fact, reliability ranks the top research priorities by the Semiconductor Research Corporation (SRC).

1.1 Motivation

Benefiting from the constant technology scaling, integrated circuit (IC) is able to achieve smaller device dimensions and higher operating frequency. However, problems arise at the mean time since devices are exposed to higher operating temperature (due to thermal stress) and higher electric field, which leads to faster device aging [1]. These reliability issues influence both transistors and circuits: at the transistor level, devices become slower owing to parameter changes such as the creation of interface states (N_{IT}); since the circuit delay is determined by the charging/discharging speed of the transistors, slower transistors lead to delay degradation at the gate level. As is known, a circuit is restricted with time constraints to meet the frequency requirements; the delay degradation may violate these constraints, resulting in malfunctions of the circuit.

Electronic Design Automation (EDA) tools (such as SPICE) are of great importance for IC designers in that they provide analyses and synthesis at the circuit level. Simulations could be done to predict the circuit performance. Since transistors are the basic components of circuits, transistor level reliability study will help gate level reliability analysis.

A transistor-level static compact model is the foundation of the research work. The V_{th}-based models requires large numbers of mathematical smoothing functions and empirical model parameters; besides, models like the BSIM [2] encounters Gummel symmetry problems [3]. The charge-based models mainly serve the analog application since they are able to describe the weak inversion region precisely [4, 5, 6]; however, they still rely on numbers of mathematical smoothing functions. The surface potential (ϕ_s)-based model method is physic-based. It avoids problems in the aforementioned two modelling methods. Besides, it is scalable and accurate. Therefore, ϕ_s-based compact model is adopted for the reliability study.

The two reliability problems-Negative Bias Temperature Instability (NBTI) and Hot Carrier Injection (HCI) are both $stress$ and $time$ related: the stressing voltages applied to the devices terminals trigger the aging process, the seriousness of which depends on the stressing time (t_{stress}). Device modelling parameters such as V_{th} and mobility (μ) will change as a result.

These parameter shifts lead to transistor performance (such as I_{on}) degradation. Compared with a "fresh" device, a degraded one consumes more time to transfer charges of the same amount, which means it is "slower". Slower devices requires longer time to charge/discharge the digital circuit. Gate delay is thus longer. With time going, circuit timing specification may be violated [7, 8, 9].

Simulation tools such as SPICE are capable of handling small size circuits with high accuracy. However, for circuits with tens of thousands of gates, SPICE simulation becomes very time consuming. For complex circuits, the timing analysis could only be done at the gate level with methods such as Static Timing Analysis (STA). STA improves the time efficiency at the sacrifice of a slight trade-off of accuracy. As is mentioned in [10], STA achieves a three orders of magnitude speed-up on average, while the mean error is below 7%. Hence, STA is the state-of-art for circuit timing analysis. SPICE simulation is adopted for the verification of the critical path delay obtained from STA.

As could be seen later in Chapter 2, some gate level reliability models build up the simple relation (some are even linear) between gate-level delay degradation (ΔD) and transistor-level parameter (current, threshold voltage, and etc) degradation (Δp); the accuracies of those models have been proven. An important point is that the accuracy and the efficiency of the gate-level models rely on those of the transistor-level models. A key point is that Δp is time (t)-related. As will be described in Chapter 3, NBTI is either described with a simple DC model (neglecting the fact that a device is turned on/off alternately such that it is actually an AC case) or with an iterative AC model; the former results in overestimation while the latter suffers from an efficiency problem. Besides, Chapter 3 and Chapter 4 will also show that the transistor-level models require the aid of subcircuits (voltage source/current source are added into the original circuit), which leads to the circuit complexity. Our motivation is to build a transistor level reliability model such that it could accurately and efficiently capture the $\Delta p - t$ relation. Only in this way will the accuracy and efficiency of the gate level reliability model be secured. We are going to take the advantages of both the ϕ_s based device-level model and the STA based gate model. As a new parameter, time (t) is incorporated into the ϕ_s based core model. At the transistor level, compact model equations are to be derived to model

the parameter shift (Δp). Since most gate level models directly link these shifts with the gate delay degradation (ΔD), accurate transistor level reliability models are required.

1.2 Objective

The objectives of this research includes:

- Develop a ϕ_s based compact model **of** ΔV_{th} due to NBTI. The model must be concise, precise, and scalable. Meanwhile, it must be highly time-efficient.
- Develop a ϕ_s based compact model **of** ΔV_{th} owing to HCI. Accuracy and Scalability must be guaranteed.

1.3 Major Contributions of the Thesis

The major contributions of the thesis are as follows:

- Proposed a new compact modelling method for the AC-case NBTI to replace the time-consuming iterative method. The new model is analytical and generates the same results as the iterative method, but the time efficiency is improved by orders.
- Proposed a ϕ_s based compact modelling method for HCI. Compared with the old-style ΔI_{on} model, it links t directly with ΔV_{th}. The new model is scalable and more precise. Besides, as a model parameter, ΔV_{th} is easer to manipulate in gate level modelling.
- Simplified the AgingGate model with a single device model parameter shift ΔV_{th} by eliminating ΔI_{on} (since I_{on} is a performance parameter rather than model parameter, and it does not appear in the device model card). The recovery effect of NBTI is added in without affecting the time efficiency; in addition, the overlapping part of NBTI and HCI is removed. The overall precision of the model might be improved.

1.4 Organization of the report

The thesis comes in 5 chapters. Apart from the introduction in Chapter 1, the thesis is organised in the following style:

- Chapter 2 briefly reviews the modelling methods of NBTI and HCI. Then the reliability-related background knowledge of the ϕ_s-based transistor level model is introduced. In addition, fundamentals of the timing analysis at the gate level, including the state-of-the-art of the reliability-relevant timing analysis is brought in. The problems arises from models at both levels will be discussed, the solution of which are the targets of this thesis.

- Chapter 3 firstly describes the reaction-diffusion (RD) based models of NBTI for both DC and AC cases. Their advantages and disadvantages are discussed. Then the simulation results for both scenarios are compared, based on which a compact model for AC case is obtained. In addition, the role that NBTI plays in gate level modelling is introduced, according to the AgingGate model (mainly for t_{stress} determination).

- Chapter 4 starts with the I_{on} based HCI models and their drawbacks due to which they are not suitable as the state-of-the-art. Then the 2-D RD model is introduced to explain the time-coefficient dependence of HCI. ϕ_s is finally inserted to build up the new model, in which ΔV_{th} is modelled instead of ΔI_{on}. Again we take advantage of the AgingGate model to determine the stressing time of HCI. Though beyond the scope of this thesis, we propose a methodology to simplify the AgingGate model. Meanwhile, since the ϕ_s based model does not include the parameter V_{th}, which is a key parameter for transistor/gate level reliability modelling, alternative methods are discussed to incorporate the ϕ_s-based model into transistor/gate levels.

- Chapter 5 concludes the thesis and gives some perspectives for the future work.

Chapter 2

Fundamentals of reliability modelling at transistor/gate levels

2.1 Fundamentals of reliability modelling

The CMOS technology advancement greatly improves the integration density of electronic systems. However, the reliability issues, including NBTI and HCI, turn to limit this density. The designed lifetime of a circuit must be guaranteed with the correct functional operation [11]; therefore, reliability modelling of a circuit has to be carried out at the design phase, taking the clock frequency, operating voltage, and temperature into consideration. The reliability of circuit depends on the performance of the transistors since they are the basic building blocks of a digital circuit. So it is reasonable to starts the reliability modelling at the transistor level. This section briefly review the models of two reliability issues: NBTI and HCI.

2.1.1 ΔV_{th} modelling due to NBTI

NBTI occurs in a PMOSFET when it is turned on such that the potential at its gate terminal is lower than those at the source and drain terminals. It becomes more severe at higher temperature. The physical mechanism of NBTI is still under debate, and a widely accepted hypothesis is the dissociation of Silicon-Hydrogen **(**$Si - H$**)** bonds located at the gate oxide and bulk interface, with the subsequent diffusion of the hydrogen species into the oxide [12]. This hypothesis leads to the so-called Reaction-Diffusion (RD) model [13, 14, 15].

Transistors in the nano-regime has thinner gate oxides, while the applied voltages do not scale equivalently. The electric field strength in the gate oxide (E_{ox}) is thus increasing, which is one of the main reasons for NBTI. To alleviate boron diffusion, nitrogen is "inserted" in gate oxide; but as shown in [16], the insertion of nitrogen deteriorates the NBTI effects.

Electrical performances such as the saturation current (I_{sat}) and the transconductance (g_m) degrade due to NBTI. NBTI does not require the current flow through the device, and it degrades the PMOSFET as long as $|V_{gs}| > |V_{th}|$ [17, 18]. The dissociation of $Si - H$ bonds leaves interface states (N_{IT}) behinds, leading to V_{th} shift, which plays the role of the monitor that characterises the degradation process. ΔV_{th} is time-dependent, experimental observations and mathematical solutions demonstrate that for a stressing time t, $\Delta V_{th}(t)$ could be related to t as:

$$\Delta V_{th}(t) \propto t^n \qquad \text{(Eq. 2.1)}$$

n is the time-coefficient ranging from 0.16 to 0.25 [18, 9]. According to [19], 0.16 corresponds to H_2 diffusion while 0.25 is explained as the diffusion of neutral H atoms [13]. Apart from time, NBTI relies on the voltage stress and the thermal stress. [20, 21, 22] describe an exponential law to explain the NBTI dependence on V_g. And the temperature (T) dependence is modelled with the Arrhenius law. Altogether:

$$\Delta V_{th}(t) \propto exp(\beta V_g) \cdot exp(-\frac{E_a}{k_B T}) \cdot t^n \qquad \text{(Eq. 2.2)}$$

In Eq. 2.2:

- β: coefficient describing the V_g dependence;
- E_a: activation energy of NBTI;
- k_B: Boltzmanns constant.

The RD model attributes NBTI to E_{ox} rather than V_g, there for $\mathrm{Eq.}\ 2.2$ should provide more details. As could be seen later, the ϕ_s-based model is able to describe E_{ox} as a function of device structure parameters and the input voltages (including V_g).

2.1.2 HCI modelling

When electrons/holes at the drain end of a MOSFET gain sufficient energy, they are able to be injected into the gate oxide, leading to device parameter degradation (including V_{th} and the channel mobility μ). Lateral electric field [23] plays the key role of HCI, as charge carriers are accelerated along the channel from the source to the drain. In addition, impact ionisation occurs near the drain region. Under the influence of the vertical electric field (like NBTI), some of those ”hot” charge carriers get ”lucky” to surmount the surface energy barrier and they are trapped in the gate oxide. Similar to NBTI, HCI also degrades I_{sat} and g_m.

According to [24], ΔV_{th} caused by HCI is linearly related to the density of interface states and inversely proportional to the effective channel length L_{eff}. Consequently, L_{eff} scaling makes HCI more severe. Despite the fact that the future technology will reduce the supply voltage to below $1V$, HCI influence could not be ignored.

[25] proposed a power-low equation that related N_{IT} **to** the various parameters:

$$N_{IT} = C_0 \cdot [\frac{I_{ds}}{W} \cdot exp(-\frac{\phi_{it}}{q\lambda_e E_m})]^n \qquad \text{(Eq. 2.3)}$$

In which

- C_0: process constant ranging from 1.9-2.0 [26];

- I_{ds}: drain-source current;
- W: channel width;
- ϕ_{it}: critical energy to create interface states, with a value of $3.7eV$
- λ_e: channel mean free path of an electron, [26] gives a value of $6.7nm$;
- E_m: lateral electric field strength near the drain region

Unlike NBTI that is singly influenced by the vertical electric field, HCI is also affected by the lateral field, E_m. [27] gives an analytical expression of E_m:

$$E_m = \frac{V_{ds} - V_{dsat}}{\sqrt{3t_{ox}x_j}} \quad \text{(Eq. 2.4)}$$

In $\mathrm{Eq.\ 2.4}$, t_{ox} is the gate oxide thickness and x_j is the drain-junction depth. [28] states that $\sqrt{3t_{ox}x_j}$ is the effective thickness of the pinch-off region. "3" is the ratio of Si permittivity and that of SiO_2. V_{dsat} is the saturation voltage. For the V_{th}-based device models, V_{ds} for long channel transistors is simply described as $V_{gs} - V_{th}$; a more complicated V_{dsat} model for short channel devices in [27] says:

$$V_{dsat} = \frac{(V_{gs} - V_{th}) \cdot L \cdot E_{sat}}{(V_{gs} - V_{th}) + L \cdot E_{sat}} \quad \text{(Eq. 2.5)}$$

E_{sat} the lateral field strength for electron velocity saturation.

Apart from N_{IT}, there are also models that relate HCI to the stressing time t_{stress}. Most models depict HCI as a power law function of t_{stress}:

$$HCI \propto t_{stress}^{n} \quad \text{(Eq. 2.6)}$$

[29, 30] describes HCI as a function of channel length, the stressing voltage, temperature, and t_{stress}. Those equations have similar forms to $\mathrm{Eq.\ 2.6}$; however, they do not give a physical explanation of the derivation of n. Later in Chapter 4, we will adopt the reaction-diffusion model to solve this problem.

2.2 ϕ_s-based transistor model

2.2.1 ϕ_s modelling

As could be seen from the last section, transistor level reliability models are not just simple functions of time. Voltages (or more precisely, the electric field strength), and temperature are also involved. These parameters could be captured by the ϕ_s-based transistor model. The advantage of the ϕ_s-based model over those based on V_{th} or the charge has been discussed in Chapter 1. This model is able to analytically describe the terminal electric characteristics of transistors based on physic, and it precisely predicts the behaviours of real devices. Apart from accuracy, it is faster compared with the conventional iterative methods [31]. A short summary of the advantages of the ϕ_s-based model is listed below:

- Gummel Symmetry Test (GST) is a benchmark test to qualify a compact model [32, 33]. GST requires that the drain-source current (I_{ds}) to be a strict **odd** function of V_{ds}. The ϕ_s based model passes this test [34]; however, the V_{th} based method fails this test [35];
- The V_{th} based model uses smoothing functions to join the sub-V_{th} region and the strong inversion region with regional approximations in each region [36, 37]. By nature, it could not model the moderate inversion region (the region in between the sub-V_{th} region and the strong inversion region). ϕ_s model overcomes this difficulty;
- The V_{th} based method pins ϕ_s to a constant value when the gate voltage is higher than V_{th} (but in reality ϕ_s changes slightly with the gate voltage). In this sense, the V_{th} based model loses accuracy in the strong inversion region;
- Simplicity: the V_{th} based model requires large numbers of empirical model parameters to capture the short-channel effects [38]. These parameters are neither physical nor scalable. But this is not an issue for the ϕ_s based model.

This section reviews parts of the ϕ_s-based model that are relevant to reliability modelling, including the analytical modelling of the surface potential, the inversion charge modelling, and

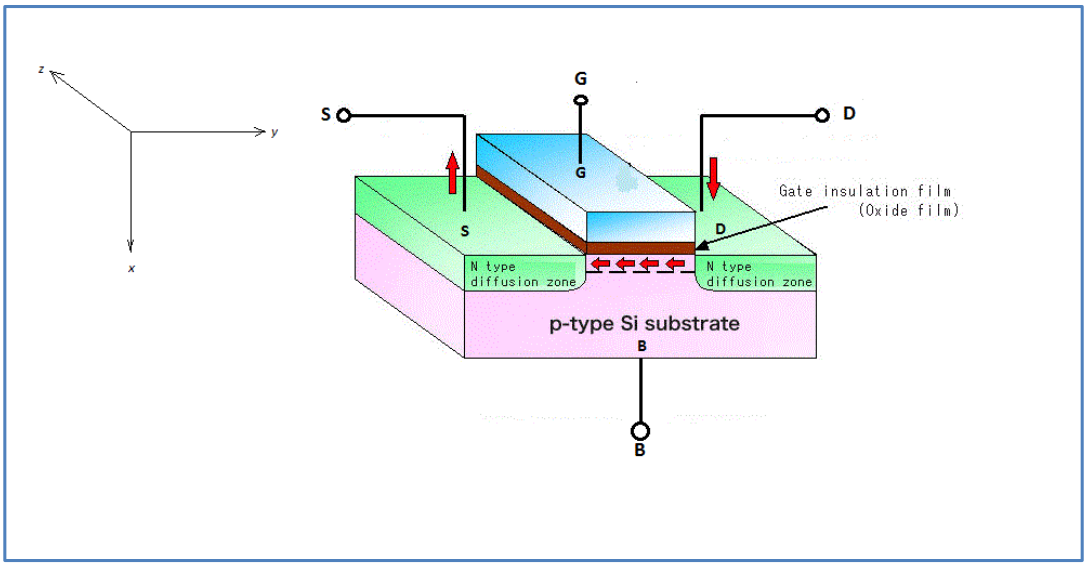


Figure 2.1: 3-D NMOSFET

the saturation voltage determination.

Figure 2.1 shows a conventional n-type bulk MOSFET. The four labels, G, B, S, and D represent the gate, the bulk, the source, and the drain terminals, via which the device communicates with other circuit components. It is a three-dimensional (3D) structure. Projected in the Cartesian coordinate system, the direction from G to B is along the *x-axis*, while that from S to D (along the channel) falls on the *y-axis*, and the channel width is on the *z-axis*. Regardless of the direction along the channel width, a 2D problem is to be solved to describe the behaviour of the device, which could only be done numerically. In order to facilitate the development of the compact model, the scenario is simplified with proper assumptions. Approximations of the analytical solutions are thus obtained and adopted in circuit simulators:

- Gradual Channel Approximation(GCA):

  To start with, the 2D problem is decomposed into two 1D problems: without considering the source and drain, a MOS capacitor remains, consisting of the gate, the oxide, and the bulk. The two voltages inputs applied to the gate (V_g) and bulk (V_b) control the polarities and amount of charge located in the Si bulk close to the Si-SiO_2 interface.

For an NMOS, inversion charge (electrons) is able to flow in the y-direction in case a potential difference is applied between the source (V_s) and the drain (V_d). A 1D Poisson equation is formed with the voltages applied to the gate (V_g) and the bulk (V_b), based on which ϕ_s is obtained. ϕ_s determines the density of the charge, which is driven to follow along the channel due to the difference between V_d and V_s to give the output current. To decompose the 2-D problem, the potential along the channel is assumed to change gradually. As a result, it does not cause perturbation of the charge distribution controlled by the Poisson equation. The 2-D problem decomposition is the so-called Graduation Channel Approximation (GCA).

- Charge Sheet Approximation(CSA):

  The work done by Pao and Sah [39] represents a criterion for other models even today. The model is physically accurate while computationally inefficient in that it treated the inversion charge to be located in a layer with a certain thickness. To meet the demand for fast simulation, the inversion charge is assumed to be on an infinitesimally thin sheet such that the potential drop on the inversion charge could be neglected. Calculation could thus be greatly simplified.

With the gate-bulk bias (V_{gb}) varying, a MOSFET could be operated in three different regions, namely, accumulation, depletion, and strong inversion. Every region has one type of dominating charge (taking an NMOSFET as an example): positive holes dominates in the accumulation region; in the depletion region, the depleted negative acceptors are the main charge carriers; while in strong inversion, the negative inversion charge is the major charge carriers.

The three regions are separated by two particular voltages: the flatband voltage (V_{FB}) and V_{th} are the boundaries of them. V_{FB} separates the accumulation and depletion regions while V_{th} locates between the depletion and strong inversion regions. ϕ_s in these regions could be derived explicitly according to the input voltage equation. The idea is, only the dominating type of charge is considered in its respective region while the other two are ignored. As ϕ_s is a continuous function of V_{gb}, solutions in the three regions are stitched together to give a single piece ϕ_s solution, which is called the unified regional surface potential

(URSP). At the boundary of two regions where there is no "dominating" charge, ϕ_s is tuned asymptotically to the "actual" value with the aid of transition/smoothing functions [40]. This method is of great significance in that it simplifies the computation complexity; besides, it is suitable for future devices.

If the source and drain in Figure 2.1 are ignored, only an MOS capacitor remains. When a potential difference is applied between the poly gate and the p-type Si bulk, it is balanced by three parts:

- ϕ_{MS}: work functions' difference between bulk-Si and the gate;
- V_{ox}: voltage drop on the capacitor;
- ϕ_s: surface potential denoting the band bending from Si-SiO_2 interface into the bulk.

The potential balance equation could be established as $\mathrm{Eq.\ 2.7}$:

$$V_{gb} = \phi_{MS} + \phi_s + V_{ox} \tag{Eq. 2.7}$$

The value of ϕ_{MS} is determined by:

$$\phi_{MS} \approx -v_{tm} \cdot ln[\frac{N_{ch}N_g}{n_i^2}] \tag{Eq. 2.8}$$

In $\mathrm{Eq.\ 2.8}$, N_{ch} stands for the channel doping concentration, while N_g is the gate doping concentration; n_i, on the other hand, is the intrinsic carrier concentration. Thermal voltage is defined in $\mathrm{Eq.\ 2.9}$:

$$v_{tm} = \frac{k_B T}{q} \tag{Eq. 2.9}$$

In $\mathrm{Eq.\ 2.9}$, k_B is the Boltzmann's constant and T is temperature in Kelvin; q represents the fundamental charge. The oxide capacitance per unit area, C_{ox}, is related the gate oxide thickness t_{ox}:

$$C_{ox} = \frac{\epsilon_{ox}}{t_{ox}} \tag{Eq. 2.10}$$

The voltage drop across the gate oxide V_{ox} is related to the gate charge Q_g by:

$$Q_g = V_{ox} \cdot C_{ox} \tag{Eq. 2.11}$$

Apart from Q_g, another two types of charge are involved in the device during operation: the effective oxide charge (Q_{ox}) due to the crystal and technology imperfection, and the induced channel & depletion charge (Q_{sc}). The charge balance requires [41]:

$$Q_g + Q_{ox} + Q_{sc} = 0 \tag{Eq. 2.12}$$

The combination of $\mathrm{Eq.\ 2.7}$ and $\mathrm{Eq.\ 2.12}$ leads to the following relation:

$$V_{gb} = \underbrace{\phi_{MS} - \frac{Q_{ox}}{C_{ox}}}_{V_{FB}} + \phi_s - \frac{Q_{sc}}{C_{ox}} \tag{Eq. 2.13}$$

The first two terms on the RHS of $\mathrm{Eq.\ 2.13}$ is the flatband voltage, V_{FB}. The equations above neglects the surface potential of the poly-S_i gate (ϕ_p) because of the high gate doping concentration.

Charges in the bulk comes from four parts: holes, electrons, ionised acceptors and donors; and their densities are p, n, N_A, and N_D respectively. For an NMOSFET, the bulk Si possesses a Fermi potential, ϕ_F, the value of which is:

$$\phi_F = v_{tm} \cdot arcsinh(\frac{N_A - N_D}{2n_i}) \tag{Eq. 2.14}$$

The electron and hole concentrations are related to ϕ_F as[40]:

$$n = n_i \cdot exp(\frac{\phi - V_{cb} - \phi_F}{v_{tm}}) \tag{Eq. 2.15}$$

and

$$p = n_i \cdot exp(\frac{-\phi + \phi_F}{v_{tm}}) \tag{Eq. 2.16}$$

In $\mathrm{Eq.\ 2.15}$ and $\mathrm{Eq.\ 2.16}$, ϕ is the potential representing the band bending at a random point in

the bulk, while V_{cb} is a position-dependent channel-to-bulk voltage. N_A and N_D are determined by the hole and electron concentrations far in the bulk (the neutral bulk) where ϕ approaches 0. Consequently [41, 42]:

$$N_A = p_\infty|_{\phi\to 0} = n_i \cdot exp(\frac{\phi_F}{v_{tm}}) \quad \text{(Eq. 2.17)}$$

While

$$N_D = n_\infty|_{\phi\to 0} = n_i \cdot exp(-\frac{V_{cb}+\phi_F}{v_{tm}}) \quad \text{(Eq. 2.18)}$$

The charge density ρ is established as sum of n, p, N_A, and N_D. Charges in the bulk is related to the electric field strength (E_{si}) at the $S_i - S_iO_2$ interface according to Gauss' law. Considering the relation between potential and electric field strength:

$$Q_{sc} = -\epsilon_{si}E_{si} = \epsilon_{si}\frac{d\phi}{dx}|_{x=0} \quad \text{(Eq. 2.19)}$$

Poisson-Boltzmann equation links the potential to the charge density:

$$\frac{d^2\phi}{dx^2} = -\frac{\rho}{\epsilon_{si}} \quad \text{(Eq. 2.20)}$$

The total charge in Si is obtained by integrating $\mathrm{Eq.\ 2.20}$ once with the following boundary conditions:

$$\begin{cases} \phi|_{x=0} = \phi_s \\ \frac{d\phi}{dx}|_{x\to\infty} = 0 \\ \phi|_{x\to\infty} = 0 \end{cases} \quad \text{(Eq. 2.21)}$$

Then the analytical solution for Q_{sc} is:

$$Q_{sc} = -C_{ox}\cdot\gamma\cdot sgn(\phi_s)\cdot \sqrt{v_{tm}(exp(\frac{-\phi_s}{v_{tm}})-1)+\phi_s+exp(-\frac{V_{cb}+2\phi_F}{v_{tm}})(v_{tm}exp(\frac{\phi_s}{v_{tm}})-v_{tm}-\phi_s)} \quad \text{(Eq. 2.22)}$$

γ is the body factor. N_{ch} is the effective channel doping concentration.$sgn(\phi_s)$ stands for the sign of ϕ_s, which is 1 if $\phi_s > 0$ and -1 otherwise. For $N_A \gg N_D$:

$$N_A = n_i exp(\frac{\phi_F}{v_{tm}}) \quad \text{(Eq. 2.23)}$$

The implicit ϕ_s solution described by [40]:

$$V_{gb} - V_{FB} - \phi_s = \gamma sgn(\phi_s) \cdot \sqrt{\underbrace{v_{tm}(exp(\frac{-\phi_s}{v_{tm}}) - 1)}_{\text{holes}} + \underbrace{\phi_s(1 - exp(-\frac{V_{cb} + 2\phi_F}{v_{tm}}))}_{\text{ionised impurities}} + \underbrace{v_{tm}exp(-\frac{V_{cb} + 2\phi_F}{v_{tm}})(exp(\frac{\phi_s}{v_{tm}}) - 1)}_{\text{electrons}}} \quad \text{(Eq. 2.24)}$$

The contributions from different types of charges are shown in $\mathrm{Eq.\ 2.24}$. As aforementioned, ϕ_s in every region is derived by considering the dominating charge. Therefore, in the accumulation region (V_{gb}¡V_{FB}), $\mathrm{Eq.\ 2.24}$ is simplified:

$$V_{gb} - V_{FB} - \phi_s = -\gamma\sqrt{v_{tm}(exp(-\frac{\phi_s}{v_{tm}}) - 1)} \quad \text{(Eq. 2.25)}$$

This equation has a solution in the form of *Lambert W* function (denoted as $\mathcal{L}\{w\}$):

$$\phi_{s,acc} = V_{gb} - V_{FB} - 2v_{tm}\mathcal{L}\{\frac{\gamma}{2\sqrt{v_{tm}}}exp(-\frac{V_{gb} - V_{FB}}{2v_{tm}})\} \quad \text{(Eq. 2.26)}$$

In the depletion region where only the depletion charge is considered, the voltage equation is modified as:

$$V_{gb} - V_{FB} - \phi_s = \gamma\sqrt{\phi_s} \quad \text{(Eq. 2.27)}$$

The solution in this region is then:

$$\phi_{s,dep} = \{-\frac{\gamma}{2} + \sqrt{\frac{\gamma^2}{4} + V_{gb} - V_{FB}}\}^2 \quad \text{(Eq. 2.28)}$$

With only electron counted in the strong inversion region, the potential equation has a similar form to that in the accumulation region:

$$V_{gb} - V_{FB} - \phi_s = \gamma\sqrt{v_{tm}exp(-\frac{V_{cb} + 2\phi_F}{v_{tm}})(exp(\frac{\phi_s}{v_{tm}}) - 1)} \quad \text{(Eq. 2.29)}$$

The solution of this equation is:

$$\phi_{s,str} = V_{gb} - V_{FB} - 2v_{tm}\mathcal{L}\{\frac{\gamma}{2\sqrt{v_{tm}}}exp(\frac{V_{gb} - V_{FB} - 2\phi_F - V_{cb}}{2v_{tm}})\} \quad \text{(Eq. 2.30)}$$

It should be notices that $\mathrm{Eq.\ 2.26}$, $\mathrm{Eq.\ 2.28}$, and $\mathrm{Eq.\ 2.30}$ are all mathematical expressions for ϕ_s. They should be restricted in their respective regions such that they are physically meaningful. Mathematical smoothing functions are adopted to join them together [31].

The ϕ_s solutions in the three regions (curves) are plotted together with the one-piece unified ϕ_s (dots) in Figure 2.2. The overlapping part of the red curve and the dots represents ϕ_s in the accumulation region ($\phi_{s,acc}$); ϕ_s in the depletion region ($\phi_{s,dep}$) is shown as the portion of the blue curve that covered by the dots, ϕ_s is almost linearly related to V_{gb} in this region; the reliability issues are closely related to ϕ_s in the strong inversion region ($\phi_{s,str}$), denoted as the dots crossed by the green curve. A very important property of $\phi_{s,str}$ is that it is almost constant, with only a slight increment as V_{gb} increases. This could be explained by $\mathrm{Eq.\ 2.29}$: the ratio of $\frac{\phi_s}{v_{tm}}$ appears in an exponential term, representing the electron density; therefore a smaller change in ϕ_s leads to a huge increment of electron density. The potential drop across the gate oxide is obtained according the potential balance:

$$V_{ox} = V_{gb} - \phi_{MS} - \phi_{s,str} \quad \text{(Eq. 2.31)}$$

Since $\phi_{s,str}$ is almost constant, there is a nearly linear relation between V_{ox} and V_{gb} in the strong inversion region. Further, E_{ox} is related to V_{ox} by:

$$E_{ox} = \frac{V_{ox}}{t_{ox}} \quad \text{(Eq. 2.32)}$$

This is the ϕ_s-based electric field strength modelling. Both NBTI and HCI requires the modelling of the channel charge density. Many compact models model the charge with charge sheet approximation [43]. Besides, the depletion approximation assumes that there exists no mobile charge in the depletion region beneath the $Si - SiO_2$ interface. Along the y direction from source to drain, the position-dependent ϕ_s is denoted as $\phi_s(y)$. The depletion charge

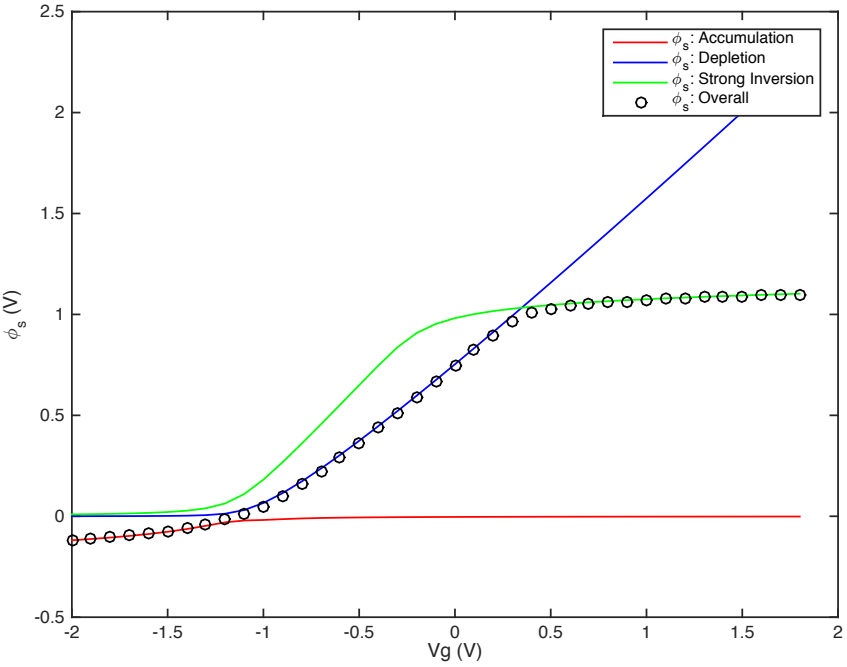


Figure 2.2: ϕ_s solutions in the three regions and the unified ϕ_s

density is approximately:

$$Q_b = -\gamma C_{ox}\sqrt{\phi_s(y)} \quad \text{(Eq. 2.33)}$$

The total charge in S_i consists of the depletion charge and the inversion charge (Q_i), and according to charge balance by $\mathrm{Eq.\ 2.12}$:

$$Q_i = -(Q_g + Q_{ox}) - Q_b = -C_o x(V_{gb} - V_{FB} - \phi_s(y) - \gamma\sqrt{\phi_s(y)}) \quad \text{(Eq. 2.34)}$$

2.2.2 V_{dsat} modelling

The demand for V_{dsat} modelling is in that it is related to HCI. It could be modelled with the current flow through the device. The potential difference between the drain and the source leads to the drift current I_{drift}; as $\phi_s(y)$ in $\mathrm{Eq.\ 2.34}$ is position-dependent, Q_i decreases monotonously from the source to the drain. Electrons tend to diffuse from the high concentration to the low, leading to the diffusion current (I_{diff}). The overall current is the sum of the two. An NMOSFET could be operated in either the linear region or the saturation region. In the linear region, the linear current is described as [42]:

$$I_{dlinear} = \mu_{eff0} C_{ox} \frac{W}{L}(V_{gt,s} - \frac{A_{b,s}\Delta\phi}{2} + v_{tm}A_{b,s})\Delta\phi \quad \text{(Eq. 2.35)}$$

β is the gain factor defined as $\beta = \frac{\mu_n W C_{ox}}{L}$. $\Delta\phi$ is the surface potential difference between the source and the drain, which is approximately V_{ds}. $A_{b,s}$ is the source side bulk charge factor, defined as:

$$A_{b,s} = 1 + \frac{\gamma}{2\sqrt{\phi_{s,s}}} \quad \text{(Eq. 2.36)}$$

In the saturation region, I_{drift} dominates the current, and I_{diff} could be omitted. According to $\mathrm{Eq.\ 2.34}$, the charge density at the source side (normalised to C_{ox}) is:

$$V_{gt,s} = V_{gb} - V_{FB} - \phi_{s,s} - \gamma\sqrt{\phi_{s,s}} \quad \text{(Eq. 2.37)}$$

And the saturation current is[42]:

$$I_{dsat} = WC_{ox}(V_{gt,s} - A_{b,s}\Delta\phi)\nu_{sat} \quad \text{(Eq. 2.38)}$$

ν_{sat} in $\mathrm{Eq.}\ 2.38$ is the saturation velocity of the electrons. It is related to the channel mobility and the saturation electric field strength (lateral)[44]:

$$\nu_{sat} = 2\mu_0 E_{sat} \quad \text{(Eq. 2.39)}$$

By equating the RHS of $\mathrm{Eq.}\ 2.35$ and that of $\mathrm{Eq.}\ 2.38$, $Vdsat$ is expressed as:

$$V_{ds,sat} = \frac{V_{gt,s}E_{sat}L}{V_{gt,s} + A_{b,s}E_{sat}L + 2A_{b,s}v_{tm}} \quad \text{(Eq. 2.40)}$$

$\mathrm{Eq.}\ 2.40$ has a similar form to $\mathrm{Eq.}\ 2.5$.

2.3 Reliability analysis at the circuit and gate levels

2.3.1 Circuit level reliability models

At the circuit level, three steps are involved in the analysis of circuit performance degradation:

- SPICE simulation of the fresh circuit with the storage of transistor terminal current and voltage waveforms;
- For every single transistor, the degradation model is generated from those waveforms;
- The degradation models are applied to the second SPICE simulation to give the circuit performance degradation.

The Berkeley reliability tools (BERT)[45] is the first reliability simulator published. It could be used for the determination of HCI caused circuit performance degradation; in addition,

the probability of circuit failure owing to time dependent dielectric breakdown (TDDB) and electromigration (EM) could be computed from BERT.

BERT first determines current components from the device channel, gate, and substrate- $I_d(t)$, $I_g(t)$, and $I_{sub}(t)$. Then a parameter called "AGE" is computed for every single transistor according to these currents. "AGE" measures the severity of degradation. To determine the device degradation after a time period of t_{deg}:

$$AGE_n = \int_0^{t_{deg}} \frac{I_d(t)}{WH_n} (\frac{I_{sub}(t)}{I_d(t)})^{m_n} dt \qquad \text{(Eq. 2.41)}$$

$$AGE_p = \int_0^{t_{deg}} \frac{1}{H_p} (\frac{I_g(t)}{W})^{m_p} dt \qquad \text{(Eq. 2.42)}$$

The subscripts n and p are for NMOSFET and PMOSFET respectively. For a given technology, H and m are experimentally determined. It is impractical for very long time (for example, 10 years) circuit simulation. Therefore, "AGE" is extrapolated from short time circuit simulation. With BERT, the device degradation model could be determined such that the parameter degradation Δp is an "AGE"-dependent function. The circuit performance degradation could be obtained as the final step simulation.

Apart from BERT, there are commercial reliability simulators available such as RelXpert [46]. Reliability analysis has been integrated to HSPICE[47] and ELDO [48]. Both NBTI and HCI are involved in RelXpert. ELDO applies an iteration method to determine Δp: a certain t_{deg} is equally divided into N intervals. In every interval, the first two steps for degradation analysis are conducted. In this case, the influence of degraded waveforms on Δp is taken into account. [49] takes the influences of both process variation and aging into consideration. Similar to ELDO, it is based on iteration.

The degradation equations are confidential in those commercial tools so that users do not have access to the verification about the calculation of the degradation. They have to trust those tools anyway. [50] claims that when user defined equations are adopted in RelXpert, it suffers from accuracy problems.

Circuit level reliability simulators have the advantage of precision. However, like other circuit level simulations, it is stunted with time efficiency problems. Besides, as a prerequisite, input vector is another problem: both the first and third steps of degradation analysis rely on the input vector. It is needed to cause circuit degradation in the first step; while in the third step it is used for the measurement of circuit performance degradation. Similar to SPICE, those simulators are not suitable for simulations of complex circuits.

2.3.2 Gate level reliability models

To cope with the time efficiency problem, timing analysis at the gate level is done for complex circuits, which is called STA. It is much faster with a simplified gate model and no requirement for input vectors. A gate model is required for gate delay computation in STA. It gives a delay caused by rising/falling transition for a timing arc, which is defined from the input to the output of a gate. SPICE simulations are used to pre-characterise the delays of gates from the standard cell library. The gate model is thus generated. STA only does the evaluation at the gate level rather than the transistor level. Since no SPICE simulation is done for the entire circuit, STA is more time efficient.

[51] models the gate delay with the simple formula:

$$D = C_1 + C_2 \cdot C_L \qquad \text{(Eq. 2.43)}$$

The gate delay D consists of two components: intrinsic part denoted as C_1 and the (C_L)-dependent part with coefficient C_2. C_L is the output load capacitance, which is composed of the succeeding gates capacitance and the interconnect capacitance. This model is quite rough since it does not take the impact of the input slope s_{in} into consideration.

In STA, a signal is represented as a ramp with a certain slope, which is tracked as the transition time. Together with C_L, a popular gate model relies on the 2D look-up table (LUT).

The LUT contains the gate delays as a function of both s_{in} and C_L:

$$D = f(s_{in}, C_L) \quad \text{(Eq. 2.44)}$$

For values of s_{in} and C_L that are not on the LUT, interpolation is adopted to generate the respective D. Since s_{in} is required in this model, and the s_{in} of a succeeding gate is actually the output slope (s_{out}) of its preceding gate. Therefore, s_{out} is also stored in the LUT as a function of s_{in} and C_L:

$$s_{out} = g(s_{in}, C_L) \quad \text{(Eq. 2.45)}$$

The accuracy of the LUT based model could be improved with more (s_{in}, C_L) points characterised. The technology improvement leads to smaller gate capacitance but larger interconnect resistance. As a result, a load could not be treated as a pure capacitance any more. [52] proposed the ”effective capacitance” method in which only one value is needed for the interconnect network. With the application of this method, the LUT-based model could still be used.

The LUT based gate model has been expanded for reliability study. [53] introduced the analysis flow of the gate level reliability modelling and [54] is able to generate multiple LUTs according to the predefined $time$, T, and SP(signal probability, which will be discussed in details in Chapter 3)conditions. An obvious drawback of this method is the huge number of LUTs, considering the workload at the gate input. For an inverter with only one input, 4 LUTs are to be generated with 4 different SPs, as is shown in Figure 2.3. In case it is a NOR gate, 4x4=16 LUTs are needed since there are two inputs. In a standard cell library, some gate could possess three or even more inputs. This means much more LUTs.

[55] proposed a reliability model called GLACIER, which describes the influence of HCI on gate delay. It defines a factor α:

$$\alpha(s_{in}, C_L, TD) = \frac{D_{aged}}{D_{fresh}} \quad \text{(Eq. 2.46)}$$

TD is the transition density, which will be covered in details in Chapter 4. D_{fresh} is gate delay

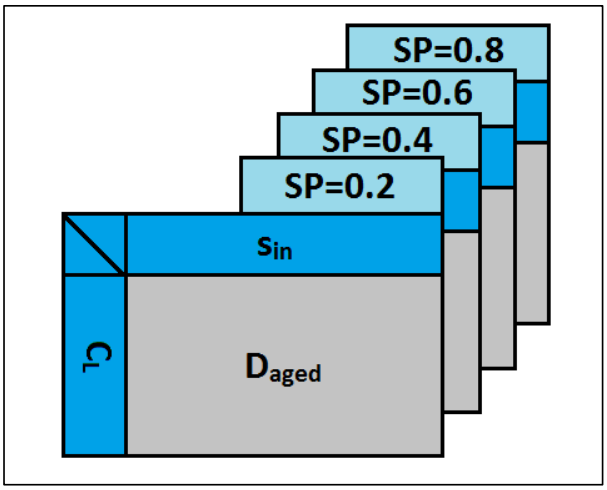


Figure 2.3: LUTs for gate level reliability modelling

before reliability problems occur while D_{aged} is that after t_{deg}. D_{fresh} has been described by Eq. 2.44, while D_{aged} is also a function of TD apart from s_{in} and C_L. For a gate with multiple inputs, every input has its own TD. Then D_{aged} is a function of all TDs. For N transistors in series:

$$\alpha = \sum_{n=1}^{N} \alpha_n - (N-1) \qquad \text{(Eq. 2.47)}$$

α_n comes from input n when it is the only switching input. This method does not consider the influence from other inputs, neither does it take the parameter shift due to internal gate structure into account. The gate models are extracted from circuit level reliability simulations, which are subject to certain conditions (also called use profile) such as T, and SP. When there is a condition change, a re-characterisation is required for the whole gate library.

A more general gate model directly related D_{aged} to Δp. For examples, when the impact of NBTI on gate delay is discussed, D_{aged} is related to ΔV_{th}:

$$D_{aged} = D_{fresh} + \Delta D(\Delta V_{th}) \qquad \text{(Eq. 2.48)}$$

D_{aged} is represent as the sum of fresh gate delay and a "delay shift" (ΔD) caused by parameter shift. In this model, parameter shift is calculated independently according to the stressing

conditions but not during the gate model characterisation. Therefore, it avoids the dependency on the use profile. Linear approximations are used to related D_{aged} and a parameter shift:

$$D_{aged} = D_{fresh} + \frac{\partial D}{\partial p} \cdot \Delta p \qquad \text{(Eq. 2.49)}$$

$\frac{\partial D}{\partial p}$ is the sensitivity of gate delay D on parameter p. In case of NBTI, $p = V_{th}$. [56] introduces the α-power methodology that relates the drain current $I_{on}$with V_{th}:

$$I_{on} \propto (V_g - V_s - V_{th})^{\alpha} \qquad \text{(Eq. 2.50)}$$

[57] takes advantage of this method. D is assumed to be purely controlled by C_L recharging:

$$D = \frac{V_{dd} \cdot C_L}{I_{on}} = \frac{C_3}{(V_g - V_s - V_{th})^{\alpha}} \qquad \text{(Eq. 2.51)}$$

C_3 in $\mathrm{Eq.\ 2.51}$ is a certain constant. As a result, the sensitivity is:

$$\frac{\partial D}{\partial V_{th}} = \frac{\alpha \cdot D}{V_g - V_s - V_{th}} \qquad \text{(Eq. 2.52)}$$

[58] also derives the gate model according to the α-power methodology. It mainly discusses the impact of temperature on device degradation. The model in [59] considers the stack effect, which mean every transistor in series has it own V_{gs} other than V_{dd}.

Δp for every transistor in a gate may be different, since the total time when it is under stress may be different for every one of them. But there is usually only one ΔV_{th} used in all those gate models. A common case is, ΔV_{th} for every individual transistor is calculated, but only the Maximum is used in the model. This might introduce in large error.

[60] shows that for a NOR gate, the SPs at both inputs will influence ΔV_{th} of either transistor. Nevertheless, it just points out this phenomenon but no algorithm is derived. [61] provides the method to derive the time that a transistor is under stress such that ΔV_{th} could be calculated; but it does not build up a gate model.

2.4 Problems and Targets

In the first section of this chapter, the transistor level reliability models are introduced. Most of these models are based on the accelerated experiment results: a device is under continuous stress for a certain period of time. We call them DC models. At the gate level, the voltage applied to the gate terminal of a MOSFET is switching; besides, for devices in series, the drain/source terminal voltages are influenced by neighbours. A PMOSFET is not always "ON"; and there is no continuous current flow through a MOSFET. Therefore, The DC reliability models could not be directly inserted into the gate models.

The ϕ_s-based transistor model is able to accurately model the stressing condition of a devices with the solution of the electric field strength and saturation voltage. But it is a static model by nature, t must be introduced into this model for reliability modelling.

The gate level models treat t as a statistical parameter (unless historical data is available, which tracks the terminal voltage of a device over time), which could be used at the transistor level. The gate delay degradation is directly related to Δp from the transistor level model. Therefore, the main target of this thesis is to provide compact yet accurate reliability models at the transistor level such that they could be used for gate level modelling. This thesis takes advantage of the gate level reliability model built by Dr. Dominik Lorenz from TUM [29], and details of this model will be covered in Chapter 3 and 4, along with the NBTI and HCI models. The definitions and algorithms for SP and TD will be introduced, which are key parameters for ΔV_{th} (at the transistor level) determination.

Chapter 3

Negative Bias Temperature Instability Modelling

Presently, NBTI is considered as the most severe reliability issue. In the past 40 years, it has been a research topic [62]. The modern semiconductor technologies suffer more from NBTI, make these research more and more popular.

Only PMOSFETs are influenced by NBTI. When the device is biased in strong inversion where its gate terminal is at a lower potential compared with the source/drain, it is under stress. NBTI tends to create interface states at the $Si - SiO_2$ interface. Because of this, there is a V_{th} shift (ΔV_{th}) for the device such that it is more difficult to turn the device on. Or in other words, an aged device has a performance degradation (such as I_{on}) compared with a fresh one. Since a PMOSFET possesses a negative V_{th}, conventionally people say the absolute value of V_{th} ($|V_{th}|$) increases due to NBTI. NBTI is temperature-dependent and voltage-dependent; an increase in T or V_{supply} makes NBTI more severe.

3.1 NBTI mechanism and DC case NBTI Modelling

The physical mechanism behind NBTI is still controversial, which leads to different models. A commonly adopted reaction-diffusion (RD) model [13] is the foundation of the model to be constructed, although there also exists the hole trapping model [63].

The structure of a MOSFET is in such a way that there is an isolation SiO_2 layer in between the Si substrate and the poly-Si gate. At the $Si - SiO_2$ interface, there exists dangling bonds [64], which are also called interface states. They are actually Si atoms with an unsatisfied valence. The performance of a device is affected by these bonds since they are able to capture charges. When fabricated, hydrogen is incorporated into the device to passivate the interface states by forming $Si - H$ bonds. Under stress, these bonds will be broken with the generation of interface states. As is mentioned, this leads to performance degradation of both the device and the circuit.

[13] has attributed NBTI to the $S_i - H$ bounds breaking (Reaction) and the resultant hydrogen species diffusing (Diffusion). These bonds locate at the $Si - SiO_2$ interface. Figure 3.1 taken from [65] explains the mechanism of NBTI:

- Stressed in the strong inversion, holes accumulate in the channel close to the $S_i - S_iO_2$ interface;
- Holes are captured by the $S_i - H$ bonds with release of hydrogen species, which corresponds to the reaction phase;
- Hydrogen species diffuse away from the $S_i - S_iO_2$ interface, leaving behind the S_i-dangling bonds; this corresponds to the diffusion phase.

The following two equations mathematically model this process:

$$\frac{dN_{IT}}{dt} = \underbrace{k_F(N_0 - N_{IT})}_{\text{forward reaction}} - \underbrace{k_r N_H N_{IT}}_{\text{backward reaction}} \quad (x = 0) \qquad \text{(Eq. 3.1)}$$

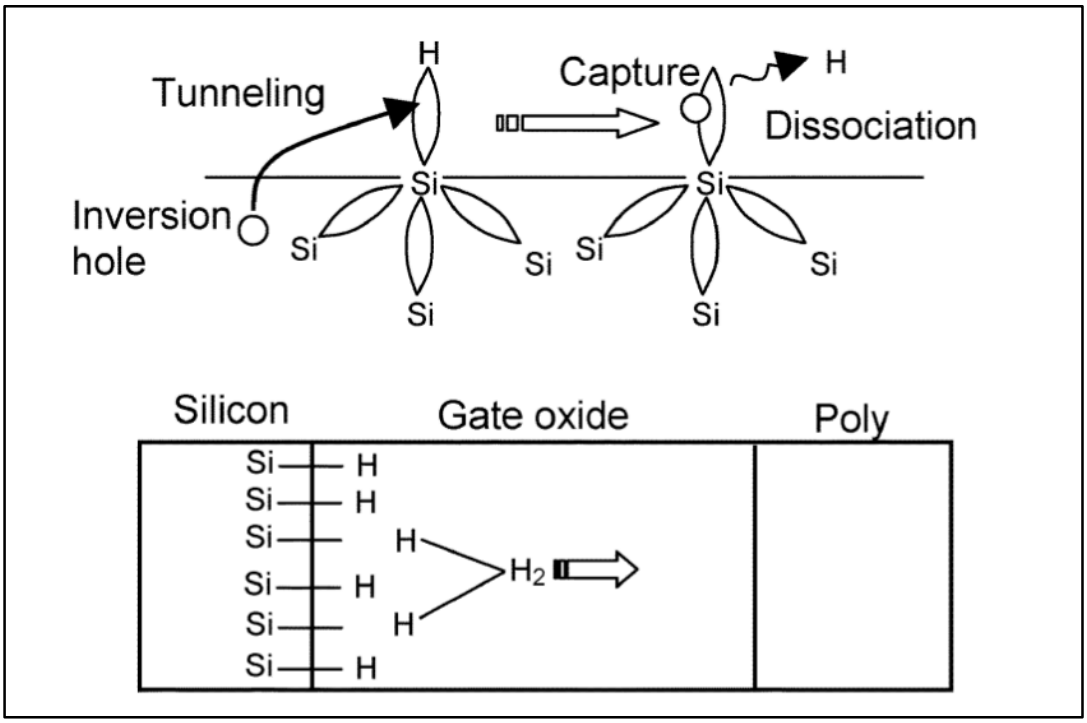


Figure 3.1: Hole tunneling, capture & $S_i - H$ bond dissociation & Hydrogen diffusion

$$\frac{dN_{IT}}{dt} = D_H(\frac{dN_H}{dx}) + \frac{\delta}{2}\frac{dN_H}{dt}(0 < x < \delta) \qquad \text{(Eq. 3.2)}$$

Physical parameters in $\mathrm{Eq.\ 3.1}$, $\mathrm{Eq.\ 3.2}$ are summarised below:

- N_{IT}: interface trap density per area;
- k_f: forward dissociation rate constant;
- N_0: initial Si-H bond density;
- k_r: annealing rate constant;
- N_H: hydrogen concentration;
- δ: reaction interface thickness;
- D_H: hydrogen diffusion coefficient;
- t: stressing time.

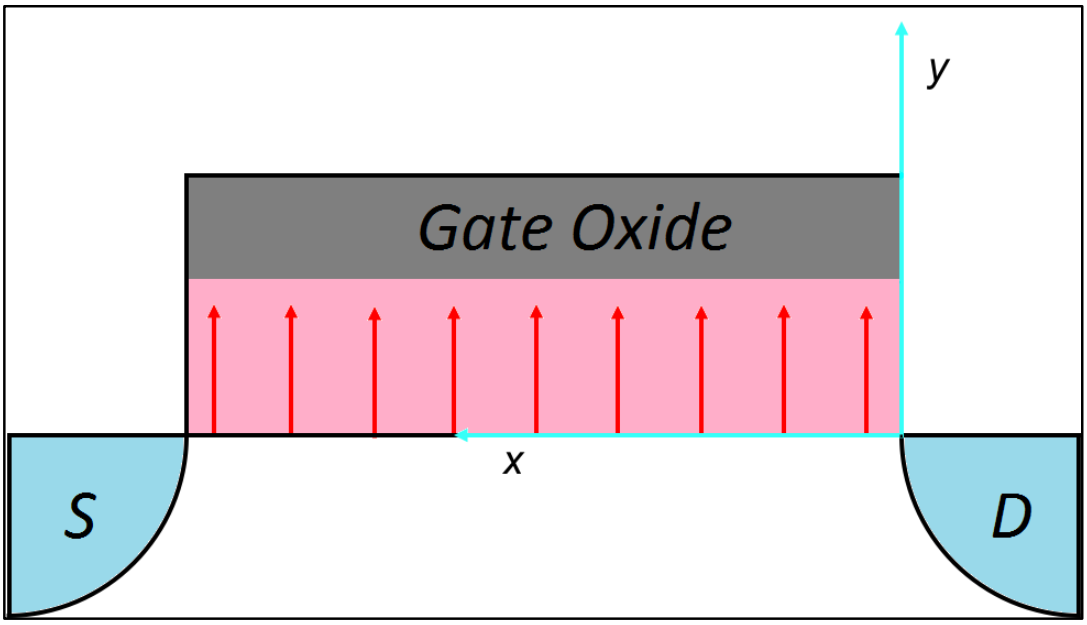


Figure 3.2: 1D hydrogen diffusion during NBTI

Eq. 3.1 says that NBTI is a bidirectional process. The first term on the RHS is controlled by the stress, which vanishes once the stress is removed; while the second term on the RHS continues even if there is no stress any more. As a result, NBTI presents a special "recovery" phenomenon: with the stress continuing, N_{IT} increases with time; when the stress is removed, N_{IT} is not a constant but decreases. This "recovery" is a unique property of NBTI (HCI, however, does not possesses this property)leads to the complexity of the AC case NBTI modelling where the device is turned on and off from time to time, and this will be covered later.

Figure 3.2 states that the diffusion of the generated hydrogen species during NBTI is one-dimensional (1D): once generated, the hydrogen species diffuse in the direction from the $Si - SiO_2$ interface into the gate oxide, and then they arrive at the poly-Si gate. In fact, at the gate oxide edge close to the source and drain region, the diffusion is 2D; but this part could be ignored since the lengths of these regions are much smaller compared with the channel length.

[66] also supports this 1D diffusion theory. Like what is stated in [13], the diffusing hydrogen density decreases approximately linearly with the distance into the oxide, as is shown

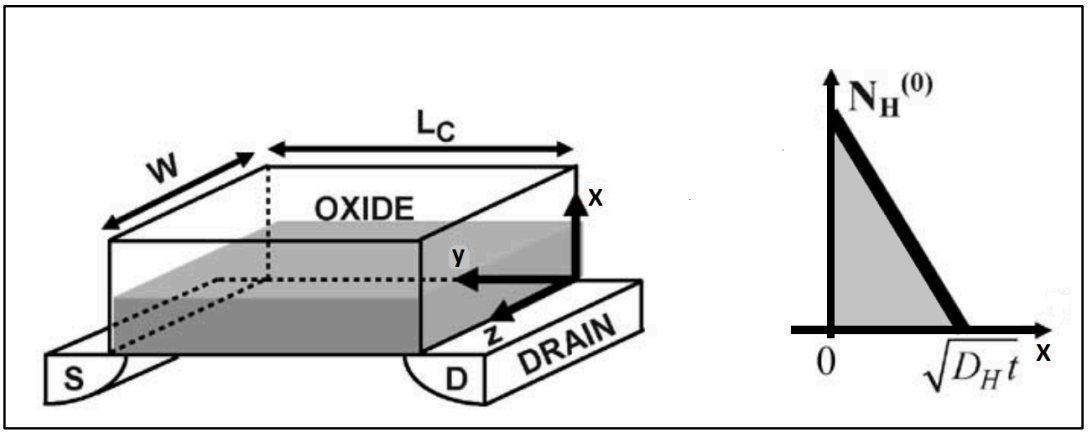


Figure 3.3: linear relation between the diffusing hydrogen density and the distance into the oxide

in Figure 3.3 [66]. The left part of Figure 3.3 is a regeneration of Figure 3.2, and the diffusing species are confined in the shaded rectangular soild. The right part of Figure 3.3 shows the relation between the hydrogen density and the distance into the gate oxide. $N_H^{(0)}$ is the hydrogen density at the $Si-SiO_2$ interface and $\sqrt{D_H t}$ is the diffusion front where the hydrogen density is 0. At a random location x inside the oxide with $0 \leqslant x \leqslant \sqrt{D_H t}$, the hydrogen density $N_H^{(x)}$ follows:

$$N_H^{(x)} = N_H^{(0)} - \frac{N_H^{(0)}}{\sqrt{D_H t}} \cdot x (0 \leqslant x \leqslant \sqrt{D_H t}) \qquad \text{(Eq. 3.3)}$$

A one-step integration of $\mathrm{Eq.\ 3.3}$ over distance gives the total amount of hydrogen; since hydrogen and the interface states follow a one-to-one relation, this results in the interface states per area:

$$\begin{aligned} N_{IT}(t) &= \int_0^{\sqrt{D_H t}} N_H^{(0)} (1 - \frac{x}{\sqrt{D_H t}}) dx \\ &= \frac{1}{2} N_H^{(0)} \sqrt{D_H t} \end{aligned} \qquad \text{(Eq. 3.4)}$$

Based on proper assumptions, the combination of $\mathrm{Eq.\ 3.1}$, $\mathrm{Eq.\ 3.2}$, and $\mathrm{Eq.\ 3.3}$ gives the rela-

tion between $N_IT(t)$ and the stressing time t:

$$N_IT(t) = \sqrt{\frac{k_f N_0}{2k_r}} \cdot (D_H t)^{\frac{1}{4}} \qquad \text{(Eq. 3.5)}$$

The diffusing species is another controversial matter. [67] shows that H^+ could be the resultant, while [19] explained that H atoms are generated in the first place and they then combine to H_2 and diffuse away. In this thesis, hydrogen atom is taken as the diffusing species such that the time coefficient is $\frac{1}{4}$. If H_2 has to be incorporated anyway, the model could be easily modified by shifting the time coefficient to $\frac{1}{6}$.

It should be noticed that the time coefficient n is not a fixed value especially for long period stress, which is indicated in [13]: in the reaction phase, n could take the value of 1 and then 0 with time increasing; however, this does not contribute to the final NBTI. For a "very long" time, n could be $\frac{1}{2}$, which corresponds to the phase when Hydrogen front reaches the oxide/poly interface; but this is highly impossible for the AC case. $n = \frac{1}{4}$ corresponds to the diffusion phase that is the most readily observed. As a result, only $n = \frac{1}{4}$ is chosen for NBTI modelling (in case of Hydrogen atom diffusion).

$\mathrm{Eq.\ 3.5}$ says that from the time point 0, a PMOSFET is under stress for a time period of t. As has been mentioned in $\mathrm{Eq.\ 3.1}$, the backward reaction leads to recovery after stress. In [68], the final N_{IT} is related to the stress time t and the recovery time t_0:

$$N_{IT}(t + t_0) = \frac{N_{IT}(t)}{1 + \sqrt{\frac{\xi t_0}{t + t_0}}} \qquad \text{(Eq. 3.6)}$$

In $\mathrm{Eq.\ 3.6}$, ξ has a value of 0.5 to represent the double-sided hydrogen diffusion. Meanwhile, to make the thesis more readable, the stress time is denoted as t_{stress} while the recovery time is $t_{recovery}$. The total time t_{total}, starting from the beginning of t_{stress} and terminating at the end of $t_{recovery}$, is then:

$$t_{total} = t_{stress} + t_{recovery} \qquad \text{(Eq. 3.7)}$$

In addition, two more parameters are defined. One is called stress factor (SF):

$$SF = \frac{t_{stress}}{t_{total}} \tag{Eq. 3.8}$$

Another is the signal probability (SP):

$$SP = \frac{t_{recovery}}{t_{total}} = 1 - SF \tag{Eq. 3.9}$$

So far the NBTI is discussed for the DC case with a single stress and recovery phases. But the definitions for SP and SF are also suitable for the AC case: SF is actually the duty cycle and SP is the complementary part. The threshold voltage shift (ΔV_{th}) is directly related to N_{IT}:

$$\Delta V_{th} = \frac{qN_{IT}}{C_{ox}} \tag{Eq. 3.10}$$

With the rearrangement of $\mathrm{Eq.\ 3.5}$-$\mathrm{Eq.\ 3.10}$, $V_{th}(t_{total})$ is expressed as:

$$\Delta V_{th}(t_{total}) = \frac{\mathbf{1}}{\mathbf{1} + \sqrt{\boldsymbol{\xi} \cdot SP}} \cdot \frac{\mathbf{q}}{\mathbf{C_{ox}}} \sqrt{\frac{\mathbf{k_f N_0}}{\mathbf{2k_r}}} \cdot \mathbf{D_H}^{\frac{\mathbf{1}}{\mathbf{4}}} \cdot \mathbf{SF}^{\frac{\mathbf{1}}{\mathbf{4}}} \cdot t_{total}^{\frac{\mathbf{1}}{\mathbf{4}}} \tag{Eq. 3.11}$$

In case the device is under stress all the way with $SP = 0$, $SF = 1$, and $t_{total} = t_{stress}$, $\mathrm{Eq.\ 3.11}$ could be simplified:

$$\Delta V_{th}(t_{stress}) = \frac{\mathbf{q}}{\mathbf{C_{ox}}} \sqrt{\frac{\mathbf{k_f N_0}}{\mathbf{2k_r}}} \cdot (\mathbf{D_H} t_{stress})^{\frac{\mathbf{1}}{\mathbf{4}}} \tag{Eq. 3.12}$$

Figure 3.4 is plotted according to $\mathrm{Eq.\ 3.11}$ and $\mathrm{Eq.\ 3.12}$. The solid dot at the peak of the curve corresponds to ΔV_{th} at the end of the stress phase (also the beginning of the recovery phase); while the empty dot corresponds ΔV_{th} at the end of the recovery phase, and it represents the net increase of V_{th}. One thing to be mentioned is that physical parameters in $\mathrm{Eq.\ 3.11}$ and $\mathrm{Eq.\ 3.12}$ are not adjusted for Figure 3.4 since they do not influence the shape of the curve. The DC case NBTI modelling is relatively simple, since there is only one stress phase, followed by a single recovery phase. However, scenarios like the DC case is highly impossible in a digital circuit: with the input signal switched between ”0” and ”1”, a PMOSFET undergoes the stress and recovery phases alternately. Therefore, ΔV_{th} obtained from $\mathrm{Eq.\ 3.11}$ could not be simply transferred to the gate level for timing analysis; larger errors would be introduced otherwise.

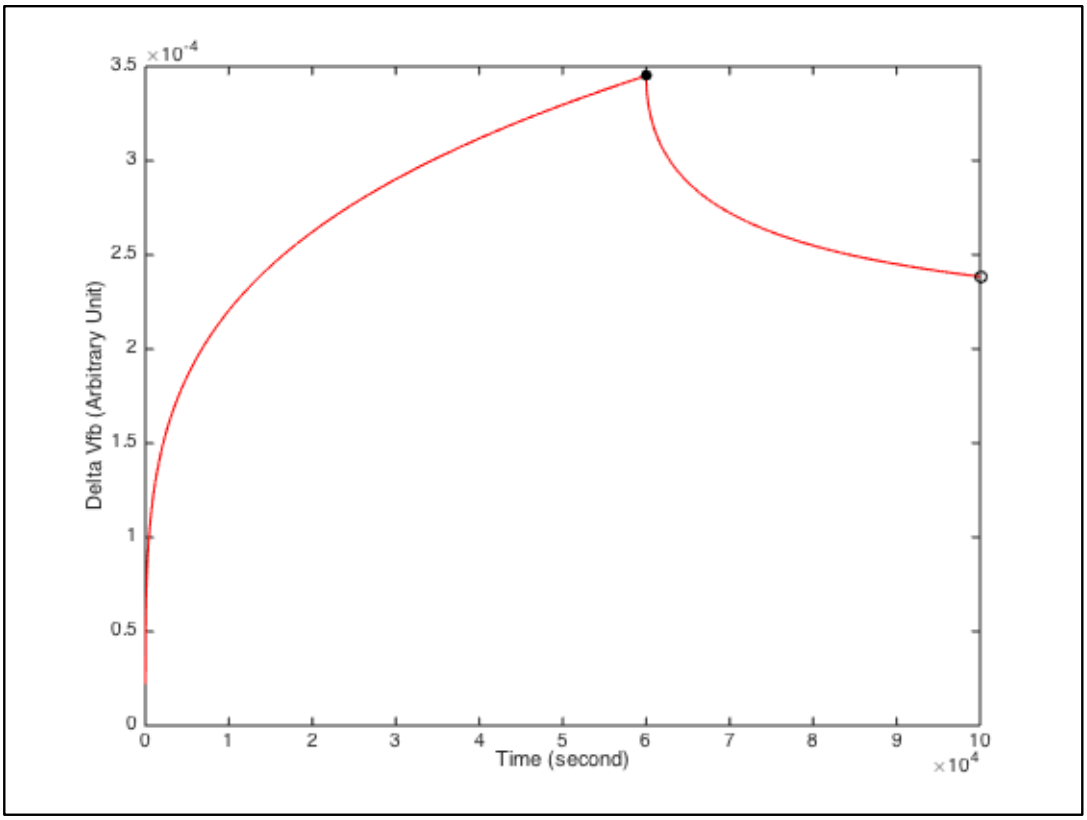


Figure 3.4: DC mode NBTI with $t_{stress} = 6e4s$ and $t_{recovery} = 4e4s$

As a consequence, AC case (or more precisely, for a random signal pattern) NBTI modelling needs more attention.

3.2 AC case NBTI Modelling

It seems that if the historical data (the signal pattern) is available for a PMOSFET, Eq. 3.11 could be adopted iteratively to precisely capture the NBTI influence. However, this is quite impractical for two reasons:

- This thesis focuses on the long term impact of NBTI (for example, 10 years); and the high frequency of the circuits means a signal could alternate multiple time even with in 1

second. This requires a huge storage space for the historical data, even for a single device.

- The long term time period and the high frequency also means a large iteration of $\mathrm{Eq.\ 3.11}$, which is computationally expensive (low time efficiency). In addition, $\mathrm{Eq.\ 3.11}$ could not be applied directly for the AC case due to the complex physics.

Therefore alternative methods are needed for AC case NBTI modelling. In spite of the complexity, some experimental results help counteract the two points above:

- Instead of the historical data, NBTI depends the duty cycle (SF) of the signal applied to the device gate; in case if the signal is aperiodic (a random alternating signal), SF is redefined as the portion of t_{total} that V_g is negative with respective to V_s/V_d.
- NBTI is frequency-independent. Given some t_{total}, signals with different frequencies result in identical ΔV_{th} [69].

Figure 3.5 show two signals: the upper one is periodic with frequency $f = 1Hz$ and $SF = 0.6$; while the lower one is a random alternating signal (aperiodic) with $t_{stress}/t_{total} = 0.6$. These two signals results in the same ΔV_{th} in the long term. This property suggests that it is possible to convert an aperiodic signal into periodic, which may bring about the convenience of calculation. From now on, the focus is shifted to periodic signals only.

Experimental results in [69] show that with the signal frequency $f \geq 1Hz$, NBTI is frequency independent.[68] gives a mathematical proof of this independency. The significance of this property is that a high frequency signal could be treated as a low frequency one with the same duty cycle such that the requirement of calculation is reduced by orders. For example, for a time period of $1,000s$, a $100Hz$ signals mean the "stress-recovery" cycle repeats for $100,000$ times while that for a $1Hz$ signal is only 1000.

[68] and [60] proposed an iterative method for AC case NBTI modelling, called *Signal Probability and Activity Factor (SPAF)* method. It takes a periodic signal with $f = 1Hz$ as the

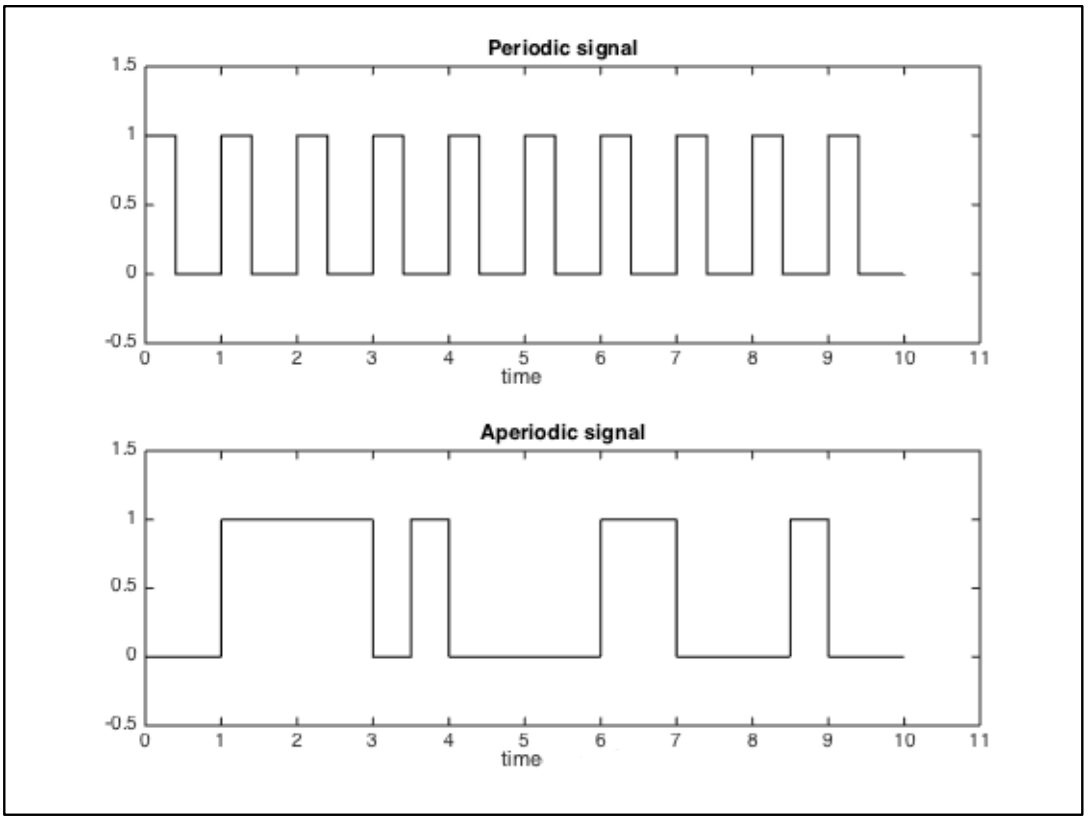


Figure 3.5: Periodic and Aperiodic signals with the same SF

input and gives the analytical expression for N_{IT} at any time point. $t = 0$ is marked as the beginning of the stress and it continues for SF(seconds) within $1s$(1 stress-recovery cycle). Therefore, for an integer t, $t + SF$ is the end of a stress phase, followed by the recovery phase which ends at $t + 1$. For the simplicity, ΔV_{th} derived from the *SPAF* method is only provided at time points $t + k$ where t is an integer and $k \in \{SF, 1\}$.

First, according to $\mathrm{Eq.\ 3.12}$, a base value is calculated for $t_{stress} = 1s$. With a constant stress condition (voltage, temperature), the bold part of $\mathrm{Eq.\ 3.12}$ is a constant, which is represented as **C**. Then:

$$\Delta V_{th}(t_{stress} = 1s) = \mathbf{C} \cdot 1^{\frac{1}{4}} = \mathbf{C} \qquad \text{(Eq. 3.13)}$$

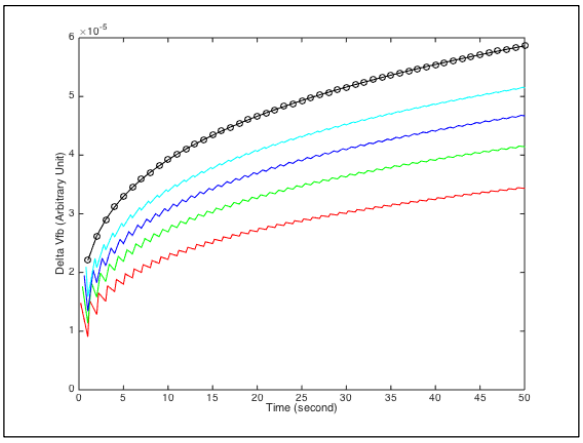


Figure 3.6: NBTI with $t_{total} = 50s$

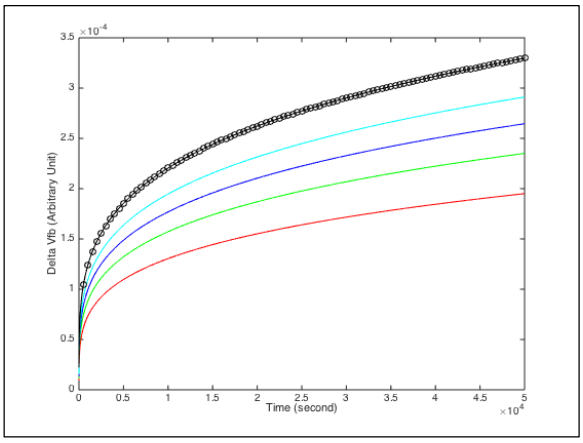


Figure 3.7: NBTI with $t_{total} = 50000s$

Then for $t \in \mathbb{Z}$ and $k \in \{SF, 1\}$, a ratio named $s(t+k)$ is defined as follows:

$$s(t+k) = \frac{\Delta V_{th}(t_{total} = t+k)}{\Delta V_{th}(t_{stress} = 1s)} = \frac{\Delta V_{th}(t_{total} = t+k)}{\mathbf{C}} \qquad \text{(Eq. 3.14)}$$

Based on the results from [60]:

$$s(t+k) = \begin{cases} (SF + s(t)^4)^{\frac{1}{4}} (k = SF) & \text{(Eq. 3.15a)} \\ \dfrac{s(t+SF) + s(t) \cdot \sqrt{\xi \cdot SP}}{1 + \sqrt{\xi \cdot SP}} (k = 1) & \text{(Eq. 3.15b)} \end{cases}$$

As is shown in Figure 3.6, the 4 zigzag-shaped curves are obtained based on Eq. 3.15a and Eq. 3.15b with $t_{total} = 50s$; from the bottom up, they correspond to SF with value of 0.2, 0.4, 0.6, and 0.8 respectively. The black-colour curve has $SF = 1$. Eq. 3.15a and Eq. 3.15b will merge into Eq. 3.12 with $SF = 1$. That's why the black curve fully overlaps with the black dots obtained according to Eq. 3.12. Figure 3.7 shows the simulation results with $t_{total} = 50000s$. The SPs are the same as Figure 3.6. It is noticed that these curves are "smooth" compared with Figure 3.6. This is just a scaling problem. In fact, data corresponding to the first $50s$ on 3.7 are exactly the same as those on Figure 3.6. Again, physical parameters are not calibrated: for one thing, they do not impact the shapes of the curves; for another, so far only the algorithm is discussed but not the physics. The calibration will be covered at the end of this chapter.

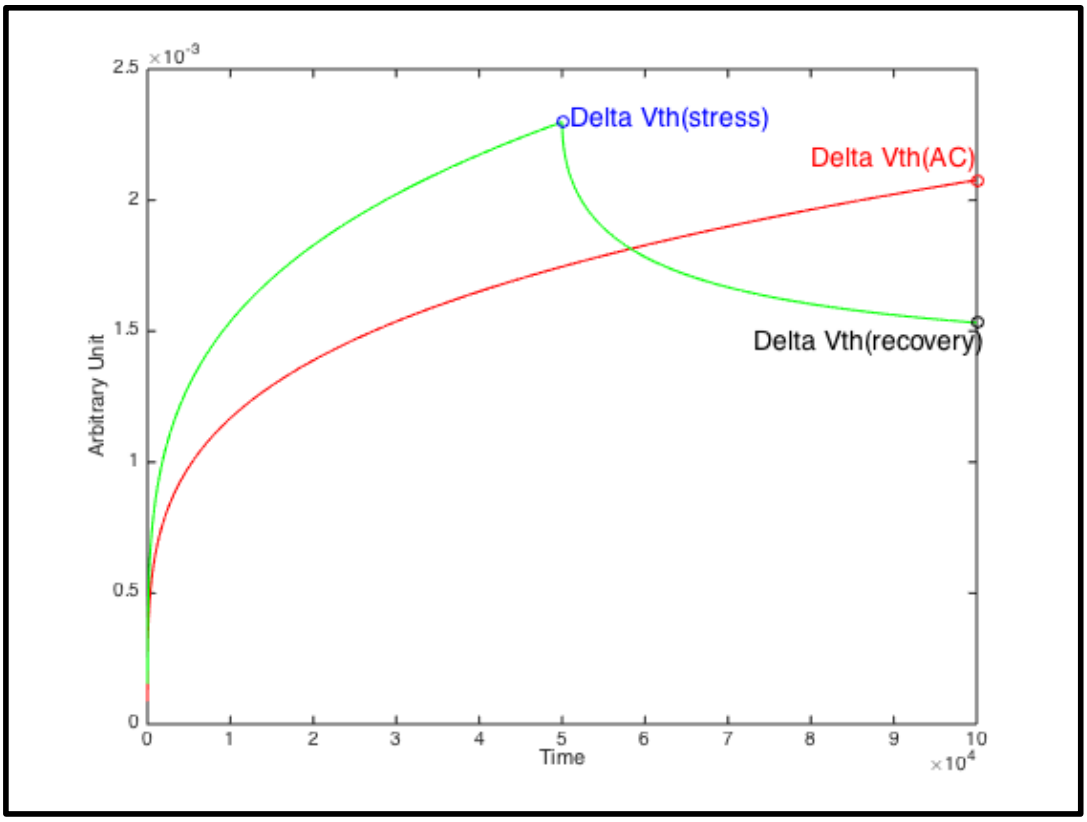


Figure 3.8: DC-NBTI VS AC-NBTI for $t_{total} = 1e5s$ and $SF = 0.5$

3.3 Comparison of DC-NBTI and AC-NBTI

The reliability modelling, be it at the device level or the gate level, requires both accuracy and time efficiency. Accuracy means the modelling result should not deviate much from the ”real” case; while time efficiency demands that the simulation result could be obtained as fast as possible. This section carries out a comparison between the DC and AC case models.

Figure 3.8 plots the curves for both the DC and AC cases with $t_{total} = 1e5s$ and $SF = 0.5$. As has been discussed, the AC mode NBTI modelling is closer to the ”real scenario” where a device is turned on and off from time to time. Therefore, the dot marked with ”Delta Vth(AC)” is the ”real” ΔV_{th} for this time period; the dot marked by ”Delta Vth(stress)” on the DC curve does not take the recovery effect into account, it leads to overestimation compared

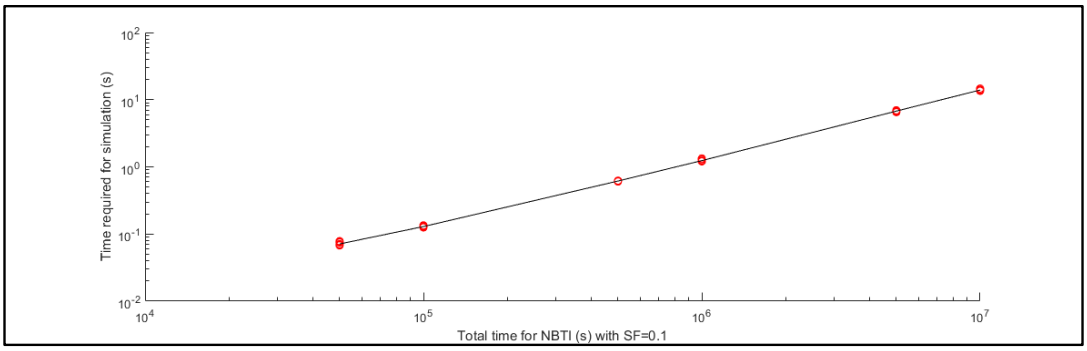


Figure 3.9: Simulation time required VS NBTI stress duration with $SF = 0.1$

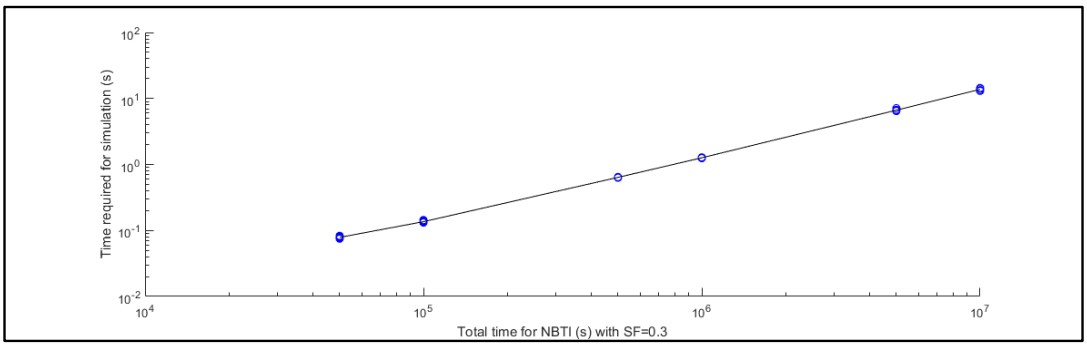


Figure 3.10: Simulation time required VS NBTI stress duration with $SF = 0.1$

with the "real" one; the dot marked by "Delta Vth(recovery)", on the other hand, results in underestimation. This is the reason that the DC NBTI model could not be directly applied to circuit simulation. The $SPAF$ method for the AC NBTI modelling, in spite of the accuracy, suffers from the time efficiency problem.

Simulations based on the $SPAF$ method are done with MatLab codes running on a quad-core 2.2GHz Macbook Pro. The stress factor SFs are chosen from $\{$0.1, 0.3, 0.4, 0.7, 0.9$\}$, and t_{total}s are from $\{$5e4, 1e5, 5e5, 1e6, 5e6, 1e7$\}$ (unit: second). For every t_{total}, the simulation repeats 10 times. The simulation time (t_{sim}) required for each SF is plotted against its respective t_{total}. As is shown in Figure 3.9-Figure 3.14:

- Once all conditions are set, t_{sim} is very stable since the 10 simulation results are almost the same for every (SF, t_{total}) combination;

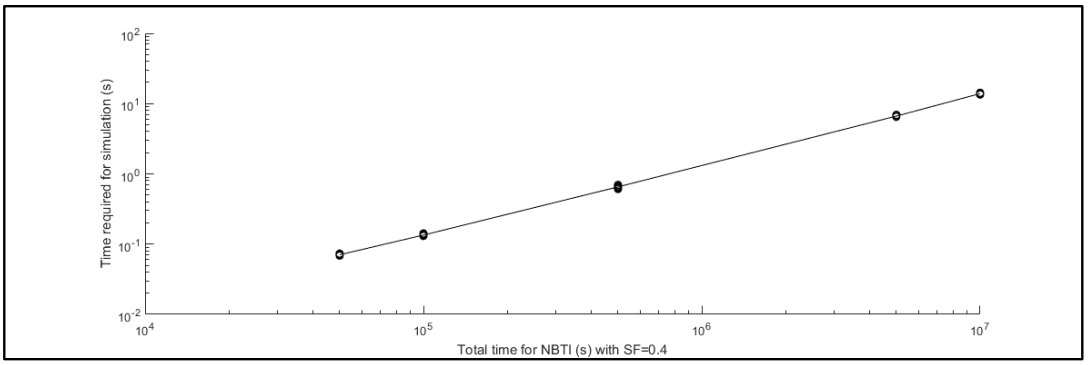


Figure 3.11: Simulation time required VS NBTI stress duration with $SF = 0.4$

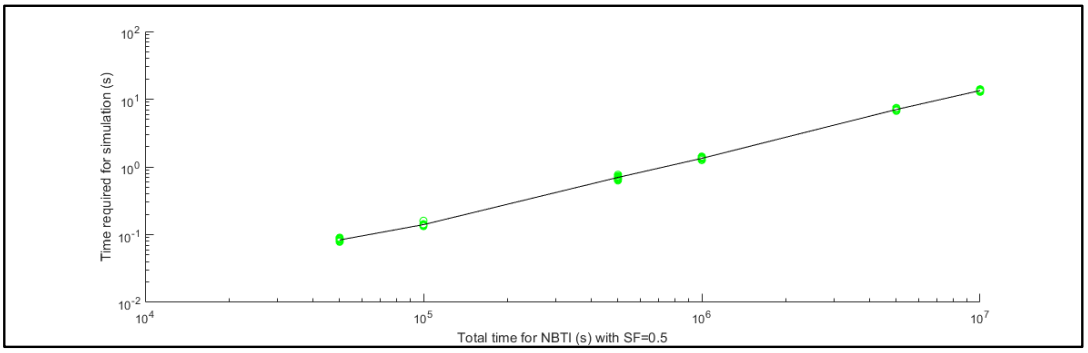


Figure 3.12: Simulation time required VS NBTI stress duration with $SF = 0.5$

■ t_{sim} is approximately a linear function of t_{total}.

This is not surprising since the code just executes $\mathrm{Eq.\ 3.15a}$ and $\mathrm{Eq.\ 3.15b}$ iteratively. Also t_{sim} is independent of SF, as is shown in Figure 3.15.

Since the $SPAF$ method possesses an $O(n)$ time complexity, it suffers from the time efficiency problem as aforementioned. For instance, to obtain the pull-up delay of a NOR gate in which two PMOSFETs are connected in series, ΔV_{th}s for both PMOSFETs are to be determined(ΔV_{th}s may differ for the two, and the reason will be covered in Chapter 5). Based on our simulation results, it takes 10mins (if the codes are in parallel, otherwise t_{sim} is doubled). This is why in [68], the author built an LUT to store ΔV_{th} according to the stressing conditions. Even so, it is not as straightforward as the DC model, where ΔV_{th} is obtained on

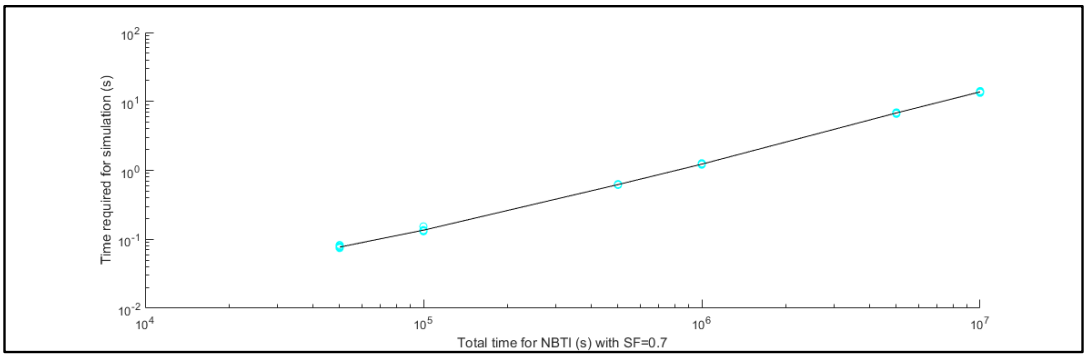


Figure 3.13: Simulation time required VS NBTI stress duration with $SF = 0.7$

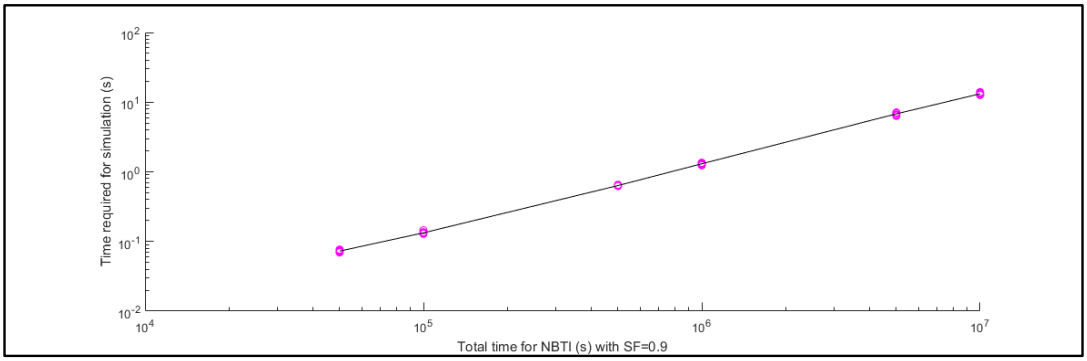


Figure 3.14: Simulation time required VS NBTI stress duration with $SF = 0.9$

”one click”.

3.4 Compact Modeling for AC NBTI

In the previous two sections, the NBTI models for both DC and AC cases are reviewed with their advantages and disadvantages discussed: the DC model generate ΔV_{th} fast, but the accuracy could not be ensured; the AC model, on the other hand, provides precise data at the sacrifice of time efficiency. To meet the requirement of the gate level NBTI modelling, the model at the device level should satisfy:

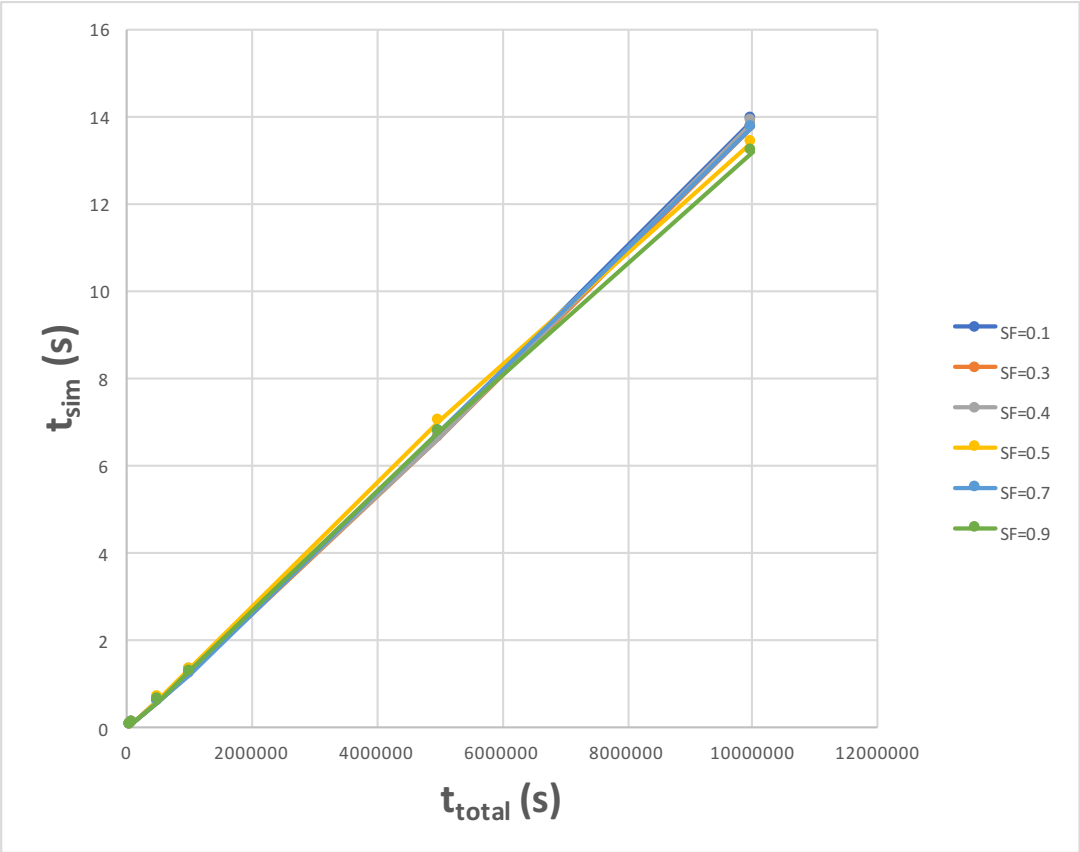


Figure 3.15: t_{sim} VS t_{total} for multiple SFs

- Equations should be simple and compact, for time efficiency;
- When calibrated with the $SPAF$ method, the simulated results should at least show equivalent accuracy.

In this section, we propose the compact modelling method that satisfies both. The inspiration of the new model comes from the comparison of the simulation results of the DC and AC models from the previous two sections.

For some t_{total} and SF, the threshold voltage shift, denoted as $\Delta V_{th,DC}(t_{total})$, could be easily obtained according to $\mathrm{Eq.\ 3.11}$; under the same circumstance, $\Delta V_{th,AC}(t_{total})$ is obtained based on $\mathrm{Eq.\ 3.13}$-$\mathrm{Eq.\ 3.15b}$. The two of them form a pair ($\Delta V_{th,DC}(t_{total})$, $\Delta V_{th,AC}(t_{total})$). This pair of ΔV_{th}s is then projected on the $x-y$ plane, where the x-axis is for the DC data and

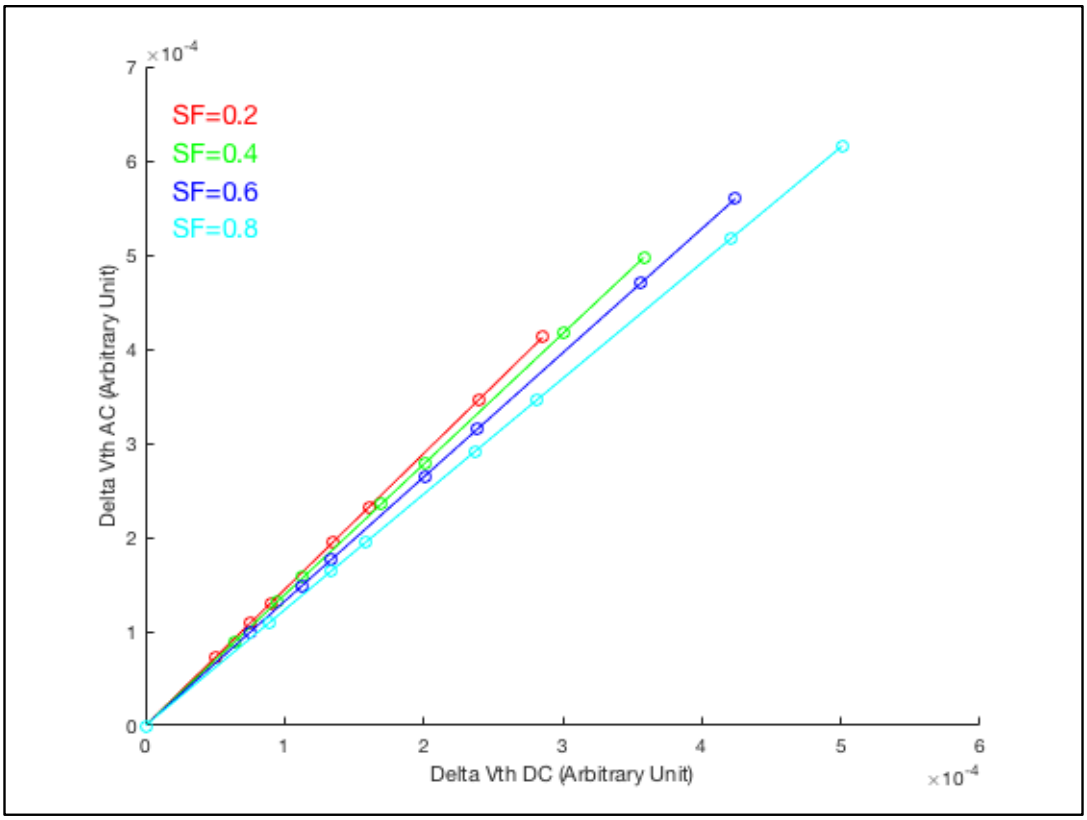


Figure 3.16: $\Delta V_{th,AC}$ VS $\Delta V_{th,DC}$

the y-axis for the AC data. For a fixed SF, 8 t_{total}s are selected to generate 8 data pairs, and they are on some "curve" on the plane. 4 SFs are chosen so that 4 curves are obtained, with $SF \in \{0.2, 0.4, 0.6, 0.8\}$.

Based on the observation of Figure 3.16 and data analysis, it is found that for a every SF, the $(\Delta V_{th,DC}(t_{total}),\ \Delta V_{th,AC}(t_{total}))$ pairs lie on straight lie crossing the original point; since these straight lines are not parallel and SF is the only varying parameter, the slope of the lines should be SF-related. More data for other SFs show that this property always holds, which means the observation in Figure 3.16 is not a coincidence, but it should be general enough. This generality has to be proven. To better carry out the demonstration, the time coefficient $\frac{1}{4}$ in $\mathrm{Eq.}\ 3.11$ and $\mathrm{Eq.}\ 3.15$a is replaced with n.

The rightmost line in 3.16 is taken as an example. Its DC data are obtained from

Eq. 3.11, in which the bold part is a constant. Together with $SF = 0.8$, they are represented by a single parameter C'.Therefore $\Delta V_{th,DC}$ could be simplified:

$$\Delta V_{th,DC}(t_{total}) = C' \cdot t_{total}^{n} \tag{Eq. 3.16}$$

The slope for a straight line is a constant that could be determined by randomly selecting two points on it. By choosing two t_{total}s ($t_{total,1}$ and $t_{total,2}$), the slope is calculated as:

$$\begin{aligned} slope &= \frac{\Delta V_{th,AC}(t_{total,2}) - \Delta V_{th,AC}(t_{total,1})}{\Delta V_{th,DC}(t_{total,2}) - \Delta V_{th,DC}(t_{total,1})} \\ &= \frac{\Delta V_{th,AC}(t_{total,2}) - \Delta V_{th,AC}(t_{total,1})}{C' \cdot (t_{total,2}^{\mathbf{n}}) - C' \cdot (t_{total,1}^{\mathbf{n}})} \\ &= \frac{\Delta V_{th,AC}(t_{total,2}) - \Delta V_{th,AC}(t_{total,1})}{C' \cdot (t_{total,2}^{\mathbf{n}} - t_{total,1}^{\mathbf{n}})} \end{aligned} \tag{Eq. 3.17}$$

The value of the slope is independent of the choices of $t_{total,1}$ and $t_{total,2}$. Therefore, a good guess is that $\Delta V_{th,AC}$ takes the form of Eq. 3.18, in which X is a SF-related parameter:

$$\Delta V_{th,AC}(t_{total}) = X \cdot t_{total}^{n} \tag{Eq. 3.18}$$

The expression in Eq. 3.18 might be the only possible solution for $\Delta V_{th,AC}$ since it could successfully explain the fact that the ($\Delta V_{th,DC}(t_{total})$, $\Delta V_{th,AC}(t_{total})$) pairs fall on a straight line. Eq. 3.18 convert Eq. 3.17 into:

$$\begin{aligned} slope &= \frac{\Delta V_{th,AC}(t_{total,2}) - \Delta V_{th,AC}(t_{total,1})}{C' \cdot (t_{total,2}^{\mathbf{n}} - t_{total,1}^{\mathbf{n}})} \\ &= \frac{X(t_{total,2}^{n} - t_{total,1}^{n})}{C'(t_{total,2}^{\mathbf{n}} - t_{total,1}^{\mathbf{n}})} = \frac{X}{C'} \end{aligned} \tag{Eq. 3.19}$$

For one thing, the final expression $\frac{X}{C'}$ is irrelevant to t_{total}; for another, as X is SF-related, the straight lines in Figure 3.16 are not parallel. In the following part, the expression of X is given.

According to Eq. 3.13, Eq. 3.14, and Eq. 3.15a, for a sufficiently large integer t ($t \gg$

1), $s(t) \gg SF$. Then $s(t)^{\frac{1}{n}} \gg SF$ since $n = \frac{1}{4}$. Eq. 3.15a could be expanded:

$$\begin{aligned} s(t+k) &= (s(t)^{\frac{1}{n}} + SF)^n \\ &= (s(t)^{\frac{1}{n}})^n + n \cdot (s(t)^{\frac{1}{n}})^{n-1} \cdot SF \\ &= s(t) + n \cdot s(t)^{1-\frac{1}{n}} \cdot SF \end{aligned} \tag{Eq. 3.20}$$

By substituting Eq. 3.20 into Eq. 3.15b:

$$\begin{aligned} s(t+1) &= \frac{s(t) + n \cdot s(t)^{1-\frac{1}{n}} \cdot SF + s(t) \cdot \sqrt{\xi \cdot SP}}{1 + \sqrt{\xi \cdot SP}} \\ &= s(t) + \frac{n \cdot SF}{1 + \sqrt{\xi \cdot SF}} \cdot s(t)^{1-\frac{1}{n}} \end{aligned} \tag{Eq. 3.21}$$

The relation between $s(t+1)$ and $s(t)$ could be described with Taylor Expansion. With $t \gg 1$, by reasonably truncating the high order terms:

$$s(t+1) = s(t) + \frac{s'(t)}{1!} \cdot 1 \tag{Eq. 3.22}$$

In Eq. 3.22, $s'(t)$ is the derivative of $s(t)$ with respective to t. Eq. 3.21 and Eq. 3.22 are the same since they both describes $s(t+1)$. By equating their right hand sides (RHS):

$$s'(t) = \frac{n \cdot SF}{1 + \sqrt{\xi \cdot SF}} \cdot s(t)^{1-\frac{1}{n}} \tag{Eq. 3.23}$$

Eq. 3.23 is an ordinary differential equation. It could be solved with the following solution:

$$s(t) = (\frac{1}{1 + \sqrt{\xi \cdot SP}})^n \cdot SF^n \cdot (t + C_0)^n \tag{Eq. 3.24}$$

C_0 is a random constant in Eq. 3.24 if there is no boundary conditions. To satisfy the condition that $s(t) = 0$ at $t = 0$, C_0 is set as 0. Then the expression for $s(t)$ becomes:

$$s(t) = (\frac{1}{1 + \sqrt{\xi \cdot SP}})^n \cdot SF^n \cdot t^n \tag{Eq. 3.25}$$

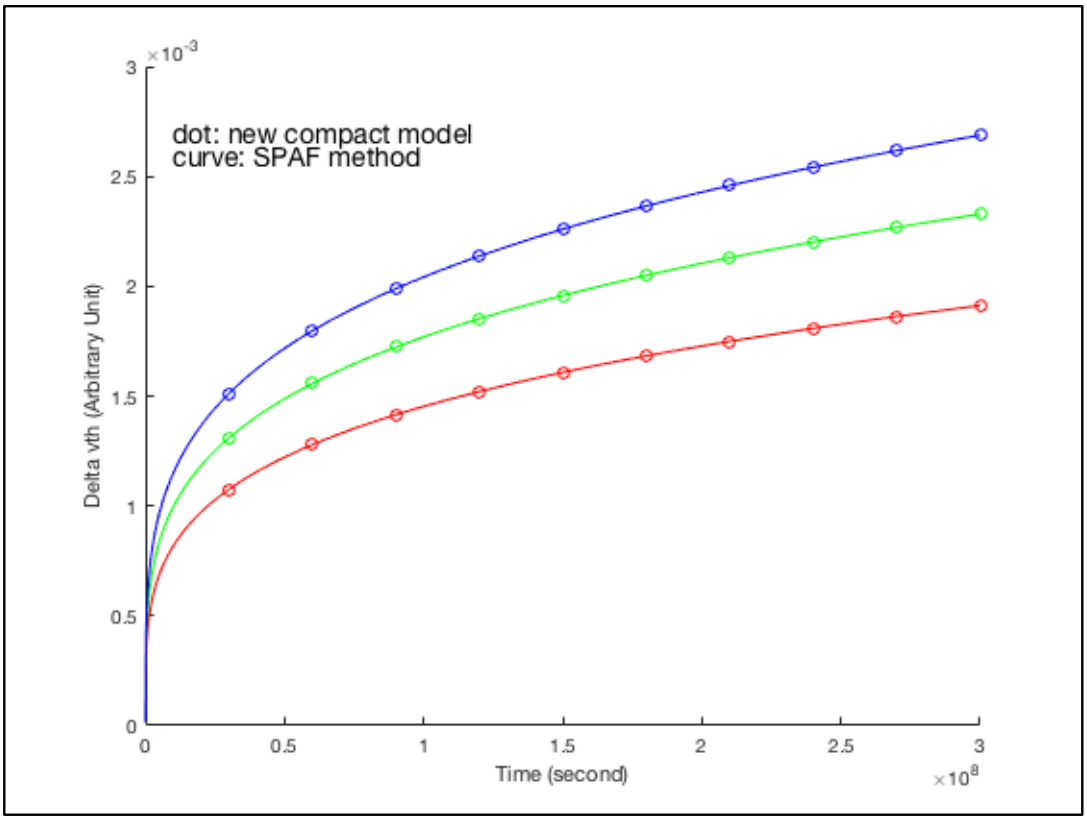


Figure 3.17: $\Delta V_{th,AC}$ generated from the $SPAF$ method and the new compact model

The combination of Eq. 3.11, Eq. 3.13, Eq. 3.14, and Eq. 3.25 results in:

$$\Delta V_{th,AC}(t_{total}) = (\frac{1}{1+\sqrt{\xi \cdot SP}})^n \cdot \frac{q}{C_{ox}} \sqrt{\frac{k_f N_0}{2k_r}} \cdot D_H^n \cdot SF^n \cdot t_{total}^n \qquad \text{(Eq. 3.26)}$$

Eq. 3.26 is a reconstruction of the $SPAF$ method. It takes the similar form as the DC model in Eq. 3.11, therefore ΔV_{th} could be obtained on "one click" instead of the iterative method. t_{sim} could therefore be significantly reduced to achieve the time efficiency. In addition, mathematically it generates data that are the same as that from the $SPAF$ method, so the accuracy could be guaranteed. As is shown in Figure 3.17, for multiple SFs, the dots, obtained from the compact model Eq. 3.26, all fall on the curves plotted according to the $SPAF$ method. The former only takes 10 "clicks", while it takes 10mins to generate the latter.

Table 3.1: A comparison of the slopes obtained from DC & AC simulation and calculation

SP	Simulated Results	Calculated Result (Eq. 3.28)
1	1	1
0.9	1.1635	1.1634
0.7	1.2784	1.2783
0.5	1.3555	1.3554
0.3	1.4172	1.4170
0.1	1.4697	1.4696

Apart from ΔV_{th}, the "guessed" parameter X is also derived from Eq. 3.26:

$$X = (\frac{1}{1+\sqrt{\xi \cdot SP}})^n \cdot \frac{q}{C_{ox}} \sqrt{\frac{k_f N_0}{2k_r}} \cdot D_H^n \cdot SF^n \qquad \text{(Eq. 3.27)}$$

It could be seen from Eq. 3.27, with all other stressing conditions set, X is only related to SF (since $SP = 1 - SF$). Meanwhile, the slopes of the straight lines could be analytically expressed as:

$$slope = \frac{X}{C'} = (\frac{1}{1+\sqrt{\xi \cdot SP}})^{n-1} \qquad \text{(Eq. 3.28)}$$

Then the DC model and the $SPAF$ method are adopted again with $SF \in \{$1, 0.9, 0.7, 0.5, 0.3, 0.1$\}$, and 6 slopes are obtained from simulation. Next, the SP values are substituted into Eq. 3.28 to generate 6 calculated results. As is shown in Table 3.1, the results from the simulation and those from the calculation are highly consistent with each other, which proves the correctness of the new model equations.

[13] provides the values of the parameters related to NBTI, which are listed below:

□ $k_f : 10^{-2}$;

□ $k_r : 10^{-18}$;

□ $N_0 : 1.24 \cdot 10^{14}$;

□ $C_{ox} : 2.6562 \cdot 10^{-6}$(since $t_{ox} = 1.3nm$);

□ D_H : $1.13 \cdot 10^{-16}$([13] does not explicitly give the value of D_H, this value is obtained by backward retrieval based on Figure 4 of [13]).

Then a graph showing the relation between ΔV_{th} and SF is given:

Taylor Expansion is adopted several times during the derivation of the new model equations. High order terms are truncated since they are less significant: this is only reasonable when the time t is sufficiently large such that $s(t) \gg SF$. As a result, for a small t_{total}, $\mathrm{Eq.}\ 3.26$ may bring in error.

In Figure 3.19 and Figure 3.20, a comparison is carried out for the $SPAF$ method and the new model. It could be seen in Figure 3.19 that for $t_{total} < 100s$ there is a slight difference of the simulation results for the two models; but the curves quickly merge and overlap with large t_{total}, as is shown in Figure 3.20.

But this difference is not a big issue. For one thing, when focused on the long term, the error vanishes, which is shown in Figure 3.20 and Figure 3.17; for another, in case t_{total} is small, the $SPAF$ method could still be applied since the iteration times is not that large and t_{sim} required is small, and the time efficiency could still be guaranteed.

In case of H_2 diffusion, $\mathrm{Eq.}\ 3.11$ is modified:

$$\Delta V_{th}(t_{total}) = \frac{\mathbf{1}}{\mathbf{1} + \sqrt{\boldsymbol{\xi} \cdot SP}} \cdot \frac{\mathbf{q}}{\mathbf{C_{ox}}} \left(\frac{\mathbf{k_f N_0}}{\mathbf{k_r}}\right)^{\frac{2}{3}} \cdot \mathbf{D_H}^{\frac{1}{6}} \cdot \mathbf{SF}^{\frac{1}{6}} \cdot t_{total}^{\frac{1}{6}} \qquad \text{(Eq. 3.29)}$$

In addition, in $\mathrm{Eq.}\ 3.15$a, "4" is changed to "6" while $\frac{1}{4}$ is modified to $\frac{1}{6}$. And according to $\mathrm{Eq.}\ 3.26$, the new compact equation for AC-case NBTI modelling is:

$$\Delta V_{th,AC}(t_{total}) = \left(\frac{1}{1 + \sqrt{\xi \cdot SP}}\right)^{\frac{1}{6}} \cdot \frac{q}{C_{ox}} \left(\frac{k_f N_0}{k_r}\right)^{\frac{2}{3}} \cdot D_H^{\frac{1}{6}} \cdot SF^{\frac{1}{6}} \cdot t_{total}^{\frac{1}{6}} \qquad \text{(Eq. 3.30)}$$

As is shown in Figure 3.21, even with the modification of the time coefficient, the new compact model equation gives consistent results as the $SPAF$ method. In Figure 3.21, the curves are generated from the $SPAF$ method while the dots are from the compact model equation. From low to high, the four curves correspond to SF values of 0.2, 0.4, 0.6, and 0.8 respectively.

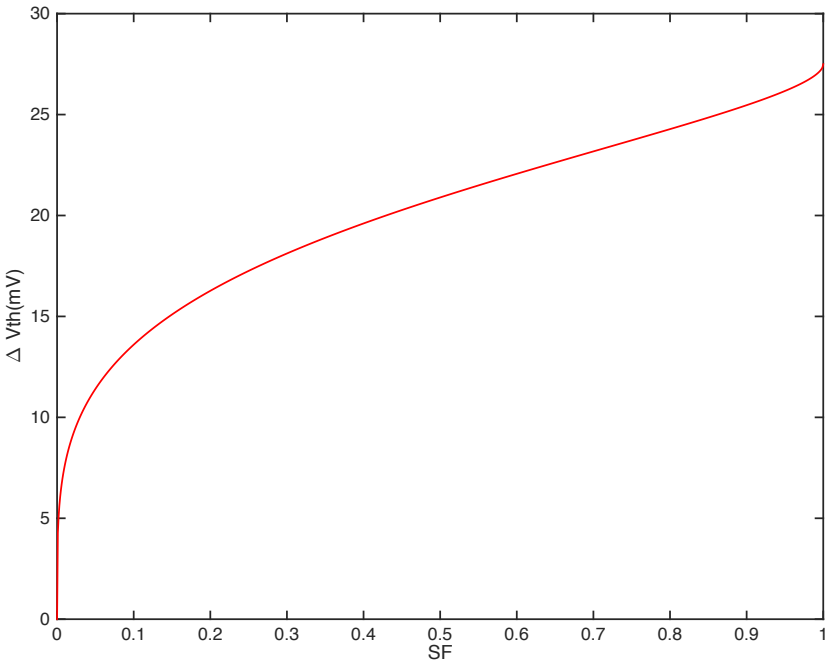


Figure 3.18: ΔV_{th} vs SF

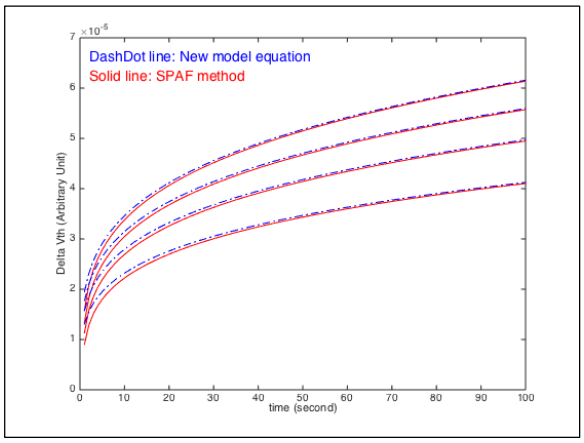


Figure 3.19: Old VS New AC NBTI models for $t_{total} = 100s$

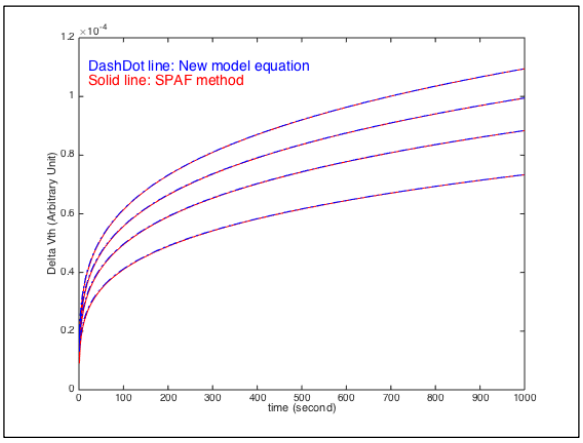


Figure 3.20: Old VS New AC NBTI models for $t_{total} = 1000s$

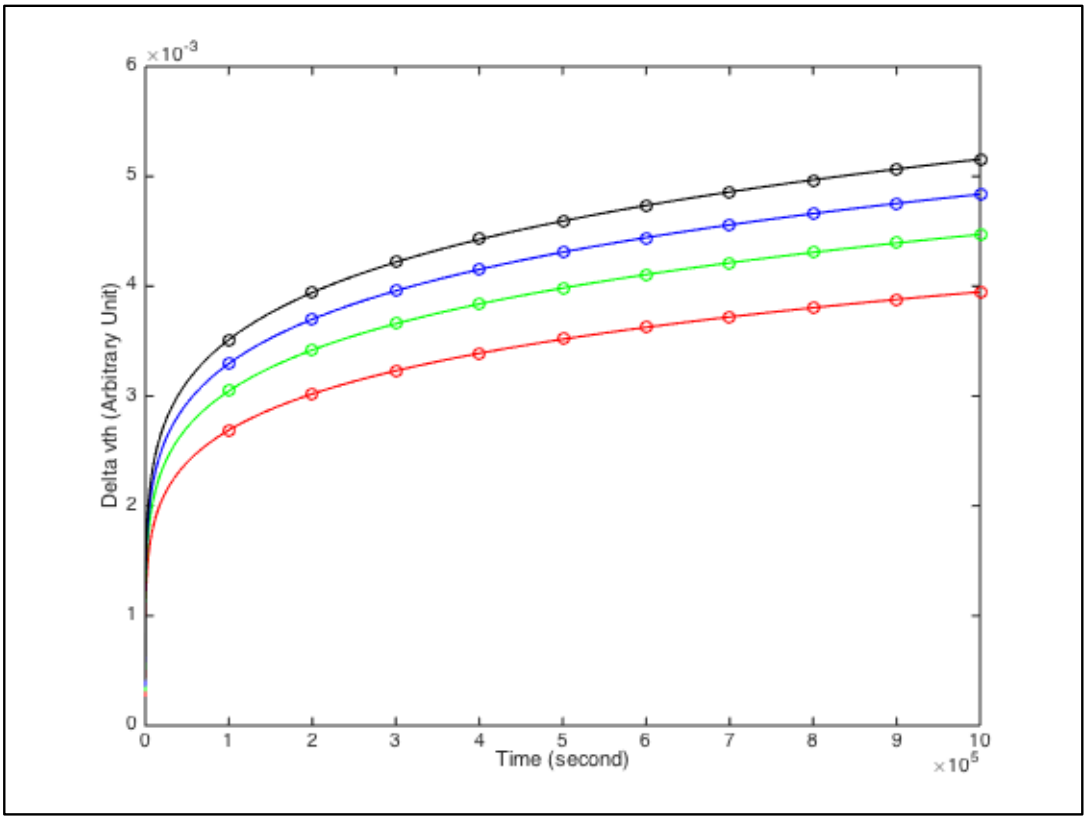


Figure 3.21: $\Delta v_{th,AC}$ generated from the $SPAF$ method and the new compact model with $n = \frac{1}{6}$

Table 3.2: Relation between simulation time (t_{sim}) and (t_{total}, SF) combination for the iterative method

	$SF = 0.1$	$SF = 0.2$	$SF = 0.3$	$SF = 0.4$
$t_{total} = 1e6(s)$	4.35	4.36	4.38	4.38
$t_{total} = 2e6(s)$	8.73	8.75	8.75	8.91
$t_{total} = 3e6(s)$	14.18	13.30	13.59	13.80
$t_{total} = 4e6(s)$	17.73	17.69	18.07	18.48
$t_{total} = 5e6(s)$	23.38	21.87	21.94	21.89
$t_{total} = 6e6(s)$	26.12	26.15	26.16	26.20

Table 3.3: Relation between simulation time (t_{sim}) and (t_{total}, SF) combination for the compact method

	$SF = 0.1$	$SF = 0.2$	$SF = 0.3$	$SF = 0.4$
$t_{total} = 1e6(s)$	7.25e-3	6.49e-3	7.39e-3	7.50e-3
$t_{total} = 2e6(s)$	7.13e-3	6.56e-3	6.63e-3	7.29e-3
$t_{total} = 3e6(s)$	7.25e-3	6.99e-3	7.24e-3	7.27e-3
$t_{total} = 4e6(s)$	7.39e-3	6.91e-3	7.19e-3	7.28e-3
$t_{total} = 5e6(s)$	6.49e-3	7.02e-3	7.55e-3	6.91e-3
$t_{total} = 6e6(s)$	7.27e-3	7.06e-3	6.74e-3	7.35e-3

Both the $SPAF$ method and the compact model have been implemented with Matlab code. Figure 3.22 is the simple code for both methods:

The proposed compact model is simpler and faster compared with the $SPAF$ method: it obtains the result directly instead of iteratively. Both methods have been programmed in Matlab. We have run both programs on a quad-core Macbook pro, and the simulation times (t_{sim}) for multiple t_{total} and SF combinations are tracked and stored in Table 3.2. Every table entry stores the t_{sim} for the respective (t_{total}, SF) combination. While those for the compact method are stored in Table 3.3.

Table 3.2 shows that for a certain t_{total}, t_{sim} is almost irrelevant to SF; on the other hand, t_{sim} is approximately linearly related to t_{total}. This is not surprising: since the number of calculations (n_{sim}) required to accomplish the simulation has a simple relation with only t_{total}:

8/3/18 6:43 PM /Users/xuliu/.../iterative_compact_thesis.m 1 of 1

```
function NBTI_iteration(time,stress_factor,kesi,time_coefficient)

t_stress(1)=stress_factor;
t_recovery(1)=1;
Sm(1)=power(t_stress(1),time_coefficient);
S1(1)=Sm(1)./(1+sqrt(kesi.*(1-stress_factor)));
for j=2:time
    t_stress(j)=stress_factor+j-1;
    t_recovery(j)=j;
    Sm(j)=power(stress_factor+power(S1(j-1),1./time_coefficient),time_coefficient);
    S1(j)=(Sm(j)+S1(j-1).*sqrt((1-stress_factor)./2./1))./(1+sqrt((1-stress_factor)./2.↙
/1));
end
for k=1:time
    timing(2.*k-1)=t_stress(k);
    timing(2.*k)=t_recovery(k);
    S(2.*k-1)=Sm(k);
    S(2.*k)=S1(k);
end
plot(timing, S, 'r')
end
function COMPACT(time,stress_factor,kesi,time_coefficient)
for i=1: time
    t_recovery_compact(i)=i;
    compact_value(i)=power(1./(1+sqrt(kesi.*(1-stress_factor))),time_coefficient)...
.*power(stress_factor,time_coefficient).*t_recovery_compact(i).^time_coefficient;
end
plot(t_recovery_compact,compact_value,'b')
hold on
save('t_recovery_compact')
save('compact_value')
end
%time: unit is second;
%stress_factor: 0<=stress_factor<=1;
%kesi: kesi has a value of 0.5 for double-sided diffusion during recovery;
%time_coefficient: possessing a value of 0.25 for Hydrogen atom diffusion;
%t_stress(1): time point of the end of the first stress cycle;
%t_recovery(1): time point of the end of the first recovery cycle;
%t_stress(j): time point of the end of the j-th stress cycle;
%t_recovery(j): time point of the end of the j-th recovery cycle;
%Sm(1): Vth shift at the end of the first stress cycle;
%S1(1): Vth shift at the end of the first recovery cycle;
%Sm(j): Vth shift at the end of the j-th stress cycle;
%S1(j): Vth shift at the end of the j-th recovery cycle;
%timing(): an array of time consisting of t_stress(j) and t_recovery(j);
%S(): an array of Vth shift values consisting of Sm(j) followed by S1(j);
%t_recovery_compact(): the array of time points for the compact model;
%compact_value(): the array of Vth shifts for the compact model.
```

Figure 3.22: NBTI code for both the iterative model and the compact model

$n_{sim} = 2 \cdot t_{total}$.

Table 3.3 shows that t_{sim} for the compact method is neither related to SF nor t_{total}, and t_{sim} is around 7e-3s. This is reasonable since the number of calculations required for the compact method is always 1. The efficiency improvement could be observed by simply reading both tables.

While the proposed compact model achieves the simplicity and time efficiency, it is meaningful to carry out a further comparison between the $SPAF$ method and the compact model:

The $SPAF$ method is a continuous "real time" model: given an integer t and the stress factor SF, it not only gives $s(t)$ and $s(t + SF)$, but also any $s(t + \Delta t)$ with $\Delta t \in (0, 1)$. [60] provides the following equation:

$$s(t+\Delta t)_{iterative} = \begin{cases} (\Delta t + s(t)^{\frac{1}{n}})^n, & 0 < \Delta t \leq SF \\ \frac{s(t+SF)+s(t)\cdot\sqrt{\xi\cdot\frac{\Delta t - SF}{\Delta t}}}{1+\sqrt{\xi\cdot\frac{\Delta t - SF}{\Delta t}}}, & SF < \Delta t \leq 1 \end{cases} \quad \text{(Eq. 3.31)}$$

Based on Equation $\mathrm{Eq.\ 3.34}$, a graph is plotted for the first 10 stress-recovery cycles with $SF = 0.4$: our compact method is a discrete method, which is only able to capture ΔV_{th} for the integer t. From this sense, the iterative method contains more abundant information. Meanwhile, our compact model evolves from the $SPAF$ method, where $\mathrm{Eq.\ 3.34}$ is manipulated with mathematical tools (including Taylor Expansion). Only the significant terms are kept while those insignificant are dropped: errors are thus introduced especially for small t, where the contributions of those insignificant terms could not be ignored.

As is shown in Figure 3.24 where t_{total} is only $10s$, results from the compact method deviate from those of the iterative method with a significant overestimation; nevertheless, with t_{total} increasing, results from the two methods gradually approach each other and merge, as is shown in Figure 3.25. Quantitively, the ratio of $\frac{s(t)_{iterative}}{s(t)_{compact}}$ ($s(t)_{iterative}$ is the result from the iterative method and $s(t)_{compact}$ is that from the compact method) is calculated and plotted for

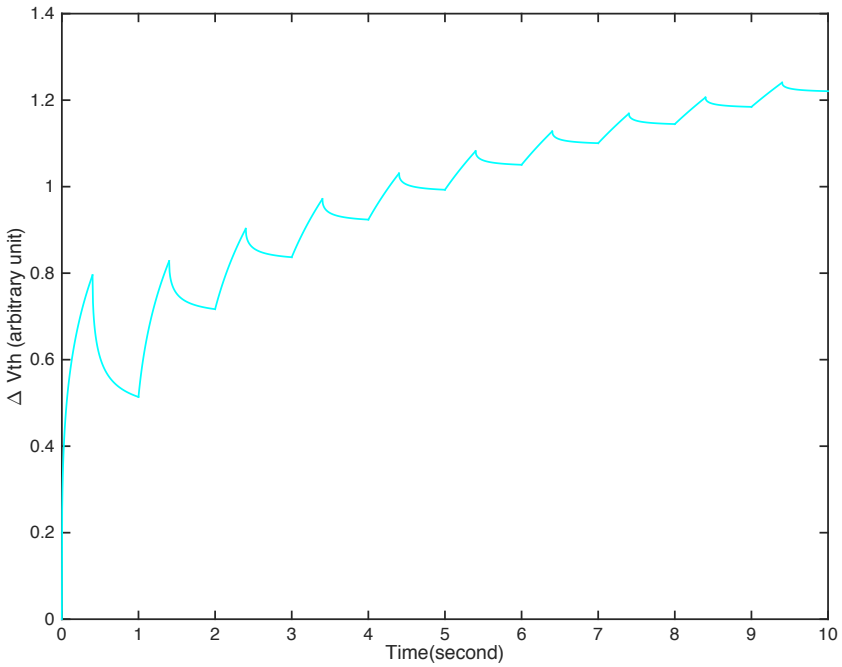


Figure 3.23: ΔV_{th} vs time for the SPAF method

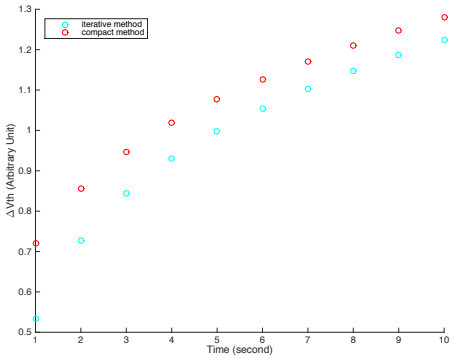


Figure 3.24: Iterative VS compact NBTI models for $t_{total} = 10s$

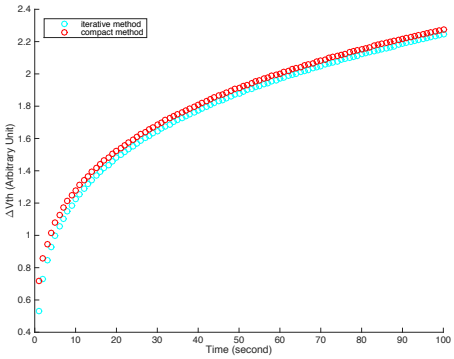


Figure 3.25: Iterative VS compact NBTI models for $t_{total} = 100s$

multiple SFs:

Figure 3.26 shows that for $t_{total,0} < 80s$, the compact method overestimates the result. Beyond $t_{total,0}$, results from the two methods differ little. For $t_{total} = t_{total,0}$, simulation time required to obtain $s(t)_{iterative}$ is 5e-3s (based on our simulation), which is even less than those of the compact method shown in Table 3.3.

Therefore, the iterative method and the compact method could be reasonably combined to improve both the accuracy and the time efficiency. The pseudo code is as follows:

---

if $(t_{total} \leq t_{total,0})$

$s(t) = s(t)_{iterative}$ //comment: for accuracy

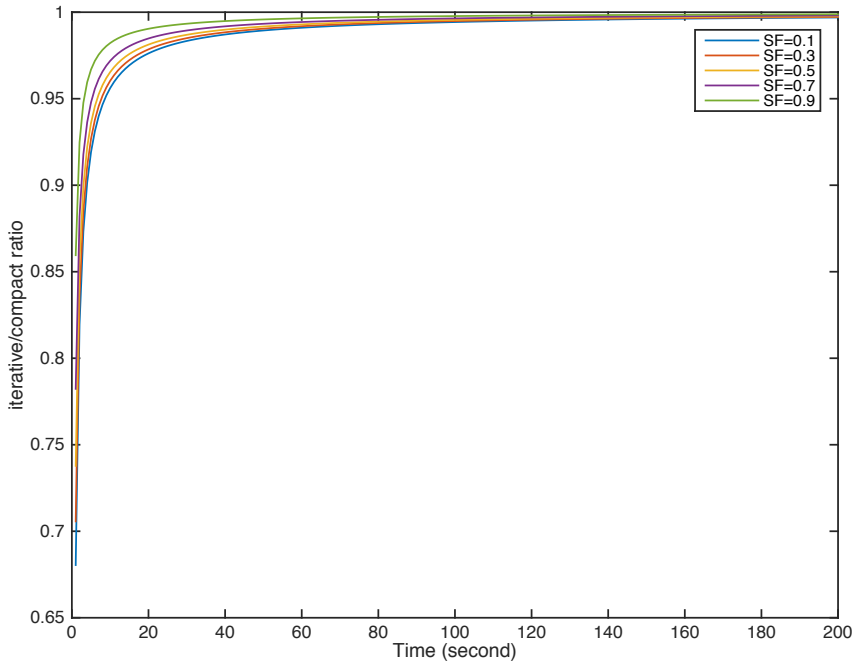


Figure 3.26: $\frac{s(t)_{iterative}}{s(t)_{compact}}$ for multiple SFs

else

$s(t) = s(t)_{compact}$ //comment: for time efficiency

As we mentioned earlier, our compact model is not continuous, and it is not capable of generating the result for a non-integer t_{total}. But with the aid of the iterative method, this could be realised:

For an integer t, the compact method quickly generates $s(t)_{compact}$ (which is also the disadvantage of the iterative method-long simulation time required to obtain $s(t)_{i}terative$), the substitution of $s(t)_{compact}$ into the first part of Eq. 3.34 gives $s(t + \Delta t)$ for $\Delta t \in (0, SF]$:

$$s(t + \Delta t) = (\Delta t + s(t)_{compact}^{\frac{1}{n}})^n \quad \text{(Eq. 3.32)}$$

Eq. 3.32 will also give $s(t + SF)$:

$$s(t + SF) = (SF + s(t)_{compact}^{\frac{1}{n}})^n \quad \text{(Eq. 3.33)}$$

Then for $\Delta t \in (SF, 1)$, according to the second part of Eq. 3.34:

$$s(t + \Delta t) = \frac{s(t + SF) + s(t)_{compact} \cdot \sqrt{\xi \cdot \frac{\Delta t - SF}{\Delta t}}}{1 + \sqrt{\xi \cdot \frac{\Delta t - SF}{\Delta t}}} \quad \text{(Eq. 3.34)}$$

The pseudo code for this process (for a random time point t_{random}) is as follows:

---

t_{int}=(**int**) t_{random}; //comment: find the integer no larger than t_{random}

$t_{delta} = t_{random} - t_{int}$; //comment: find the fractional part

if ($t_{delta} == 0$)

$s(t_{random}) = s(t_{int})_{compact}$; //comment: t_{random} happens to be an integer

else if($t_{delta} \leq SF$)

$$s(t_{random}) = (t_{delta} + s(t_{int})_{compact}^{\frac{1}{n}})^n$$

else

$$s(t_{random}) = \frac{s(t_{int}+SF)+s(t_{int})_{compact}\cdot\sqrt{\xi.\frac{t_{delta}-SF}{t_{delta}}}}{1+\sqrt{\xi.\frac{t_{delta}-SF}{t_{delta}}}}$$

In such a way, the iterative method and the compact method complement each other:

- For a short t, the iterative method is adopted to avoid the overestimation in the compact method, and the simulation time required is small;

- For a long t(t is an integer), the compact method is adopted to avoid the long simulation time, with the accuracy of the simulation result secured;

- For a long t_{random}, the compact method is first adopted to obtain $s(t_{int})$ (and therefore bypass the huge number of iterative cycles); then the iterative method is adopted for at most one stress-recovery cycle to obtain $s(t_{random})$. Both the time efficiency and the accuracy could be secured.

We ran the simulation with $t_{random} = 1000000.7s$, first with the iterative method, which takes 3.72s; and then the combination of the iterative and compact methods is adopted, which takes 0.32s. Figure 3.27 shows that the result from the iterative method and that from the combination are the same. But the time efficiency is greatly improved.

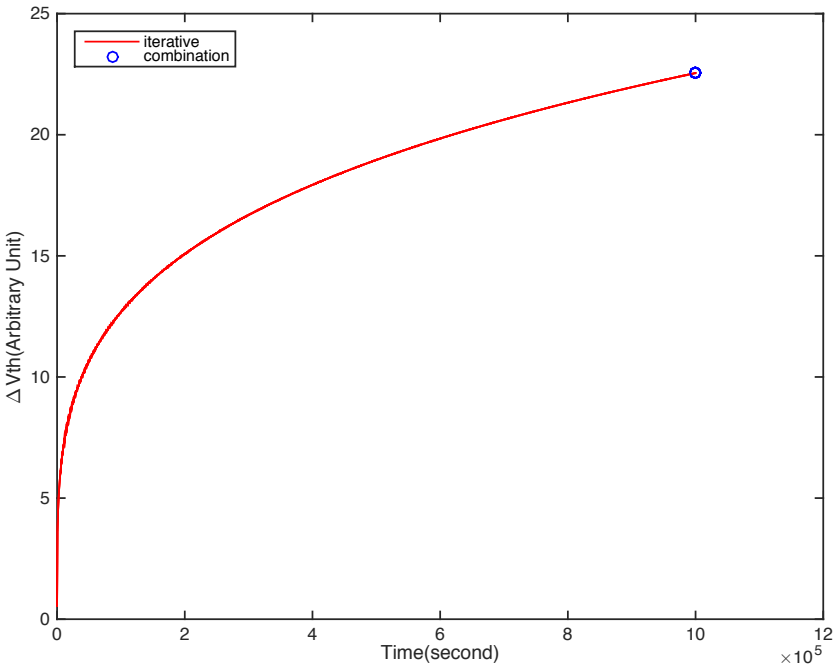


Figure 3.27: $s(t_{random})$ with the combination of the iterative and the compact methods

3.5 The "AgeGate" model and SF determination

The "AgeGate" is a gate level reliability model proposed by Dr. Dominik Lorenz in [29] [70]. It is a canonical gate model that describes the relation between gate delay degradation(ΔD and D_{aged}) and the parameter shift (Δp) of every single transistor. This model relies much on accurate description of Δp, which is also our ultimate target. The model could be described by the following simple equation:

$$D_{aged} = D_{fresh} + \Delta D = D_{fresh} + \sum_{m \in G} \sum_{p \in P} \frac{\partial D}{\partial p_{m,P}} \cdot \Delta p_m \qquad \text{(Eq. 3.35)}$$

In $\mathrm{Eq.\ 3.35}$:

- m: a transistor of the gate;
- G: the set of all transistors of the gate;
- P: parameters of interest that are influenced by aging (V_{th} for NBTI and I_{on} for HCI);
- $\frac{\partial D}{\partial p_{m,P}}$: sensitivity of D with respect to a certain parameter drift (for example, V_{th}) from transistor m.

The AgeGate model takes both NBTI and HCI into consideration. To make the thesis more readable, this Chapter only covers NBTI in the AgeGate, while that of HCI is added in in next chapter. The $\mathrm{Eq.\ 3.35}$ is simplified to:

$$D_{aged}(NBTI) = D_{fresh} + \Delta D = D_{fresh} + \sum_{m \in G} \frac{\partial D}{\partial V_{th,m}} \cdot \Delta V_{th,m} \qquad \text{(Eq. 3.36)}$$

Since the AgeGate model takes the parameter shift of every single transistor into account, the pessimism caused by using only a "worst case" parameter shift could be avoided. However, this model uses $\mathrm{Eq.\ 3.5}$ and $\mathrm{Eq.\ 3.10}$ to obtain ΔV_{th}, and the recovery effect is ignored. This is regarded as a trade-off for time efficiency. The stress time in $\mathrm{Eq.\ 3.5}$ is calculated as $(1 - SP) \cdot t_{total}$. The a ratio of $\Delta V_{th,o}$ (V_{th} shift without consideration of the recovery effect) over

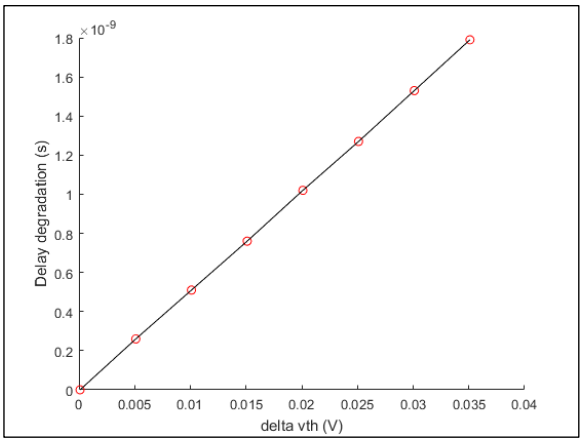


Figure 3.28: Linear relation between inverter rising delay degradation and threshold voltage shift

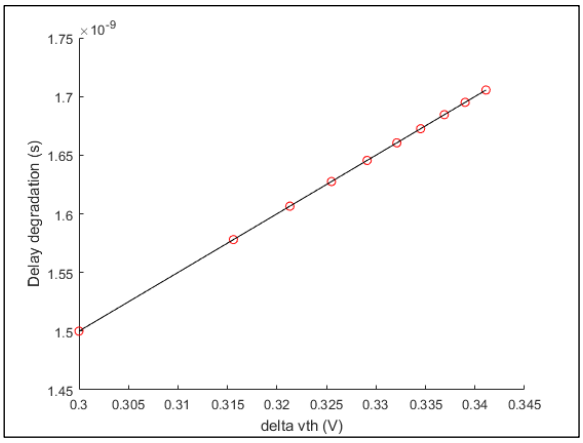


Figure 3.29: Delay degradation overestimation for $SP \sim (0.1 - 0.9)$ without the recovery effect

$\Delta V_{th,w}$(V_{th} shift with the recovery effect considered) is:

$$\frac{V_{th,o}}{V_{th,w}} = (1 + \sqrt{\xi \cdot SP})^{\frac{1}{4}} \qquad \text{(Eq. 3.37)}$$

The negligence of the recovery effect leads to overestimation of ΔV_{th}. Since $\mathrm{Eq.\ 3.36}$ is a linear equation with respective to ΔV_{th}, this leads to overestimation of delay degradation. $\mathrm{Eq.\ 3.37}$ also indicates that the larger SP is, the larger will the overestimation will be. Figure 3.28 shows the linear relation between ΔD and ΔV_{th}, and the sensitivity could be obtained. In Figure 3.29, the leftmost dot corresponds to the delay degradation due to $\Delta V_{th} = 30mV$(the recovery effect considered); when the recovery effect is ignored, ΔV_{th} will be overestimated according to $\mathrm{Eq.\ 3.37}$. Figure 3.29 clearly shows that ΔD is overestimated. With $SP = 0.9$, the overestimation could be 13.7%. The overestimation will accumulate over the stages of the gates, which may result in the wrong determination of the critical path; even if the critical path is the correct one, the overestimation of ΔD requires more design margin, resulting in lower operating frequency. The new modelling methodology proposed in the previous section helps alleviate this problem, without time-efficiency issue.

The correct determination of SP is of significant importance for ΔV_{th} calculation. Early methods for SP determination still relies on the signal pattern and logic simulation, as is

proposed in [71]. It gives accurate SP at the sacrifice of simulation time. [72] uses a recursive method to calculate the gate out SP based on SPs at the inputs. This method requires a binary decision digram (BDD) [73], which describes the dependency of the gate output on inputs. However, it is impractical to build BDD for complex circuits, although a compromise is introduced in [74].

In a circuit, a PMOSFET suffers from NBTI on condition that:

- Its gate terminal is connected to a logic "low";
- Its source/drain terminal is connected to a logic "high".

The device is biased to strong inversion only both conditions are fulfilled. For the first condition, the probability that the gate terminal is biased to low is defined as P_{on}. As the second condition states, the path from V_{dd} to the source/drain of the PMOSFET must be conducting [61]. Therefore, there must exist at least on path such that from V_{dd} to the source/drain of the PMOSFET of interest (marked as P_i), all PMOSFETs in between are with their gate terminals connected to logic "low".

For independent signals, SF of P_i (marked as SF_i) could be calculated as the product of all SFs of every PMOSFET on the conducting path (marked as $PATH_j$):

$$SF_i = \prod_{PMOSFET \in PATH_j} P_{on,k} \qquad \text{(Eq. 3.38)}$$

k is the mark of the device on the conducting path. In case there are multiple paths from V_{dd} to P_i, at least one conducting path is required and the SF is finally calculated as:

$$SF = 1 - (\prod_i (1 - SF_i)) \qquad \text{(Eq. 3.39)}$$

Figure 3.30 is a 2-input NOR gate. Since the source of P_1 is directly connected to V_{dd}, $SF_1 = P_{on,1} = 0.6$; If P_2 is to be stressed, both gate terminals of P_1 and P_2 should be connected to "low". In case the two signals are independent, according to $\mathrm{Eq.\ 3.38}$, $SF_2 = 0.6 \cdot 0.4 = 0.24$.

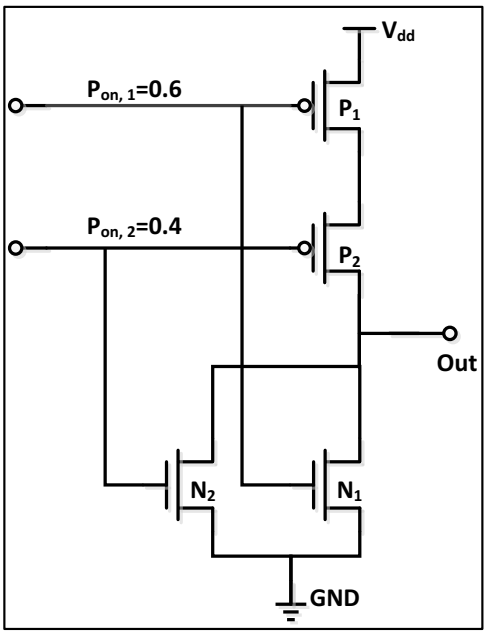


Figure 3.30: NOR gate

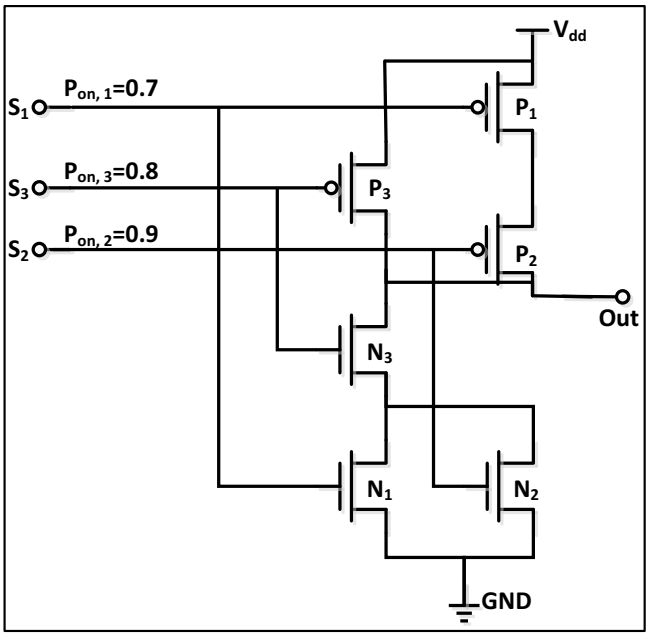


Figure 3.31: Circuit with three inputs

In case the two signals are inter-dependent, a worst-case analysis is adopted. Again P_1 is not affected, but the algorithm for SP_2 is changed: it is determined by the "maximum" probability that both PMOSFETs are on at the same time, which is the "minimum" P_{on} on the conducting path (0.4 in this case). In general, for a single conducting path, SF for the PMOSFET that is farthest from V_{dd} is:

$$SF_i = min(P_{on,k}) \qquad \text{(Eq. 3.40)}$$

Figure 3.31 is a 3-input circuit that realises the function of $\overline{(S_1 + S_2) \cdot S_3}$. There could be two conducting paths to P_2: from P_1 to the source of P_2 or from P_3 to the drain of P_2. For independent signals, SF_2 is obtained according to Eq. 3.39, with $SF_2 = 1 - (1 - 0.7 \cdot 0.9)(1 - 0.8 \cdot 0.9) = 0.8964$. But if these signal are inter-dependent, the worst case is that the common time that all paths are conducting is the least(which means the time duration that P_i is on is as long as possible, but $SF_i \leqslant 1$). For the case in Figure 3.31, $SF_{P_1-to-P_2} = min(0.7, 0.9) = 0.7$; $SF_{P_3-to-P_2} = min(0.8, 0.9) = 0.8$. As $0.7 + 0.8 = 1.5 > 1$, SF_2 is taken as 1, which means P_2 is under stress all the time. In general, the worst case SF for a PMOFSET on multiple

conducting paths is calculated as:

$$SF = min(\sum_i SF_i, 1) \quad \text{(Eq. 3.41)}$$

In Eq. 3.41, SF_i is derived based on Eq. 3.40.

3.6 ϕ_s based NBTI modelling

The previous section introduces a new methodology for AC NBTI modelling which significantly reduces the simulation time. It is a mathematical method. However, so far the physics behind has not yet been involved in the model. In this section, the ϕ_s based method is proposed such that the model is related to input parameters such as V_g, V_d/V_s and the environmental parameter T.

Chapter 2 introduces the way to solve for ϕ_s of an NMOSFET. Similarly, ϕ_s of a PMOSFET could be obtained.

$$\begin{cases} \phi_{cc} = V_g - V_s - V_{FB} - 2v_{tm}\mathcal{L}\{\frac{\gamma}{2\sqrt{v_{tm}}} \cdot exp(\frac{V_g - V_s - V_{FB}}{2v_{tm}})\} & \text{(Eq. 3.42a)} \\ \phi_{dd=} - [\sqrt{\frac{\gamma^2}{4} - (V_g - V_s - VFB)} - \frac{\gamma}{2}]^2 & \text{(Eq. 3.42b)} \\ \phi_{ss} = V_g - V_s - V_{FB} + 2v_{tm}\mathcal{L}\{\frac{\gamma}{2\sqrt{v_{tm}}} \cdot exp(-\frac{V_g - V_s - V_{FB} + 2\phi_f}{2v_{tm}})\} & \text{(Eq. 3.42c)} \end{cases}$$

Parameters in Eq. 3.42a and their physical meanings have been explained in Chapter 2. NBTI occurs when a PMOSFET is at the ”on” state when its gate is connected to a logic ”0” while both the source and the drain are connected to V_{dd}. The device is biased in the strong inversion region. Therefore only Eq. 3.42c is adopted in this section.

According to [13], apart from time, NBTI is also a function of the channel hole density, the oxide field strength, and T. Those parameters influence the forward dissociation rate

constant k_f most:

$$k_f = f(E_{ox}, N_{hole}, T) = C_1 \cdot N_{hole} \cdot exp(\frac{E_{ox}}{E_0}) \cdot exp(-\frac{E_{a1}}{kT}) \quad \text{(Eq. 3.43)}$$

Parameters in the equation above are:

- C_1: constant to be calibrated;
- N_{hole}: inversion hole density;
- E_{ox}: oxide electric field strength;
- E_0: field dependent activation energy;
- E_{a1}: Arrhenius activation energy.

Based on the the ϕ_s based model, the inversion hole density is related to ϕ_s as:

$$N_{hole} = \underbrace{\gamma\sqrt{v_{tm} \cdot exp(-\frac{2\phi_f}{v_{tm}}) \cdot [exp(-\frac{\phi_s}{v_{tm}}) - 1] - \phi_s}}_{\text{total charge density (depletion charge + inversion hole)}} - \underbrace{\gamma\sqrt{-\phi_s}}_{\text{depletion charge density}} \quad \text{(Eq. 3.44)}$$

Besides, the potential balance equation and the voltage-electric field relation determines E_{ox}:

$$E_{ox} = |\frac{V_g - V_b - \phi_{MS} - \phi_s}{t_{ox}}| \quad \text{(Eq. 3.45)}$$

ϕ_{ms} is the work function difference between the poly-gate and the bulk. To better illuminate the derivation of k_f, we summarise the 7-step procedure followed with a flowchart:

- ☐ Step (1): based on the PMOS parameters, the $gate - bulk$ work function difference (ϕ_{MS}), the gate capacitance (C_{ox}), and the Fermi potential (ϕ_F) could be determined, according to equations in Chapter 2;
- ☐ Step (2): the flatband voltage V_{FB} is then determined based on ϕ_{MS}, Q_{ox}, and C_{ox};
- ☐ Step (3): the surface potential (ϕ_s) is thus derived according to Equation 3.38c;

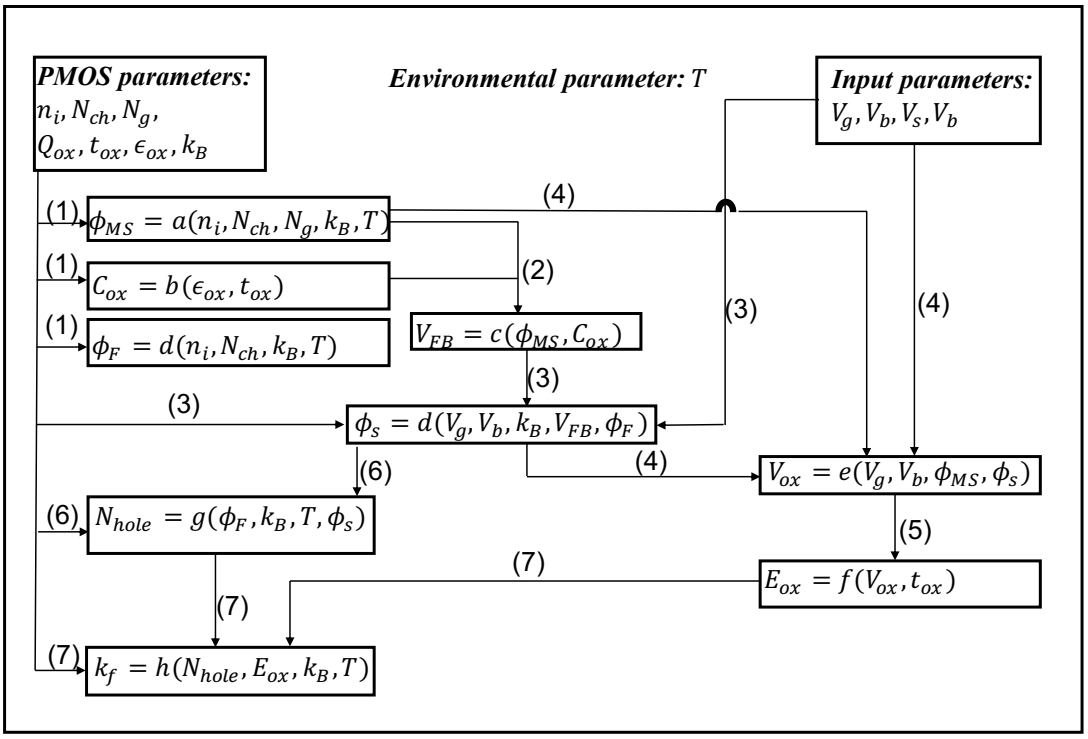


Figure 3.32: Procedures for k_f determination

□ Step (4): the oxide voltage is determined according to potential balance;

□ Step (5): the oxide field strength could be easily obtained from the oxide voltage and the oxide thickness;

□ Step (6): The hole density is obtained according to Equation 3.40;

□ Step (7): k_f is determined as the final step based on parameters previously obtained.

Figure 3.32 is the flowchart showing the 7 steps, in which the parameter to be determined is written as functions of the parameters known(in the form of $Y = X(n_1, n_2, ...)$, where Y is the parameter to be determined; X is the name of the function and n_1, n_2... are the known parameters).

The annealing rate constant k_r and the hydrogen diffusion coefficient are described

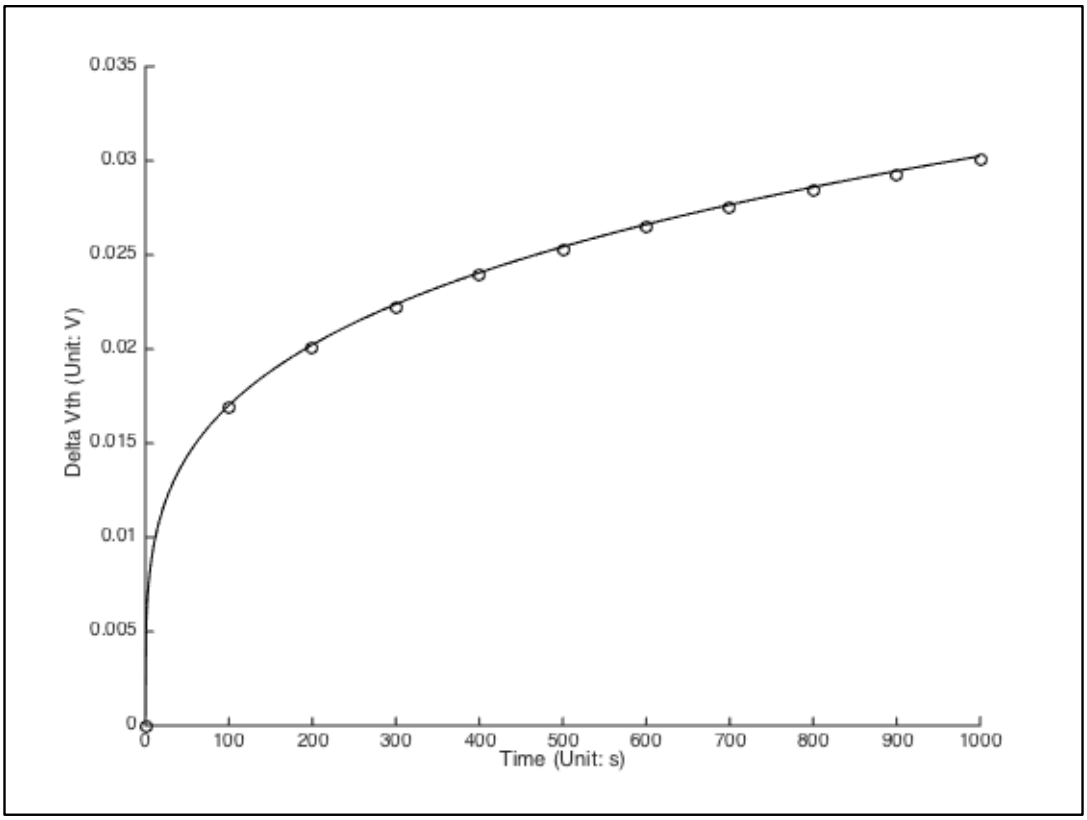


Figure 3.33: Comparison between NBTI data from [75] and that from our model

with the Arrhenius equations:

$$k_r = k_{r0} \cdot exp(-\frac{Ea2}{kT}) \quad \text{(Eq. 3.46)}$$

$$D_H = D_{H0} \cdot exp(\frac{Ea3}{kT}) \quad \text{(Eq. 3.47)}$$

E_{a2} and E_{a3} are activation energies for k_r and D_H respectively. Based on $\mathrm{Eq.\ 3.26}$ in which $n = \frac{1}{4}$, there is a net activation energy for NBTI:

$$E_a = \frac{2E_{a1} - 2E_{a2} + E_{a3}}{4} \quad \text{(Eq. 3.48)}$$

A set of NBTI data is available in [75]: the oxide thickness t_{ox} is $1.3nm$ with the bias $|V_g| = 2.7V$, and the temperature is controlled at $375K$. A comparison is carrier between these data and the simulation results from this ϕ_s based model: The dots in Figure 3.33 are from

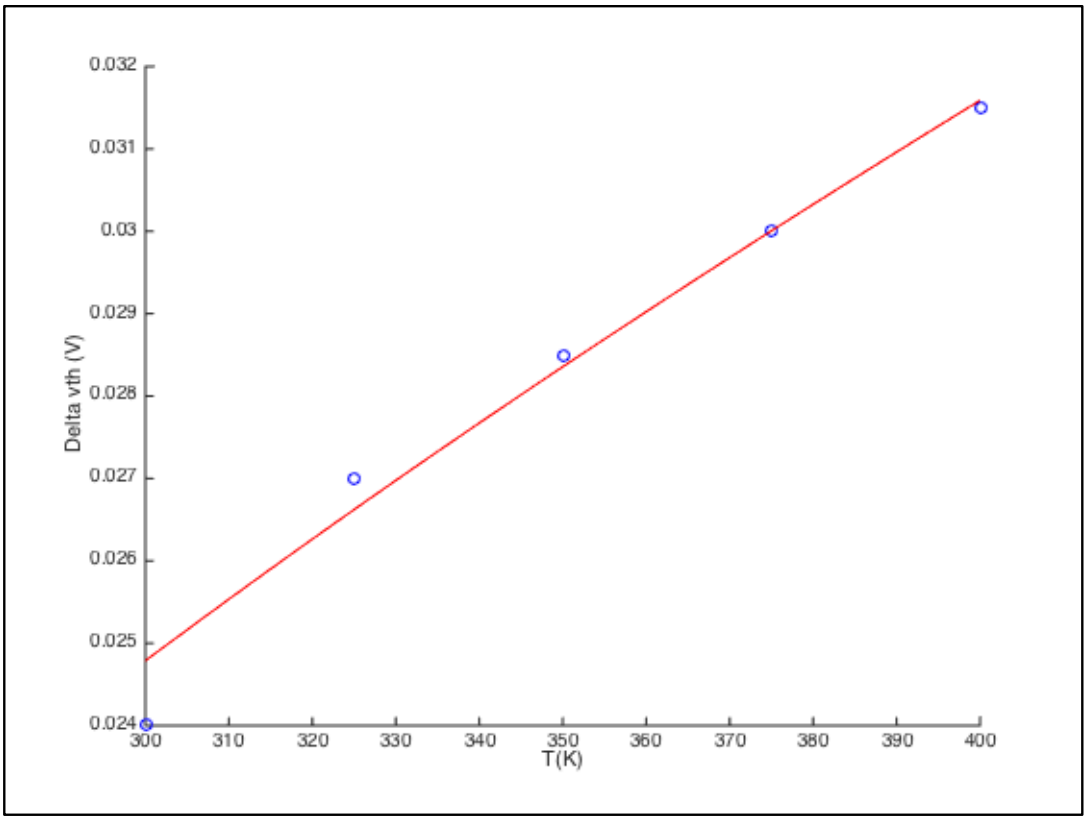


Figure 3.34: Temperature influence on NBTI

[75] while the curve is based on our model. The two sets of data match each other well.

Meanwhile, [75] also investigates into the temperature influence on NBTI. Five data points are collect for a total time of 1000 seconds, with $T(K) \in \{300, 325, 350, 375, 400\}$. With these limited number of data, E_a is approximately $0.09eV$ such that data from our model have a good match with these data, as is shown in Figure 3.34. It is seen from Figure 3.34, at $E_a = 0.09eV$, the simulation results match the measurement data well at higher temperature; at smaller T, a 3% difference is observed, which is acceptable.

This thesis has adopted the RD model for NBTI. In fact, there also exists the so-called “trapping-detrapping” model for NBTI. The aim of this thesis is to provide the transistor level NBTI model for the gate level, which has been created in [29]. In Chapter 3 of [29], the author mentioned that the simple DC model (Reaction-Diffusion based) was adopted, which has

the tradeoff in accuracy (the author also mentioned this in [76] where even the recovery was ignored). Naturally it is a breakthrough point to take the recovery into consideration; meanwhile, to be closer to "reality", an AC model is in demand rather than DC. [29] mentioned the disadvantage of the AC models (also Reaction-Diffusion based). Then we got the motivation for this thesis: creating a compact AC model such that it absorbs both the advantages of the DC model (compact, fast) and the AC model (accuracy), while avoiding their disadvantages. This is the reason that this thesis adopts the Reaction-Diffusion model instead of using the trapping-detrapping model.

In this thesis, we have focused on NBTI of PMOS, while PBTI is not discussed much. Here we give a short discussion of the NBTI and PBTI trend based on the others' research work:

[77] carries out experiments on both NBTI and PBTI for T_iN triple-gate FinFET: the FinFET sidewall (110) has a higher N_{it} than that from the top surface (100). Therefore, the narrower the FinFET, the more degradation the FinFET suffers from- the overall "contribution to NBTI" from the sidewall is larger for a narrower FinFET. PBTI on the other hand, becomes less severe for narrow FinFET (according to the experiment results), the reason is unclear so far, but it may be related to the different mechanism of PBTI from NBTI- PBTI results from the filling of pre-existing traps in the dielectric rather than the creation of interface states. The experiments results show that NBTI is more serious than PBTI in FinFETs, which is also supported by [78].

[79] focuses on the influence of the gate dielectric material on NBTI and PBTI. The experiment results show that for High-k MOSFET, NBTI is almost irrelevant to the choice of the gate dielectric ($SiO_2/NiSi$ and $SiO_2/HfO_2/NiSi$) are the same as the conventional $SiON/poly-Si$ MOSFETs. PBTI, however, is sensitive to the dielectric: for the high-k layer, PBTI increases significantly with increment of Hf content; in consistent with [77], PBTI in TiN gated devices is much smaller.

Chapter 4

Hot Carrier Injection Modelling

Near the drain end of a device, charge carriers (either electrons or holes) may gain enough energy such that they could surpass the potential barrier at the $S_i - S_iO_2$ interface to be injected into the gate oxide. As a result, device parameters such as V_{th} and μ will suffer. This phenomenon is called Hot Carrier Injection. It occurs in both NMOSFET and PMOSFET. The word "hot" describes the fact that charge carriers get accelerated along the channel from the source to the drain, and the high electric field near the drain region triggers impact ionisation[23]. Though NBTI is the dominating reliability issue and HCI contributes less to device/circuit degradation[80], the latter could not be ignored since the parameter shifts caused by HCI is inversely proportional to device channel length L[24]. L is decreasing with technology scaling[81]. Therefore, HCI modelling is still needed for device and circuit simulations.

4.1 HCI mechanism and the HCI models

According to the source of the injected charge carrier, two popular mechanisms have been proposed to explain the mechanisms of HCI:

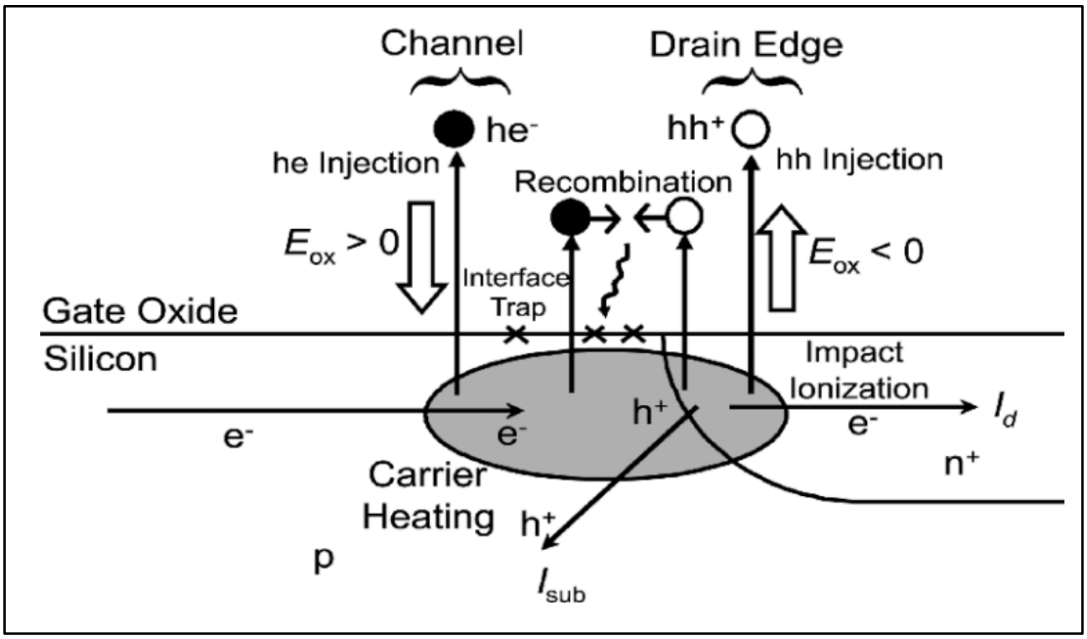


Figure 4.1: DAHC degradation mechanism due to recombination of hot electrons and hole in the gate oxide (for NMOSFET)[82]

- The Drain Avalanche Hot Carrier (DAHC)(NMOSFET taken as an example)[82, 83]: inside the channel, impact ionisation creates both electrons and holes, which are heated by the high electric field near the drain. They further produce hot electrons and hot holes. The oxide electric field (E_{ox}) directs from the gate into the channel ($E_{ox} > 0$), leading to hot electron injection; but inside the drain depletion layer, the direction of E_{ox} reverses ($E_{ox} < 0$), resulting in hot hole injection. The condition for maximum DAHC is:

$$V_{gs} = \frac{V_{ds}}{2} \qquad \text{(Eq. 4.1)}$$

  In Figure 4.1, e^- and h^+ represent channel electrons and hole respectively due to impact ionisation near the drain region; he^- and hh^+ are hot electrons and holes; and I_{sub} is the substrate current dominated by the holes due to impact ionization near the drain region.

- The Channel Hot Carrier(CHC)[84, 85]: instead of impact ionisation, hot carriers are those originally inside the channel and accelerated by the lateral channel electric field. E_{ox}

makes it possible that some charge carriers are "lucky" enough to overcome the $S_i - S_iO_2$ interface potential barrier and "invade" into the oxide. The condition for maximum CHC is:

$$V_{gs} = V_{ds} \tag{Eq. 4.2}$$

Several HCI models focus on I_{on} degradation, denoted as ΔI_{on}. ΔI_{on} is defined as:

$$\Delta I_{on} = \frac{I_{on,fresh} - I_{on,aged}}{I_{on,fresh}} \tag{Eq. 4.3}$$

$I_{on,fresh}$ and $I_{on,aged}$ are the values of the on current before and after the HCI stress. These models relate ΔI_{on} to input parameters (V_d), device geometric parameters (L), T, and t in a simple way. For example, in [30]:

$$\Delta I_{on} \propto L^{\alpha} \cdot exp(-\frac{\beta}{V_d}) \cdot exp(-\frac{E_{a,HCI}}{k_B T}) \cdot t^n \tag{Eq. 4.4}$$

In Eq. 4.4:

- α: gate length scaling factor;
- β: acceleration factor of voltage;
- $E_{a,HCI}$: activation energy of HCI.

While in [29], an equation with a similar form is found:

$$\Delta I_{on} = Aexp(-\frac{E_a}{kT}) \cdot V_{ds}^b \cdot L^{-m} \cdot t^n \tag{Eq. 4.5}$$

In Eq. 4.5, A is a proportional parameter, b, m, and n are the relevant parameter coefficients. The difference between Eq. 4.4 and Eq. 4.5 to some extent exposes the empirical nature of the I_{on}-based modelling method. But the main disadvantage of this model is its complexity for both circuit simulation and gate level modelling. When ΔI_{on} is used to model to influence of HCI on circuit delay, a device has to be replaced with a subcircuit, as is shown in Figure 4.2:

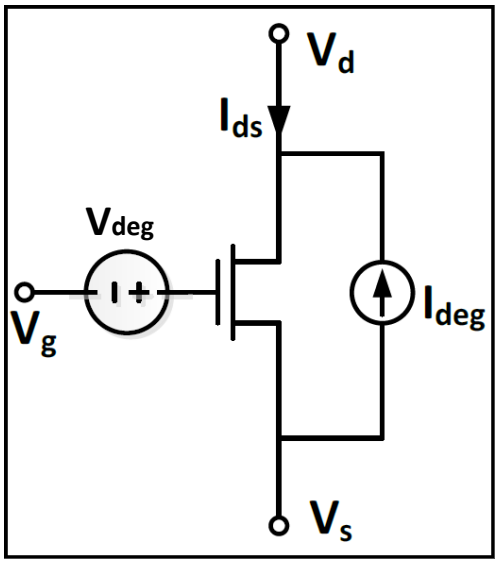


Figure 4.2: Equivalent subcircuit of a transistor due to HCI

The voltage source V_{deg} connected to V_g is to model the impact of HCI on V_{th} since HCI results in interface states like NBTI, and the value of ΔV_{th} is ΔI_{on}-related. The current source I_{deg} is also a function of ΔI_{on}, and it is used to model the mobility degradation caused by interface roughness; it is in the opposite direction from I_{ds}.

(Eq. 2.49) connects circuit delay degradation directly to device parameter shift. This is convenient for NBTI; however, since I_{on} is not a parameter of the transistor model card [86], the sensitivity $\frac{\partial D}{\partial I_{on}}$ could not be obtained from the adjoint network analysis [87], which has been integrated to the SPICE simulator for sensitivity calculation. In stead, the sensitivity is obtained in an indirect way:

$$\frac{\partial D}{\partial I_{on}} = \frac{\partial D}{\partial \Delta v_{th}} \cdot \frac{\partial \Delta v_{th}}{\partial \Delta I_{on}} + \frac{\partial D}{\partial I_{deg}} \cdot \frac{\partial I_{deg}}{\partial \Delta I_{on}} \qquad \text{(Eq. 4.6)}$$

In Eq. 4.6, $\frac{\partial D}{\partial \Delta V_{th}}$ and $\frac{\partial D}{\partial I_{deg}}$ are determined with the replacement of transistors by the equivalent circuit in Figure 4.2, while $\frac{\partial \Delta V_{th}}{\partial \Delta I_{on}}$ and $\frac{\partial I_{deg}}{\partial \Delta I_{on}}$ could be obtained from the equations with respect to ΔI_{on}.

Channel hot carriers could be characterised by the substrate current I_{sub} degradation[27, 88]. The severity of HCI could be represented by the value of I_{sub}. But with the technology scaling, I_{sub} is not dominating anymore, other leakage current components, such as the junction current and the gate induced leakage current introduces in large error, since the model-predicted value and the measurement data deviate from each other [88].

[89] provides an equation that models V_{th} degradation due to HCI, this model has a similar form as Eq. 4.4:

$$\Delta V_{th} = C_1 \cdot (\frac{1}{L})^b \cdot exp(C_2 V_{ds}) \cdot exp(-C_3(V_{ds} - V_{gs})) \cdot t^n \qquad \text{(Eq. 4.7)}$$

The similarity between Eq. 4.4 and Eq. 4.7 implies that there exists a simple relation between ΔV_{th} and ΔI_{on}. Figure 4.3 shows that ΔI_{on} is proportional toΔV_{th}. It is favourable to take V_{th} as the modelling parameter for HCI since it is easier to manipulate, like what is done for NBTI modelling.

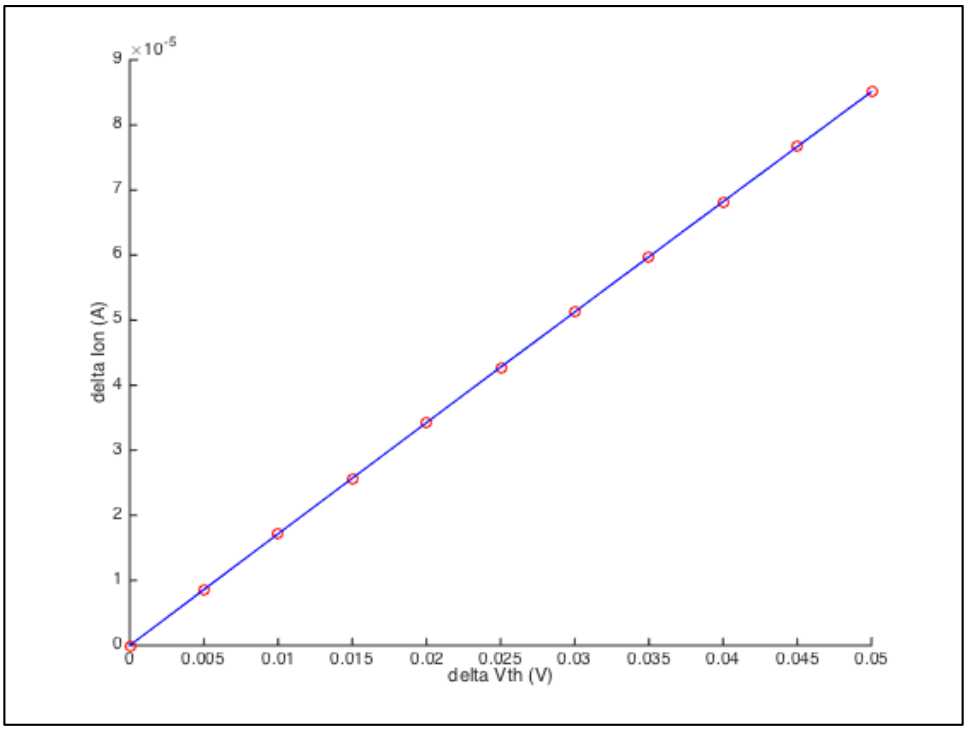


Figure 4.3: ΔI_{on} VS ΔV_{FB}

4.2 RD HCI model

It has been claimed by many authors [88, 66, 90, 91] that the physical progress of HCI is similar to that of NBTI: the Silicon dangling bonds lead to N_{IT} at the $S_i - S_iO_2$ interface. [92] says that $S_i - O$ bonds also join the HCI progress apart from the $S_i - H$ bonds; meanwhile, the relation between the speed of bond breaking and input voltages (V_d, V_g, and V_b) is discussed.

The HCI model is also based on the RD model, the details of which has been discussed in Chapter 3. For NBTI, bond breaking occurs in the whole channel region as long as the device is on. The diffusing species leaves the $S_i - S_iO_2$ interface and diffuse towards the gate. As aforementioned, this is a 1D diffusion process. HCI, however, is more complex: it happens mainly in the velocity saturation region (from the pinch-off point to the drain, for an NMOSFET) with a smaller length of ΔL(length of the velocity saturation region); the diffusing species also diffuse from the velocity saturation region towards the source. This is a 2D

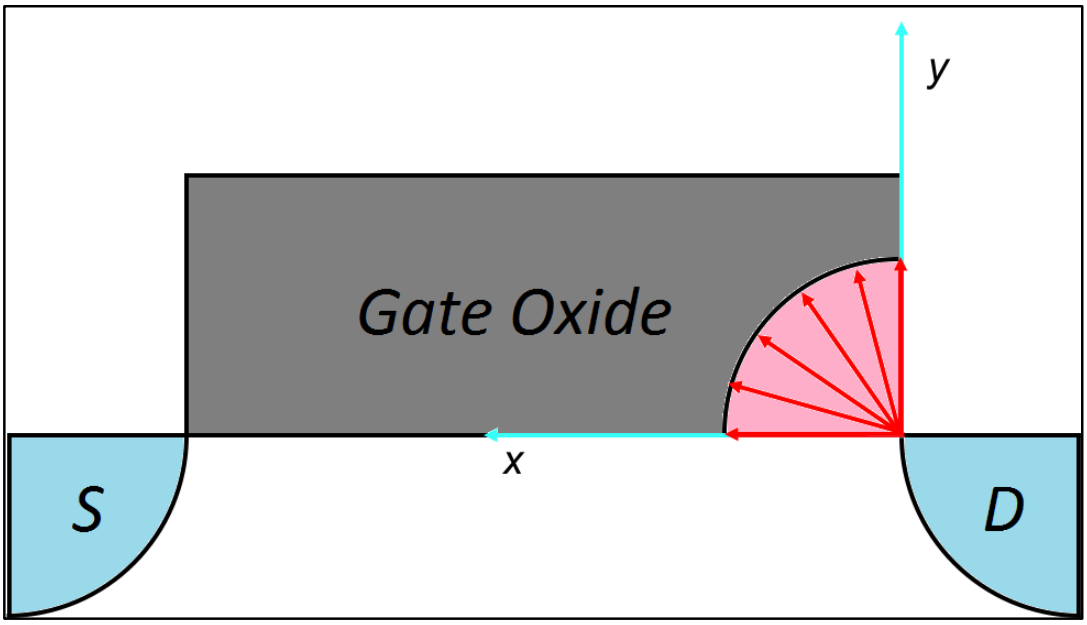


Figure 4.4: 2D diffusion of HCI process

process, which reasonably explains the time coefficient of HCI.

Figure 4.4 is a 2D structure of the HCI process. The diffusing species will diffuse in the directions of x, y, and all the directions in between equally. All diffusing species are confined in the shaded area (pink color).

The diffusion is assumes to start from the drain "point" with the density of diffusing species $N_H(0)$. The distance between the diffusion front (where the density is 0) and the drain point is $\sqrt{D_H t}$. Similar to NBTI, the density of the diffusing species is assumed to decrease linearly. Therefore, at the distance r from the drain, the density of the diffusing species is:

$$N_H = N_H(0) - \frac{N_H(0)}{\sqrt{D_H t}} \cdot r \qquad \text{(Eq. 4.8)}$$

Shown in Figure 4.5, for a infinitesimal distance dr, the total area in the shade is:

$$A = \frac{1}{4} \cdot 2\pi r \cdot dr \qquad \text{(Eq. 4.9)}$$

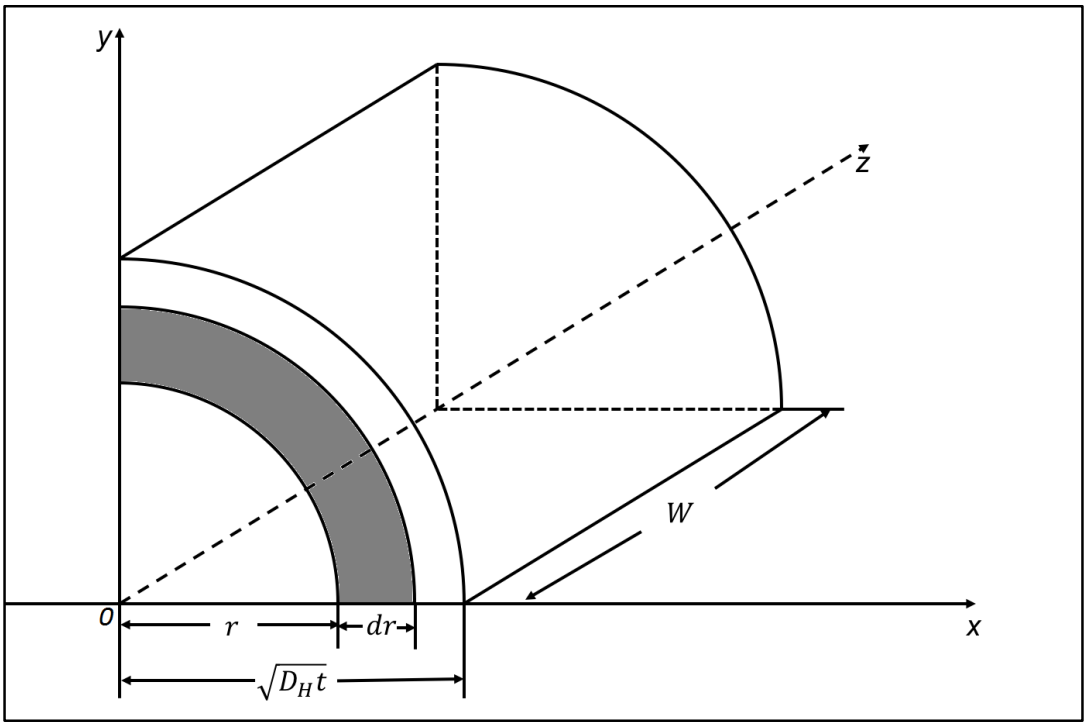


Figure 4.5: $\frac{1}{4}$ column of diffusion space

The the total number of the diffusing species is:

$$N_{2D} = \int_0^{\sqrt{D_H t}} \underbrace{(N_H(0) - \frac{N_H(0)}{\sqrt{D_H t}} \cdot r)}_{N_H} \cdot \underbrace{\frac{1}{4} \cdot 2\pi r \cdot dr}_{A} \qquad \text{(Eq. 4.10)}$$

Taking the width W of the device into consideration, the diffusing species are actually locate in a quarter of a column, as is shown in Figure 4.5. Then the total number of the diffusing

species is:

$$\begin{aligned} N_H(total) &= W \cdot N_{2D} \\ &= \mathbf{W} \cdot \int_0^{\sqrt{D_H t}} (N_H(0) - \frac{N_H(0)}{\sqrt{D_H t}} \cdot r) \cdot \frac{1}{4} \cdot 2\pi r \cdot dr \\ &= \frac{\pi W N_H(0)}{2} \cdot \int_0^{\sqrt{D_H t}} (r - \frac{r^2}{\sqrt{D_H t}}) dr \\ &= \frac{\pi W}{2} N_H(0) \cdot (\frac{r^2}{2} - \frac{r^3}{3\sqrt{D_H t}})|_0^{\sqrt{D_H t}} \\ &= \frac{\pi W \cdot N_H(0) \cdot D_H t}{12} \end{aligned} \tag{Eq. 4.11}$$

$N_H(total)$ is the same as the total number of interface states created if the diffusing species is H atom: $N_H(total) = N_{IT}(total)$. $N_{IT}(total)$ locates in the rectangle with area of $L_{eff} \cdot W$. Then the density of interface states (per area) is:

$$N_{IT} \cdot L_{eff} \cdot W = \frac{\pi W \cdot N_H(0) \cdot D_H t}{12} \tag{Eq. 4.12}$$

And:

$$N_{IT} = \frac{\pi N_H(0) \cdot D_H t}{12 L_{eff}} \tag{Eq. 4.13}$$

[13] highlights the relation between N_{IT} and the initial $S_i - H$ bond density N_0:

$$k_f \cdot N_0 = k_r \cdot N_{IT} \cdot N_H(0) \tag{Eq. 4.14}$$

Physical meanings of the parameters in $\mathrm{Eq.\ 4.14}$ has been introduced in Chapter 3. The combination of $\mathrm{Eq.\ 4.14}$ and $\mathrm{Eq.\ 4.14}$ leads to the final expression of N_{IT}:

$$N_{IT} = \sqrt{\frac{k_f \cdot \pi \cdot N_0}{12 \cdot k_r \cdot L_{eff}}} (D_H t)^{\frac{1}{2}} \tag{Eq. 4.15}$$

$\mathrm{Eq.\ 4.15}$ physically explains the HCI time coefficient $\frac{1}{2}$. Then ΔV_{th} is obtained via $(\mathrm{Eq.\ 3.10})$.Since HCI and NBTI are subject to the same mechanism, theoretically there should be the ”recovery” effect for HCI. [93] proves that the recovery in HCI is so insignificant that it could be reasonably neglected: there exists localised degradation near the drain region, the reaction surface

for annealing is thus reduced significantly; as a consequence, most of the generated Hydrogen species diffuse away rather than diffuse back. This to a great extent reduces the complexity for transistor/gate level HCI modelling since HCI is treated as a "one-direction" process. Therefore, there is no need for the time-consuming iterative method.

To better serve the gate level modelling, the most important parameter in $\mathrm{Eq.}\ 4.15$ is t (t_{stress}), which must be calculated as accurate as possible. Unlike NBTI that happens even at the standby mode, HCI occurs only when the MOSFET is switched from the "off" state to the "on" state: lateral electric field is established between the source and the drain and current flows through the channel. The current either charges C_L via the pull-up path (PMOSFETs) or discharge it via the pull-down path (NMOSFETs). In case like Figure 3.30, P_1 suffers from HCI only when S_2 is a logic "low" and S_1 switches from "high" to "low" and C_L is charged by V_{dd}. In case the former is not fulfilled while the latter is satisfied, there will be only a charging of the internal capacitance, which is ignored since it is substantially small.

The definition of transition density (TD) is introduced in for HCI modelling: according to [94], TD is the mean number of signal transitions per clock period. In this thesis the pull-down path (for NMOSFETs) is taken as an example for t_{stress} derivation, while that for PMOSFETs could be obtained accordingly. The NMOSFET of interest is represented by N_i; all NMOSFETs but N_i must be "on" when it is to be switched to the "on"state, therefore a subset of NMOSFETs is obtained from the pull-down path ($PATH_j$): $PATH_j \backslash \{N_i\}$, the probability that an NMOSFET (with subscript k, any NMOSFET but N_i) is "on" is $N_{on,k}$. The probability that N_i undergoes HCI is SF_i:

- For a single path with independent signals:

$$SF_i = \prod_{NMOSFET_k \in PATH_j \backslash \{N_i\}} N_{on,k} \qquad (\mathrm{Eq.}\ 4.16)$$

- For a single path with dependent signals, worst case is assumed and a formula similar to

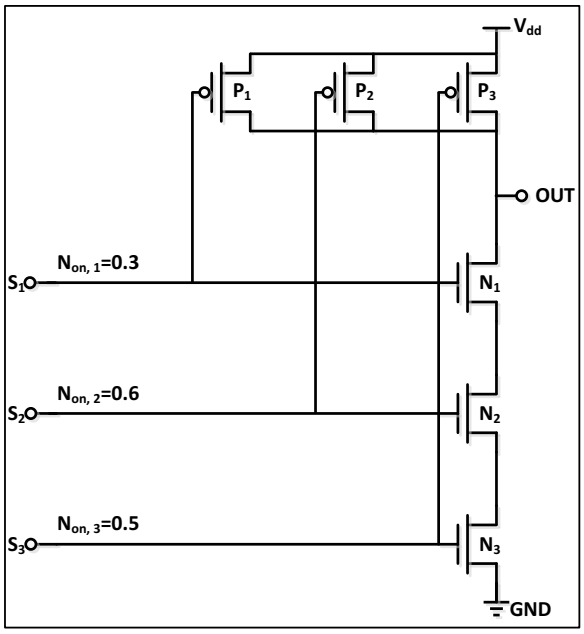


Figure 4.6: Single pull-down conducting path

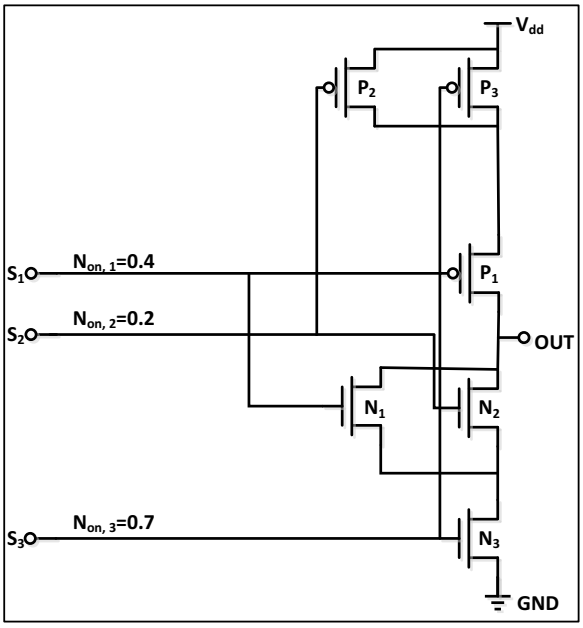


Figure 4.7: Multiple pull-down conducting paths

(Eq. 3.40) is obtained. For $NMOSFET_k \in PATH_j \backslash \{N_i\}$:

$$SF_i = min(N_{on,k}) \tag{Eq. 4.17}$$

■ For multi-paths with independent signals:

$$SF = 1 - (\prod_i (1 - SF_i)) \tag{Eq. 4.18}$$

■ But if the signals are dependent for the multi-paths:

$$SF = min(\sum_i SF_i, 1) \tag{Eq. 4.19}$$

In Figure 4.6, there is single pull-down path. For independent signals, when N_1 switches: OFF $\rightarrow$ ON, the probability that a conducting path exist from C_L to ground is $SF_1 = 0.6 \cdot 0.5 = 0.3$; and that for N_2 is $SF_2 = 0.3 \cdot 0.5 = 0.15$; for N_3 it is $SF_3 = 0.3 \cdot 0.6 = 0.18$. If these signals are dependent, the worst case of SF_1 is limited by the minimum N_{on} and $SF_1 = N_{on,3} = 0.5$; while

those for N_2 and N_3 are $SF_2 = N_{on,1} = 0.3$ and $SF_3 = N_{on,1} = 0.3$.

N_3 in Figure 4.7 locates on two pull-down paths: $N_1 \rightarrow N_3$ and $N_2 \rightarrow N_3$. For independent signals, the probability that at least one conducting path exists when N_3 switches from OFF to ON is: $SF_3 = 1 - (1 - 0.4)(1 - 0.2) = 0.52$; while that for dependent signals is simply the sum of $N_{on,1}$ and $N_{on,2}$, which is 0.6.

A significant property of HCI that is different from NBTI is that HCI is frequency-dependent. This is not surprising since the total time that a transistor suffers from HCI relies on the total number of times that there is current flow in the conducting path, which has a value of:

$$N_{total,i} = \frac{TD}{2} \cdot f \cdot SF_i \qquad \text{(Eq. 4.20)}$$

There is a "2" in $\mathrm{Eq.\ 4.20}$ because only half of the transitions are from OFF to ON state.

Then the total HCI stress time t in $\mathrm{Eq.\ 4.15}$ is then represented as the product of $N_{total,i}$ and the input slope s_{in}:

$$t = N_{total,i} \cdot s_{in} \qquad \text{(Eq. 4.21)}$$

Apart from V_{th}, the creation of N_{IT} increases the $S_i - S_iO_2$ interface roughness and reduces the channel mobility μ; this is why in Figure 4.2 the current source I_{deg} in the opposite direction of I_{ds} is added. As has been discussed before, the insertion of I_{deg} complicates the circuit for SPICE simulation, and it also leads to the complexity of the sensitivity calculation by $\mathrm{Eq.\ 4.6}$. A more favourable method is to "bury" this influence in the device model card by extra modelling of μ. [95] gives the empirical equation that relates μ and Δv_{th}:

$$\mu_{aged} = \frac{\mu_{fresh}}{1 + \alpha \Delta V_{th}} \qquad \text{(Eq. 4.22)}$$

α is a fitting parameter that has a typical value of 1. As is described by $(\mathrm{Eq.\ 2.35})$, $(\mathrm{Eq.\ 2.38})$, and $(\mathrm{Eq.\ 2.39})$, $I_{ds} \propto \mu$. The combinations of those equations with $\mathrm{Eq.\ 4.22}$ represents an I_{ds} degradation due to V_{th} degradation, stemming from HCI. Then I_{deg} in Figure 4.2 is not compulsory any more. Meanwhile, as aforementioned, the voltage source ΔV_{th} could also

be replaced by direct modification of V_{th} inside the model card. Then the SPICE simulation does not require structure change of the circuit.

For the pull-down path of a gate, the delay degradation only considers HCI; while for the pull-up path, NBTI is also taken into account. The original "AgeGate" model uses I_{on} for HCI modelling. Then for the pull-up path, with the addition of HCI influence:

$$D_{aged}(NBTI\&HCI) = D_{fresh} + \sum_{m \in G} \frac{\partial D}{\partial V_{th,m}} \cdot \Delta V_{th,m} + \sum_{m \in G} \frac{\partial D}{\partial I_{on,m}} \cdot \Delta I_{on,m} \qquad \text{(Eq. 4.23)}$$

The sensitivity with respect to I_{on} is described by $\mathrm{Eq.\ 4.6}$, in which $\frac{\partial D}{\partial \Delta v_{th}}$ and $\frac{\partial D}{\partial I_{deg}}$ are another two sensitivities- they could be determined only by SPICE characterisation with the subcircuit in Figure 4.2. Therefore, three sensitivities are required in this gate model: $\frac{\partial D}{\partial v_{th}}$, $\frac{\partial I_{on}}{\partial I_{deg}}$, and $\frac{\partial D}{\partial I_{deg}}$. With V_{th} chosen as the modelling parameter of HCI, the model could be simplified:

- Both NBTI and HCI are modelled as ΔV_{th}, then only a single sensitivity $\frac{\partial D}{\partial V_{th}}$ is required for the gate model;
- The subcircuits, be it Figure 4.12 for NBTI or Figure 4.2 for HCI, are not required for transistor level simulation.

The gate model equation could be simplified:

$$D_{aged}(NBTI\&HCI) = D_{fresh} + \sum_{m \in G} \frac{\partial D}{\partial V_{th,m}} \cdot (\Delta V_{th,NBTI} + \Delta V_{th,HCI}) \qquad \text{(Eq. 4.24)}$$

The new model equation requires V_{th} shifts caused by NBTI ($\Delta V_{th,NBTI}$) and HCI ($V_{th,HCI}$) respectively. $V_{th,HCI}$ is to be modelled based on ϕ_s in the next section.

4.3 ϕ_s-based HCI modelling

In this section, the ϕ_s-based model is introduced in to represent the parameters in $\mathrm{Eq.\ 4.15}$, and HCI in NMOSFETs is taken as an example. k_r and D_H are T-related, and they are described

by Arrhenius equations, like the NBTI case ((Eq. 3.46) and (Eq. 3.47)). k_f, however, is more complex: apart from T, it is also influenced by two electric field strengths:

- E_{ox}: vertical electric field strength in the gate oxide;
- E_m: lateral electric field strength in the velocity saturation region.

HCI occurs when there is current flow through the channel, which means the device is biased in the strong inversion region. Therefore, (Eq. 2.30) is adopted to obtain ϕ_s. Then according to potential balance and the voltage-field relation, E_{ox} could be derived:

$$E_{ox} = \frac{V_g - V_s - \phi_{MS} - \phi_s}{t_{ox}} \tag{Eq. 4.25}$$

It is noticed that Eq. 4.16 and (Eq. 3.38) share the same form. But (Eq. 3.45) is specially for PMOSFETs, in which ϕ_s is derived from (Eq. 3.42c).

To determine E_m, the length of the velocity saturation region (Δl) is needed. Our group has proposed the method for Δl modelling in [96]. In (Eq. 2.4) there is the effective thickness of the pinch-off region. This technology-dependent parameter is repeated here as:

$$l = \sqrt{\frac{\epsilon_{si}}{\epsilon_{ox}} t_{ox} x_j} \tag{Eq. 4.26}$$

In addition, the saturation field strength E_{sat} is determined by the saturation velocity ν_{sat} and the channel mobility μ, which is described by (Eq. 2.39); while the saturation voltage $V_{ds,sat}$ is depicted by (Eq. 2.40). As is described by [96], the maximum electric field strength (E_{max}) in the velocity saturation region is derived as:

$$E_{max} = \sqrt{\frac{(V_{ds} - V_{ds,sat})^2}{l^2} + E_{sat}^2} \tag{Eq. 4.27}$$

And Δl could be derived:

$$\Delta l = l \cdot ln(\frac{\frac{V_{ds} - V_{ds,sat}}{l} + E_{max}}{E_{sat}}) \tag{Eq. 4.28}$$

The "effective" lateral electric field strength E_m is defined as[90]:

$$E_m = \frac{V_{ds} - V_{ds,sat}}{\Delta l} \quad \text{(Eq. 4.29)}$$

The dependency of k_f on E_{ox} and E_m is then described as:

$$k_f \propto exp(\frac{E_{ox}}{E_{0,HCI}}) \cdot exp(-\frac{\phi_{it}}{q\lambda E_m}) \cdot exp(-\frac{E_{k_f}}{kT}) \quad \text{(Eq. 4.30)}$$

Eq. 4.30 is divided into three parts: the first part is the k_f's dependency on E_{ox}, and $E_{0,HCI}$ is a technology-dependent parameter, which is similar to the field-dependent activation energy E_0 in NBTI; the second part describes the probability that an electron will gain enough energy such that it is injected into the gate oxide. ϕ_{it} is the impact ionisation energy with a value of $3.7eV$, and it is also the minimum energy required for an electron to be injected; λ represents the channel mean free path of an electron, which is $7.8nm$. The last part is the Arrhenius relation, describing the temperature dependency.

[88] and [90] describes HCI for the 65nm technology with published data, including model equations, parameter values. The simulation results from the model built in this chapter is compared with these data: In Figure 5.1, the three curves correspond to data collected from [88]: from high to low, the respective effective channel lengths are $65nm$, $70nm$, and $75nm$. The dots come from our model. It is notice that the two groups of data are close to yet slightly different from each other. This deviation comes from the different models for the saturation voltage. $V_{ds,sat}$ in [90] is related to the V_{th}, denoted as $V_{ds,sat,v_{th}}$:

$$V_{ds,sat,V_{th}} = \frac{(V_{gs} - V_{th})E_{sat}L_{eff}}{(V_{gs} - V_{th}) + E_{sat}L_{eff}} \quad \text{(Eq. 4.31)}$$

While for the phi_s based model, $V_{ds,sat}$ is described by (Eq. 2.40) and rewritten here (denoted as V_{ds,sat,ϕ_s}):

$$V_{ds,sat,\phi_s} = \frac{V_{gt,s}E_{sat}L}{V_{gt,s} + A_{b,s}E_{sat}L + 2A_{b,s}v_{tm}} \quad \text{(Eq. 4.32)}$$

In Eq. 4.31, $V_{gs} - V_{th}$ is the charge density normalised to C_{ox}, which is equivalent to $V_{gt,s}$ in Eq. 4.32. But the denominator of Eq. 4.32 includes more contents. This results in $V_{ds,sat,v_{th}} >$

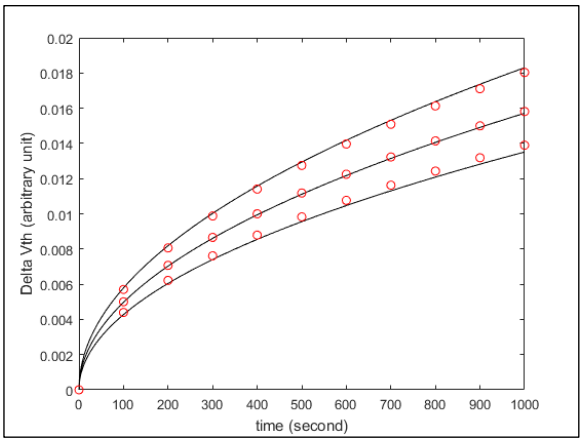


Figure 4.8: ΔV_{th} VS time for multiple channel lengths

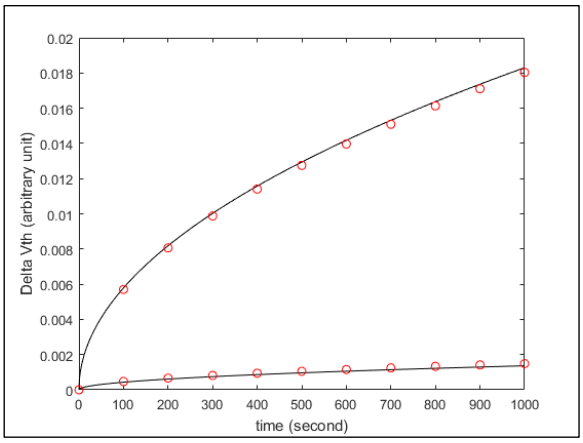


Figure 4.9: ΔV_{th} VS time for multiple V_{ds} with L_{eff}=65nm

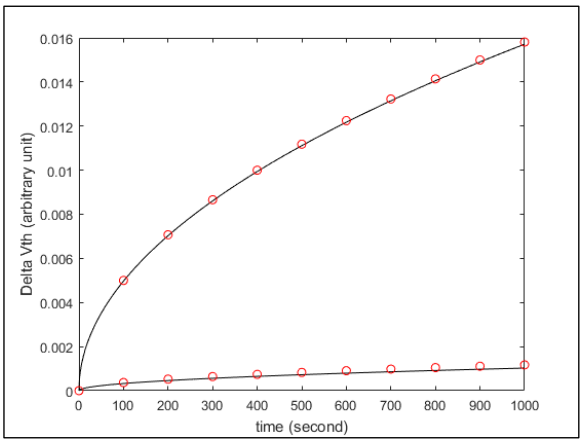


Figure 4.10: ΔV_{th} VS time for multiple V_{ds} with L_{eff}=70nm

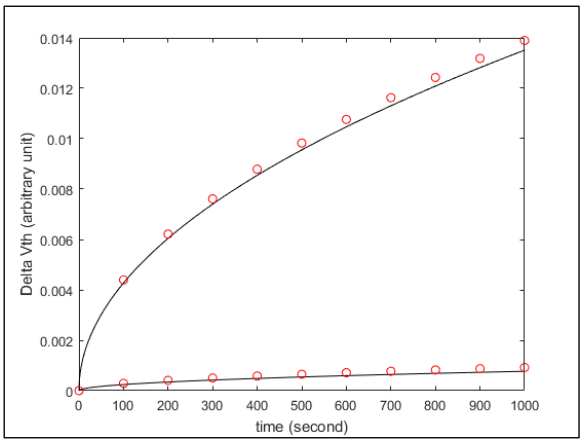


Figure 4.11: ΔV_{th} VS time for multiple V_{ds} with L_{eff}=75nm

V_{ds,sat,ϕ_s}, though the difference is small(1% for $L_{eff} = 75nm$ and 1.26% for $L_{eff} = 65nm$). However, this difference is amplified by $\mathrm{Eq.}\ 4.29$ and the exponential term in $\mathrm{Eq.}\ 4.30$. Even so, data in Figure 5.1 do no differ much (3.88% for $L_{eff} = 65nm$, 2.22% for $L_{eff} = 70nm$, and 3.75% for $L_{eff} = 75nm$).

Figure 5.2, Figure 5.3 and Figure 5.4 are obtained for two V_{ds} values: $1.3V$ and $1.6V$ with V_g fixed at $1.6V$. There three figures are for the three respective channel lengths for the sake of clearness. The data from [88] and those from our model have an acceptable match.

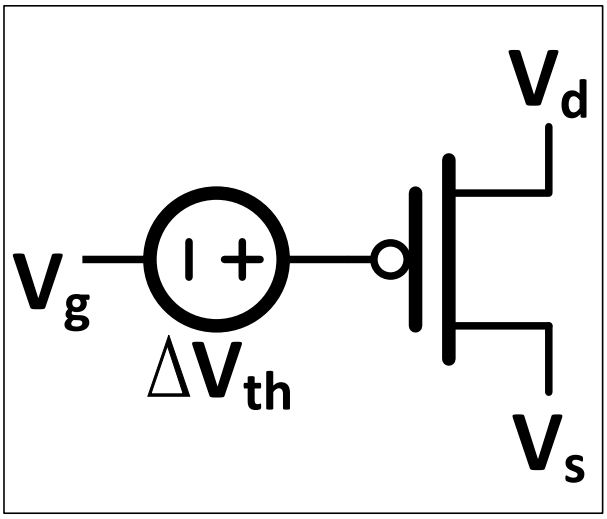


Figure 4.12: ΔV_{th}subcircuit

4.4 V_{th} modelling in the ϕ_s-based transistor model

The ϕ_s-based model does not include the parameter V_{th}; instead, most physical quantities are related to ϕ_s. However, the mainstream researches tend to integrated the influence of NBTI and HCI into ΔV_{th}. To investigate the influence of ΔV_{th} on the performance of a device or the propagation delay of a circuit, usually a subcircuit is constructed as is discussed in [97, 88] and shown in Figure 4.12: Figure 4.12 shows a PMOS rather than NMOS, though HCI is more pronounced in the latter. The reason is as follows: the aim of this thesis is to simplify the "AgingGate" model, in which both NBTI and HCI are taken into consideration for PMOS-PMOS degradation is the sum of degradation from NBTI and that from HCI, as is depicted by $\mathrm{Eq.}\ 4.23$. It is noticed that HCI in $\mathrm{Eq.}\ 4.23$ is based on I_{on} degradation. The model proposed in this chapter uses V_{th} degradation to describe HCI such that the AgingGate model could be simplified by $\mathrm{Eq.}\ 4.24$. The choice of a PMOS in Figure 4.12 is consistent with Chapter 3 (NBTI for PMOS).

The voltage source connected to the device gate terminal is equivalent to ΔV_{th}. As is shown, to maintain the performance of a PMOSFET, a more negative V_g is required to offset the influence of ΔV_{th}. The insertion of the voltage source turns to make the circuit

more complicate, especially for complex circuits. Another method is to change the structure description file of the device, where V_{th} is directly modified. V_{th} is not available in the ϕ_s-based model. In case a circuit is built up with devices described by this model, a "substitution" parameter is needed for the direct modification. It is found that the flatband voltage V_{FB} is good candidate. There are two reasons:

- The definition for V_{FB} shown in ($\mathrm{Eq.}\ 2.13$) includes two components. The second one $\frac{Q_{ox}}{C_{ox}}$ is the influence of the equivalent charges located at the $S_i - S_iO_2$ interface [41]. The aging effects tend to create interface states N_{IT}, which could be treated as a "net" increase of Q_{ox}. In addition, $\Delta V_{th} = \frac{qN_{IT}}{C_{ox}}$ has the same form as $\frac{Q_{ox}}{C_{ox}}$.

- The classic definition of V_{th} is:

$$V_{th} = V_{FB} + 2\phi_F + \gamma\sqrt{2\phi_F} \qquad (\mathrm{Eq.}\ 4.33)$$

  The linear relation between V_{th} and V_{FB} described by $\mathrm{Eq.}\ 4.33$ implies that a change in V_{FB} causes the same amount of change in V_{th}:

$$\Delta V_{FB} = \frac{qN_{IT}}{C_{ox}} = \Delta V_{th} \qquad (\mathrm{Eq.}\ 4.34)$$

  This is consistent with the first point.

The definition for V_{th} in $\mathrm{Eq.}\ 4.33$ is quite rough: it just set V_{th} as the value of V_g when it drives ϕ_s to the value of twice of ϕ_F. In reality several methods are proposed to extract V_{th}:

- The constant-current (CC) method [98, 99, 100, 101, 102] is proposed for its simplicity. It determines V_{th} as the V_g value which drives the device to a predefined constant drain-source current $I_{ds,0}$. However, the arbitrary choice of $I_{ds,0}$ makes V_{th} quite vulnerable.

- The match-point (MP) method is introduced in [103]: when the exponential subthreshold I_{ds} semi-log extrapolation deviates 5% from the measured I_{ds}, the corresponding V_g is set as V_{th}. Like the CC method, when the deviation of extrapolation is set to other

values than 5%, different V_{th} values are created. Meanwhile, this method focus on weak inversion but strong inversion is underemphasised.

- Introduced in [104, 105, 106], the linear extrapolation (LE) method determines V_{th} with the V_g axis intercept (which means $I_{ds} = 0$) of the linear extrapolation of the $I_{ds} - V_g$ curve at the maximum $\frac{\partial I_{ds}}{\partial V_g}$ point. This value is then added with $\frac{V_d}{2}$ to give V_{th}. In spite of its popularity, this method is affected by mobility degradation and the drain and source resistances R_{sd}, as is proved by [107].

- The transconductance change method is proposed in [108], where V_{th} is the value of V_g when $\frac{\partial^2 I_{ds}}{\partial V_g^2}$ is at its maximum. This method may be contaminated by measurement errors and noises.

In [109, 110], our group introduces the method called "critical current at linear threshold ($I_{critical}$@$V_{th,linear}$)": the maximum-g_m definition in the linear region is firstly adopted to derive the linear region threshold voltage $V_{th,linear}$; the critical current is then set as $I_{critical} = I_{ds}(V_g = V_{th,linear})$. Only the linear $I_{ds} - V_g$ curve is enough for $V_{th,linear}$ and $I_{critical}$ extraction. V_{th} at any other bias is defined/measured as V_g that leads to $I_{ds} = I_{critical}$. This method avoids the arbitrary choice of $I_{critical}$ in the CC method while maintains the simplicity of V_{th} definition and measurement for different biases.

In this thesis, V_{th} for the ϕ_s-based model is determined according to the critical current at linear threshold method. The relation between the extracted V_{th} and V_{FB} is investigated and shown in Figure 4.13. It is noticed that V_{th} and V_{FB} are linearly related with the slope of 1. Therefore, $\mathrm{Eq.}\ 4.34$ still holds for the ϕ_s-based model.

Then the subcircuit in Figure 4.12 should be equivalent to a direct change of V_{FB} of the device, with $\Delta V_{FB} = \Delta V_{th}$. Figure 4.14 show multiple $I_d - V_g$ relations of a devices: V_g is fixed at 1.8V; V_d is swept from 0 to 1.8V. The curves are obtained with the voltage source insertion, the value of which is $|\Delta V_{th}|$; while the dot plots are generated with direct V_{FB} modification. It is found that with $|\Delta V_{th}| = \Delta V_{FB}$, the dots all fall on the curves. This demonstrates the claim at the beginning of this paragraph. Meanwhile, it provides an alternative for ϕ_s-based

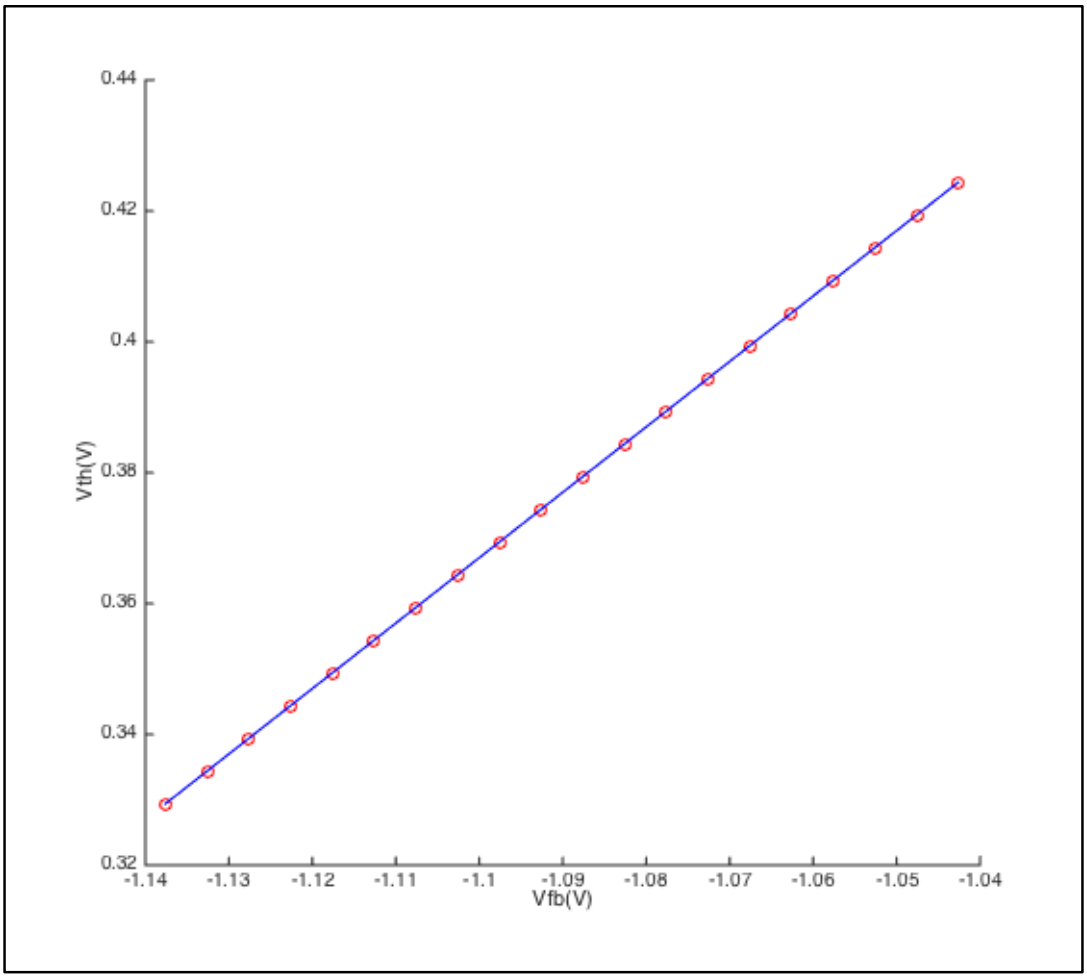


Figure 4.13: Relation between V_{th} and V_{FB}

reliability modelling: apart from the subcircuit, V_{FB} could be chosen as the model parameter.

The choice of V_{FB} is then checked at the transistor level with the simplest circuit, an inverter. The relation between circuit delay D and ΔV_{FB} is tracked. As a comparison, method in Figure 4.12 is again adopted to describe how D is related to the value of the inserted voltage source ΔV_{th}. Data are shown in Table 4.1.

Columns 2 and 3 of Table 4.1 are highly consistent with each other, meaning V_{FB} could be an option for transistor level reliability modelling. When plotted, the relation between D and ΔV_{FB} is shown in Figure 4.15:

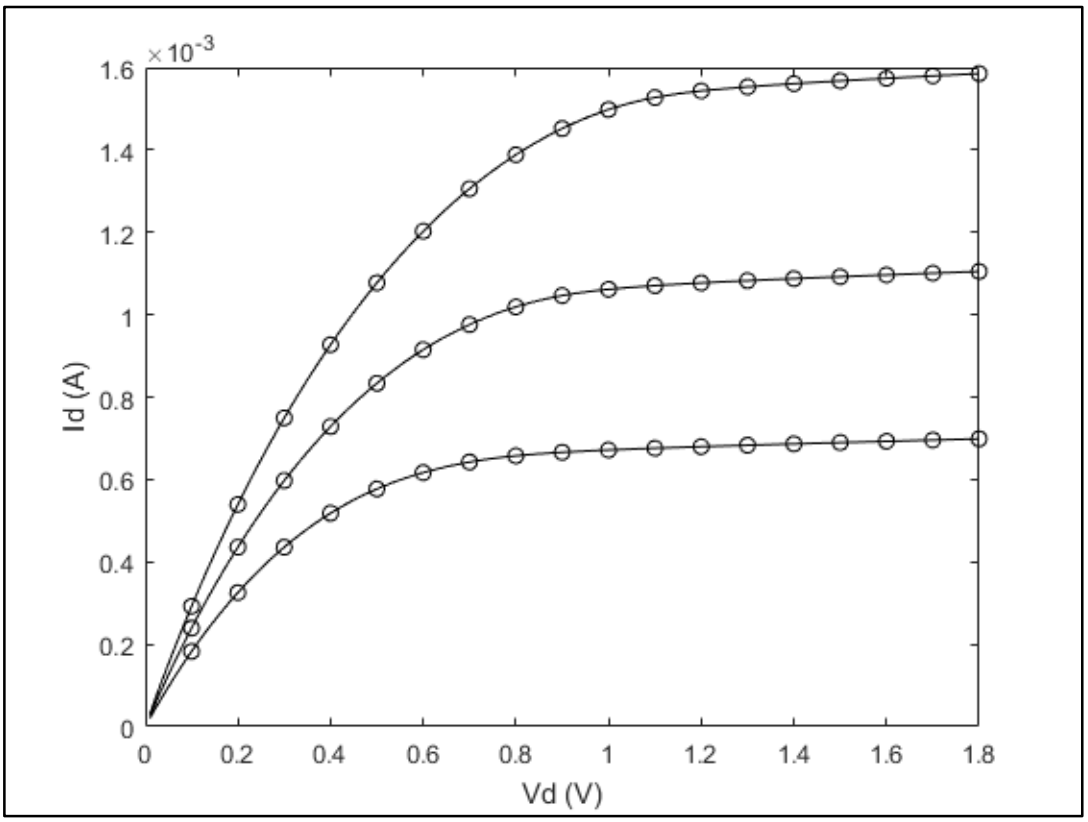


Figure 4.14: Device performance equivalence of Δv_{th} voltage source insertion and direct V_{FB} modification

Table 4.1: Comparison of the influences of voltage source insertion and direct V_{FB} modification on ΔD

ΔV_{th} (mV)	D(s) due to ΔV_{th}	D(s) due to ΔV_{FB}	ΔV_{FB} (mV)
0	2.40E-08	2.40E-08	0
5	2.41E-08	2.41E-08	5
10	2.42E-08	2.42E-08	10
15	2.43E-08	2.43E-08	15
20	2.44E-08	2.44E-08	20
25	2.45E-08	2.45E-08	25
30	2.46E-08	2.46E-08	30
35	2.48E-08	2.47E-08	35
40	2.49E-08	2.49E-08	40
45	2.50E-08	2.50E-08	45
50	2.51E-08	2.51E-08	50

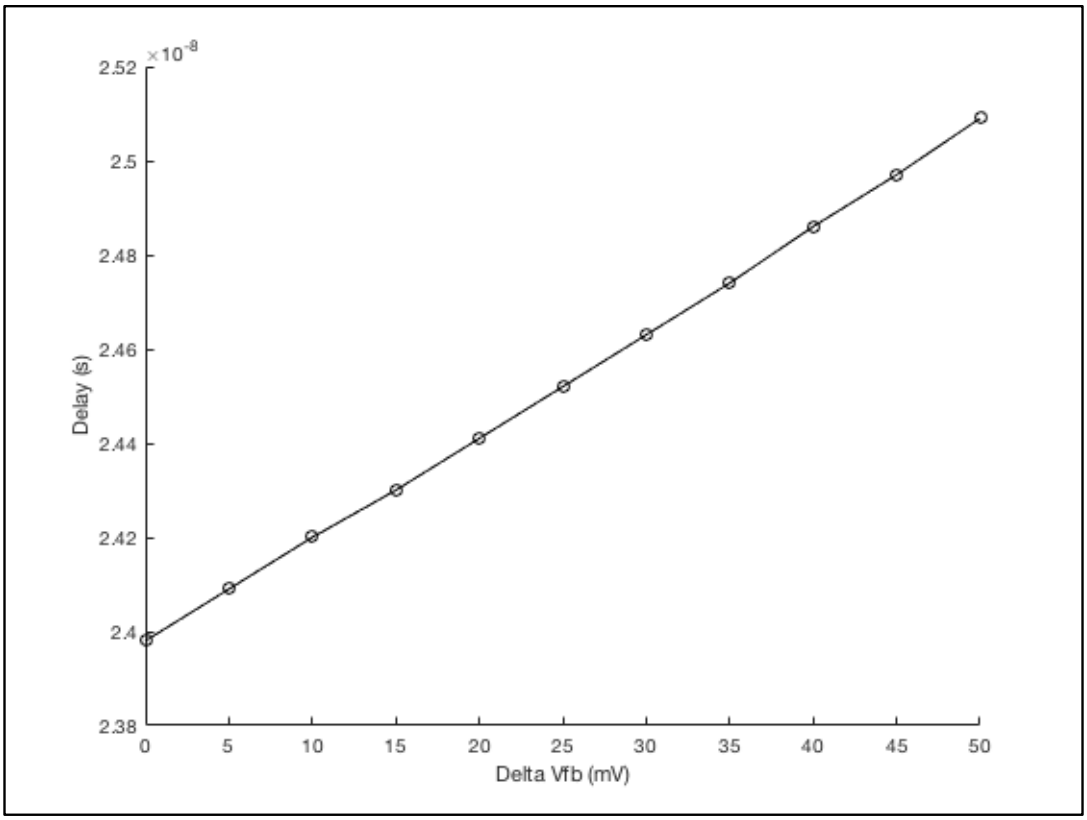


Figure 4.15: Relation between circuit delay and ΔV_{FB}

Figure 4.15 show that the delay degradation is almost linearly related to ΔV_{FB}. This might be an inspiration for the future work- for circuits consisting of MOSFETs described by the ϕ_s-based model, V_{FB} rather than V_{th} could be adopted as the reliability modelling parameter.

4.5 A comparison of HCI in NMOS and PMOS

We have found a comparison between NMOS HCI and PMOS HCI in [29]. The comparison focuses on the stressing voltage and temperature for the $90nm$ technology:

Figure 4.16 shows the dependence of HCI on the stressing voltage follows a power law for both NMOS and PMOS; besides, the severity of I_{on} degradation is similar for both

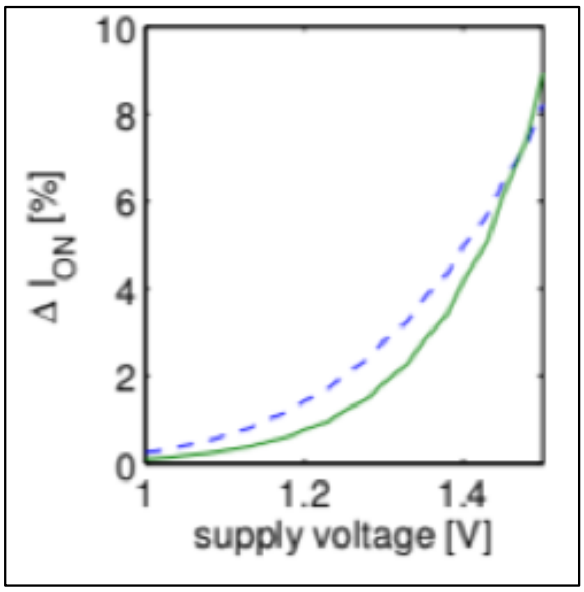


Figure 4.16: Dependence of HCI on the stressing voltage for NMOS/PMOS[29]

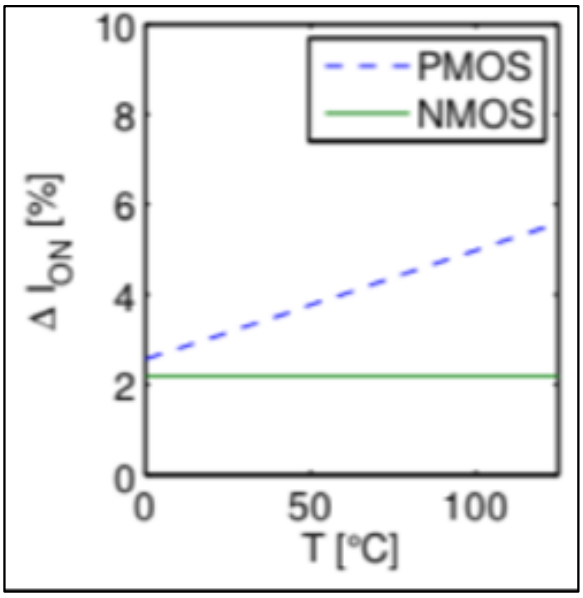


Figure 4.17: Dependence of HCI on temperature for NMOS/PMOS[29]

devices. In Figure 4.17, the stressing voltage is fixed at $1.32V$. It shows that NMOS HCI is quite immune to the temperature variation; PMOS, on the other hand, increases linearly with temperature increment.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

The reliability modelling at the transistor level and gate level are interrelated: at the transistor level, the device parameter shift (Δp) is modelled as a time (t)-related function, where t could only be determined at the gate level since a specific device always switches between the "ON" and "OFF" states; at the gate level, the timing (delay degradation) is a function of Δp, which is determined at the transistor level, where the physical geometry and the input voltages to the specific device are involved. The aim of this thesis is to provide degradation equations at the transistor level such that they could be adopted at the gate level.

Two reliability issues are covered in this thesis- NBTI and HCI. Both models are built based on the R-D model: NBTI corresponds to 1D diffusion, while that for HCI is 2D. Chapter 3 of this thesis introduces in a new algorithm for NBTI modelling, which is suitable for gate level modelling. The inspiration of this algorithm comes from the comparison of data from both the DC and AC simulations. This algorithm is compact, taking the unique recovery effect of NBTI into consideration. It possesses the following advantages compared with the existing DC and AC case NBTI models:

- It has a compact form like the DC case. Parameter shift could be obtained once the stressing parameters (voltages, time, device geometry) are known. It avoids the iterative computations for ΔV_{th} derivation. The time efficiency is improved by orders;

- It achieves identical accuracy as the iterative method for long term, where the recovery effect is taken into account. The DC model without considering the recovery effect leads to overestimation of ΔV_{th}; while the one with a single recovery phase results in underestimation.

The modelling parameters are determined according to the ϕ_s-based model. The simulation results match well with the measurement data, with the temperature influence considered.

Chapter 4 of the thesis models HCI as a threshold voltage shift rather than the on current shift. By doing so the following simplicities are achieved:

- V_{th} is a device parameter in the model card. ΔV_{th} could be directly inserted into the circuit to investigate its impact on circuit delay degradation, and the complex subcircuit is not compulsory; should it be ΔI_{on}, not only is a subcircuit is required, but also more relations: I_{on} and V_{th}, and I_{on} and degradation current (for mobility modelling) are required. Since every step may involve error, this complexity might leads to much error.

- The AgingGate model considers both NBTI (ΔV_{th})and HCI (ΔI_{on}), where 6 sensitivities are required. With this ΔV_{th}-based HCI model, the AgingGate model is unified as a ΔV_{th} based gate model. Then only 1 sensitivity ($\frac{\partial Delay}{\partial V_{th}}$) is needed. The overall ΔV_{th} is the sum of those from NBTI and HCI. The simplification of gate model could be achieved.

Like NBTI, the modelling parameters are also determined based on the ϕ_s-based device model. The simulation results are compared with the published data for the $65nm$ technology, the two of them match well with an $error < 5\%$.

5.2 Future Work

Apart from NBTI and HCI, there exists other reliability issues, among which TDDB received more and more attention. Three successful models, namely, the thermochemical model [111, 112], the anode hole injection (AHI) model [23, 113], and the voltage-driven model [114, 115].

So far almost all circuit simulations of TDDB require subcircuits, where current sources or resistances are added into the circuit. This may result in complicated gate models. With the experience in NBTI and HCI modelling, a bridge is needed between TDDB and ΔV_{th} such that the gate model could be simplified. Or in the worst case, TDDB could be related to a parameter residing in the device model card. In this case there is no need for subcircuits, even though one more sensitivity is required. Apart from TDDB modeling on single devices, statistical influence of TDDB on large Application Specific Integrated Circuit (ASIC) block may be another interesting research direction.

This thesis provides degradation model equations at the transistor level for gate level use, but the gate model is beyond the scope of this thesis. Since now these equations are available, they could be adopted in a random gate model, as long as they are NBTI/HCI related.

The advantage of the ϕ_s-based model has been aforementioned. Currently this device model has been built into Cadence for circuit simulation. Though V_{th} is not in this model card, Chapter 4 demonstrates that V_{FB} could take the place for reliability modelling. If the ϕ_s-based model could become a standard in the future, reliability models should be related to parameters in its model card for simplicity.

5.3 Reliability Issues in High-k Devices

This thesis focuses on conventional MOSFET. With the conventional SiO_2/SiO_xN_y gate dielectric scaling, the gate leakage current density increases dramatically, which limits the scaling. High-k dielectrics are thus adopted. Owing to the increased **physical thickness**, high-k

gate dielectric is able to reduce the leakage current significantly [116, 117, 118]. Despite the advantages of reducing the gate leakage current, high-k dielectrics (such as HfO_2) suffer from some fundamental issues: fixed charge, trapped charge, for example. In terms of NBTI, there have been reports showing that for pMOS with ultrathin $HfO2$ dielectric and tungsten (W) gate, it shows comparable V_{th} shift to those in SiO_2-poly Si stacks [119, 14]. The mechanism of PBTI in high-k MOSFET is the same as that in SiO_2 based MOSFET, which is a consequence of charging trapping. It is believed that the pre-existing bulk traps are filled will charges. Due to the structure of the high-k dielectric, it suffers from even more severe PBTI. This has been demonstrated in [120, 121, 122].

5.4 Reliability Issues in FinFET

The FinFET has a structure difference compared with the planar MOSFET. Figure 5.1 shows the 3D structure of a double-gate FinFET. The physical parameters are as follows:

- L: channel length;
- W_{Si}: channel width;
- H_{Si}: height;
- t_{ox}: gate oxide thickness.

Figure 5.2 is a cross section perpendicular to the x direction, which gives a better view of the FinFET. Figure 5.3 is the cross section perpendicular to the z direction: it provides a good view for NBTI analysis. In Figure 3.2 of Chapter 3 we depict the NBTI for the planar MOSFET. Figure 5.3 is like two of Figure 3.2 lying together symmetrically. The $N_{it}(t)$ and ΔV_{th} equations derived in Chapter 3 are also suitable for describing a FinFET. This kind of modelling could also be found in [91].

Figure 5.4 is the cross section of FinFET HCI. It is like two of Figure 4.4 lying together symmetrically. So the equations derived in Chapter 4 also fit in with FinFET HCI modelling.

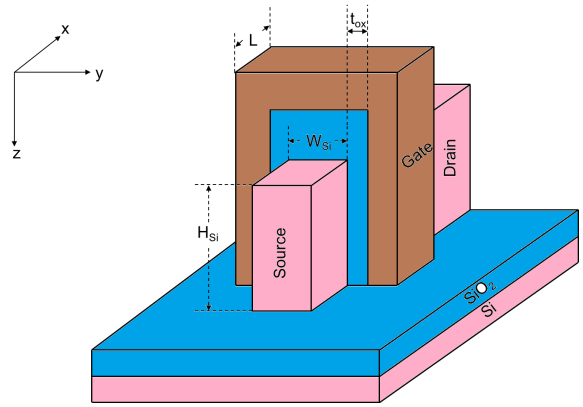


Figure 5.1: 3D structure of FinFET

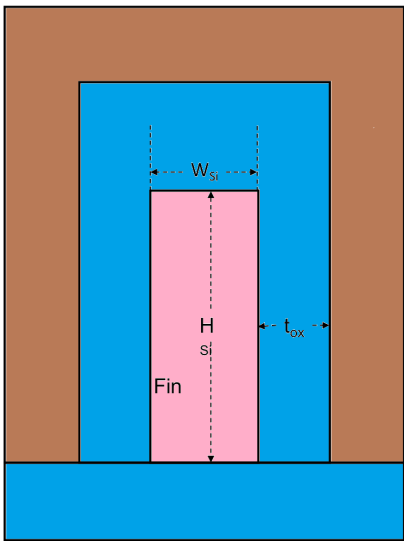


Figure 5.2: FinFET cross section

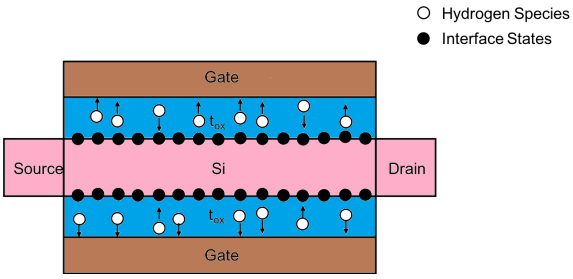


Figure 5.3: NBTI cross section of FinFET

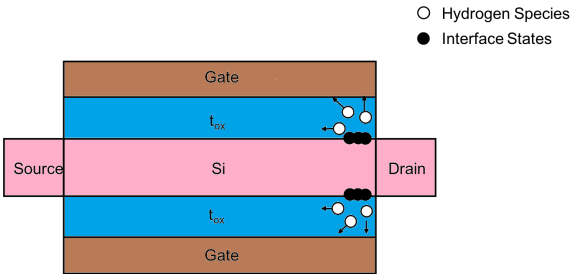


Figure 5.4: HCI cross section of FinFET

[77] mentions PBTI in FinFET, where the dielectric is high-k HfO_2. The author also attributes PBTI to the filling of the pre-existing trap sites, so this should be the same as stated in Question 13. In [123], a comparison of PBTIs in planar MOSFET and FinFET is carried out. Both FETs have HfO_2 gate dielectric. The result shows with the same electric field in the gate dielectric, PBTIs for planar MOSFET and FinFET are comparable (which means PBTI is irrelevant to the device structure).

Compared with the planar MOSFET, FinFET presents better short channel behaviour; meanwhile, it has potential area benefits [124, 125, 126]. In terms of TDDB, the relation between time-to-breakdown (t_{BD}) and the gate stress voltage (V_G) is described with a simple power law model [127]:

$$t_{BD} = B \cdot V_G^{-n} \qquad \text{(Eq. 5.1)}$$

B is a constant and n is the acceleration factor. Statistically, t_{BD} follows the Weibull distribution; with the maximum-likelihood estimation, the Weibull slope β could be fitted according to the TDDB data. At the very beginning, a voltage breakdown model (V_G model) [128] is adopted to explain TDDB mechanism; then later in [129, 130] it attributes TDDB to the hydrogen release; with the high-k gate stack replacing the conventional SiO_2, studies start to move to the influence of the oxygen vacancy traps [131, 132]: the high-k layer contains oxygen vacancy defects that induce the Stress Induced Leakage Current (SILC) during oxide degradation. [133] interprets the power law to be correlated to deep trap generation for high-k n-type FinFET: the deep oxide trap results in permanent damage; it also contributes to the percolation path formation.

[134] carries out a comparison of TDDB between FinFET and the planar MOSFET. The results of the experiments show that both types of devices possess almost the same Weibull slope and similar acceleration factors. This indicates that no new mechanisms are introduced for the FinFET structure. There has been a major concern that the concentration of the electric field around the fin corners may result in preferential dielectric breakdown. The results in [134] overturn this suspect. Beside, [134] shows that soft breakdown occurs when there is a three-trap path through the dielectric.

Author's Publications

Conferences

- **X. Liu**, A. Bernardini, U. Schlichtmann, X. Zhou, "A Compact Model of Negative Bias Temperature Instability Suitable for Gate-Level Circuit Simulation," accepted by *International Symposium on Quality Electronic Design*, 2019.

Co-Author

- X. Zhou, S. B. Chiah, A. Ajaykumar, B. Syamal, H. T. Zhou, and **X. Liu**, "Unified HEMT/CMOS Compact Models for Future Heterogeneous III-V/Si Co-integrated Technology," (***Invited Paper***), *Proc. of the 13th International Conference on Solid-State and Integrated-Circuit Technology (ICSICT2006)*, Hangzhou, China, Oct. 2016, S37 - 2.
- X. Zhou, S. B. Chiah, B. Syamal, A. Ajaykumar, **X. Liu**, and H. T. Zhou, "Compact Modeling of III-V/Si FETs," (***Invited Paper***), *Proc. of the 12th International Conference on Solid-State and Integrated-Circuit Technology (ICSICT2014)*, Guilin, China, Oct. 2014, pp. 1012 - 1015.
- X. Zhou, S. B. Chiah, B. Syamal, H. T. Zhou, A. Ajaykumar, and **X. Liu**, "Compact Model Characteristics for Generic MIS-HEMTs," (***Invited Paper***), *Proc. of the NSTI Nanotech (WCM-Nanotech2014)*, Washington DC, Jun. 2014, vol. 2, pp. 495-498.

- X. Zhou, S. B. Chiah, B. Syamal, H. T. Zhou, A. Ajaykumar, and **X. Liu**, "Challenges and Prospects of Compact Modeling for Future Generation III-V/Si Co-integrated ULSI Circuit Design," *Proc. of the 11th International Workshop on Compact Modeling (IWCM-2014) at the Asia and South Pacific Design Automation Conference (ASP - DAC2014)*, Singapore, Jan. 23, 2014, pp. 1 - 2.

Bibliography

[1] C. C. Chen, Soonyoung Cha, Taizhi Liu, and L. Milor. System-level modeling of microprocessor reliability degradation due to bti and hci. In *2014 IEEE International Reliability Physics Symposium*, pages CA.8.1–CA.8.9, June 2014.

[2] M Dunga, WM Yang, XJ Xi, J He, W Liu, KM Cao, X Jin, JJ Ou, M Chan, AM Niknejad, et al. Bsim 4.6. 0 mosfet model-user?s manual. department of electrical engineering and computer sciences. *University of California, Berkeley*, 2006.

[3] K. Joardar, K. K. Gullapalli, C. C. McAndrew, M. E. Burnham, and A. Wild. *IEEE Transactions on Electron Devices*.

[4] Jin He, Xuernei Xi, Hui Wan, Mansun Chan, A. Niknejad, and Chenming Hu. Surface-potential-plus approach for next generation cmos device modeling. In *Junction Technology, 2004. IWJT '04. The Fourth International Workshop on*, pages 321–324, March 2004.

[5] Carlos Galup-Montorogi, Márcio C Schneider, Ana IA Cunhaw, and Oscar C Gouveia-Filhom. Theory, development, and applications of the advanced compact mosfet (acm) model. 2003.

[6] Matthias Bucher, Christophe Lallement, Christian Enz, Fabien Théodoloz, and François Krummenacher. The epfl-ekv mosfet model equations for simulation.

[7] A. T. Krishnan, V. Reddy, S. Chakravarthi, J. Rodriguez, S. John, and S. Krishnan. Nbti impact on transistor and circuit: models, mechanisms and scaling effects [mosfets]. In *IEEE International Electron Devices Meeting 2003*, pages 14.5.1–14.5.4, Dec 2003.

[8] B. C. Paul, Kunhyuk Kang, H. Kufluoglu, M. A. Alam, and K. Roy. Impact of nbti on

the temporal performance degradation of digital circuits. *IEEE Electron Device Letters*, 26(8):560–562, Aug 2005.

[9] S. Chakravarthi, A. Krishnan, V. Reddy, C. F. Machala, and S. Krishnan. A comprehensive framework for predictive modeling of negative bias temperature instability. In *2004 IEEE International Reliability Physics Symposium. Proceedings*, pages 273–282, April 2004.

[10] Manuel J Bellido, Jorge Juan Chico, and Manuel Valencia. *Logic-timing simulation and the degradation delay model*. Imperial College Press, 2006.

[11] Xiaojun Li, Jin Qin, and Joseph B Bernstein. Compact modeling of mosfet wearout mechanisms for circuit-reliability simulation. *IEEE Transactions on Device and Materials Reliability*, 8(1):98–121, 2008.

[12] Dieter K Schroder and Jeff A Babcock. Negative bias temperature instability: Road to cross in deep submicron silicon semiconductor manufacturing. *Journal of applied Physics*, 94(1):1–18, 2003.

[13] Muhammad Ashraful Alam and S Mahapatra. A comprehensive model of pmos nbti degradation. *Microelectronics Reliability*, 45(1):71–81, 2005.

[14] Sufi Zafar, Byoung H Lee, James Stathis, Allesandro Callegari, and Tak Ning. A model for negative bias temperature instability (nbti) in oxide and high/spl kappa/pfets 13/spl times/-c6d8c7f5f2. In *VLSI Technology, 2004. Digest of Technical Papers. 2004 Symposium on*, pages 208–209. IEEE, 2004.

[15] Sufi Zafar. Statistical mechanics based model for negative bias temperature instability induced degradation. *Journal of Applied Physics*, 97(10):103709, 2005.

[16] V Huard, M Denais, F Perrier, N Revil, C Parthasarathy, Alain Bravaix, and E Vincent. A thorough investigation of mosfets nbti degradation. *Microelectronics Reliability*, 45(1):83–98, 2005.

[17] Vijay Reddy, John Carulli, Anand Krishnan, William Bosch, and Brendan Burgess. Impact of negative bias temperature instability on product parametric drift. In *Test Conference, 2004. Proceedings. ITC 2004. International*, pages 148–155. IEEE, 2004.

[18] Vijay Reddy, Anand T Krishnan, Andrew Marshall, John Rodriguez, Sreedhar Natarajan, Tim Rost, and Srikanth Krishnan. Impact of negative bias temperature instability on

digital circuit reliability. *Microelectronics Reliability*, 45(1):31–38, 2005.

[19] Muhammad Ashraful Alam, Haldun Kufluoglu, Dhanoop Varghese, and Souvik Mahapatra. A comprehensive model for pmos nbti degradation: Recent progress. *Microelectronics Reliability*, 47(6):853–862, 2007.

[20] Maxim Ershov, Sharad Saxena, Sean Minehane, P Clifton, Mark Redford, R Lindley, H Karbasi, S Graves, and S Winters. Degradation dynamics, recovery, and characterization of negative bias temperature instability. *Microelectronics Reliability*, 45(1):99–105, 2005.

[21] Hideki Aono, Eiichi Murakami, Kousuke Okuyama, A Nishida, Masataka Minami, Y Ooji, and Katsuhiko Kubota. Modeling of nbti saturation effect and its impact on electric field dependence of the lifetime. *Microelectronics Reliability*, 45(7):1109–1114, 2005.

[22] A Narr and A Lill. Lifetime prediction for pmos and nmos devices based on a degradation model for gate-bias-stress. *Microelectronics Reliability*, 37(10-11):1433–1436, 1997.

[23] JEDEC Solid State Technology Association et al. Failure mechanisms and models for semiconductor devices. *JEDEC Publication JEP122-B*, 2003.

[24] F-C Hsu and Simon Tam. Relationship between mosfet degradation and hot-electron-induced interface-state generation. *IEEE Electron Device Letters*, 5(2):50–52, 1984.

[25] Guido Groeseneken, Robin Degraeve, Tanya Nigam, HE Maes, et al. Hot carrier degradation and time-dependent dielectric breakdown in oxides. *Microelectronic Engineering*, 49(1-2):27–40, 1999.

[26] Yusuf Leblebici and Sung-Mo Steve Kang. *Hot-carrier reliability of MOS VLSI circuits*, volume 227. Springer Science & Business Media, 2012.

[27] Chenming Hu, Simon C Tam, Fu-Chieh Hsu, Ping-Keung Ko, Tung-Yi Chan, and Kyle W Terrill. Hot-electron-induced mosfet degradation-model, monitor, and improvement. *IEEE Journal of Solid-State Circuits*, 20(1):295–305, 1985.

[28] Yuan Taur and Tak H Ning. *Fundamentals of modern VLSI devices*. Cambridge university press, 2013.

[29] Dominik Lorenz. *Aging analysis of digital integrated circuits*. PhD thesis, Technical University of Munich, 2012.

[30] L. Ma, Z. Chen, X. l. Ji, F. Yan, and Y. Shi. Multi-parameter model for hci lifetime predic-

tion. In *2012 IEEE 11th International Conference on Solid-State and Integrated Circuit Technology*, pages 1–3, Oct 2012.

[31] Guan Huei See. *Scalable compact modeling for nanometer CMOS technology.* PhD thesis, 2009.

[32] Oscar Camps, Rodrigo Picos, Miquel Roca, Benjamin Iniguez, and Eugeni Garcia-Moreno. A web application for mosfet compact model validation using cmc tests. In *Devices, Circuits and Systems, 2008. ICCDCS 2008. 7th International Caribbean Conference on*, pages 1–6. IEEE, 2008.

[33] Colin C McAndrew. Validation of mosfet model source–drain symmetry. *IEEE transactions on electron devices*, 53(9):2202–2206, 2006.

[34] Guan Huei See, Xing Zhou, Karthik Chandrasekaran, Siau Ben Chiah, Zhaomin Zhu, Chengqing Wei, Shihuan Lin, Guojun Zhu, and Guan Hui Lim. A compact model satisfying gummel symmetry in higher order derivatives and applicable to asymmetric mosfets. *IEEE Transactions on Electron Devices*, 55(2):624–631, 2008.

[35] Liu Weidong, Jin Xiaodong, Xi Xuemei, et al. Bsim3v3. 3 mosfet model users' manual. *Berkeley, CA: The Regents of the University of California*, 2005.

[36] Bing J Sheu, Donald L Scharfetter, P-K Ko, and M-C Jeng. Bsim: Berkeley short-channel igfet model for mos transistors. *IEEE Journal of Solid-State Circuits*, 22(4):558–566, 1987.

[37] Min-Chie Jeng, Zhihong Liu, and Yuhua Cheng. Deep-submicron mosfet modeling for circuit simulation. In *Microelectronics and VLSI, 1995. TENCON'95., IEEE Region 10 International Conference on*, pages 204–209. IEEE, 1995.

[38] W Yang, MV Dunga, X Xi, J He, W Liu, MC Kanyu, X Jin, JJ Ou, M Chan, AM Niknejad, et al. Bsim4. 6.2 mosfet model–user's manual. *Department of Electrical Engineering and Computer Sciences, http://www-device. eecs. berkley. edu/bsim3/BSIM4/BSIM462/doc/BSIM462 Manual. pdf*, 2008.

[39] H.C. Pao and C.T. Sah. Effects of diffusion current on characteristics of metal-oxide (insulator)-semiconductor transistors. *Solid-State Electronics*, 9(10):927 – 937, 1966.

[40] Xing Zhou, Guojun Zhu, Guan Huei See, Karthik Chandrasekaran, Siau Ben Chiah, and Khee Yong Lim. Unification of mos compact models with the unified regional modeling

approach. *Journal of computational electronics*, 10(1):121–135, 2011.

[41] Yannis Tsividis and Colin McAndrew. *Operation and Modeling of the MOS Transistor*. Oxford Univ. Press, 2011.

[42] GJ Zhu. *Compact modeling of non-classical MOSFETs for circuit simulation*. PhD thesis, Ph. D. thesis, Nanyang Technological Univ., Singapore, 2011.

[43] J. R. Brews. A charge-sheet model of the MOSFET. *Solid State Electronics*, 21:345–355, February 1978.

[44] C. G. Sodini, Ping-Keung Ko, and J. L. Moll. The effect of high fields on mos device and circuit performance. *IEEE Transactions on Electron Devices*, 31(10):1386–1393, Oct 1984.

[45] R. H. Tu, E. Rosenbaum, W. Y. Chan, C. C. Li, E. Minami, K. Quader, P. K. Ko, and C. Hu. Berkeley reliability tools-bert. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 12(10):1524–1534, Oct 1993.

[46] C RelXpert. Users manuals bsimpro+/relxpert/ultrasim. *Cadence Design Systems Inc*, 2010.

[47] Star-HSPICE Manual. Release 2001.2, 2001.

[48] M. Karam, W. Fikry, H. Haddara, and H. Ragai. Implementation of hot-carrier reliability simulation in eldo. In *ISCAS 2001. The 2001 IEEE International Symposium on Circuits and Systems (Cat. No.01CH37196)*, volume 5, pages 515–518 vol. 5, 2001.

[49] E. Maricau and G. Gielen. Efficient variability-aware nbti and hot carrier circuit reliability analysis. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 29(12):1884–1893, Dec 2010.

[50] H. Kufluoglu, V. Reddy, A. Marshall, J. Krick, T. Ragheb, C. Cirba, A. Krishnan, and C. Chancellor. An extensive and improved circuit simulation methodology for nbti recovery. In *2010 IEEE International Reliability Physics Symposium*, pages 670–675, May 2010.

[51] Sachin Sapatnekar. *Timing*. Springer Science & Business Media, 2004.

[52] J. Qian, S. Pullela, and L. Pillage. Modeling the ldquo;effective capacitance rdquo; for the rc interconnect of cmos gates. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 13(12):1526–1535, Dec 1994.

[53] Jifeng Chen, Shuo Wang, Nemat Bidokhti, and Mohammad Tehranipoor. A framework for fast and accurate critical-reliability paths identification. In *Proceedings of the IEEE North Atlantic Test Workshop (NATW)*, 2011.

[54] I Synopsys. Hspice user?s manual: Simulation and analysis. 2010.

[55] Lifeng Wu, Jingkun Fang, Hirokazu Yonezawa, Yoshiyuki Kawakami, Nobufusa Iwanishi, Heting Yan, Ping Chen, Alvin I-Hsien Chen, Norio Koike, Yoshifumi Okamoto, Chune-Sin Yeh, and Zhihong Liu. Glacier: a hot carrier gate level circuit characterization and simulation system for vlsi design. In *Proceedings IEEE 2000 First International Symposium on Quality Electronic Design (Cat. No. PR00525)*, pages 73–79, 2000.

[56] T. Sakurai and A. R. Newton. Alpha-power law mosfet model and its applications to cmos inverter delay and other formulas. *IEEE Journal of Solid-State Circuits*, 25(2):584–594, Apr 1990.

[57] Bipul C Paul, Kunhyuk Kang, Haldun Kufluoglu, Muhammad Ashraful Alam, and Kaushik Roy. Temporal performance degradation under nbti: estimation and design for improved reliability of nanoscale circuits. In *Proceedings of the conference on Design, automation and test in Europe: Proceedings*, pages 780–785. European Design and Automation Association, 2006.

[58] Hong Luo, Yu Wang, Ku He, Rong Luo, Huazhong Yang, and Yuan Xie. Modeling of pmos nbti effect considering temperature variation. In *Quality Electronic Design, 2007. ISQED'07. 8th International Symposium on*, pages 139–144. IEEE, 2007.

[59] Hong Luo, Yu Wang, Ku He, Rong Luo, Huazhong Yang, and Yuan Xie. A novel gate-level nbti delay degradation model with stacking effect. In *PATMOS*, volume 4644, pages 160–170. Springer, 2007.

[60] S. V. Kumar, C. H. Kim, and S. S. Sapatnekar. Nbti-aware synthesis of digital circuits. In *2007 44th ACM/IEEE Design Automation Conference*, pages 370–375, June 2007.

[61] Alexander Stempkovsky, Alexey Glebov, and Sergey Gavrilov. Calculation of stress probability for nbti-aware timing analysis. In *Quality of Electronic Design, 2009. ISQED 2009. Quality Electronic Design*, pages 714–718. IEEE, 2009.

[62] Yoshio Miura and Yasuo Matukura. Investigation of silicon-silicon dioxide interface using mos structure. *Japanese Journal of Applied Physics*, 5(2):180, 1966.

[63] D. Ielmini, M. Manigrasso, F. Gattel, and M. G. Valentini. A new nbti model based on hole trapping and structural relaxation in mos dielectrics. *IEEE Transactions on Electron Devices*, 56(9):1943–1952, Sept 2009.

[64] E Cartier, JH Stathis, and DA Buchanan. Passivation and depassivation of silicon dangling bonds at the si/sio2 interface by atomic hydrogen. *Applied physics letters*, 63(11):1510–1512, 1993.

[65] S. Mahapatra, P. B. Kumar, and M. A. Alam. Investigation and modeling of interface and bulk trap generation during negative bias temperature instability of p-mosfets. *IEEE Transactions on Electron Devices*, 51(9):1371–1379, Sept 2004.

[66] H. Kufluoglu and M. A. Alam. Theory of interface-trap-induced nbti degradation for reduced cross section mosfets. *IEEE Transactions on Electron Devices*, 53(5):1120–1130, May 2006.

[67] S. N. Rashkeev, D. M. Fleetwood, R. D. Schrimpf, and S. T. Pantelides. Defect Generation by Hydrogen at the Si- SiO_2 Interface. *Physical Review Letters*, 87(16):165506, October 2001.

[68] S. V. Kumar, C. H. Kim, and S. S. Sapatnekar. An analytical model for negative bias temperature instability. In *2006 IEEE/ACM International Conference on Computer Aided Design*, pages 493–496, Nov 2006.

[69] M. A. Alam. A critical examination of the mechanics of dynamic nbti for pmosfets. In *IEEE International Electron Devices Meeting 2003*, pages 14.4.1–14.4.4, Dec 2003.

[70] Dominik Lorenz, Martin Barke, and Ulf Schlichtmann. Efficiently analyzing the impact of aging effects on large integrated circuits. *Microelectronics Reliability*, 52(8):1546–1552, 2012.

[71] Michael G Xakellis and Farid N Najm. Statistical estimation of the switching activity in digital circuits. In *Proceedings of the 31st annual Design Automation Conference*, pages 728–733. ACM, 1994.

[72] Mehmet A Cirit. Estimating dynamic power consumption of cmos circuits. In *IEEE International Conference on Computer-Aided Design*, pages 534–537, 1987.

[73] Farid N Najm. Transition density, a stochastic measure of activity in digital circuits. In *Proceedings of the 28th ACM/IEEE Design Automation Conference*, pages 644–649.

ACM, 1991.

[74] Farid N Najm. Transition density: A new measure of activity in digital circuits. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 12(2):310–323, 1993.

[75] GCKY Chen, KY Chuah, MF Li, Daniel SH Chan, CH Ang, JZ Zheng, Y Jin, and DL Kwong. Dynamic nbti of pmos transistors and its impact on device lifetime. In *Reliability Physics Symposium Proceedings, 2003. 41st Annual. 2003 IEEE International*, pages 196–202. IEEE, 2003.

[76] Dominik Lorenz, Martin Barke, and Ulf Schlichtmann. Aging analysis at gate and macro cell level. In *Proceedings of the International Conference on Computer-Aided Design*, pages 77–84. IEEE Press, 2010.

[77] Jin Ju Kim, Moonju Cho, Luigi Pantisano, Ukjin Jung, Young Gon Lee, Thomas Chiarella, Mitsuhiro Togo, Naoto Horiguchi, Guido Groeseneken, and Byoung Hun Lee. Process-dependent n/pbti characteristics of tin gate finfets. *IEEE Electron Device Letters*, 33(7):937–939, 2012.

[78] Tony Tae-Hyoung Kim, Pong-Fei Lu, Keith A Jenkins, and Chris H Kim. A ring-oscillator-based reliability monitor for isolated measurement of nbti and pbti in high-k/metal gate technology. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 23(7):1360–1364, 2015.

[79] S Zafar, Y Kim, V Narayanan, C Cabral, V Paruchuri, B Doris, J Stathis, A Callegari, and M Chudzik. A comparative study of nbti and pbti (charge trapping) in sio2/hfo2 stacks with fusi, tin, re gates. In *VLSI Technology, 2006. Digest of Technical Papers. 2006 Symposium on*, pages 23–25. IEEE, 2006.

[80] Shushanik Karapetyan and Ulf Schlichtmann. Integrating aging aware timing analysis into a commercial sta tool. In *VLSI Design, Automation and Test (VLSI-DAT), 2015 International Symposium on*, pages 1–4. IEEE, 2015.

[81] Dustin K Slisher, RG Filippi, Daniel W Storaska, and Alberto H Gay. Scaling of si mosfets for digital applications. *Rensselaer Polytechnic Institute*, 1999.

[82] N. Koike and K. Tatsuuma. A drain avalanche hot carrier lifetime model for n- and p-channel mosfets. *IEEE Transactions on Device and Materials Reliability*, 4(3):457–466,

Sept 2004.

[83] E. Takeda, A. Shimizu, and T. Hagiwara. Role of hot-hole injection in hot-carrier effects and the small degraded channel region in mosfet's. *IEEE Electron Device Letters*, 4(9):329–331, Sep 1983.

[84] P. Heremans, R. Bellens, G. Groeseneken, and H. E. Maes. Consistent model for the hot-carrier degradation in n-channel and p-channel mosfets. *IEEE Transactions on Electron Devices*, 35(12):2194–2209, Dec 1988.

[85] Chen Ih-Chin, Choi Jeong Yeol, and Hu Chenming. The effect of channel hot-carrier stressing on gate-oxide integrity in mosfets. *IEEE Transactions on Electron Devices*, 35(12):2253–2258, Dec 1988.

[86] D. Lorenz, G. Georgakos, and U. Schlichtmann. Aging analysis of circuit timing considering nbti and hci. In *2009 15th IEEE International On-Line Testing Symposium*, pages 3–8, June 2009.

[87] Lawrence T.. Pillage, Ronald A Rohrer, and Chandramouli Visweswariah. *Electronic Circuit and System Simulation Methods*. McGraw-Hill, 1995.

[88] W. Wang, V. Reddy, A. T. Krishnan, R. Vattikonda, S. Krishnan, and Y. Cao. Compact modeling and simulation of circuit reliability for 65-nm cmos technology. *IEEE Transactions on Device and Materials Reliability*, 7(4):509–517, Dec 2007.

[89] I Messaris, N Fasarakis, TA Karatsori, A Tsormpatzoglou, G Ghibaudo, and CA Dimitriadis. Hot carrier degradation modeling of short-channel n-finfets. In *Device Research Conference (DRC), 2015 73rd Annual*, pages 183–184. IEEE, 2015.

[90] Elie Maricau, Pieter De Wit, and Georges Gielen. An analytical model for hot carrier degradation in nanoscale cmos suitable for the simulation of degradation in analog ic applications. *Microelectronics Reliability*, 48(8):1576–1580, 2008.

[91] Yao Wang, Sorin Cotofana, and Liang Fang. A unified aging model of nbti and hci degradation towards lifetime reliability management for nanoscale mosfet circuits. In *Proceedings of the 2011 IEEE/ACM International Symposium on Nanoscale Architectures*, pages 175–180. IEEE Computer Society, 2011.

[92] Souvik Mahapatra, Dipankar Saha, Dhanoop Varghese, and P Bharath Kumar. On the generation and recovery of interface traps in mosfets subjected to nbti, fn, and hci stress.

IEEE Transactions on Electron Devices, 53(7):1583–1592, 2006.

[93] Haldun Kufluoglu and M Ashraful Alam. A geometrical unification of the theories of nbti and hci time-exponents and its implications for ultra-scaled planar and surround-gate mosfets. In *Electron Devices Meeting, 2004. IEDM Technical Digest. IEEE International*, pages 113–116. IEEE, 2004.

[94] Farid N Najm. A survey of power estimation techniques in vlsi circuits. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 2(4):446–455, 1994.

[95] SC Sun and James D Plummer. Electron mobility in inversion and accumulation layers on thermally oxidized silicon surfaces. *IEEE Journal of Solid-State Circuits*, 15(4):562–573, 1980.

[96] K. Y. Lim and X. Zhou. An analytical effective channel-length modulation model for velocity overshoot in submicron mosfets based on energy-balance formulation. *Microelectronics Reliability*, 42(12):1857–1864, 2002.

[97] Y. Cao, J. Velamala, K. Sutaria, M. S. W. Chen, J. Ahlbin, I. S. Esqueda, M. Bajura, and M. Fritze. Cross-layer modeling and simulation of circuit reliability. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 33(1):8–23, Jan 2014.

[98] Juin Jei Liou, Adelmo Ortiz-Conde, and Francisco Garcia-Sanchez. *Analysis and design of MOSFETs: modeling, simulation, and parameter extraction*. Springer Science & Business Media, 2012.

[99] Kazuo Terada, Katsuhiko Nishiyama, and Kei-Ichi Hatanaka. Comparison of mosfet-threshold-voltage extraction methods. *Solid-State Electronics*, 45(1):35–40, 2001.

[100] Dieter K Schroder. *Semiconductor material and device characterization*. John Wiley & Sons, 2006.

[101] M. Tsuno, M. Suga, M. Tanaka, K. Shibahara, M. Miura-Mattausch, and M. Hirose. Physically-based threshold voltage determination for mosfet's of all gate lengths. *IEEE Transactions on Electron Devices*, 46(7):1429–1434, Jul 1999.

[102] A. Bazigos, M. Bucher, J. Assenmacher, S. Decker, W. Grabinski, and Y. Papananos. An adjusted constant-current method to determine saturated and linear mode threshold voltage of mosfets. *IEEE Transactions on Electron Devices*, 58(11):3751–3758, Nov 2011.

[103] B El-Kareh, WR Tonti, and SL Titcomb. A submicron mosfet parameter extraction technique. *IBM journal of research and development*, 34(2.3):243–249, 1990.

[104] M. B. Das. Physical limitations of MOS structures. *Solid State Electronics*, 12:305–336, May 1969.

[105] Chih-Tang Sah. Characteristics of the metal-oxide-semiconductor transistors. *IEEE Transactions on Electron Devices*, 11(7):324–345, 1964.

[106] L Vadasz and AS Grove. Temperature dependence of mos transistor characteristics below saturation. *IEEE Transactions on Electron Devices*, 13(12):863–866, 1966.

[107] Adelmo Ortiz-Conde, FJ Garcıa Sánchez, Juin J Liou, Antonio Cerdeira, Magali Estrada, and Y Yue. A review of recent mosfet threshold voltage extraction methods. *Microelectronics Reliability*, 42(4):583–596, 2002.

[108] Hon-Sum Wong, Marvin H White, Thomas J Krutsick, and Richard V Booth. Modeling of transconductance degradation and extraction of threshold voltage in thin oxide mosfet's. *Solid-State Electronics*, 30(9):953–968, 1987.

[109] X. Zhou, K. Y. Lim, and D. Lim. A simple and unambiguous definition of threshold voltage and its implications in deep-submicron mos device modeling. *IEEE Transactions on Electron Devices*, 46(4):807–809, Apr 1999.

[110] Xing Zhou, KY Lim, and W Qian. Threshold voltage definition and extraction for deep-submicron mosfets. *Solid-State Electronics*, 45(3):507–510, 2001.

[111] Yuan Chen, John S Suehle, C-C Shen, J Bernstein, C Messick, and P Chaparala. The correlation of highly accelerated q/sub bd/tests to tddb life tests for ultra-thin gate oxides. In *Reliability Physics Symposium Proceedings, 1998. 36th Annual. 1998 IEEE International*, pages 87–91. IEEE, 1998.

[112] John S Suehle. Ultrathin gate oxide reliability: Physical models, statistics, and characterization. *IEEE Transactions on Electron Devices*, 49(6):958–971, 2002.

[113] Yee-Chia Yeo, Qiang Lu, and Chenming Hu. Mosfet gate oxide reliability: Anode hole injection model and its applications. *International journal of high speed electronics and systems*, 11(03):849–886, 2001.

[114] Ernest Y Wu, Edward J Nowak, Alex Vayshenker, Wing L Lai, and David L Harmon. Cmos scaling beyond the 100-nm node with silicon-dioxide-based gate dielectrics. *IBM*

Journal of Research and Development, 46(2.3):287–298, 2002.

[115] E Wu, J Sune, W Lai, E Nowak, J McKenna, A Vayshenker, and D Harmon. Interplay of voltage and temperature acceleration of oxide breakdown for ultra-thin gate oxides. *Solid-State Electronics*, 46(11):1787–1798, 2002.

[116] C Chaneliere, JL Autran, RAB Devine, and B Balland. Tantalum pentoxide (ta2o5) thin films for advanced dielectric applications. *Materials Science and Engineering: R: Reports*, 22(6):269–322, 1998.

[117] M Gurvitch, L Manchanda, and JM Gibson. Study of thermally oxidized yttrium films on silicon. *Applied physics letters*, 51(12):919–921, 1987.

[118] W Tsai, Lars-Ake Ragnarsson, Tom Schram, Stefan Degendt, Marc Heyns, MC Ozturk, LJ Chen, PJ Timans, F Roozeboom, EP Gusev, et al. Challenges in integration of metal gate high-k dielectrics gate stacks. *Advanced short-time thermal processing for Si-based CMOS devices II, Proc. ECS*, pages 321–327, 2004.

[119] Sufi Zafar, Byoung H Lee, and James Stathis. Evaluation of nbti in hfo/sub 2/gate-dielectric stacks with tungsten gates. *IEEE Electron Device Letters*, 25(3):153–155, 2004.

[120] EP Gusev, DA Buchanan, E Cartier, A Kumar, D DiMaria, S Guha, A Callegari, S Zafar, PC Jamison, DA Neumayer, et al. Ultrathin high-k gate stacks for advanced cmos devices. In *Electron Devices Meeting, 2001. IEDM'01. Technical Digest. International*, pages 20–1. IEEE, 2001.

[121] Michel Houssa, Andre Stesmans, M Naili, and MM Heyns. Charge trapping in very thin high-permittivity gate dielectric layers. *Applied Physics Letters*, 77(9):1381–1383, 2000.

[122] Charles M Perkins, Baylor B Triplett, Paul C McIntyre, Krishna C Saraswat, Suvi Haukka, and Marko Tuominen. Electrical and materials properties of zro 2 gate dielectrics grown by atomic layer chemical vapor deposition. *Applied Physics Letters*, 78(16):2357–2359, 2001.

[123] Miaomiao Wang, Ramachandran Muralidhar, James H Stathis, Barry Paul Linder, Hemanth Jagannathan, and Jonathan Faltermeier. Superior pbti reliability for soi finfet technologies and its physical understanding. *IEEE Electron Device Letters*, 34(7):837–839, 2013.

[124] Jong-Tae Park and J-P Colinge. Multiple-gate soi mosfets: device design guidelines. *IEEE transactions on electron devices*, 49(12):2222–2229, 2002.

[125] BS Doyle, S Datta, M Doczy, S Hareland, B Jin, J Kavalieros, T Linton, A Murthy, R Rios, and R Chau. High performance fully-depleted tri-gate cmos transistors. *IEEE Electron Device Letters*, 24(4):263–265, 2003.

[126] Colinge Jean-Pierre. Silicon-on-insulator technology: materials to vlsi. *Edition, by*, 2, 1997.

[127] T Kauerauf. *Degradation and breakdown of MOS gate stacks with high permittivity dielectrics*. PhD thesis, Ph. D. dissertation, Katholieke Universiteit Leuven, Leuven, Belgium, 2007.

[128] Paul E Nicollian, William R Hunter, and Jerry C Hu. Experimental evidence for voltage driven breakdown models in ultrathin gate oxides. In *IEEE International Reliability Physics Symposium Proceedings*, pages 7–15. IEEE; 1999, 2000.

[129] K Ohgata, M Ogasawara, K Shiga, S Tsujikawa, E Murakami, H Kato, H Umeda, and K Kubota. Universality of power-law voltage dependence for tddb lifetime in thin gate oxide pmosfets. In *Reliability Physics Symposium, 2005. Proceedings. 43rd Annual. 2005 IEEE International*, pages 372–376. IEEE, 2005.

[130] Paul E Nicollian, Anand T Krishnan, Cathy A Chancellor, Rajesh B Khamankar, Srinivasan Chakravarthi, Chris Bowen, and Vijay K Reddy. The current understanding of the trap generation mechanisms that lead to the power law model for gate dielectric breakdown. In *Reliability physics symposium, 2007. proceedings. 45th annual. ieee international*, pages 197–208. IEEE, 2007.

[131] K Torii, H Kitajima, T Arikado, K Shiraishi, S Miyazaki, K Yamabe, et al. Physical model of bti, tddb and silc in hfo2-based high-k gate dielectrics. electron devices meeting, 2004. *IEDM technical digest. IEEE international*, pages 13–15, 2004.

[132] X Wu, DB Migas, X Li, M Bosman, N Raghavan, VE Borisenko, and KL Pey. Role of oxygen vacancies in hfo 2-based gate stack breakdown. *Applied Physics Letters*, 96(17):172901, 2010.

[133] CH Yang, SC Chen, YS Tsai, R Lu, and Y-H Lee. The physical explanation of tddb power law lifetime model through oxygen vacancy trap investigations in hkmg nmos

finfet devices. In *Reliability Physics Symposium (IRPS), 2017 IEEE International*, pages 3C–4. IEEE, 2017.

[134] Pedro C Feijoo, Thomas Kauerauf, María Toledano-Luque, Mitsuhiro Togo, Enrique San Andrés, and Guido Groeseneken. Time-dependent dielectric breakdown on sub-nanometer eot nmos finfets. *IEEE Transactions on Device and Materials Reliability*, 12(1):166–170, 2012.