Variability in BTI-Induced Device Degradation: from Silicon Measurement to SRAM Yield Prediction

Hiromitsu AWANO

Abstract

Large scale integration (LSI) circuits are now recognized as indispensable infrastructures for our daily lives. With advancements of semiconductor manufacturing technologies, LSIs keep providing increasingly advanced functions while the device cost is maintained or even reduced, which will further promotes the adoption of LSIs. Applications of LSIs include not only consumer electronics products but also mission critical systems such as automotive or medical equipments. Hence, ensuring LSI longevity is recognized as a critical concern.

Toward sub-ten nanometer eras, ensuring the reliability of LSIs has emerged as an increasingly challenging task. One good example is the gate oxide film whose thickness is now approaching few nano-meters at the cutting edge technology node. Due to the extremely thinned film, the gate electric field becomes increasingly strong, causing various reliability problems. In fact, major degradation modes of modern LSIs have relation to the gate insulator films or nearby components.

One of the major reliability concern is the device degradation induced by bias temperature instability (BTI). In the conventional studies on BTI-induced device degradation, researchers had usually considered the degradation of a single transistor. However, as the transistor size shrinks, increasing device variability causes a lot of problem. For example, in the early 2000s, we have witnessed that vast amount of effort has been paid to develop statistical timing analysis tools or yield estimation tools. Similarly, in the highly scaled transistors, the variability in the device degradation is expected to become dominant. Therefore, neglecting the variability in the degradation may enforce designers to maintain either too pessimistic or too optimistic design margin, which should be avoided.

This dissertation first proposes the circuit structure suitable for the statistical measurement of the BTI-induced degradation on a silicon chip. In order to invoke the noticeable threshold voltage (V_{TH}) shift, hours or even days of stress time is required, making the serial measurements of the degradation on many transistors almost impossible. Hence, the proposed circuit structure, which is named as "BTIarray," employs a parallel measurement scheme in which the

stress is applied to transistors in parallel and their V_{TH} are measured in a pipeline manner. In the degradation measurement, the stress time is typically magnitudes longer than that required for acquiring the V_{TH} and hence almost N-fold reduction of total measurement time can be achieved if N transistors are integrated on a chip in an array fashion. Through the measurements of the fabricated circuits, it is found that the statistical distribution of the power-law exponent follows the log-normal distribution.

In the measurement of the BTI-induced device degradation, stair-like V_{TH} shifts, which is known to be random telegraph noise (RTN), are commonly observed. In the formation of the BTI-induced V_{TH} shifts, electrically active traps located at the silicon-dielectric interface are considered to play an important role: these traps capture or emit carriers, which in turn causes the fluctuation of the inversion charge, resulting in the V_{TH} shift. The observation of the stair-like V_{TH} shift supports this theory. For the better understandings of the physics behind BTI and RTN, the stair-like V_{TH} shifts are examined closely in this dissertation. Usually, a single transistor contains two or more number of traps, and the V_{TH} shift is determined as the combination of each trap state, resulting in the multi-level V_{TH} fluctuation. In order to closely examine the effect of each trap, the trap activities must be extracted from the measured V_{TH} sequence. This problem is known to be the under determined problem because only the mixture of the fluctuations can be observed. In this dissertation, this problem is solved by utilizing the statistical machine learning technique. The generation process of the stair-like V_{TH} shift is represented as a statistical generative model and the parameters of the model are adjusted so that the model best describes the observed V_{TH} time series. Through the numerical experiments with synthetic and with the measured V_{TH} time series observed on BTIarray, it is shown that the proposed method successfully decomposes the multi-trap activities from the single V_{TH} sequence.

The measurement or the analysis of the degradation are not always sufficient to enhance the reliability of LSI. Circuit designers are usually interested not just in the physical mechanism behind the device degradation but in how their design suffers from the degradation. Hence, a method that can predict the yield degradation of an SRAM bit cell is proposed. Because an SRAM bit cell is highly replicated on a chip, each bit cell is required to satisfy the high level of reliability, e.g. typical failure probability required for an SRAM bit cell is 10^{-6} or lower. In order to estimate such rare failure probability, importance sampling technique is usually introduced to accelerate the convergence of Monte Carlo estimation. However, the construction of the alternative distribution arises another concern. Moreover, due to the device degradation, shape of the suitable alternative distribution changes over time. Hence, in this dissertation, particle filter is introduced to sequentially track the temporal change of the alternative

distribution. Through numerical experiment, it is shown that the proposed method achieves 9.76× speed up compared to the repeated yield estimation at each aging time step in which the alternative distribution is constructed independently.

This dissertation addresses the BTI-induced reliability issues in multiple layers, i.e. from the silicon measurement, analysis, and the SRAM yield prediction. The problems in each layer are discovered and formulated, and corresponding techniques suitable to solve respective problems are proposed. To the best of my knowledge, this dissertation is the first attempt to comprehensively handle the variability in BTI-induced degradation, i.e. from the BTI measurement to the yield prediction. In the advanced technology nodes, the impact of the variability in device degradations is expected to increase and hence the techniques proposed in this dissertation are expected to be vital for enhancing the reliability of the LSI systems.

Acknowledgments

This work was accomplished at Takashi Sato Laboratory, Graduate School of Informatics, Kyoto University. I express my gratitude to all people who helped me.

Firstly, I would like to express my sincere gratitude to my advisor Prof. Takashi Sato for the continuous support of my Ph.D study and related research. He reviewed all of my papers and provided hundreds of valuable suggestion. His advise helped me in all the time of my long student life. I have been amazingly fortunate to have the advisor who gave me the opportunity to explore on my own, and at the same time, the guidance to recover when my steps faltered. I could not have imagined having a better advisor and mentor for my Ph.D study.

Besides my advisor, I would like to express my deep gratitude to the rest of my dissertation committee, Prof. Hidetoshi Onodera and Prof. Naofumi Takagi for carefully reading of and commenting on this dissertation. Their comments were absolutely imperative to widen my research from various view perspective.

I would like to express my gratitude to Prof. Hiroshi Okuno for his support of my bachelor's research. He taught me a fundamental knowledge to be a good researcher, which forms the platform on which my carrier as a researcher has been building. I am grateful to Prof. Tetsuya Ogata for providing a new perspective on robotics and intelligence.

My sincere thanks also go to Prof. Hiroyuki Ochi, Prof. Hiroshi Tsutsui, and Dr. Masayuki Hiromoto for reviewing my papers and for providing access to the laboratory and research facilities. Without their precious support, it would be not possible to complete this dissertation.

I deeply thank Takashi Sato Laboratory members, who are good colleagues and friends. Discussions with Dr. Michihiro Shintani and Dr. Takashi Imagawa were highly suggestive and helped me to come up with a lot of new ideas. I also thank Mr. Yuto Takagaki, Mr. Satoshi Konishi, Mr. Tadaaki Oni, Mr. Motoki Yoshinaga, Mr. Song Bian, and the other students. Chatting and enjoying sweets with them gave me a chance to let off steam. I would like to express my gratitude to excellent secretaries, Ms. Akiko Ishii and Ms. Shuko Nishiyama.

I express my gratitude to the members of frog research team, Dr. Ikkyu

Aihara, Dr. Kazutoshi Itoyama, Dr. Takeshi Mizumoto, and Mr. Yoshiaki Bando. Thanks to Dr. Aihara, I had an opportunity to pursue the quite exciting project of frog chorus. With their support, I could gain a lot of amazing experiences including field experiments at Okinoshima-island or at Springbrook National Park in Australia.

I am grateful to the Japan Society for the Promotion and Science (JSPS) for their financial support as a Fellowship for Young Scientists (DC1).

Last but not least, I am truly grateful to my parents, Hikaru Awano and Akiko Awano for their support of my long student life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Large scale integration (LSI) circuit has been emerged as an indispensable infrastructure for our daily lives. With the advancement of semiconductor manufacturing technologies, increasing number of transistors can be integrated into a small silicon chip. Hence, a single chip can provide a wide variety of complex functions while the device cost is remained or even reduced, which has been contributed to the wide adoption of LSIs. Example applications of LSIs include mission critical systems such as for medical use, aircraft, defence system, electric power system, traffic control system, and so on. This wide adoption means that LSIs must shoulder increasingly important responsibilities. Failures of LSIs may lead to a great disservice to society and hence ensuring the reliability of LSIs have been emerged as a critical concern.

As we enter into the deca-nanometer era, the increasing current density and the high operational temperature have originated many reliability issues. One example is the gate oxide film whose thickness is now approaching few nano-meters at the cutting edge technology node. Strong gate electric field damages the gate dielectric, which leads to the performance degradation or the circuit failure. In fact, major degradation modes of modern LSIs have relation to the gate insulator films or nearby components and hence the quality of the gate insulator film has emerged as a major factor that determines the reliability of LSI.

The major degradation modes of a metal-oxide-semiconductor (MOS) transistor are time-dependent dielectric breakdown (TDDB) [1], hot carrier injection (HCI) [2], and bias temperature instability (BTI) [3]. TDDB is the breakdown of the gate oxide, which is caused by the formation of the conducting path between the gate electrode and the silicon surface. The defects located

inside the gate insulator films are considered to be the origin of the formation of the conducting path [1]. While TDDB causes complete breakdown of the transistor, HCI and BTI cause the threshold voltage (V_{TH}) shift, which slow down the switching speed of the transistor. Due to the shrinkage of the channel length, the carriers are accelerated by the increasingly strong drain electric field and injected into the gate insulator films, which causes HCI-induced degradation. BTI-induced V_{TH} shift is promoted by the application of the strong gate electric field and the elevated temperature [4]. What makes BTI-induced degradation unique is the recovery effect. Most part of the degraded V_{TH} recovers as soon as the transistor is released from the stress condition, making the measurement and prediction of the BTI-induced degradation as difficult challenges. Due to the unrecoverable component in the degraded V_{TH}, repetitive application of the stress and the recovery eventually leads to the large V_{TH} shift in the long-term circuit operation, which finally leads to the malfunction of the LSI. The physical mechanism of BTI is still under intensive research but many researchers agree that traps located at the silicon-dielectric interface is an important key to explain the V_{TH} shift. These electrically active traps capture or emit carriers, which causes the fluctuation of the inversion charge and the V_{TH} shift. Traditionally, negative BTI (NBTI) observed on pMOS transistors has attracted major concern. However, due to the application of new materials such as high-k gate insulators and metal gates, positive BTI (PBTI) observed on nMOS transistors also attracts increasing concern [5]. Usually, BTI refers to the long-term evolution of V_{TH} shifts. However, fluctuation of V_{TH} within relatively short period of time is also observed, which is known to be random telegraph noise (RTN) [6].

BTI and RTN-induced V_{TH} shifts are expected to increase rapidly as the transistor shrinks [7] because the impact of single electron increases inversely with the channel area. Hence, in this dissertation, BTI and RTN are the topic of interest but note that the techniques proposed in this dissertation are general and thus they can be applied to other degradation modes with slight modifications.

In order to consider the device degradation in the deca-nanometer era, it is definitely required to take into account the variability in the device degradation. Due to the process shrinkage, even an atomic level bump on the gate electrode or the fluctuations in the number of dopant ions under the channel area have a large impact on the characteristics of the transistor. Therefore, the large variation of the device degradation is expected in the deca-nanometer era. Neglecting this variability in the degradation may lead to a pessimistic or optimistic design margin, which should be avoided.

In this dissertation, a circuit structure suitable for the BTI measurement on the large number of transistors is first proposed. Note again that the

physical mechanism of the BTI-induced degradation is still under the intensive research and hence the observation of the BTI on the actual silicon is practically important not only for the reliability estimation or prediction but also for the better understandings of the device physics. Through the measurement of the BTI, the stair-like V_{TH} shifts is frequently observed, which implies that the discrete event of carriers' capture and emission at the silicon-dielectric interface is strongly associated with the formation of the BTI-induced degradation. Hence, in this dissertation, the stair-like V_{TH} is closely examined and a method that can automatically extract the model parameters of carriers from the measured V_{TH} sequence is proposed. However, the measurement or the examination of BTI is not sufficient for enhancing the circuit's reliability because we can not predict the impact of BTI on the actual circuit component. Hence, this dissertation finally proposes an efficient method for the circuit failure probability estimation which can take into account the BTI-induced degradation and their variability.

1.2 Issues and Solutions

In this section, details of the above mentioned issues and their solutions are provided.

1.2.1 BTI measurement

The physical mechanism behind the BTI degradation is still the target of intensive researches. Hence, measurements of BTI on silicon devices and acquiring experimental data on V_{TH} shifts are definitely required to understand the BTI phenomenon. Developing the measurement circuit for the efficient characterization of BTI on silicon is thus an important component not only for the academic community to study the degradation physics but also for the industrial community to ensure reliabilities of their products. In order to measure BTI-induced degradation, two major problems should be addressed. First, the BTI-induced V_{TH} shift progresses very slowly and hence it takes several hours or even days to invoke the noticeable V_{TH} shift even under the accelerated condition, i.e. at a high stress voltage and at an elevated temperature. Second, similar to the "static" variability originated from the manufacturing process variability, BTI-induced V_{TH} shifts are also expected to vary from transistor to transistor. In order to capture the statistical aspect of the BTI-induced degradation, BTI-induced V_{TH} shift must be measured on thousands or on even large number of transistors. Because even a measurement of a single transistor takes hours or days, serial measurements of

BTI on thousands of transistors are obviously unrealistic and hence a parallel measurement technique is definitely required.

In this dissertation, BTIarray, a circuit structure suitable for the parallel BTI measurement is proposed. In typical BTI measurements, almost all of the measurement time is used for applying the stress to invoke the degradation and the V_{TH} measurement occupies only a small part of the total measurement time. BTIarray utilizes this property: by overlapping stress time over all devices under test (DUTs), the time required for BTI measurement on N DUTs becomes comparable to that for a single DUT. Moreover, because BTIarray employs an array structure, PAD terminals can be shared among all DUTs, contributing the small layout area. Moreover, a single semiconductor parametric analyzer (PA) can be time-shared, which reduces the measurement cost.

An automated measurement environment is also developed. In order to characterize BTI-related parameters, duration in which the DUTs are in stress or in recovery mode must be controlled accurately. Moreover, same measurement scenarios may be repeatedly used to see the hysteresis effect of the degradation. Hence, the automated measurement system based on an FPGA-based pattern generator and scripting environment are developed. The control signals for BTIarray is generated using the FPGA-based pattern generator to reduce the timing jitter. The measurement scenario can be written using Python, which contributes to enhance the productivity and readability of the measurement scenario.

1.2.2 Automated data analysis

In the V_{TH} measurement waveforms of highly scaled transistors, stair-like changes in the V_{TH} shift are commonly observed, leading to an increasing attention to the trap-detrap (TD) theory [8]. According to the TD theory, the electrically active traps located inside the gate oxide film are considered to be the origin of the BTI-induced temporal V_{TH} shift. When the stress is applied to the transistor, the carriers are captured by the traps, leading to the temporal change of the inversion charge. As a result, the V_{TH} shift is observed. The trapped carriers are emitted randomly when the transistor is released from the stress condition, leading to the recovery of the degraded V_{TH}.

Usually, a single transistor contains several traps and hence only the summation of the contributions from each trap included in the transistor can be observed as the total V_{TH} shift. In order to understand the mechanism behind BTI, the activities of respective traps need to be analyzed separately. However, manual separation of the trap activities from hundreds or even from thousands of measurement data is obviously unrealistic. Automated separation of the trap activities is thus required.

The difficulty of this problem is that only the degenerated information, i.e. total V_{TH} shifts, can be obtained. Reconstruction of each trap activity is thus an underdetermined problem, i.e. there are fewer observations to fully reconstruct the degenerated trap activities. In this dissertation, this problem is solved by introducing a statistical machine learning technique. The generation process of the stair-like V_{TH} waveform is statistically modeled and it is used as the supplement for the degenerated information. Model parameters are estimated so that the model best describes the observed V_{TH} fluctuation. The separated trap activities are finally obtained as one of the estimated parameters. With the proposed parameter extraction method, time constants and the magnitudes of V_{TH} shifts for each trap, which are mutually dependent each other, can be simultaneously estimated. Hence, the proposed extraction method can achieve the higher estimation accuracy compared to the conventional method which extracts the magnitudes of V_{TH} shifts and the time constants separately.

1.2.3 Circuit reliability analysis

Ensuring the reliability of LSIs is not satisfied only by the BTI measurement and the data analysis. A method to analyze the impact of BTI-induced device degradation on the actual circuit component is required as an elemental technology to connect the device modeling community and the circuit design community.

This dissertation specifically focuses on the reliability issue of static random access memory (SRAM). Due to the data-storage usage of an SRAM cell, the switching rate of the transistors used in an SRAM cell is usually lower than that of the transistors in combinational circuits. Hence, transistors used in an SRAM cell experience a long stress time, which accelerates the BTI-induced degradation. Moreover, heat dissipated by other circuit components further complicates the problem. Instruction dispatcher or reorder buffer is equipped with large logic circuits to realize out-of-order executions. Due to large circuit and high switching activities, these components produce a large amount of heat that spreads to nearby components such as register files or first level instruction caches. Therefore, circuit designers must be extremely careful to design an SRAM cell and its layout to ensure the circuit reliability.

In order to analyze the impact of BTI on the SRAM cell, the BTI-induced temporal V_{TH} shift must be taken into consideration. Conventionally, multiple reliability analyses with slightly different chip ages are required to see the time-development of the SRAM cell stability, which consumes a lot of computational resources. In this dissertation, an sequential approach based on particle filter is proposed. Particle filter is first proposed in the statistical community [9, 10] for a sequential approximation of a non-Gaussian distribution. One of the most

successful application of the particle filter is the object tracking in the image processing [11]. Particle positions are iteratively updated so that each particle best approximates the position of the target object. Once the particles are allocated near the position of the target object, the temporal change in its position can be sequentially tracked by searching the only limited region, i.e. the region near the current position of the particles.

In this dissertation, this characteristics of particle filter – it can sequentially track a temporally changing variables – is utilized. Specifically, the temporal change in the SRAM cell stability due to BTI is tracked using the particles. In order to take into account the impact of BTI with conventional methods, the exploration of the variability space has to be repeated at each aging time step by changing parameters of device aging. With the aid of the particle filter, the proposed method can be conducted in one simulation run and hence the total calculation time required to capture the temporal change of the cell stability is greatly reduced. With the aid of the proposed method, circuit designers can predict how their designs suffer from BTI-induced degradation, thus contributing to enhance the circuit reliability.

1.3 Organization

The organization of this dissertation is as follows. In Chapter 2, backgrounds that forms the basis of this dissertation and reviews of related works are provided to make the contribution of this dissertation clear. The reviews of three sub-problems are provided, respectively. The contributions of this dissertation are summarized at the end of this chapter.

Chapters 3 and 4 correspond to BTIarray and its measurement results. The concept and example implementation of BTIarray are described in Chapter 3. The performances of the fabricated BTIarray, such as the measurement precision, are also evaluated. Then, in Chapter 4, BTI-induced device degradations observed on a large amount of transistors are provided. Discussions about the experimental results are given at the end of Chapter 4.

In Chapter 5, the automated parameter extraction method is described. First, the problem in the automated parameter analysis is again summarized. The overview of the proposed method and the details of the statistical parameter estimation methodology that includes Markov chain Monte Carlo are also provided. Then, using the artificially synthesized V_{TH} sequences, the performance of the method is analyzed. Finally, the parameters extracted from the V_{TH} measured on BTIarray is provided.

Chapter 6 describes the SRAM yield estimation considering the BTI-induced device degradation. Backgrounds that form the basis of the proposed yield

estimation method are first described, which include particle filter and support vector machine. Then, details of the proposed method and the experimental results are provided.

In Chapter 7, summary of this dissertation and some remaining problems are discussed.

Chapter 2

Backgrounds and Literature Review

This chapter provides backgrounds and reviews of the conventional work to make clear the contributions in this dissertation. Basics of BTI and RTN-induced temporal change in the V_{TH} will be provided in Section 2.1. Then, literature reviews on each separated topics will be given in Section 2.2. Finally, Section 2.3 summarizes the contributions of this dissertation.

2.1 Basics of BTI and RTN

2.1.1 BTI-induced degradation and its physics

NBTI and PBTI are the gradual shift of V_{TH} observed in pMOS and nMOS transistors, respectively. The phenomena are promoted by strong vertical electrical field applied to the gate insulator and the elevated temperature. The degraded V_{TH} partially recovers when the electrical field is removed, which makes BTI-induced degradation unique from the other degradation modes. Due to this recovery effect, a prediction model of BTI-induced degradation naturally becomes time-dependent. Hence, prediction of the future V_{TH} shift given device usage scenario has been found to be a difficult task.

When a negative bias (logic "0") is applied to the gate terminal of the pMOS transistor, its V_{TH} starts to increase gradually, which is known to be NBTI. The increased V_{TH} recovers as soon as the transistor is released from the stress condition. However, there is an unrecoverable component in the degraded V_{TH}. Hence, repeatedly applying stress will gradually increase V_{TH} of the pMOS transistor and will finally cause a malfunction of the circuit. Recently, the degradation on nMOS transistors (PBTI) also attracts an increasing concern.

In order to reduce the leakage currents, recent technology nodes adopt new materials such as high-k gate insulators and metal gates, which is considered to be the origin of the PBTI-induced degradation.

In spite of the intensive researches on NBTI, its physical mechanism is still a controversial topic. Currently proposed physical models of NBTI are divided into two groups: one based on a reaction diffusion (RD) theory [12] and the other based on a trap detrap (TD) theory [8]. RD-theory explains the NBTI induced V_{TH} shift as follows. First, when a negative bias is applied to the gate terminal, silicon-hydrogen (Si-H) bounds at the silicon-oxide interface are broken by the vertical gate electric field and H atoms are released. The fixed positive charge is formed by the migration of the released H atoms into the gate oxide, which contributes to the V_{TH} shift. The remaining dangling bond (Si-) captures or emits an electron, which also contributes to the V_{TH} shift. Based on this model, the relationship between device age and V_{TH} shift can be written using a power law model [13]:

$$\Delta V_{\mathrm{TH}}^{\mathrm{NBTI}} = k \cdot t_{\mathrm{age}}^{n}. \tag{2.1}$$

Here, k is a model parameter which reflects the stress condition, operational temperature, and fabricated process. t_{age} is an age of a transistor. n is also a model parameter but it varies transistor to transistor [13].

Meanwhile, in small transistors, researchers noticed a stair-like recovery of V_{TH} when a transistor is released from the negative stress condition [14]. This observation leads to a TD-theory, in which pre-existing defects (traps) located inside the gate oxide film (Fig. 2.1) are considered to be the origin of the V_{TH} shift [8]. When a negative bias is applied, the defects capture electrons, causing V_{TH} to increase. The defects then release the electrons when the transistor is released from the stress condition and the degraded V_{TH} starts to recover. The following is a compact model derived from the TD-theory [13]:

$$\Delta V_{\mathrm{TH}}^{\mathrm{NBTI}} = \phi \left[A + \log(1 + C \cdot t_{\mathrm{age}}) \right]. \tag{2.2}$$

Here, A and C are parameters that reflect the device usage condition or a manufacturing process, and thus they are relatively constant for the transistors on a same chip. ϕ is also a model parameter which reflects the number of defects included in the transistor, and hence ϕ is unique to each transistor.

2.1.2 Random telegraph noise (RTN)

Random telegraph noise (RTN) is the temporal fluctuation in the V_{TH}. Among analog circuit designers, this phenomenon is also known as flicker noise which deteriorates the signal-to-noise ratio of the op-amp or other analog circuits.

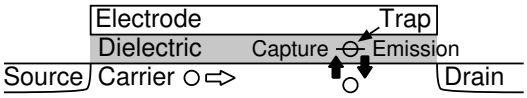


Figure 2.1: Physics assumed in the trap-detrap model.

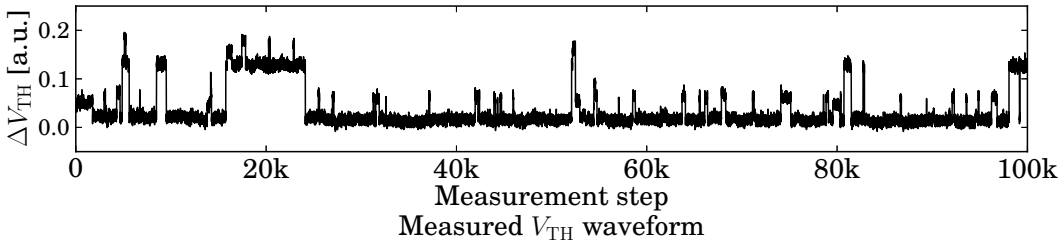


Figure 2.2: Example of RTN-induced V_{TH} shift.

Traditionally, the physical mechanism of RTN is seemingly unrelated to that of BTI because the chemical reaction assumed in the RD-model is too slow to explain the RTN-induced V_{TH} fluctuation. Recently, due to the observation of the stair-like V_{TH} recovery in the BTI measurement, many researchers now agree that RTN is originated from the same physical mechanism as BTI. When the stress is applied to the transistor, more number of defects are activated. This leads to large V_{TH} shift which is known as the BTI-induced degradation. Even if the transistor is in the recovery bias condition, defects located close to the silicon-insulator interface are still activated, causing the discrete V_{TH} fluctuation that is known as RTN. Hence, a close examination of either BTI or RTN but not both is insufficient to fully understand these degradation mechanism.

2.2 Literature review

The impact of BTI and RTN is expected to increase rapidly as transistor dimension shrinks and hence the development of design methodology, which can take into account the BTI and RTN-induced reliability degradation, has been emerged as an urgent issue. Considering this kind of background, following three problems are targeted in this dissertation: (1) how to measure BTI and RTN induced V_{TH} shift efficiently on silicon, (2) how to examine the trap activity from the observable behaviour of a transistor, and (3) how to predict the impact of BTI on the circuit component such as an SRAM cell. In order to clarify the motivations of this dissertation, literature reviews on three topics are provided in the followings.

2.2.1 BTI measurement

The measurement circuit can roughly divided into two groups: the circuit that require off-chip equipments such as parametric analyzers with source-measurement units (SMUs) and the circuits that enables on-chip measurements.

Measurements using off-chip equipment

The most straightforward approach to measure the BTI-induced V_{TH} shift is to measure I-V curves of the transistor. The advantage of this measurement is that once full I-V curve is obtained, most of the device parameters such as V_{TH} or carrier mobility can be extracted. However, acquiring a single I-V curve takes a long time because it requires multiple voltage measurement while sweeping the gate voltage. Due to stress interruption during the measurement, part of the degraded V_{TH} may recover, which leads to an unreliable measurement result.

Another approach is to use constant current method in which V_{TH} is measured as the gate-to-source voltage when a particular current flows from a source to a drain terminal [15]. Although the stress interruption is magnitudes shorter than that in the I-V curve measurement, partial recovery of V_{TH} is still unavoidable. In [16], op-amp-based current source is proposed to shorten settling time of the output voltage and to reduce the impact of the stress interruption.

On-the-fly (OTF) measurement [17] is developed to fully eliminate the impact of the stress interruption. In OTF measurement, instead of acquiring the V_{TH}, the drain current is measured while the stress is constantly applied to the transistor. The measured drain current is then converted to the V_{TH} of the transistor. The conversion model should be constructed prior to the OTF measurement. With an ultra-fast OTF (UFOTF) method [18], timing resolution of $1\,\mu$s can be achieved. In the UFOTF method, current-to-voltage converters are used to isolate the target transistor from proving PADs. Thanks to the current-to-voltage converters, the weak drain current is magnified so as to enable quick steering of the voltages at probing PADs or at an input terminal of an oscilloscope. The drawback of this method is the weak sensitivity of the drain current to the V_{TH}.

There are also various circuit structures for the characterization of a single transistor [19, 20]. For example, the circuit proposed by Matsumoto et al. [20] utilizes the leakage current of a transistor. The leakage current has a very high sensitivity to the V_{TH} and hence an accurate V_{TH} measurement can be achieved.

On-chip measurements

Various structures have been proposed for a on-chip degradation monitoring circuit [19, 21–23]. Kim et al., for example, proposed a sensor circuit that captures device degradation accurately by measuring the beat frequencies of two ring oscillators [23]. This sensor is suitable for monitoring on-chip path-delay degradation. However, ring-oscillator-based sensors are not suitable for NBTI modeling because they are composed of many transistors, and the observed degradation is the sum of the degradations caused by the transistors. Because the contribution of a single device is important in the construction of degradation models, it is impossible to use this kind of structure.

In order to focus on the BTI degradation of small subset of transistors in the ring-oscillator-based sensor, an inhomogeneous ring oscillator is proposed [24]. In the inhomogeneous ring oscillator, a pass gate transistor is introduced for connecting the input of the inverter stage of interest and the output of the preceding inverter stage. Thanks to the inhomogeneous circuit structure, the inverter of interest has significantly larger delay than the other inverter delays, making the total oscillation frequency dominated by the delay of the two transistors, i.e. the pass gate transistor and the transistor used in the subsequent inverter. Hence, the degradation of the transistors included in the single inverter stage can be measured as the degradation of the oscillation frequency.

2.2.2 Automated analysis of stair-like V_{TH} waveform

Stair-like V_{TH} shifts, such as the one shown in Fig. 2.2, are usually observed on highly miniaturized transistors. The origin of the stair-like V_{TH} shift is considered to be electrically active traps located in the gate dielectric. Physics behind the stair-like V_{TH} is now considered to be an important key for better understanding of the BTI and RTN. In an effort to help better understanding of these phenomena, some techniques were proposed to analyze the stair-like V_{TH} waveforms, which will be reviewed in this section.

Nagumo et al. introduced a time lag plot (TLP) [25] to investigate the magnitude of the V_{TH} shift. TLP is a graphical method to magnify the steps contained in the stair-like V_{TH}. A point (x,y) on TLP corresponds to the V_{TH} at current time step ($V_{\mathrm{TH}}(t)$) and that at the previous time step ($V_{\mathrm{TH}}(t-1)$). When the V_{TH} remains unchanged, the points concentrate on the diagonal of TLP. Hence, the number of traps can be extracted by counting the number of clusters on the diagonal of TLP. The authors of [26] further analyzed other parameters such as trap position by measuring devices with a single observable trap.

Miki et al. [27] and Realov et al. [28] applied a hidden Markov model (HMM)

for investigating the stair-like V_{TH} waveforms. HMM is a popular method for modeling temporal signals. One of the well known application of HMM is a speech recognition [29]. HMM can be considered to be a statistical generative model, in which the observations $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N)$ is considered to be generated by a sequence of internal states $\boldsymbol{Z} = (\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_N)$. Here, N is the length of the observed sequence. The sequence of the internal states cannot be observed directly, i.e. the sequence of states is "hidden." An emission probability $P(\boldsymbol{x}_t|\boldsymbol{z}_t)$ is assigned to each internal state. It represents the probability that $\boldsymbol{x}_t$ is observed when the internal state is $\boldsymbol{z}_t$. In the context of the analysis of the stair-like V_{TH} waveform, the internal state $\boldsymbol{Z}$ corresponds to the state of the trap, i.e. the trap is empty or occupied. For example, if three traps are involved in the formation the stair-like V_{TH}, an HMM with eight ($2^3 = 8$) internal states is required. To estimate the parameters of the HMM such as the state transition probability, an expectation-maximization (EM) algorithm [30] is used. Once the transition probability is obtained, the temporal characteristics of the stair-like V_{TH} can be extracted.

2.2.3 SRAM yield estimation considering BTI-induced degradation

Using the observations obtained from the statistical BTI measurement, a method that can estimate the impact of BTI on actual circuit components is definitely required. This dissertation specifically focuses on the impact of BTI on the SRAM cell yield and proposes an efficient yield estimation method.

Several frameworks have been proposed to analyze the impact of device degradations on circuit components. In [31] and [32], the methods that can consider both static variability (e.g. variability that originates from manufacturing process) and dynamic variability (e.g. NBTI or hot carrier injection) are proposed. In order to reduce the computational effort, the circuit response to the process variability is approximated by using a response surface model (RSM). In [33], an efficient method to examine the impact of NBTI on the stability of an SRAM cell is proposed, which is based on an assumption that a noise margin of an SRAM cell follows a normal distribution. A normal distribution provides a good approximation of the target distribution around its average. On the other hand, however, its rightmost tail is not so accurate. The failure probability required to the modern SRAM cell is extremely low and hence rightmost tail of the distribution must be analyzed very accurately. Introduction of such approximations adopted in [31–33] may lead to under or over estimation of the actual failure probability and can not be accepted.

Monte Carlo (MC) based methods are therefore required to accurately

analyze the failure probability. The naive MC is one of the most popular way to estimate the probability, in which random samples that correspond to variabilities of transistors are drawn from a probabilistic distribution and transistor level simulations are performed to see whether those variabilities cause malfunctions of the circuit or not. The failure probability is calculated by dividing the number of failure samples by the total number of generated samples. Theoretically, the naive MC can give an accurate estimation of the failure probability. However, due to an extremely small failure probability, millions or billions of circuit simulations are required to obtain a sufficient number of failed samples and hence the naive MC can not calculate the failure probability in a reasonable time. To solve this problem, importance sampling techniques are usually introduced [34–36]. The selection of the alternative distribution in importance sampling is crucial to the acceleration rate of the failure probability calculation. Because determination of a good alternative distribution requires significant computational efforts, many attempts are made to accelerate the construction process. Authors of [37] proposed a mean-shift method in which the alternative distribution is approximated with a distribution whose mean is shifted to the most probable failure point. In [38], a method known as "Markov chain Monte Carlo (MCMC)" is used to explore the process variability space efficiently. Authors of [39] utilized particles that move in the process variability space to automatically construct an alternative distribution.

2.3 Contributions of this dissertation

On the basis of the literature reviews, the contributions of this dissertation are summarized as follows.

2.3.1 BTI measurement

The proposed measurement circuit named BTIarray enables an efficient characterization of the BTI on vast amount of transistors by introducing a parallel measurement technique. The variability in BTI-induced degradation was not taken into account in the existing measurement methods. Over the past decade, we have witnessed that an extensive effort has been paid to tackle the problem originated from the variability of the transistors. BTI attracts increasing attention only in the resent years and hence, to the best of my knowledge, this dissertation is the first attempt to address the variabilities of BTI-induced degradation. Neglecting these variability in the degradation may lead to under or over estimation of the reliability of LSIs. With BTIarray, BTI degradations of hundreds of transistors can be measured within a reasonable time, which

Table 2.1: Summary of capabilities among the proposed and the existing methods.

	Number of traps	Amplitudes extraction	Time constants extraction
Proposed Method	Can be applied to more than two traps	OK	OK
TLP [25]	Equal to or less than two traps	OK	NG
TLP+HMM [27, 28]	Equal to or less than two traps	OK	OK (insufficient accuracy)

greatly contributes to the development of the statistical aging model. Circuit designers can predict the degradation of circuits more accurately by using the obtained statistical aging model, which contributes to the better circuit design.

2.3.2 Automated analysis of stair-like V_{TH} waveform

The target of the conventional methods, such as TLP or HMM, has been limited to devices that have equal to or less than two traps. Because devices having more than two traps in measurement data are common, it is important to develop a method that is applicable to model arbitrary number of traps. Furthermore, no conventional methods can separate each trap activity from the observed V_{TH} waveform. In the case of devices with a single trap, the magnitude of the V_{TH} shift can be simply estimated by calculating the distance between two peaks in the V_{TH} histogram. On the other hand, in the case of multiple traps, the analysis becomes much more complicated. When a transistor includes multiple traps, the summation of the contributions from each trap is observed as the temporal V_{TH} waveform of the transistor. In order to examine each trap activity separately, they should be reconstructed by using the measured V_{TH} sequence, which is known to be an ill-posed problem. Hence, the direct use of HMM, as in the conventional estimation methods, is inappropriate.

With the utilization of the statistical machine learning approach, the proposed method enables the analysis of the multi-trap cases for the first time. The capability of the proposed and the existing methods are summarized in Table 2.1.

2.3.3 SRAM yield estimation considering BTI-induced degradation

In order to see how the failure probability changes over time, multiple failure probability calculations are required. Because conventional methods [37–39] do not take into account the NBTI-induced device degradation, the alternative distribution is constructed from scratch at each repetitive calculation with different aging time step. However, the NBTI-induced device degradation is

a gradual process and hence the reconstruction of the alternative distribution is obviously inefficient. What is proposed in this dissertation is similar to [39] in that particles moving around the process variability space are used to construct an alternative distribution. In this dissertation, the sequential estimation technique [39] is further advanced so that the particles are "reused" among each failure probability calculation at different aging time steps, which enables the particles incrementally track the temporal change of the alternative distribution. Multiple explorations in the process variability space are thus eliminated, contributing the increase of the efficiency. The simulation experiment is conducted to compare the proposed method with the state-of-the-art method proposed in [39], which shows that the proposed method achieved 9.76× speed up.

Chapter 3

Circuit Structure Suitable for BTI-Measurement

3.1 Introduction

The objective of this chapter is to develop test structure array suitable for BTI measurement on silicon. As we saw, performance degradation caused by BTI is one of the main concerns [40] for long-term reliability of semiconductor devices. In spite of the intensive research on NBTI, its physical mechanism is still a controversial topic. However, many researchers agree that the two kinds of electrically active traps play an important role in the formation of the V_{TH} shift: one corresponds to the interface state generation and the other one corresponds to the pre-existing traps. When the negative bias is applied to the gate electrode, Si-H bounds are broken and H atoms are released. On top of the pre-existing traps, the remaining dangling bonds (Si-) contribute to the V_{TH} shift. These processes are based on discrete charges. Thus, the effect of BTI is larger in scaled technologies [41]. Contrary to NBTI, positive BTI (PBTI) is observed in nMOS transistors. Due to the application of new materials, such as high-k materials and metal gates, the impact of PBTI is expected to increase and hence it recently attracts increasing concern. PBTI is promoted by the positive bias (logic "1") applied to the gate terminal of the nMOS transistor.

Various models have been proposed to explain the mechanism of the V_{TH} increase [42], but none can completely explain the experimental results. Good experimental data is thus important to fully understand the BTI phenomenon and to improve circuit reliability.

The first concern in the BTI measurement is that the BTI-induced V_{TH} shift includes a very slow component. Even if the degradation is accelerated under high-temperature and high-stress-voltage conditions, it takes hours or

even days to observe a noticeable V_{TH} increase that is expected to be observed on several-years-old transistors operated under the normal condition.

The second concern is the variability in the BTI-induced V_{TH} shift. As transistor size shrinks, we have witnessed the increasing variability in the electrical property of transistors such as the V_{TH}. Similarly, the BTI-induced degradation is also expected to vary statistically among transistors. Statistical characterization of the degradation is thus becoming important in the scaled transistors. However, the long measurement time makes it almost impossible to collect statistically significant data on a large number of transistors through measurement. Therefore, variation of the BTI parameter values has not been fully investigated in spite of its importance.

In this chapter, a novel device-array structure, *BTIarray*, that is suitable for capturing the statistical behavior of BTI-related parameters is proposed. In this array, terminals of the device under test (DUT) are equipped with pass-gate switches to enable bias voltage selection. The switches are set so that the DUTs are placed into a predefined bias mode: stress, recovery, or measurement. The stress mode, in which the DUT is subjected to a stress bias condition, consumes most of the measurement time. Hence, in the proposed array, DUTs are stressed in parallel, which greatly reduces the time required for variability measurement.

This chapter is organized as follows. In Section 3.2, BTIarray for capturing statistical behavior of BTI-induced V_{TH} shift is presented. In Section 3.3, the implementation of the proposed circuit structure is described. The performances of BTIarray, such as the measurement precision, are examined in Section 3.4. Finally, Section 3.5 summarizes this chapter.

3.2 Concept of BTIarray

The requirements for a BTI-measurement circuit are summarized as follows.

- The terminal voltage of the transistor must be flexibly controllable to enable monitoring of the response of a device under various stress and recovery conditions in terms of time and bias voltage.

- The response of a single device must be accurately measurable under various conditions to develop a degradation model such as those in [43] or [44]. Moreover, because transistors used in latest logic LSIs are very small, the BTI-induced V_{TH} shifts are expected to vary widely. Therefore, the measurements of a large number of DUTs must be done efficiently to enhance the degradation model so that it can capture the statistical variations in the parameters [45].

- The degradation characteristics of transistors with different channel areas must be measured to enable the effects of scaling on the model parameters to be evaluated. Device size, i.e. the channel area of a device, greatly affects the temporal V_{TH} shifts. In the context of trap-detrap theory [46–49], device size is a critical parameter because it determines the expected number of interface traps in a device. It also determines the amplitude of the V_{TH} shift caused by capture and release of a charge in an interface trap.

The most simple approach to measure BTI is to place DUTs with each four terminals, i.e. gate, drain, source, and body, connected to respective probing pads. The drawback of this approach is the area efficiency because probing pads require physical contact and they are difficult to scale. The pads occupy much larger area than the DUT itself and hence area efficiency becomes even worse in scaled technologies. In order to increase the area efficiency, a device array structure [50] is adopted so that a set of probing pads can be time-shared with multiple DUTs by introducing control circuitry.

Another difficulty is the time required for BTI measurement. Because a long stress time is typically required to invoke a noticeable V_{TH} shift, it is almost impossible to serially measure BTI on a large amount of DUTs. One idea is to measure multiple DUTs in parallel. However, this measurement setup requires multiple source-measurement units (SMUs) with parametric analyzers and hence it is difficult to measure hundreds of DUTs in this setup.

Here, let us re-examine the typical BTI measurement scenario. In the BTI measurement, hours or days of stress is applied and the V_{TH} measurements are intermittently conducted. The time required for the V_{TH} measurement is usually at the order of few milliseconds and hence the most part of the time required for the BTI measurement is occupied by the stress period. Therefore, even if we could conduct multiple BTI measurements in parallel, the utilization efficiency of the SMU is quite low. This observation leads to the pipelined measurement adopted in the proposed device-array structure named *BTIarray*.

Fig. 3.1 illustrates the concept of stress pipelining. The thick lines represent stress periods in which the corresponding DUTs are in a stress bias condition. The small rectangular boxes represent measurement periods in which the V_{TH} of the corresponding DUT is measured. The terminals of each DUT are individually controlled so that one of three operation modes, *stress*, *recovery*, and *measurement*, can be applied. Measurements are carried out sequentially over the DUTs. With this array structure, a large number of DUTs can share a single set of probing pads, which is very area-efficient. Because only one parametric analyzer is necessary for measuring a large number of devices, the proposed method greatly reduces the equipment cost while significantly reducing

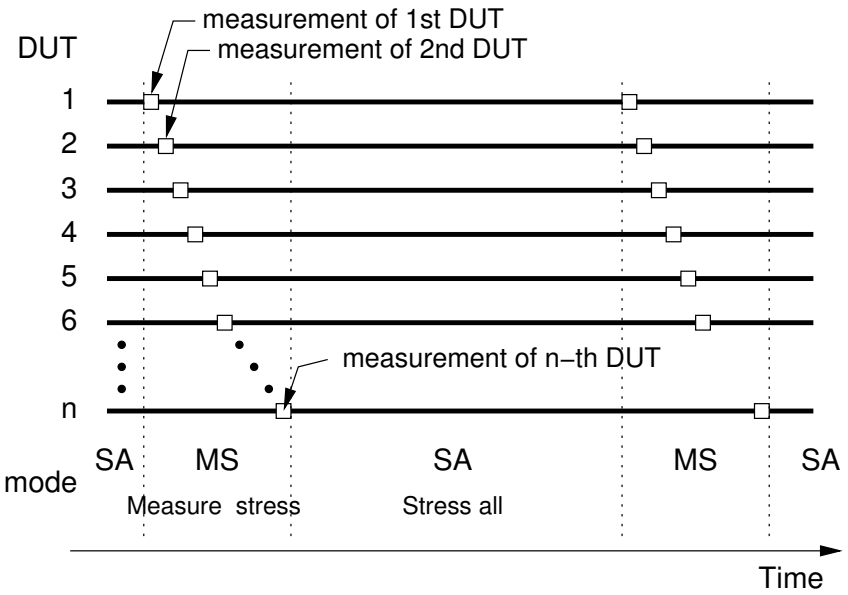


Figure 3.1: Conceptual illustration of stress pipelining. Stress periods for all DUTs are overlapped to reduce total stress time while measurements are conducted in series to simplify system.

total measurement time. With the BTIarray, which uses the concept of stress overlapping, almost n-fold reduction in the total measurement time is realized when measuring n-DUTs because each V_{TH} measurement takes a negligible amount of time, e.g. only few milliseconds, while applying stress takes hours or days.

3.3 Circuit Implementation of BTIarray

3.3.1 Circuit Structure

In order to simulate the device degradation expected in the actual environment, DUTs are required to replicate the typical bias conditions found in CMOS logic circuits. Let us take a simple inverter example shown in Fig. 3.2. When logic "0" is applied to the inverter input, the pMOS transistor becomes "ON" while the nMOS transistor becomes "OFF." Hence, the voltages at the source and drain terminals of the pMOS transistor becomes equal to the operational voltage (V_{DD}). The corresponding terminal connections of pMOS DUTs are shown in Fig. 3.2(a). Similarly, Fig. 3.2(b) shows the terminal connections of pMOS DUTs which corresponds to the situation when logic "1" is applied to the inverter input. On top of applying the stress or recovery bias voltages, it is also

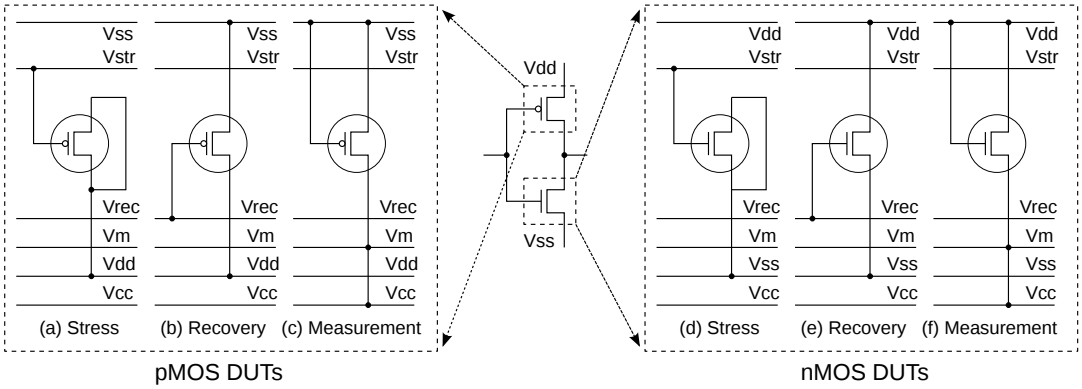


Figure 3.2: Three operational modes used in the BTI measurement.

required to measure the V_{TH} of DUTs. Hence, in the measurement mode, the constant current is forced to the source terminal of the DUT and the voltages at the source terminal is measured through V_{M} terminal as shown in Fig. 3.2(c). The operational modes for nMOS DUT are summarized in Figs. 3.2(d) to (f).

Observing Figs. 3.2(a) to (c), we notice that eight switches are sufficient to enable the on-chip mode transitions. For example, the gate terminal is connected to three different voltage supplies, i.e. V_{STR} (stress mode), V_{REC} (recovery mode), and V_{DD} (measurement mode), and hence three switches are required to control the gate terminal connection.

A simplified schematic of a pMOS-transistor DUT is shown in Fig. 3.3(a). Through pass-gate switches around the DUT, device terminals are connected to different voltages in accordance with the specified operational mode: V_{SS}, V_{STR}, V_{REC}, and V_{DD} are connected to constant voltage sources, and V_{CC} is connected to a constant current source. Fig. 3.3(b)–(d) show the switch configurations for the different modes. The control logic unit changes the configuration in accordance to the mode given from outside the chip.

For a DUT in stress mode, switches sw1, sw4, and sw6 are closed in order to apply stress bias voltages to the DUT. Here, V_{DD} is the supply voltage, and V_{STR} is the stress voltage, which is negative compared to V_{DD} when the DUT is a pMOS device. Higher stress voltages can be applied between V_{STR} and V_{DD} to accelerate degradation.

In measurement mode, the V_{TH} is measured using the constant-current method [15]. Only one DUT at a time can be in the measurement mode. Switches sw0, sw2, sw5, and sw7 are closed in this mode, which causes the measurement current to be forced to the selected DUT. The voltage at node V_{M} is measured while a constant current is applied to V_{CC}. Here, V_{SS} is zero. All the other DUTs are set to either stress or recovery mode to achieve stress overlapping. Fig. 3.3(c) shows the case when the other DUTs are in stress mode.

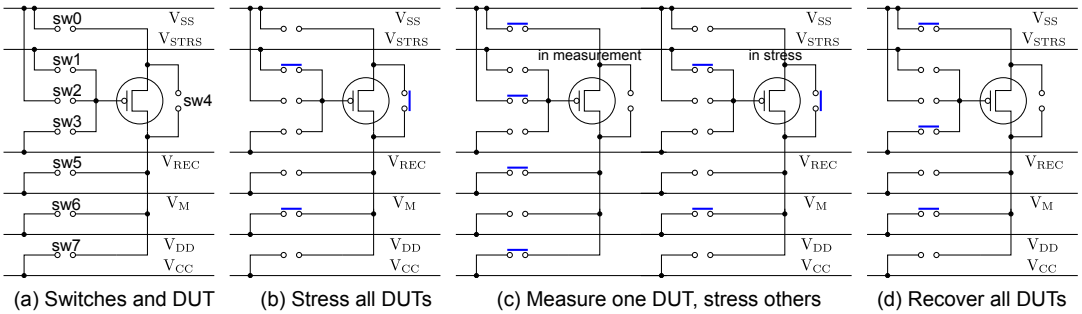


Figure 3.3: Schematic diagram of pMOS DUT unit in BTIarray.

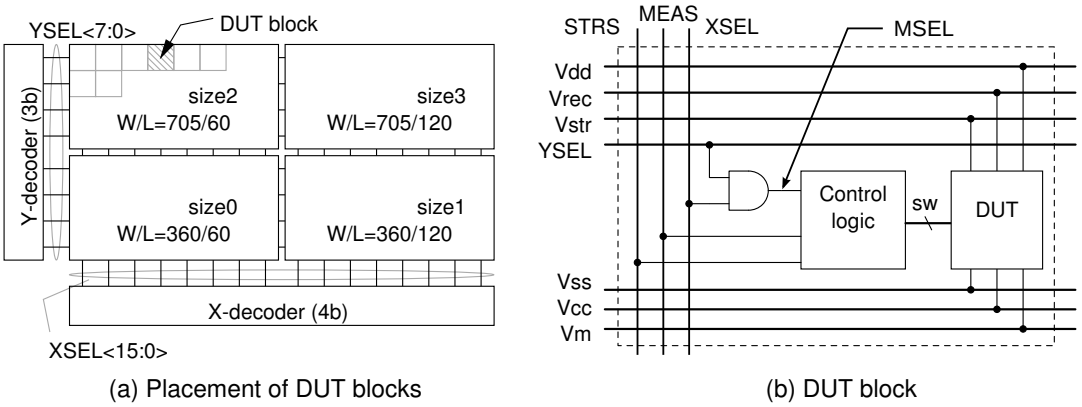


Figure 3.4: Configuration of BTIarray and DUT unit.

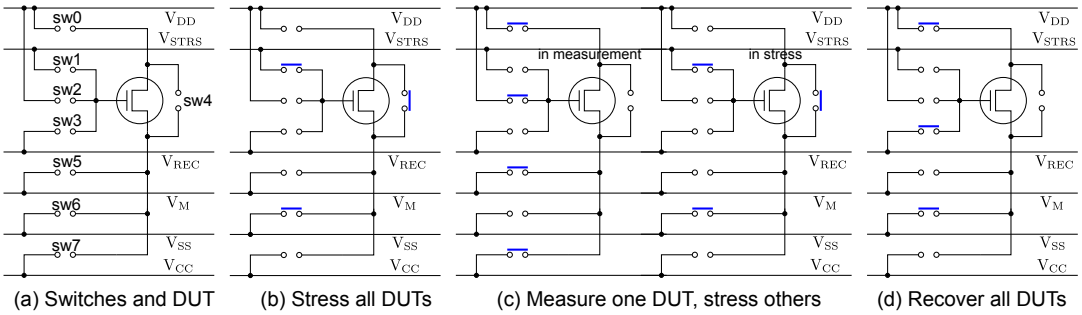


Figure 3.5: Schematic diagram of nMOS DUT unit in BTIarray.

In recovery mode, switches sw0, sw3, and sw6 are closed, as shown in Fig. 3.3(d). Here, V_{REC} is the recovery bias voltage.

Fig. 3.4(a) shows the configuration of the BTIarray. DUT blocks, shown in Fig. 3.4(b), are tiled to form the array. The DUT block, the minimum component of the array, consists of a DUT, switches, and control logic circuit.

During measurement, all voltage sources and the current source are kept constant so that the bias voltages for the DUTs are changed by switching of the

on-chip pass-gate switches. This configuration enables quick current and voltage steering due to smaller parasitic capacitances as compared to the switchings achieved by using an off-chip parametric analyzer. Hence, fast transitions between modes can be achieved. When no DUT is in the measurement mode, the current driven from V_{CC} is directed to a dummy current-drain path. The transistors for the pass-gate switches are sized relatively large to support DUT current.

Fig. 3.5 shows a schematic and the switch configurations of a DUT unit for nMOS transistors. Only a slight change is necessary, i.e. V_{DD} and V_{SS} are switched. Again, constant supply voltages are given to V_{SS}, V_{STR}, V_{REC}, and V_{DD}, and a constant current is applied to V_{CC}. Note that V_{STR} is positive compared to V_{SS} and that constant current flows from V_{DD} to V_{CC}. The operations of the pass-gate switches are the same as those of the pMOS array, so the control logic can be identical for the two types of DUTs.

3.3.2 Pass-gate Switch Design

Because we are interested in the statistical property of DUT degradation, the impact of the mismatches of any other parts of the circuit must be minimized. However, the voltage that appears at V_{M} is affected not only by the DUT itself but also by the design of pass-gate switches. Careful design of pass-gate switches is thus necessary to attain high precision.

Firstly, the impact of V_{TH} mismatch of transistors used in the pass-gate switches is investigated. Monte Carlo simulations are performed with the circuit shown in Fig. 3.3(c). Fig. 3.6 shows the variation of V_{M} in 100 k trials of Monte Carlo simulations. In order to support measurement current for the largest DUTs whose channel width is 705 nm, the pass-gate switch should be designed using a relatively large transistor. In this implementation, channel width of pMOS transistor and that of nMOS transistor are 2260 nm/1460 nm, which is three times larger than the largest DUT in terms of channel width. Studying Fig. 3.6, it can be concluded that the measurement error can be limited within $\pm 50\,\mu$V with this design.

Then, the influence of degradations of pass-gate switch is investigated. For each pass-gate switch in Fig. 3.3(c), V_{TH} of pMOS transistor is swept from 0 mV to 50 mV with a 0.1 mV step. Fig. 3.7 shows the simulated variation of voltages at V_{M}. The error bar on the "sw6L," for example, represents the voltage error at V_{M} which corresponds to V_{TH} degradation of sw6 on the left DUT in Fig. 3.3(c). Relatively large V_{TH} shift is expected for sw6 because it experiences a constant stress bias when the DUT is in the stress mode. However, according to Fig. 3.7, only a small error in the measurement results is observed, even if a very large V_{TH} shift of 50 mV is induced during the stress time. Therefore, the influence

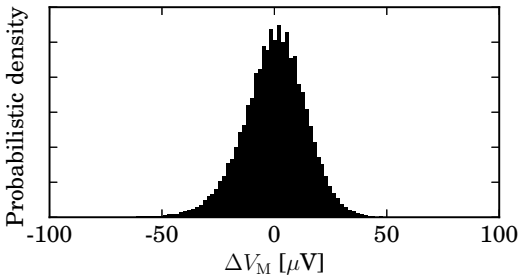


Figure 3.6: The influence of mismatches of pass-gate switches on measurement error V_{M}. The measurement error is less than $\pm 50\,\mu\mathrm{V}$.

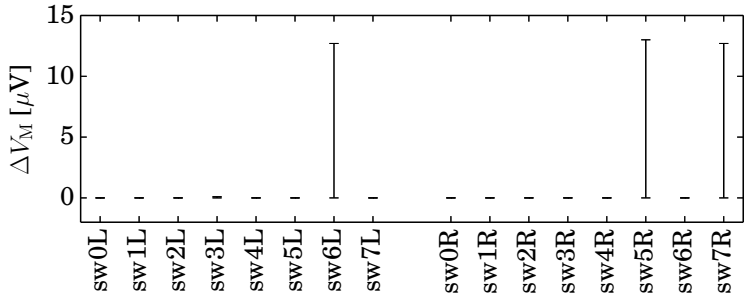


Figure 3.7: Simulated measurement error caused by the degradation of a transistor in a pass-gate switch. The range corresponds to a V_{TH} degradation of 0–50 mV.

of the degradation of pass-gate switches can safely be ignored.

3.3.3 Control Logic

The mode-switching is achieved by the control signals: "MEAS", "STRS", and "MSEL." The correspondences between the control signals and the modes are summarized in Table 3.1 as a truth table. Here, "*" indicates "don't care." A DUT is selected for the measurement only when both "MEAS" and "MSEL" are "1," i.e. the array is in the measurement mode and the DUT is selected for measurement. Otherwise, the constant bias set of either stress or recovery mode is applied to the DUT.

To realize the measurement scenario in Fig. 3.1, "STRS" is first asserted. Then, a DUT is specified by address signals and "MEAS" is asserted, which brings the selected DUT to enter into the measurement mode while constantly applying stress bias voltages to the other DUTs. V_{TH} measurements for all DUTs can be achieved by incrementing the address sequentially. After the measurement of the last DUT, "MEAS" is negated, which brings all DUTs to

Table 3.1: The truth table of control logic.

MEAS	STRS	MSEL	MODE
0	0	*	Recovery
0	1	*	Stress
1	0	0	Recovery
1	0	1	Measurement (The other DUTs are in recovery)
1	1	0	Stress
1	1	1	Measurement (The other DUTs are in stress)

be in the stress mode.

3.3.4 Test-chip Design

BTIarrays for pMOS and nMOS transistors were designed and fabricated using a 65-nm, 11-metal-layer CMOS process. The specifications of the chip are summarized in Table 3.2. Four sizes of transistors were implemented in combination with two channel lengths and two channel widths. A layout diagram of the DUT unit including the pass-gate switches and control logic circuit is shown in Fig. 3.8(a). The layout size of a unit is $10.8\,\mu$m$\times 7.4\,\mu$m. In order to maintain maximum layout regularities over the array containing four DUT sizes, the pass-gate switches are arranged so that different DUT sizes can be implemented using a single layout skeleton of the surrounding circuit. Hence, the layout designs of the switches and the control logic were identical regardless of the DUT sizes, which contributes to minimize layout-dependent variabilities. The number of DUT units having equal size is 32. The DUT units are regularly placed to form a 4×8 array. Four 4×8 arrays are tiled to form the entire array, so 128 DUTs in total are implemented in this design. The wire-routing of the core circuit was completed by using 5 metal layers or below including power distribution network, which means that BTIarray can be easily ported to most of the CMOS processes with fewer metal layers. In this implementation, an additional 6 metal layers were used for strengthening power supply network to minimize the influence of voltage drop. The layout size of the entire array is $225.2\,\mu$m$\times 62.3\,\mu$m, and the total test circuit area including the probing pads is $489.2\,\mu$m$\times 332.8\,\mu$m. A micro-photograph of the test chip is presented in Fig. 3.8(b).

The number of DUT units and layout area were the same for the pMOS and nMOS BTIarrays.

Table 3.2: Chip specifications

	pMOS Array	nMOS Array
Technology	65-nm, 11-metal layer CMOS	
Circuit area	225.2 μm×62.3 μm (DUT array), 489.2 μm×332.8 μm (incl. pads)	
Implemented DUTs (width/length)	128 DUTs	
	W/L=705/60, 705/120, 360/60, 360/120 (nm/nm), 32 each	W/L=360/60, 360/120, 180/60, 180/120 (nm/nm), 32 each
Precision @10ms/acquisition	0.07 mV rms	0.05 mV rms

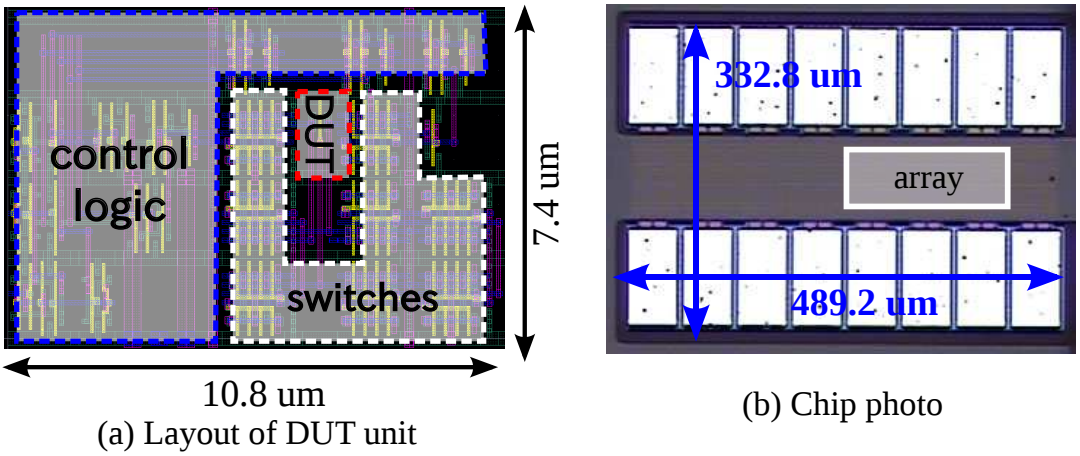


Figure 3.8: DUT unit layout and chip photograph.

3.4 Evaluation of BTIarray

3.4.1 Voltage Measurement Precision

Firstly, the precision of voltage measurements is examined through 10,000 repetitive measurements on a DUT. Example V_{TH} measurement results for a DUT on BTIarray with a fixed bias voltage at an interval of 10 ms are plotted in Fig. 3.9. They show that multi-level RTN existed in this device. Because RTN is observed in most DUTs, defining precision through repetitive measurement is difficult. Using a possible trap-free DUT having very small fluctuations, the measurement precision of BTIarray is calculated to be approximately 0.05 mV rms when the sampling frequency was 100 Hz. Because the magnitude of BTI-induced V_{TH} shift usually has the order of several millivolts, these results show that BTIarray has enough measurement precision to capture the BTI-induced V_{TH} shift.

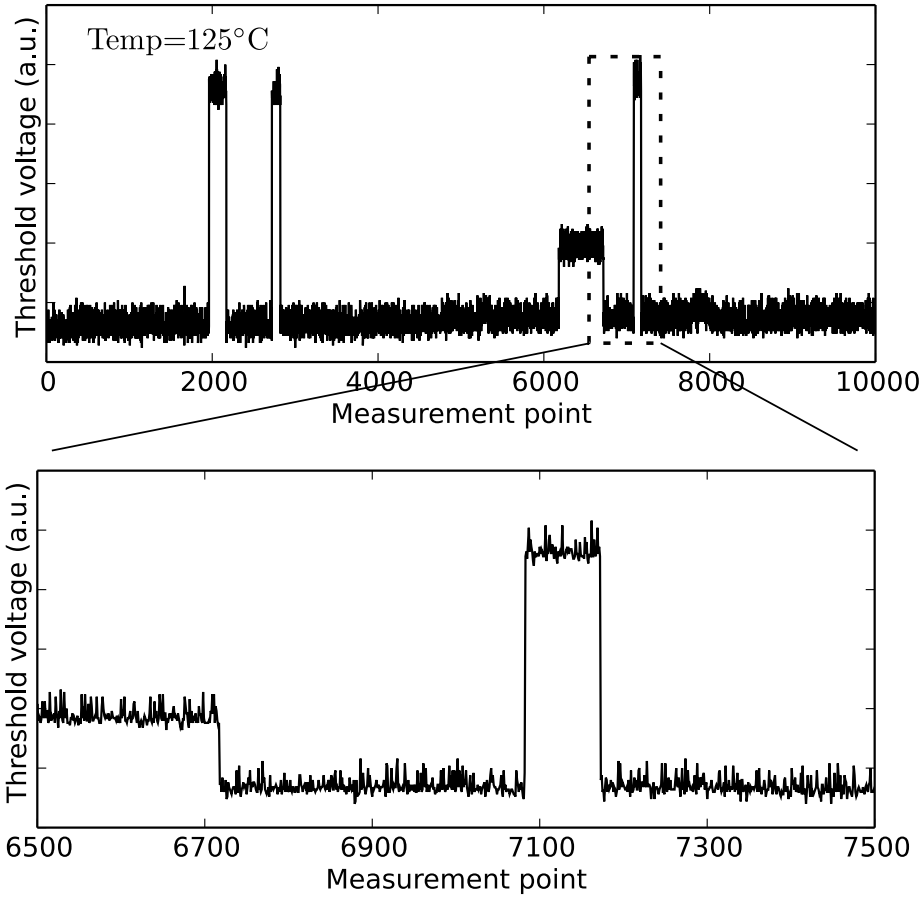
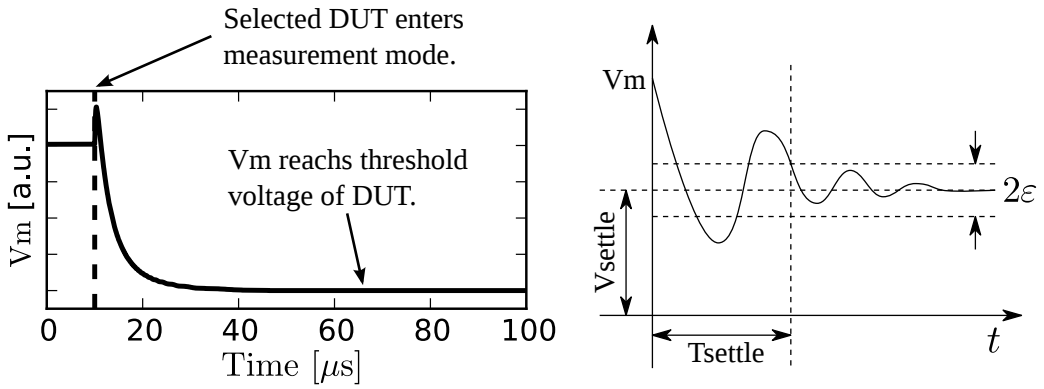


Figure 3.9: Example threshold voltage measurement result for pMOS DUT on BTIarray showing multi-level random telegraph noise (DUT size: L/W=360/60 nm).

3.4.2 Settling Time Determination

A finite time is required for the output voltage V_{M} to settle due to parasitic capacitances in the circuit and the measurement system although the current and voltage paths of the on-chip pass-gate switches can change quickly. An example transient response at V_{M} measured using an oscilloscope is shown in Fig. 3.10. The vertical dashed line at 10 μs indicates the time at which the control signal was switched from recovery to measure. From that instant, a constant current for the V_{TH} measurement was forwarded to the selected DUT. The exponential transient response of V_{M} was clearly captured. To acquire accurate V_{TH} for all DUTs, the measurements must be performed after V_{M} reaches its final voltage.



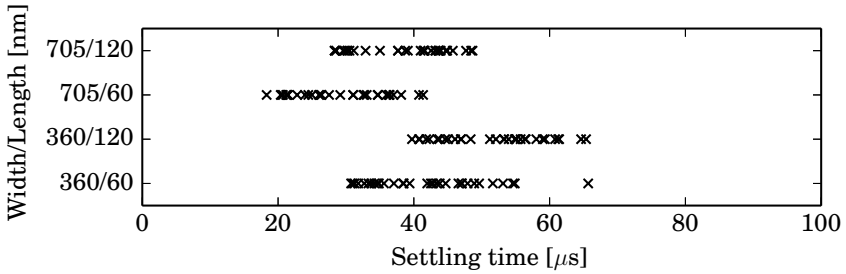
(a) Transient response of Vm (b) Definition of settling time

Figure 3.10: Example transient response of V_{M}, and definition of settling time.

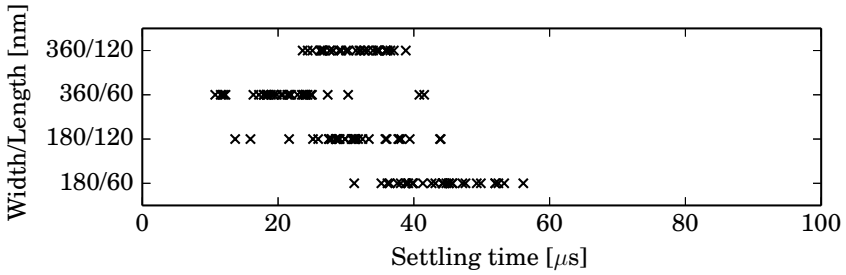
In order to determine the wait interval before measurement, the transient responses of the DUTs on the pMOS or nMOS arrays were measured. Fig. 3.11 shows the settling times measured on BTIarray. The settling time is defined as the time when V_{M} last crossed $V_{\mathrm{settle}} \pm \varepsilon$ (Fig. 3.10(b)). Here, V_{settle} is the output voltage of V_{M} after a sufficiently long time (such as 90 μs) has passed since the DUT entered the measurement mode. Margin ε was set to 0.5 mV. For almost all DUTs in the two arrays, the settling time was distributed from 10 to 60 μs. Therefore, for fabricated 128-DUT arrays, the minimum interval of time between the assertion of the measurement signal and the measurement of V_{M} can be safely set to as short as 100 μs.

3.4.3 Effect of Stress Interruption

Prior to moving on to the BTI measurement, the influence of stress interruption caused by V_{TH} measurement is evaluated. To see the influence, DUTs on BTIarray are divided into two groups labeled as "Group 1" and "Group 2." The DUTs in both groups are stressed for 20k s. DUTs in "Group 1" experience V_{TH} measurements for 16 times, while DUTs in "Group 2" experience only one measurement at the end of a 20k s of stress period. Fig. 3.12 shows the observed V_{TH} shifts at the end of stress period for four DUT sizes (W/L=360 nm/60 nm, 705 nm/60 nm, 360 nm/120 nm, and 705 nm/120 nm). The V_{TH} shifts of the DUTs in "Group 1" are indicated by cross symbols and those in "Group 2" are indicated by square symbols. From Fig. 3.12, we see very small differences between the V_{TH} shifts of two groups although "Group 1" has experienced greater number of partial recoveries. Hence, it can be concluded that the influence of stress interruption is small enough for the measurement of long term V_{TH} shift.



(a) pMOS



(b) nMOS

Figure 3.11: Distribution of settling times among DUTs on pMOS and nMOS arrays.

Further, the impact of stress interruption period is examined in detail. The duration in which constant current is forced to DUTs, i.e. the period during which "MEAS" is asserted (aperture time), is altered and the V_{TH} increases are measured by applying 5 k seconds of BTI-stress. Fig. 3.13 shows the result observed on the largest DUTs (W/L=705 nm/120 nm). Note that the V_{TH} shifts averaged over 32 DUTs are shown. Studying Fig. 3.13, we notice that the magnitude of V_{TH} shifts decreases as the aperture time increases. This phenomenon may be explained by the partial recovery due to the stress interruption. We also notice that the final V_{TH} shift at 5 k seconds observed with the 5 milliseconds of aperture time are substantially lower than the shorter aperture time configurations. Hence, it can be concluded that the aperture time should be less than 3 milliseconds in order to reduce the partial recovery during measurement. Considering these observations, the aperture time is safely set to 1 milliseconds in the BTI-measurements described in the next chapter.

3.4.4 BTI-induced V_{TH} shifts

Finally, the BTI-induced V_{TH} shifts are measured on 128 DUTs. Fig. 3.14 shows the V_{TH} increases measured on pMOS BTIarray. First, initial V_{TH} are

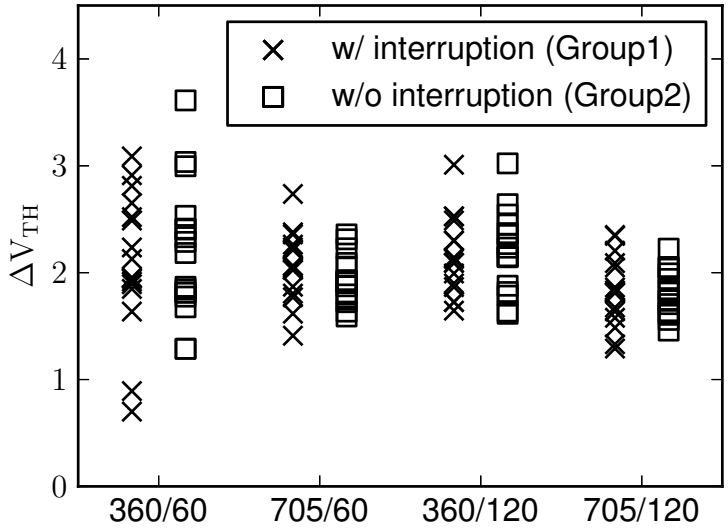


Figure 3.12: The influence of stress-interruptions due to V_{TH} measurements. The V_{TH} shift on DUTs with stress interruptions by V_{TH} measurements ("Group 1"; cross symbols) and DUTs which are stressed without interruption ("Group 2"; square symbols), both for 20k s in total, are presented.

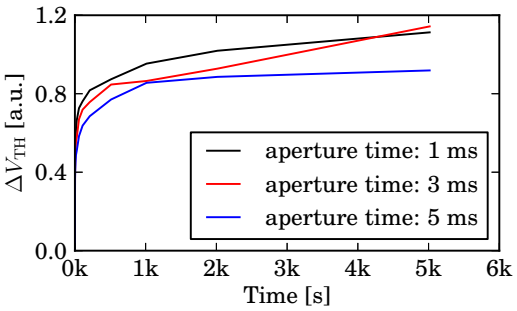


Figure 3.13: The impact of stress interruption period. The V_{TH} shift are measured by changing the aperture time from 1 milliseconds to 5 milliseconds.

measured for all DUTs. Then, negative bias voltages are applied to the DUTs for 5 k seconds and the V_{TH} measurements are intermittently conducted for 12 times during the stress period. After the application of the stress, the bias voltage is switched to the recovery mode and the extra 1,000 measurements are conducted to observe the response during recovery. As it will be examined in the next chapter, the relationship between the channel area and the variations in the V_{TH} shifts are clearly observed: i.e. the V_{TH} shifts vary widely in smaller DUTs.

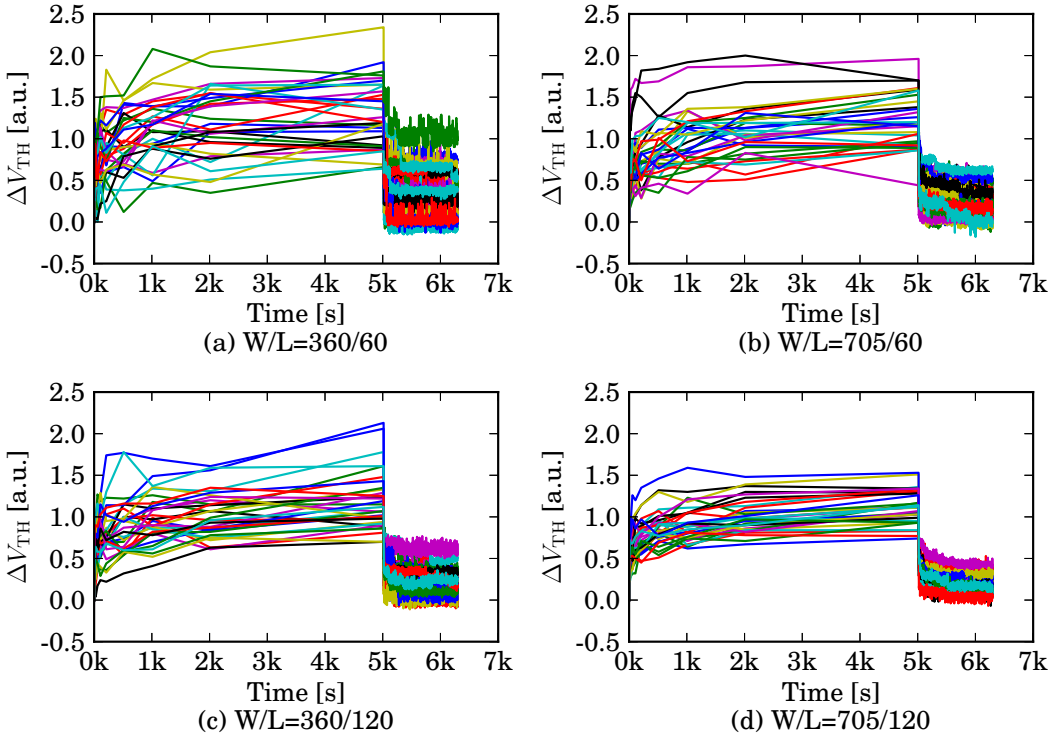


Figure 3.14: Examples of V_{TH} increases observed on pMOS BTIarray.

In this experiment, the measurement for 128 DUTs took about 1 hour and 45 minutes. If single-DUT measurement was serially conducted for 128 DUTs, measurement would take about 178 hours, or 7.4 days (not including the time required for moving the probe needles to the next DUT). BTIarray thus achieved 100× speedup with this scenario, enabling month-long measurements to be completed within a day. BTIarray did not achieve 128x speedup due to the time required for V_{TH} measurements. With the parametric analyzer used for the experiments, voltage acquisition takes 10 ms. Therefore, the acquisitions of the V_{TH} for 128 DUTs for 1 013 measurements takes about 21.6 minutes. According to the settling time measurement in Section 3.4, if an SMU with faster sampling rate is available, the measurement time can be safely reduced to 100 μs. If this was done, acquiring the V_{TH} of 128 DUTs for 1 018 times would take only 13 s, which further accelerates the BTI measurement.

3.5 Summary

In this chapter, BTIarray is proposed for efficient measurements of the temporal V_{TH} increases caused by bias temperature instability. BTIarray is designed so that the stress or recovery periods of all devices in the array are temporally overlapped, i.e. the stressing and measuring are conducted in a pipeline manner for all DUTs. The V_{TH} measurements can be carried out in series while other devices are in stress or recovery mode. With the proposed array, statistical characterization of the device degradation related to bias temperature instability are significantly accelerated.

The proposed concepts of stress overlapping and pipelined measurement were implemented on silicon using a 65 nm CMOS processes. Through the evaluation of BTIarray, it was found that the fabricated BTIarray had enough measurement accuracy to capture the BTI-induced V_{TH} shift. It was also confirmed that the impact of stress interruption due to measurement is sufficiently small for measuring the long term V_{TH} shift. In the advanced technology nodes, random mismatch among transistors has been emerged as an great concern. Statistical measurement of device degradation is thus an important challenge. BTIarray may significantly contribute not only to good understanding of the physical mechanism behind the degradation but also to construct an accurate degradation models.

Chapter 4

Measured Variability in BTI-Induced Degradation

4.1 Introduction

In this chapter, the results of BTI measurements using the fabricated BTIarray are presented. In order to invoke noticeable V_{TH} shifts, hours or days of stress application is required, which had been a barrier for the statistical V_{TH} measurement. BTIarray drastically shortens the BTI measurement on a large amount of transistors by introducing the pipelined measurement technique, which was detailed in the previous chapter.

Prior to measure the variability in the device degradation, the measurement environment which is aimed for the BTI measurement should be constructed. The requirements for the measurement environment are summarized as follows. First, the timings of control signals, such as "MEAS" or "STRS," should be accurately controlled. Because the BTI-induced degradation has been affected by the device usage history, timing variation of control signal and that of measurement timing are expected to have large impact on the measurement result. Second, in order to enable BTI measurement under various conditions, measurement scenarios, such as the stress or recovery period, should be easily modified.

Considering these requirements, an FPGA-based pattern generator and a scripting language named *BTIscript* are developed. BTIscript, which is based on Python scripting language, is firstly converted into a binary pattern file that defines the transition timings of the control signals and into GPIB commands for the measurement equipment. The binary pattern file is then transferred to the FPGA-based pattern generator that controls the measurement timing or mode switchings. This setup makes it easy to try various measurement scenario.

This chapter is organized as follows. In Section 4.2, the details of the measurement environment are presented. The accuracy of the timing control is also presented. Then, in Section 4.4, experimental results are provided. Discussions on the observed results are provided in Section 4.5. Finally, Section 4.6 summarizes this chapter.

4.2 Measurement Environment

Mode switching commands have to be continuously issued during BTIarray measurement. The order and timing of the commands strongly affects the measurement results. Hence, the automations of the parameter setup and of the operation of the measurement instruments are necessary. In addition, because we are interested in collecting statistical degradation results for a large number of DUTs, variation induced by the measurement environment should be minimized. Among many sources of the variation, timing jitter of the control signals have the largest impact on the measurement result because BTI-induced V_{TH} shift severely depends on the stress period. Although it is short, the period in which DUT is in the measurement mode also affects the measurement result because stress interruption leads to a partial recovery of the degraded V_{TH}. Therefore, minimizing timing variation of control signal and that of measurement timing are important for achieving reliable measurements.

Considering these constraints, an FPGA-based pattern generator and a scripting language interface are developed. The instrument setup and signal flow of the measurement environment are illustrated in Fig. 4.1.

4.2.1 Scripting Language to Define Measurement Scenario

The procedural steps that constitute typical measurement scenarios are defined by a scripting language, BTIscript. The measurement scenario is written using the macro commands of BTIscript, which is defined using Python scripting language. The use of Python as a macro-processing environment means that all kinds of powerful Python features, such as predefined and user-defined classes, subroutine calls, and flow controls, can be used to write the measurement procedures. This makes it easy to efficiently develop and debug measurement scripts. Without this kind of environment, lengthy, complicated, and thus error-prone commands would have to be hand-generated, leading to unreliable results. Macro commands covering typical measurement procedures for BTI degradation are listed in Table 4.1. The BTIscript is converted into a pattern

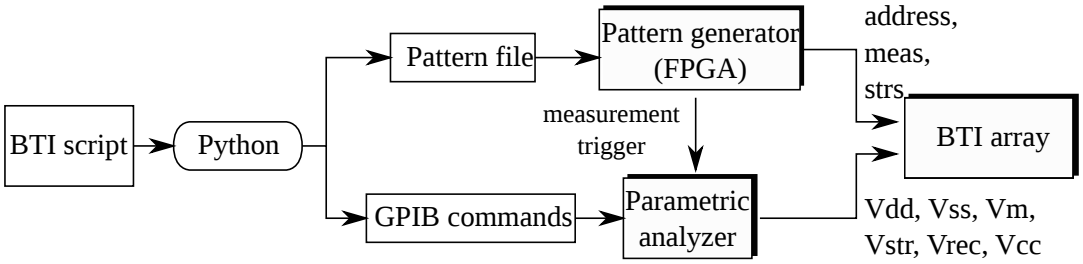


Figure 4.1: Measurement setup. Macro script is converted into a binary pattern file that defines both control signals and their output duration. Control signals and measurement trigger are generated by a pattern generator implemented on an FPGA board in order to achieve precise control of measurement timing.

Table 4.1: Subset of macro commands

Macro	Description
stress_all(*duration*)	All DUTs are biased to stress for *duration* second.
recovery_all(*duration*)	All DUTs are biased to recovery for *duration* second.
measure_stress_all(*burst*)	Measure all DUTs in ascending order. Measurement of a DUT is conducted by *burst* times in series while all other DUTs are biased to stress.
measure_stress_one(*adrs, burst*)	Measure single DUT specified by address *adrs*. Measurement of a DUT is conducted by *burst* times in series while all other DUTs are biased to stress.
measure_recovery_all(*burst*)	Measure all DUTs in ascending order. Measurement of a DUT is conducted by *burst* times in series while all other DUTs are biased to recover.
measure_recovery_one(*adrs, burst*)	Measure single DUT specified by address *adrs*. Measurement of a DUT is conducted by *burst* times in series while all other DUTs are biased to recovery.

file that defines the transition timings of the control signals and into GPIB commands that control the parametric analyzer.

4.2.2 FPGA-based Pattern Generator

Fig. 4.2 shows a block diagram of the pattern generator. The generated pattern file is transferred to the FPGA via a USB interface. Each atomic command is formatted in a 64-bit command word. The upper 40 bits and lower 24 bits correspond to the count value and the state of the control signals, respectively. The count value determines the duration for which the control signals are applied. It is used as an initial value of the 40-bit count-down counter. When

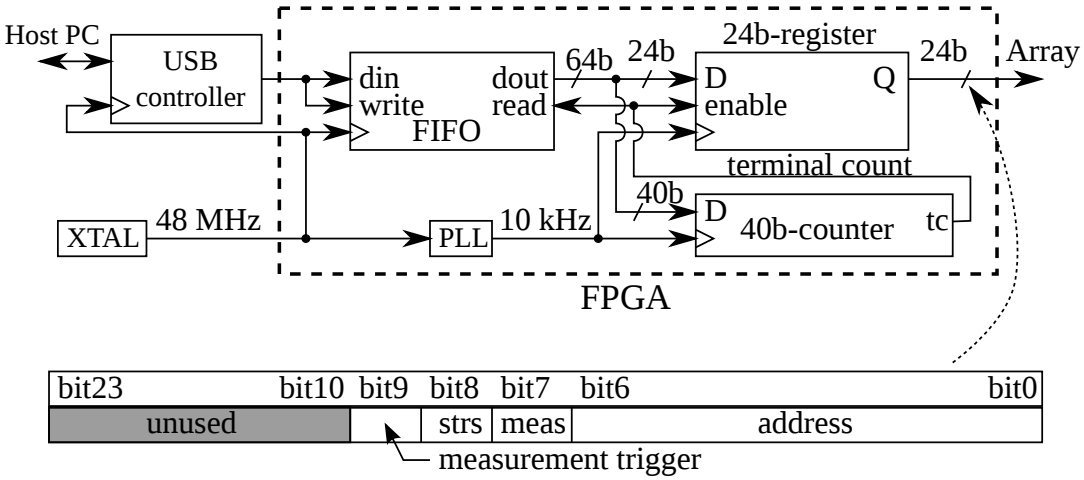


Figure 4.2: Block diagram of developed pattern generator. Sequences of bit patterns and output timings are transferred from host PC via USB. Every time the counter reaches zero, next command word is fetched from the top of the FIFO queue to reload the counter and the output register.

the counter reaches zero, the next command word is fetched from the FIFO queue to reload the counter and the output register.

The timing accuracy of measurement for this environment was evaluated. The trigger output from the parametric analyzer at the end of each measurement was monitored for the timing accuracy evaluation. Fig. 4.3 shows the result. As a comparison, the timing accuracy under a general Windows-based environment was also evaluated, in which the atomic commands were issued by a host PC running the Windows operating system. We notice that the larger fluctuation is found in the measurement timing under the Windows-based environment. The standard deviation of the measurement intervals for the Windows-based environment was about 452 μs. It was reduced to 19.0 μs by using the FPGA-based dedicated pulse generator. Note again that in this experiment, the trigger output from the semiconductor parametric analyzer was monitored. Hence, due to the delay of the measurement equipments in responding to the trigger input, there still remains the timing error of about 19.0 μs. The timing jitter of the trigger input is much more small with the timing error of sub-microseconds or smaller.

The constant voltages and currents given to the BTIarray during the measurement are summarized as follows. Constant voltages of 0.0 V and 1.8 V were used for V_{SS} and V_{DD}, respectively. For the pMOS array, stress voltage (V_{STR}) and recovery voltage (V_{REC}) were set to 0.0 V and 1.8 V, so an accelerated negative bias of 1.8 V was applied in the stress period whereas the nominal

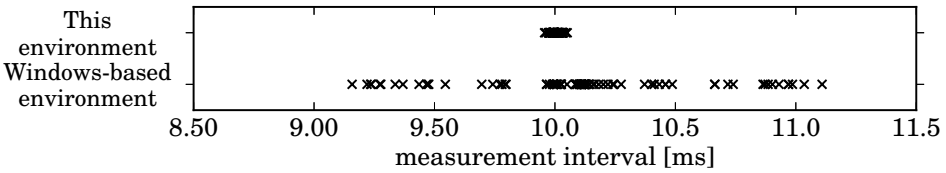


Figure 4.3: Comparison of measurement intervals. Target interval was 10 ms. Standard deviation of measurement interval was 19.0 μs in this environment and 452 μs in the Windows-based environment.

voltage of the DUT was 1.2 V. For the nMOS array, the voltage settings of V_{STR} and V_{REC} is changed, i.e. V_{STR} and V_{REC} are set to 1.8 V and 0.0 V, respectively. Hence, nMOS DUTs undergo positive stress bias. A constant current for the V_{CC} was set to 600 nA regardless of DUT size. The direction of the constant current for the nMOS array was opposite for the pMOS array.

4.3 Measurement Scenario

Fig. 4.4 shows the scenario used for BTI degradation measurement, and the corresponding script is shown in Fig. 4.5. First, the output values of the voltage sources and current source are set (lines 1–12). Next, measurement timings are defined for later use (14–17). Then, the V_{TH} of all DUTs in recovery bias mode are measured to obtain V_{TH} for the DUTs before applying stress (line 28). Here, the measurement intervals are set to 10 ms, which comes from the minimum interval of the parametric analyzer. Note that, during these measurements, the DUTs not under measurement are kept in recovery bias mode to avoid degradation. After that, all DUTs enter stress bias mode (lines 29–31). Stress is continuously applied for 20k seconds. Measurements are intermittently conducted during the stress period for all DUTs in series at 1, 2, 5, 10, 20, 50, 100, 500, 1 k, 5 k, 10 k, 12 k, 15 k, 18 k and 20 k seconds. Then, all DUTs return to recovery bias mode, and measurements are conducted 1,000 times for all DUTs (lines 32–33).

The upper part of Fig. 4.4 shows the timing of the control signals used for acquiring a single measurement. First, "MEAS" signal is asserted to measure the selected DUT in the BTIarray. In accordance with the timing margin obtained in the previous chapter (Sec. 3.4), the measurement trigger is applied several micro seconds after the assertion of the "MEAS" signal. This causes the parametric analyzer or the digital multi meter to start capturing V_{M} after it has stabilized.

The V_{TH} measurement of BTIarray is based on the constant current

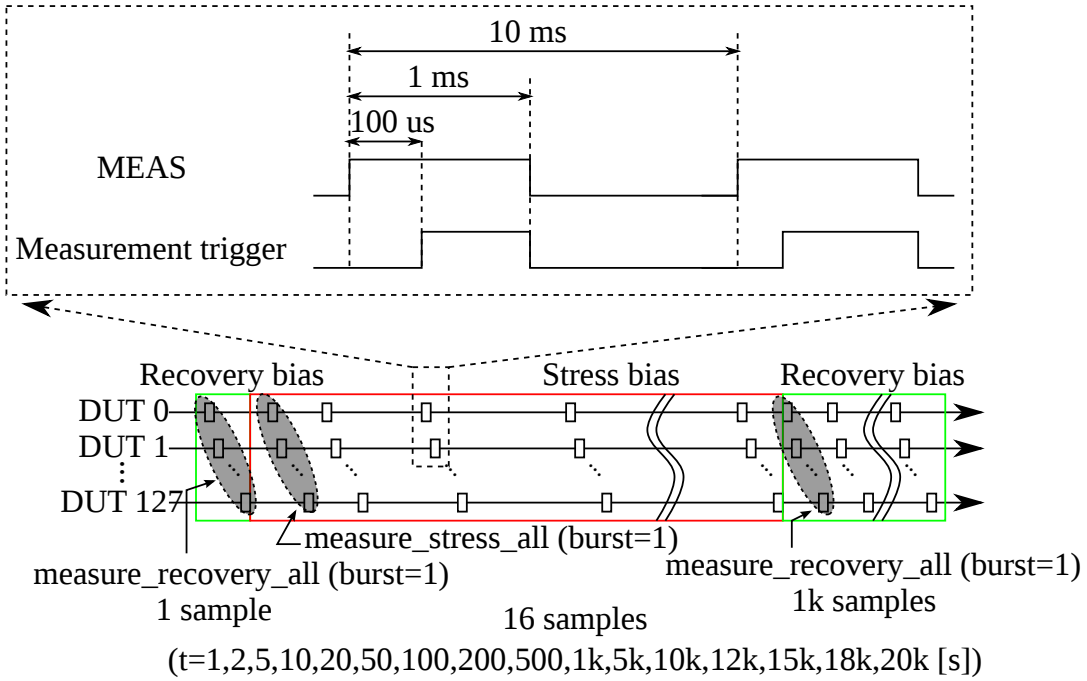


Figure 4.4: BTI degradation measurement scenario. After voltage sources and current source were initialized and measurement interval was defined, V_{TH} of DUTs in recovery bias mode were measured 1,000 times. Then, all DUTs underwent stress for 20 ks during which time V_{TH} measurements were conducted intermittently to reduce effect of interrupting stress. Finally, all DUTs returned to recovery mode, and V_{TH} were measured 1,000 times.

method [15] in which the gate voltage of the pMOS DUT is switched from V_{STR} to V_{SS} when the DUT is under measurement. The source voltage becomes V_{TH} of the selected DUT, which reduces the stress on the DUT under measurement. Because V_{TH} recovery contains a very fast component, partial recovery may occur during the measurement. To minimize this recovery, the "MEAS" signal is negated immediately after the elapse of the sampling time required by the parametric analyzer. Note again that as the evaluation in Sec. 3.4 suggests, the effect of the stress interruption can be safely ignored in the long-term BTI measurement.

4.4 Measurement Result

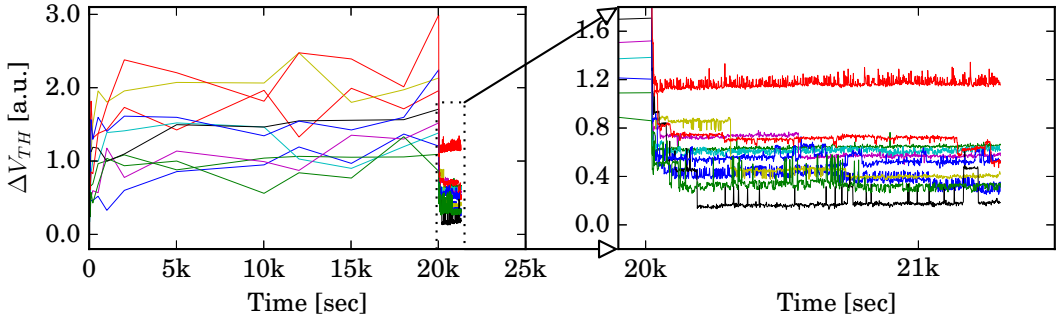
Examples of the measured V_{TH} shift on BTIarray with the scenario described in the previous section are shown in Fig. 4.6. The results for two DUT sizes are presented: W/L= 360 nm/60 nm and 705 nm/120 nm. Both instances were

```
# constant supply voltage settings
vdd = 1.8
vss = 0.0
vrec = 1.8
vstr = 0.0
# for nMOS array
# vstr = 1.8

# constant supply current setting
icc = 600e-9
# for nMOS array
# icc = -600e-9

# measurement timings
tlist = [0, 1, 2, 5, 10, 20, 50, 100, 200, 500,
         1000, 2000, 5000, 10000, 12000, 15000,
        18000, 20000]

# measurement object generation
obj = btiarray(vdd, vss, vrec, vstr, icc)

# stress period interval calculation
time_interval =
   [tlist[i+1]-tlist[i]
      for i in range(len(tlist)-1)]

# measurement scenario start
obj.measure_recovery_all(burst=1)
for t in time_interval:
    obj.stress(t)
    obj.measure_stress_all(burst=1)
for _ in range(1000):
    obj.measure_recovery_all(burst=1)
```

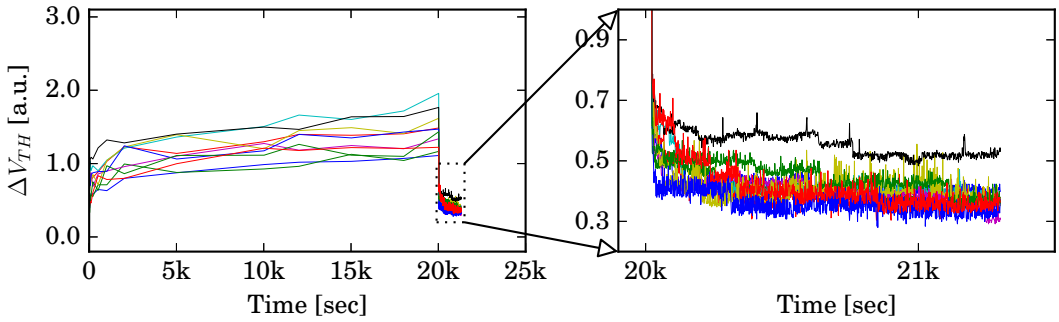
Figure 4.5: Script corresponding to measurement scenario shown in Fig. 4.4.

randomly selected from each size. In the 20k s stress period, the V_{TH} increased for all DUTs while the amounts of increases differed among DUTs. The variation in the V_{TH} increase was larger for DUTs with a smaller channel area. Because the variation is expected to increase as the transistor size decreases, statistical representations of the degradation parameters are required.

In addition, a non-monotonic increase in the V_{TH} was observed for almost all the DUTs. There are two possible causes of this non-monotonic increase; one is recovery during measurement and the other is interactions between carrier traps and emissions during the stress period. As described above, the DUTs under measurement experienced less stress voltage than the other fully stressed



(a) DUT size: W/L=360 nm/60 nm



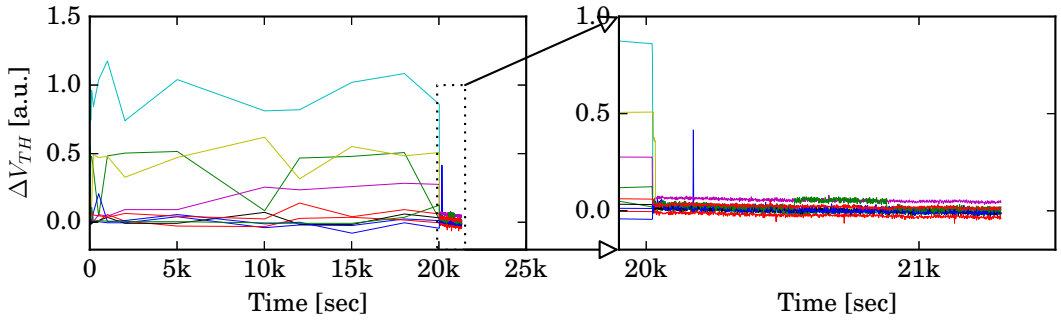
(b) DUT size: W/L=705 nm/120 nm

Figure 4.6: Temporal changes in V_{TH} (pMOS array) for ten randomly selected DUTs. Variation in V_{TH} were larger for smaller-area DUTs. In the recovery period, discrete changes in V_{TH} increases were observed for many DUTs, they were more distinct for smaller-area DUTs.

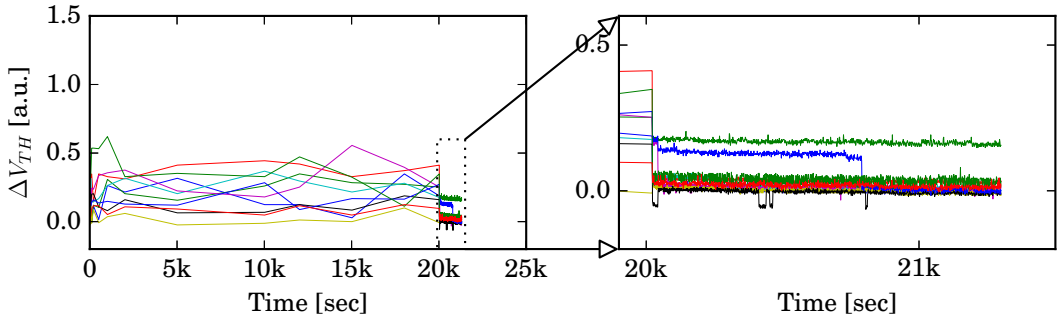
DUTs. This led to partial recovery during measurement but its effect can be considered small from the preliminary experiment in Sec. 3.4.3.

Another possible cause of the non-monotonic shift in the V_{TH} is the carrier trap and emission process at interface defects, which is considered to be the mechanism of BTI degradation [8, 51, 52]. This is a stochastic process, so the fluctuation in the V_{TH} is observed while an increase in the V_{TH} should be observed.

The time-courses of V_{TH} recovery are magnified on the right side in Fig. 4.6. Discrete changes in the V_{TH} were observed for the DUTs with a small channel area. The origin of these changes is also attributed to the carrier trap and emission process. Like in the stress period, smoother V_{TH} recovery was observed for the DUTs with a larger channel area. This observation held for the rest of the DUTs. The smooth recovery is considered to be attributed to the averaging effect. If defect densities are assumed to be uniform over all areas of



(a) DUT size: W/L=360 nm/60 nm



(b) DUT size: W/L=705 nm/120 nm

Figure 4.7: Temporal changes in threshold voltage (nMOS array), for ten randomly selected DUTs. Both the degradation and recovery occurred in a very short time.

the measured chip, more interface traps should be found in larger-channel-area DUTs than in the smaller ones, and hence trap and emission of carriers occurred more often in the larger-channel-area DUTs. Also, for the larger channel-area DUTs, the V_{TH} change caused by the trap of a single carrier is less than that for the smaller-channel-area DUTs. Therefore, a stepwise change in the V_{TH} occurred more frequently in the smaller channel-area DUTs than the larger channel-area DUTs.

Fig. 4.7 shows example V_{TH} increase for the nMOS DUTs on BTIarray measured using the same measurement scenario as for the pMOS array. A V_{TH} increase was again observed, but its magnitude was less than that for the pMOS DUTs. The increase for many of the DUTs reached a maximum relatively quickly (within 5 k seconds) and remained constant for the rest of the stress time. This degradation of the nMOS DUTs seems to have lacked "slow trap" components [53]. We also notice that the V_{TH} for both types of DUT recovered almost completely as soon as the stress was removed. These

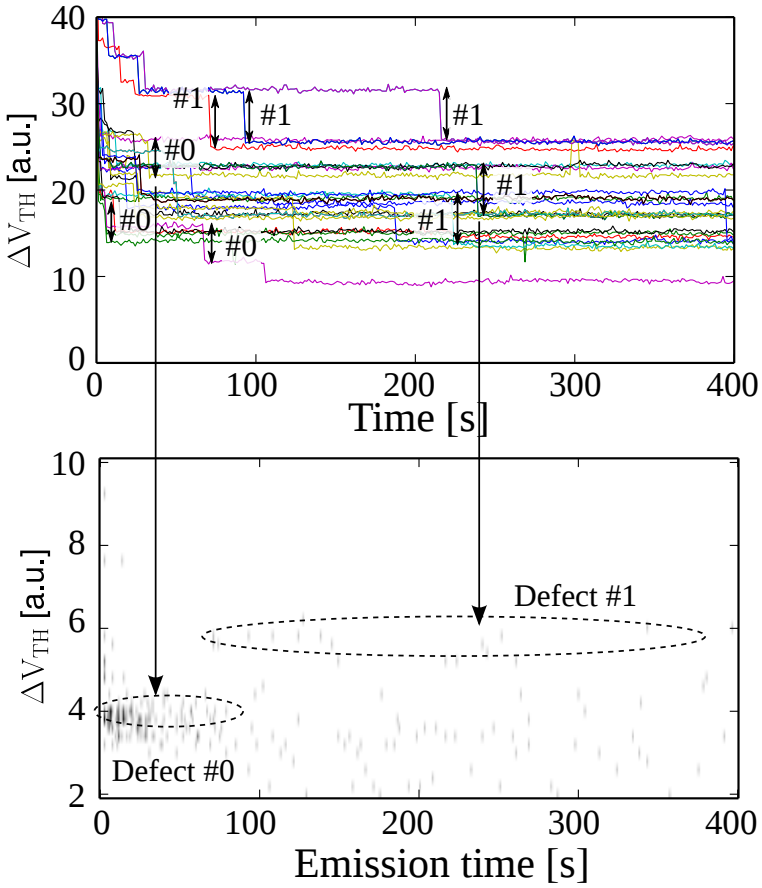


Figure 4.8: V_{TH} shift observed in repetitive stress-recovery measurements. The results of 30 trials randomly selected from 100 trials are presented. The corresponding Time Dependent Defect Spectroscopy (TDDS) plot is shown in the lower part.

phenomena are quite different from those observed for the pMOS DUTs.

4.5 Discussion

4.5.1 Stair-like V_{TH} Shift

To investigate the stair-like changes of the V_{TH} shown in Fig. 4.6(a), response to the stress and recovery bias was repeatedly measured. The DUT was first applied the stress of 1,000 s and then enters into recovery mode for 1,280s, which comprises of 1,000 measurements. This set of stress-recovery measurement was repeated for 100 times to characterize the stochastic process. The upper part of Fig. 4.8 shows the observed V_{TH} shift in the repetitive stress-recovery

measurement. The results of 30 trials randomly selected from the 100 trials are presented. We can see that there exists the same magnitude of V_{TH} shift in the results. The magnitude of V_{TH} shift is determined by the position of the defect or other random factors. Hence, the defects can be identified by their magnitude. This characterization method is known as the time dependent defect spectroscopy (TDDS) [48]. The TDDS plot is shown in the lower part of Fig. 4.8. Two major defect-clusters can be found. The emission time of Defect #0 takes the value around 10 to 50 seconds while that of Defect #1 widely distributes in the range of 100 to 400 seconds. In the lifetime extension method, such as the one that alternately uses one of two identical circuits [54], this wide range of emission time variation should be taken into account to determine when should the circuit starts operation after the sleep mode.

4.5.2 Statistical Model Parameter Extraction for NBTI

Statistical Distributions of the Power-law Exponent

As we have seen in Fig. 4.6, there was a large difference in the V_{TH} shift among DUTs of equal channel-area size especially when channel area of the DUTs was small. The V_{TH} increase during a stress period has been commonly modeled using a power law model in which the relationship between the stress period and V_{TH} increase is represented as $\Delta \mathrm{V_{th}} \propto t^n$, where t is the stress period and n is the power law exponent, which varies from transistor to transistor.

The distributions of the power law exponent for 2,048 DUTs measured using 16 pMOS BTIarrays (512 DUTs for each size) at 125°C are presented in Fig. 4.9. The variance in the exponent decreased as DUT size increased. All distributions are well approximated by a log-normal distribution. The approximated probability density functions are indicated by the red solid lines. The values of the two shape-parameters of the distributions, μ and σ, are also shown.

Confidence Intervals of the Extracted Parameters

Due to the non-monotonic increase in V_{TH} observed on small DUTs, the extracted power law exponents may contain estimation error. Hence, the 95 % confidence intervals of the shape-parameters (μ and σ) shown in Fig. 4.9 is investigated, the result of which is shown in Fig. 4.10. As expected, the confidence interval of the small DUTs was wider than that of the large DUTs due to their greater variation in V_{TH} shift. From Fig. 4.10, we can see that μ was approximately -2 regardless of DUT size. Note again that μ was an average of log-normal distribution which represents the logarithmic mean of

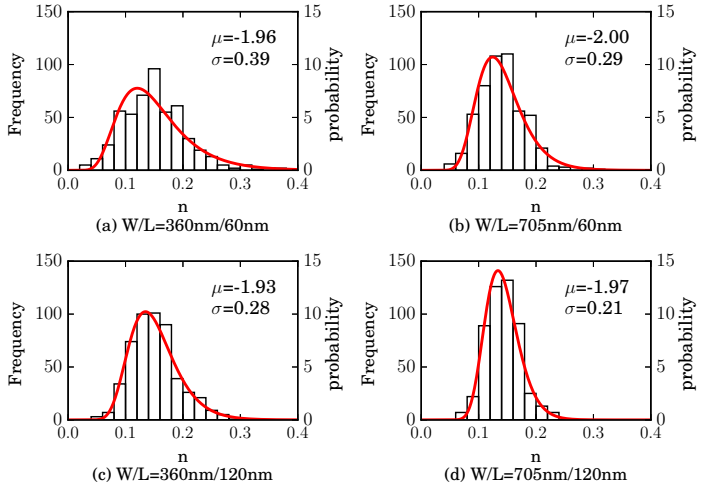


Figure 4.9: Distributions of power law exponent for DUTs of different sizes. Channel-area dependency of variation is clearly evident.

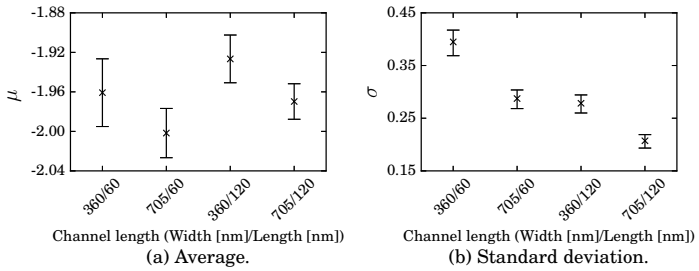


Figure 4.10: The average and the standard deviation of extracted power-law exponent for BTIarray. The 95 % confidence intervals are presented as error bars.

power-law exponent, and hence actual power-law exponents were distributed around 0.15. From Fig. 4.10, we also notice that the standard deviations of power-law exponent were about equal between the DUTs that have about the same channel-area (i.e. 705 nm/60 nm and 360 nm/120 nm DUTs), and the DUTs having smaller and larger area than those had larger and smaller variances, respectively.

As shown in Fig. 3.9, V_{TH} measurements of small DUTs are affected by RTN. This may also give influence on the accuracy of extracted power-law exponent. Here, it is quantitatively analyzed by using 1,000 V_{TH} samples $V_{\mathrm{RTN}}(t)$ while DUT is in the recovery bias. The amplitude of RTN-induced V_{TH} shift V_{A} was calculated as

$$V_{\mathrm{A}} = \max\left(V_{\mathrm{RTN}}(t)\right) - \min\left(V_{\mathrm{RTN}}(t)\right). \tag{4.1}$$

To simulate V_{TH} shift during measurement period, random samples $V_{\mathrm{NOISE}}(t)$ were generated from uniform distribution over $[-V_{\mathrm{A}}/2, V_{\mathrm{A}}/2)$. In order to simulate the temporal V_{TH} shift caused by RTN during measurement period, noise is injected to the observed V_{TH} shift as follows:

$$V_{\mathrm{SIMULATED}}(t) = V_{\mathrm{NBTI}}(t) + V_{\mathrm{RTN}}(t). \tag{4.2}$$

The power-law exponent was again calculated using $V_{\mathrm{SIMULATED}}(t)$. The above procedure is repeated for 100 times on each DUT whose channel length and width are W/L=360 nm/60 nm. As the result, it is found that the standard deviation of the extracted power-law exponent is increased by 0.029. Because the power-law exponents distributed around 0.15, the power-law exponent extracted from the smallest DUTs may have a margin of error of at most 19 %.

Correlation between Initial V_{TH} and BTI-induced V_{TH} Shift

The correlations between initial process variation and temporal degradation is finally examined. Considering such correlations, if any, in circuit designs can prevent from over design. Fig. 4.11 shows a scatter plot between V_{TH} of fresh DUTs integrated on BTIarray and their magnitudes of V_{TH} shifts during a stress period of 20,000 s. The results of the four DUT sizes are shown. From Fig. 4.11, we could see no correlation between the initial V_{TH} and their shift in this measurements.

Defects located inside the gate insulator film plays an important role in the formation of the BTI-induced V_{TH} shift while the initial V_{TH} variation is mostly caused by the variation in the positions of the dopant ions injected into the silicon substrate. Hence, initial V_{TH} variation and BTI-induced V_{TH} shift are expected to have almost no or weak correlations. However, there are some studies reporting the correlation between the initial V_{TH} and their degradation, in such as [55], and hence further investigation may be necessary.

4.6 Summary

The variability in BTI-induced device degradation was provided. Firstly, the measurement environment suitable for the BTI measurement was developed. Measurement scenario is written by Python-based scripting language named BTIscript, which drastically simplifies the development of the measurement scenario.

Fabricated BTIarrays are measured using the newly developed environment to provide new insight about the device degradation, such as the log-normally distributed power-law exponents (n). These observations can be exploited

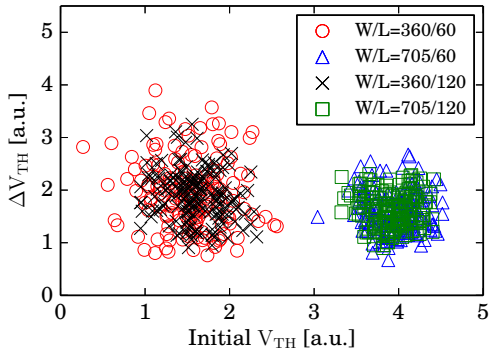


Figure 4.11: Correlations between V_{TH} of fresh DUTs and those of stressed DUTs for 20,000 s.

to predict the circuit failure probability degradation as will be described in Chapter 6.

Chapter 5

Automated Analysis of Stair-like V_{TH} Shifts

5.1 Introduction

For the good understanding of BTI and RTN, the stair-like change in V_{TH} must be carefully analyzed. For this purpose, Nagumo et al. developed a time lag plot (TLP) [25] to magnify the discrete levels of V_{TH}. TDDS [48], which was utilized in the previous chapter, is also one of the method for analyzing the discrete changes in V_{TH}. Unfortunately, however, those existing methods require a hand-extraction of vital information, such as the magnitude of V_{TH} shift caused by each trap or the time duration in which the trap captures or emits the carriers (time constants), from the analyzed result.

In the previous chapter, we saw a large variability in the V_{TH} shift caused by BTI. Hence, the characteristics of the trap such as time constants is also expected to distribute statistically. Moreover, because BTI and RTN are the stochastic process, large amount of measurement data must be analyzed even for a characterization of single transistor. However, it is obvious that the manual extraction of parameters will fail when the number of data increases. Development of an automated data analysis method is therefore an urgent issue.

In this chapter, a novel method for automated extraction of the parameters based on a machine learning method is proposed. Firstly, a statistical model that reflects the physical mechanism of the stair-like V_{TH} generation is constructed. Then, the model parameters, such as the magnitude of V_{TH} shift caused by each trap, are estimated using Markov chain Monte Carlo method (MCMC) so that the model best describes the observed V_{TH} waveform.

The proposed method has advantages over existing methods, which are again summarized as follows.

1. The extraction phase is fully automated.
2. The proposed method can cope with the complex fluctuations, i.e. V_{TH} shift caused by many traps. Specifically, V_{TH} fluctuations caused by more than two traps can be analyzed.
3. All parameters are directly estimated. No post-processing is necessary.
4. The parameters are simultaneously estimated so that estimations of the interrelated parameters become consistent.

The rest of this chapter is organized as follows. In Section 5.2, the proposed method for decomposing the statistics of traps will be explained in detail. Section 5.3 will describe the experimental validation of the method using synthetic RTN waveforms and its results. The results of a parameter extraction using measured V_{TH} waveforms and the comparison of the estimation accuracy with the conventional method are also provided. Finally, Section 5.4 summarizes this chapter.

5.2 Proposed Method

5.2.1 Problem Setting

Firstly, this section summarizes the problem setting assumed throughout this chapter. The input and the output of the proposed automated analysis method is summarized as follows.

- Input: measured V_{TH} waveform and the maximum number of traps assumed.
- Output: estimated temporal sequence of trap states, magnitudes of V_{TH} shift, and time constants of the traps.

The inputs of the proposed method are the V_{TH} as a function of time and the maximum number of traps assumed. The determination of the maximum number of traps will also be described in Section 5.2.3.

The separation of trap activities from the measured V_{TH} is an ill-posed problem. The small V_{TH} shifts caused by each trap are summed up, making the total V_{TH} shift which we can observe. To reconstruct each trap activity from the measured V_{TH} waveform, a constraint that complements the lost information is required. The proposed method utilizes the statistical model that represents the physical mechanism behind the formation of the discrete V_{TH} shift as the constraint. The model parameters such as the magnitudes of

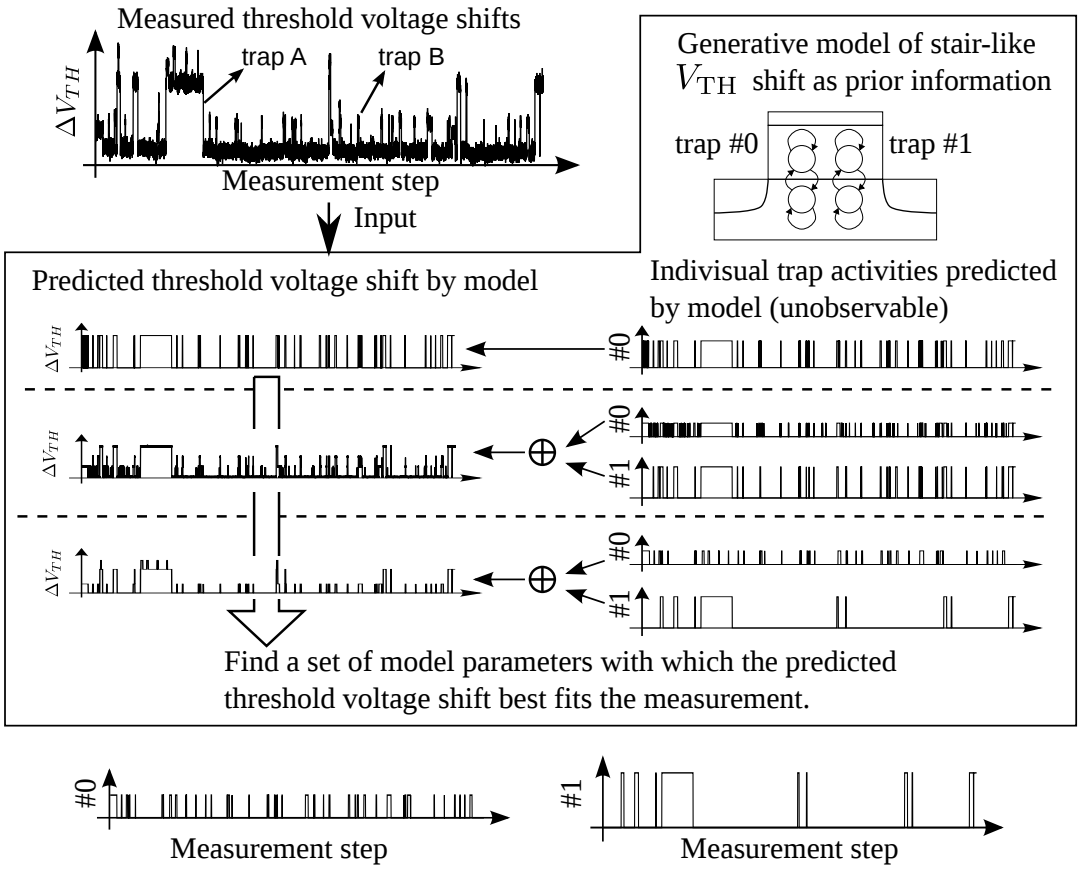


Figure 5.1: Automated analysis flow proposed in this chapter.

V_{TH} shift correspond to each trap or the trap activities are estimated so that the model prediction well fits the observed V_{TH} waveform.

Fig. 5.1 shows the automated analysis flow proposed in this chapter. The proposed method inputs the time series of measured V_{TH} waveform such as shown in the upper left of Fig. 5.1. In this particular example, the maximum number of traps assumed is set to two. The middle part of Fig. 5.1 illustrates the estimation process. Figures on the right side on the middle part of Fig. 5.1 represent the separated trap activities while those on the left side represent the reconstructed V_{TH} waveform. At the early stage of the estimation process, the proposed method tried to explain the input waveform with a single trap model. However, the input waveform is apparently composed of two components: one causes a large V_{TH} shift (trap A) and the other one causes a small V_{TH} shift (trap B) and hence the reconstruction error becomes large. Then, the method tried to fit the model with two traps. The early fitting result with the

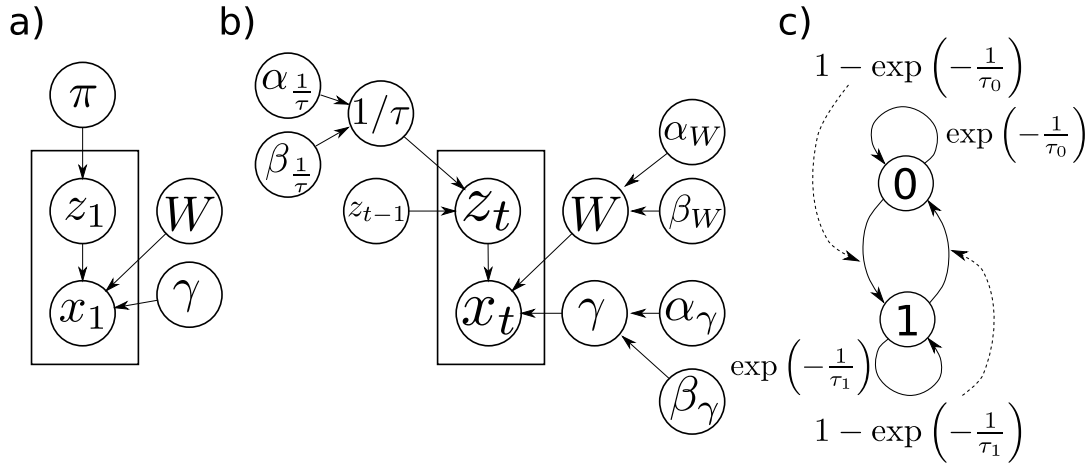


Figure 5.2: Proposed graphical model representing the generation process of the stair-like V_{TH}. a) $t = 1$, b) transition of a state from $t-1$ to t, and c) modeling of a trap.

two-trap-model is not appropriate to explain the input sequence because the separated trap activities still have high correlation, i.e. two traps capture and emit the carriers at the same time, which is unlikely to occur in the real physical system. Finally, the method successfully separates the two trap activities as shown in the bottom part of Fig. 5.1. Once we obtain the separated activities, we can calculate the time constants and magnitudes of V_{TH} shift for each trap respectively.

In the followings, the details of the statistical model used for the constraint is provided. Then, the parameter estimation method based on MCMC is described.

5.2.2 Proposed Statistical Generation Model

Fig. 5.2 shows the statistical model representation of the process from which the stair-like V_{TH} is generated. Each node represented by the circle corresponds to a random variable. The links among nodes represent the relationships among the random variables. For example, the variable $\boldsymbol{w}$ which represent the magnitudes of V_{TH} shifts depends on $\alpha_{\boldsymbol{w}}$ and $\beta_{\boldsymbol{w}}$. The trap states are represented by the binary latent variables $z_{(t,i)}$, where $\{t = 1, \cdots, N\}$ are the time steps, $\{i = 1, \cdots, K\}$ are the indices of the traps, N is the number of observations, and K is the number of traps. $X = \{x_1, x_2, \cdots, x_N\}$ is the V_{TH} at each time step.

The total V_{TH} shift is modeled as a linear summation of the V_{TH} shifts caused

by each trap as follows:

$$x_t = \sum_{i=1}^{K} x_{(t,i)} + c, \tag{5.1}$$

where $x_{(t,i)}$ is the V_{TH} shift caused by i-th trap at time t and c is the baseline of the observed V_{TH} sequence. $x_{(t,i)}$ is in turn modeled as the product of the trap state and its magnitude:

$$x_{(t,i)} = w_k \cdot z_{(t,i)}, \tag{5.2}$$

where $z_{(t,i)}$ is a binary variable representing the trap occupancy, i.e. $z_{(t,i)} = 0$ when the i-the trap is empty at time t and $z_{(t,i)} = 1$ when the trap is occupied. w_i is the magnitude of V_{TH} shift caused by the single trap. In order to handle the baseline of the V_{TH} sequence naturally, a special trap whose state $z_{(t,K+1)}$ is clamped to "1" is introduced to yield

$$x_t = \sum_{i=1}^{K+1} w_k \cdot z_{(t,i)}, \tag{5.3}$$

where $w_{K+1} = c$.

The observation error is assumed to follow a normal distribution whose standard deviation is $\sqrt{1/\gamma}$. Hence, the probability for the measured V_{TH} shift at time t (x_t) can be represented as

$$p\left(x_t \left| z_{(t,1:K+1)}, w_{1:K+1}, \gamma \right.\right) = \mathcal{N}\left(x_t \left| \sum_{i=1}^{K+1} w_i z_{(t,i)}, \sqrt{\frac{1}{\gamma}} \right.\right), \tag{5.4}$$

where $\mathcal{N}\left(x|\mu,\sigma\right)$ represents a probability density function (PDF) of the normal distribution defined as

$$\mathcal{N}\left(x|\mu,\sigma\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(x-\mu\right)^2}{2\sigma^2} \right). \tag{5.5}$$

In this dissertation, the temporal characteristics of the trap state is assumed to have Markov property, i.e. the current trap state ($z_{(t,i)}$) is fully determined by the previous trap state ($z_{(t-1,i)}$). This property is commonly assumed in many existing studies on RTN modeling [7, 56–58]. Therefore, the conditional probability distribution $p\left(z_{(t,i)}|z_{(t-1,i)}, \frac{1}{\tau}\right)$ can be written as

$$p\left(z_{(t,i)}|z_{(t-1,i)}\right) = \begin{cases} \exp\left(1/\tau_{(i,0)}\right) & z_{(t-1,i)} = 0, z_{(t,i)} = 0 \\ 1 - \exp\left(1/\tau_{(i,0)}\right) & z_{(t-1,i)} = 0, z_{(t,i)} = 1 \\ 1 - \exp\left(1/\tau_{(i,1)}\right) & z_{(t-1,i)} = 1, z_{(t,i)} = 0 \\ \exp\left(1/\tau_{(i,1)}\right) & z_{(t-1,i)} = 1, z_{(t,i)} = 1 \end{cases}. \tag{5.6}$$

The initial latent node z_1 is special because it does not depend on the previous node, so z_1 is given a marginal distribution $p\left(z_{(1,i)}|\pi_i\right)$ defined as

$$p\left(z_{(1,i)}|\pi_i\right) = \pi_i^{1-z_{(1,i)}}\left(1-\pi_i\right)^{z_{(1,i)}}. \tag{5.7}$$

Each model parameter has the following prior distribution

$$p\left(w_i\,|\theta_{w_i}\right) = \mathcal{G}\left(w_i\,|\alpha_{w_i},\beta_{w_i}\right), \tag{5.8}$$
$$p\left(1/\tau_i\,\middle|\theta_{1/\tau_i}\right) = \mathcal{G}\left(1/\tau_i\,\middle|\alpha_{1/\tau_i},\beta_{1/\tau_i}\right), \tag{5.9}$$
$$p\left(\gamma\,|\theta_\gamma\right) = \mathcal{G}\left(\gamma\,|\alpha_\gamma,\beta_\gamma\right), \tag{5.10}$$
$$p\left(\pi_i\,|\theta_\pi\right) = \mathcal{B}\left(\pi_i\,|\alpha_{\pi_i},\beta_{\pi_i}\right), \tag{5.11}$$
$$p\left(\alpha_{w_i}\,\middle|\lambda_{\alpha_{w_i}}\right) = \mathcal{L}\left(\alpha_{w_i}\,\middle|\lambda_{\alpha_{w_i}}\right), \tag{5.12}$$
$$p\left(\beta_{w_i}\,\middle|\lambda_{\beta_{w_i}}\right) = \mathcal{L}\left(\beta_{w_i}\,\middle|\lambda_{\beta_{w_i}}\right), \tag{5.13}$$
$$p\left(\alpha_{1/\tau_i}\,\middle|\lambda_{\alpha_{1/\tau_i}}\right) = \mathcal{L}\left(\alpha_{1/\tau_i}\,\middle|\lambda_{\alpha_{1/\tau_i}}\right), \tag{5.14}$$

and

$$p\left(\beta_{1/\tau_i}\,\middle|\lambda_{\beta_{1/\tau_i}}\right) = \mathcal{L}\left(\beta_{1/\tau_i}\,\middle|\lambda_{\beta_{1/\tau_i}}\right), \tag{5.15}$$

where $\mathcal{G}\left(x\,|\alpha_\mathcal{G},\beta_\mathcal{G}\right)$ $\mathcal{B}\left(x\,|\alpha_\mathcal{B},\beta_\mathcal{B}\right)$, and $\mathcal{L}\left(x\,|\lambda\right)$ are the gamma, beta, and exponential distributions, respectively. They are defined as

$$\mathcal{G}\left(x\,|\alpha,\beta\right) = \frac{1}{\Gamma(\alpha)\beta^\alpha}x^{\alpha-1}\exp\left(-\frac{x}{\beta}\right) \quad x>0, \tag{5.16}$$
$$\mathcal{B}\left(x\,|\alpha,\beta\right) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, \tag{5.17}$$

and

$$\mathcal{L}\left(x\,|\lambda\right) = \lambda\exp\left(-\lambda x\right) \quad x\geq 0. \tag{5.18}$$

Here, $B(\alpha,\beta)$ is the beta function defined as

$$B(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt. \tag{5.19}$$

5.2.3 Determination of the Number of Traps

The proposed method automatically adjusts the model complexity based on Bayesian inference. Owing to the sparse assumption placed to the prior distribution, excess traps will be degenerated, i.e. the magnitude of V_{TH} shifts

($\Delta V_{\mathrm{TH}}^{k}$), corresponding to the traps that are unnecessary to explain a given V_{TH} waveform, automatically converge around zero. By comparing the estimated magnitude of noise and the amplitude of a trap, we can determine the most appropriate number of traps. Hence, when we apply the proposed method to a temporal V_{TH} waveform having unknown number of traps, we may give a large number, such as five, as the initial number of traps. Alternatively, we can also loosely estimate the number of traps using TLP [25].

5.2.4 Parameter Estimation Algorithm

The separation of the trap activities now simplified to the problem of finding the parameters such as $\boldsymbol{w}$ and $\boldsymbol{z}$ shown in Fig. 5.2 that best describe the observed V_{TH} waveform. Mathematically, this problem is transformed to find the parameters that maximize the following posterior probability:

$$P\left(w, 1/\tau, \gamma, Z, \alpha_w, \beta_w, \alpha_{1/\tau}, \beta_{1/\tau} \,|X\right). \tag{5.20}$$

Equation (5.20) represents the probability of the model parameters given the V_{TH} waveform ($\boldsymbol{X}$). The parameters that give the higher posterior probability is considered to better describe the observed waveform. Hence, the next objective is to find a way of maximizing Eq. (5.20). However, the posterior probability of Eq. (5.20) has so complex form that it cannot be maximized analytically. Therefore, in the followings, the posterior distribution is simulated, i.e. samples are drawn from the posterior distribution, and among the generated samples, one that maximize the posterior probability is selected. Sample generation from the posterior distribution is also a challenging task because the distribution cannot be represented using well-known distributions such as Gaussian distribution. Moreover, the normalization constant for Eq. (5.20) is also difficult to obtain. Hence, Markov chain Monte Carlo (MCMC) [59] is utilized to simulate the distribution. Specifically, Gibbs sampling [60], which is a variant of MCMCs, is used.

Gibbs sampling

The purpose of Gibbs sampling is to generate a sequence of samples from a joint probability distribution of multivariate random variables [61]. Each step in the Gibbs sampling algorithm involves replacing the value of one random variable with a new sample generated from the distribution conditioned on the values of the remaining variables. Here, the procedure for applying Gibbs sampling to a Monte Carlo approximation of Eq. (5.20) is outlined. The following procedures are iterated after initialization of each random variables. In the followings, $\boldsymbol{w}^{(n)}$

indicates the samples drawn from the corresponding distribution in the n-th iteration of Gibbs sampling.

1. For $i = 1, \cdots, K+1$, sample w_i from the following conditional posterior density,
$$w_i^{(n+1)} \sim p\left(w_i \left| x_{(1:N)}, w_{(1:i-1)}^{(n+1)}, \ w_{(i+1:K+1)}^{(n)}, \boldsymbol{z}_{(1:N)}^{(n)}, \gamma^{(n)}, \alpha_{w_i}^{(n)}, \beta_{w_i}^{(n)} \right.\right). \quad (5.21)$$

2. Sample $\gamma^{(n+1)}$ from the following conditional posterior density,
$$\gamma^{(n+1)} \sim p\left(\gamma \left| x_{(1:N)}, w_{(1:K+1)}^{(n+1)}, z_{(1:N,1:K)}^{(n)}, \alpha_{\gamma}^{\text{prior}}, \beta_{\gamma}^{\text{prior}} \right.\right). \quad (5.22)$$
Note that the new sample of w_i drawn at the previous step is used.

3. Sample $z_1^{(n+1)}$ from the following posterior density,
$$z_1^{(n+1)} \sim p\left(z_1 \left| z_2^{(n)}, x_1, w_{(1:K+1)}^{(n+1)}, \gamma^{(n+1)}, 1/\tau^{(n)}, \pi^{(n)} \right.\right). \quad (5.23)$$
In this step, $w_{(1:K+1)}$ and γ are replaced with the new samples $w_{(1:K+1)}^{(n+1)}$ and $\gamma^{(n+1)}$.

4. For $i = 1, \cdots, K$, sample $\pi_i^{(n+1)}$ from the following posterior density,
$$\pi_i^{(n+1)} \sim p\left(\pi_i \left| z_{(1,k)}^{(n+1)}, \alpha_{\pi_i}^{\text{prior}}, \beta_{\pi_i}^{\text{prior}} \right.\right). \quad (5.24)$$

5. For $t = 2, \cdots, N$, sample $z_t^{(n+1)}$ from the following posterior density,
$$z_t^{(n+1)} \sim p\left(z_t \left| x_t, z_{t-1}^{(n+1)}, z_{t+1}^{(n)}, 1/\tau^{(n)}, w_{(1:K+1)}^{(n+1)}, \gamma^{(n+1)} \right.\right). \quad (5.25)$$

6. For $i = 1, \cdots, K$ and $l = 0, 1$, sample $\frac{1}{\tau_{(i,l)}}$ from the following posterior density,
$$1/\tau_{(i,l)}^{(n+1)} \sim p\left(1/\tau_{(i,l)} \left| z_{(1:N,i)}^{(n+1)}, \alpha_{1/\tau_{(i,l)}}^{(n)}, \beta_{1/\tau_{(i,l)}}^{(n)} \right.\right). \quad (5.26)$$

7. For $i = 1, \cdots, K+1$, sample $\alpha_{w_i}^{(n+1)}$ and $\beta_{w_i}^{(n+1)}$ from the following posterior densities,
$$\alpha_{w_i}^{(n+1)} \sim p\left(\alpha_{w_i} \left| w_i^{(n+1)}, \beta_{w_i}^{(n)}, \lambda_{\alpha_{w_i}} \right.\right) \quad (5.27)$$
and
$$\beta_{w_i}^{(n+1)} \sim p\left(\beta_{w_i} \left| w_i^{(n+1)}, \alpha_{w_i}^{(n+1)}, \lambda_{\beta_{w_i}} \right.\right), \quad (5.28)$$
respectively.

8. For $i = 1, \cdots, K$ and $l = 0, 1$, sample $\alpha_{1/\tau_{(i,l)}}$ and $\beta_{1/\tau_{(i,l)}}$ in the same way as sampling α_{w_i} and β_{w_i}.

Conditional posterior density

A part of developing the posterior density is derived from a paper on sound source separation using MCMC [62].

1. The posterior density corresponding to $w_i^{(n+1)}$ is

$$\begin{aligned} &p\left(w_i \left| X, w_{(1:i-1)}^{(n+1)}, w_{(i+1:K+1)}^{(n)}, z_{(1:N)}^{(n)}, \gamma^{(n)}, \alpha_{w_i}^{(n)}, \beta_{w_i}^{(n)} \right.\right) \\ &\propto p\left(x_{(1:N)} \left| z_{(1:N,1:K)}^{(n)}, w_{(1:i-1)}^{(n+1)}, w_i, w_{(i+1:K+1)}^{(n)}, \gamma^{(n)} \right.\right) p\left(w_i \left| \alpha_{w_i}, \beta_{w_i} \right.\right). \end{aligned} \tag{5.29}$$

The first term on the right-hand side of Eq. (5.29) can be written as

$$\begin{aligned} &p\left(x_{(1:N)} \left| z_{(1:N,1:K)}^{(n)}, w_{(1:i-1)}^{(n+1)}, w_i, w_{(i+1:K+1)}^{(n)}, \gamma^{(n)} \right.\right) \\ &\propto \exp\left(-\frac{\gamma_{w_i}^{\mathrm{likel}}}{2}\left(w_i - \mu_{w_i}^{\mathrm{likel}}\right)\right), \end{aligned} \tag{5.30}$$

where $\gamma_{w_i}^{\mathrm{likel}}$, $\mu_{w_i}^{\mathrm{likel}}$, and $\mathcal{E}_{(t)}^{(-i)}$ are defined as

$$\gamma_{w_i}^{\mathrm{likel}} = \gamma^{(n)}\left(\sum_{t=1}^{N} {z_{(t,i)}^{(n)}}^2\right), \tag{5.31}$$

$$\mu_{w_i}^{\mathrm{likel}} = \frac{\sum_{t=1}^{N} z_{(t,i)}^{(n)} \mathcal{E}_{(t)}^{(-i)}}{\sum_{t=1}^{N} {z_{(t,i)}^{(n)}}^2}, \text{ and} \tag{5.32}$$

$$\mathcal{E}_{(t)}^{(-i)} = x_t - \sum_{j=1}^{i-1} z_{(t,j)}^{(n)} w_j^{(n+1)} - \sum_{j=i+1}^{K+1} z_{(t,i)}^{(n)} w_j^{(n)}. \tag{5.33}$$

The second term on the right-hand side of Eq. (5.29) is the probability density function of the gamma distribution defined in Eq. (5.16). Therefore, the posterior density function can be written as

$$\begin{aligned} &p\left(w_i \left| X, w_{(1:i-1)}^{(n+1)}, w_{(i+1:K+1)}^{(n)}, z_{(1:N)}^{(n)}, \gamma^{(n)}, \alpha_{w_i}^{\mathrm{prior}}, \beta_{w_i}^{\mathrm{prior}} \right.\right) \\ &\propto {w_i}^{\alpha_{w_i}^{(n)}-1} \exp\left(-\frac{\gamma_{w_i}^{\mathrm{likel}}}{2}\left(w_i - \mu_{w_i}^{\mathrm{likel}}\right)^2 - \frac{w_i^{(n+1)}}{\beta_{w_i}^{(n)}}\right). \end{aligned} \tag{5.34}$$

The shape of this distribution is still too complex to analytically calculate the normalization constant, which requires analytical integration of the distribution. Hence, a Markov chain is again constructed using Metropolis

method [63] to sample w_i. In the Metropolis method, candidates of random samples from the target distribution are generated from the proposal distribution. The target samples are randomly accepted or rejected based on the fitness of the candidate samples to the target distribution. The sampling efficiency depends on the proposal distribution, i.e. high efficiency can be achieved when the proposal distribution is close to the target distribution. Hence, in order to improve the efficiency of sampling, the posterior distribution is approximated by a Gaussian distribution whose variance and mode is same as the unnormalized posterior distribution of Eq. (5.34). Then, a Markov chain is constructed using the approximated distribution. First, Eq. (5.34) is rewritten as follows:

$$\begin{aligned} &p\left(w_i \left| X, w_{(1:i-1)}^{(n+1)}, w_{(i+1:K+1)}^{(n)}, z_{(1:N)}^{(n)}, \gamma^{(n)}, \alpha_{w_i}^{\text{prior}}, \beta_{w_i}^{\text{prior}}\right.\right) \\ &\propto w_i^{\alpha_{w_i}^{\text{prior}}-1} \exp\left(-\frac{\left(w_i - \mu_{w_i}^{\text{post}}\right)^2 \gamma_{w_i}^{\text{post}}}{2}\right), \end{aligned} \tag{5.35}$$

where $\gamma_{w_i}^{\text{post}}$ and $\mu_{w_i}^{\text{post}}$ are defined as

$$\gamma_{w_i}^{\text{post}} = \gamma_{w_i}^{\text{likel}} \text{ and} \tag{5.36}$$

$$\mu_{w_i}^{\text{post}} = \mu_{w_i}^{\text{likel}} - \frac{1}{\beta_\gamma^{\text{prior}} \sum_{t=1}^{N} {z_{(t,i)}^{(n)}}^2}. \tag{5.37}$$

To calculate the mode of the posterior density, derivation of Eq. (5.35) with respect to w_i is calculated and the following is obtained:

$$w_i^{\alpha_{w_i}^{\text{prior}}-2} \exp\left(-\frac{\left(w_i - \mu_{w_i}^{\text{post}}\right)^2 \gamma_{w_i}^{\text{post}}}{2}\right)\left(w_i^2 - \mu_{w_i}^{\text{post}} w_i - \frac{\alpha_{w_i}-1}{\gamma_{w_i}^{\text{post}}}\right). \tag{5.38}$$

Hence, the mode of the posterior distribution is w_i that satisfies

$$w_i^2 - \mu_{w_i}^{\text{post}} w_i - \frac{\alpha_{w_i}-1}{\gamma_{w_i}^{\text{post}}} = 0. \tag{5.39}$$

Solving Eq. (5.39) yields:

$$\mu_{w_i}^{\max} = \begin{cases} 0 & D < 0 \\ \max\left\{\frac{1}{2}\left(\mu_{w_i}^{\text{post}} + \sqrt{D}\right), 0\right\} & \text{otherwise} \end{cases}, \tag{5.40}$$

where D is given by

$$D = \left(\mu_{w_i}^{\text{post}}\right)^2 + 4\frac{\alpha_{w_i}^{\text{prior}} - 1}{\gamma_{w_i}^{\text{post}}}. \tag{5.41}$$

Finally, the following proposal distribution is obtained:

$$q(w_i) = \mathcal{N}\left(w_i^{(n+1)} \middle| \mu_{w_i}^{\max}, \sqrt{1/\gamma_{w_i}^{\mathrm{post}}}\right). \tag{5.42}$$

2. The posterior density corresponding to $\gamma^{(n+1)}$ is

$$\begin{aligned} &p\left(\gamma \middle| x_{(1:N)}, z_{(1:N,1:K)}^{(n)}, w_{(1:K+1)}^{(n+1)}, \alpha_\gamma^{\mathrm{prior}}, \beta_\gamma^{\mathrm{prior}}\right) \\ &\propto p\left(x_{(1:N)} \middle| z_{(1:N,1:K)}^{(n)}, w_{(1:K+1)}^{(n+1)}, \gamma\right) p\left(\gamma | \alpha_\gamma^{\mathrm{prior}}, \beta_\gamma^{\mathrm{prior}}\right). \end{aligned} \tag{5.43}$$

The first term on the right-hand side of Eq. (5.43) can be written as

$$\begin{aligned} p\left(x_{(1:N)} | z_{(1:N,1:K)}^{(n)}, w_{(1:K+1)}^{(n+1)}, \gamma\right) &= \prod_{t=1}^{N} \mathcal{N}\left(x_t \middle| \sum_{i=1}^{K+1} z_{(t,i)}^{(n)} w_i^{(n+1)}, \gamma\right) \\ &\propto \gamma^{N/2} \exp\left\{-\frac{\gamma}{2} \sum_{t=1}^{N} \left(x_t - \sum_{i=1}^{K+1} z_{(t,i)}^{(n)} w_i^{(n+1)}\right)^2\right\}. \end{aligned} \tag{5.44}$$

Because the prior distribution $p\left(\gamma \middle| \alpha_\gamma^{\mathrm{prior}}, \beta_\gamma^{\mathrm{prior}}\right)$ is a gamma distribution, the posterior distribution can also be written as the following gamma distribution,

$$p\left(\gamma^{(n+1)} \middle| x_{(1:N)}, z_{(1:N,1:K)}^{(n)}, w_{(1:K+1)}^{(n+1)}, \alpha_\gamma^{\mathrm{prior}}, \beta_\gamma^{\mathrm{prior}}\right) = \mathcal{G}\left(\gamma^{(n+1)} | \alpha_\gamma^{\mathrm{post}}, \beta_\gamma^{\mathrm{post}}\right), \tag{5.45}$$

where

$$\alpha_\gamma^{\mathrm{post}} = \alpha_\gamma^{\mathrm{prior}} + \frac{N}{2}, \tag{5.46}$$

$$\beta_\gamma^{\mathrm{post}} = \left\{\frac{1}{\beta_\gamma^{\mathrm{prior}}} + \frac{1}{2} \sum_{t=1}^{N} \left(x_t - \sum_{j=1}^{K+1} z_{(t,j)}^{(n)} w_j^{(n)}\right)^2\right\}^{-1}. \tag{5.47}$$

3. The posterior distribution corresponding to $\boldsymbol{z}_t^{(n+1)}$ is

$$\begin{aligned} &p\left(\boldsymbol{z}_t \middle| x_t, \boldsymbol{z}_{t-1}^{(n+1)}, \boldsymbol{z}_{t+1}^{(n)}, 1/\tau^{(n)}, w_{(1:K+1)}^{(n+1)}, \gamma^{(n+1)}\right) \\ &\propto p\left(x_t \middle| \boldsymbol{z}_t, w_{(1:K+1)}^{(n+1)}, \gamma^{(n+1)}\right) p\left(\boldsymbol{z}_t \middle| \boldsymbol{z}_{t-1}^{(n+1)}, 1/\tau^{(n)}\right) p\left(\boldsymbol{z}_{t+1}^{(n)} \middle| \boldsymbol{z}_t, 1/\tau^{(n)}\right), \end{aligned} \tag{5.48}$$

where $p\left(x_t \left| \boldsymbol{z}_t, w_{(1:K+1)}, \gamma \right.\right)$ and $p\left(z_{(t,i)} \left| z_{(t-1,i)}, 1/\tau \right.\right)$ are defined in Eq. (5.4) and Eq. (5.6), respectively. Note that $\boldsymbol{z}$ is a binary variables that represents the trap state, i.e. the trap is occupied or not. Hence, the all of the possible combinations of $\boldsymbol{z}_t$ (number of combinations: 2^K) can be enumerated. Note here that the number of traps included in the modern transistor is relatively small and hence the computational cost required for the enumeration of the combinations of $\boldsymbol{z}_t$ is also small. The posterior probability for the all of the possible combinations of $\boldsymbol{z}_t$ is first calculated and $\boldsymbol{z}_t$ is sampled according to that probability.

4. The posterior distribution corresponding to $z_1^{(n+1)}$ is

$$\begin{aligned} &p\left(z_1 \left| x_1, z_2^{(n)}, 1/\tau^{(n)}, w_{(1:K+1)}^{(n+1)}, \gamma^{(n+1)} \right.\right) \\ &\propto p\left(x_1 \left| z_1, w_{(1:K+1)}^{(n+1)}, \gamma^{(n+1)} \right.\right) p\left(z_2^{(n)} \left| z_1, 1/\tau^{(n)} \right.\right) p\left(z_1 \left| \pi^{(n)} \right.\right), \end{aligned} \tag{5.49}$$

where $p\left(z_1 \left| \pi^{(n)} \right.\right)$ is defined as Eq. (5.7). Drawing samples from the posterior distribution is done in the same way as sampling of $z_t^{(n+1)}$.

5. The posterior distribution corresponding to $1/\tau_{(i,l)}^{(n+1)}$ is

$$\begin{aligned} &p\left(\frac{1}{\tau_{(i,l)}} \left| z_{1:N,i}^{(n+1)}, \alpha_{1/\tau_{(i,l)}}^{(n)}, \beta_{1/\tau_{(i,l)}}^{(n)} \right.\right) \\ &\propto p\left(z_{(1:N,i)}^{(n+1)} \left| \frac{1}{\tau_{(i,l)}} \right.\right) p\left(\frac{1}{\tau_{(i,l)}} \left| \alpha_{1/\tau_{(i,l)}}^{(n)}, \beta_{1/\tau_{(i,l)}}^{(n)} \right.\right). \end{aligned} \tag{5.50}$$

The first term on the right-hand side of Eq. (5.50) can be written as

$$p\left(z_{(1:N,i)}^{(n+1)} \left| \frac{1}{\tau_{(i,l)}} \right.\right) = \exp\left(\frac{1}{\tau_{(i,l)}}\right)^{n_{l\to l}} \left\{1 - \exp\left(\frac{1}{\tau_{(i,l)}}\right)\right\}^{n_{l\to \bar{l}}}, \tag{5.51}$$

where $n_{l\to l}$ is the number of steps whose state is the same as the previous one ($z_{(t,i)}^{(n+1)} = z_{(t+1,i)}^{(n+1)} = l$), and $n_{l\to \bar{l}}$ is the number of steps whose state is different from the previous one ($z_{(t,i)}^{(n+1)} = l$ and $z_{(t+1,i)}^{(n+1)} \neq l$). The second term on the right-hand side of Eq. (5.50) is the gamma distribution. The posterior distribution of Eq. (5.50) does not belong to well-known distributions. Therefore, Markov chain is again constructed to sample $1/\tau_{(i,l)}$ using the Metropolis method.

6. The posterior distribution corresponding to $\pi_i^{(n+1)}$ is

$$p\left(\pi_i \left| z_{(1,i)}^{(n+1)}, \alpha_{\pi_i}^{\text{prior}}, \beta_{\pi_i}^{\text{prior}} \right.\right) \propto p\left(z_{(1,i)}^{(n+1)} \left| \pi_i \right.\right) p\left(\pi_i \left| \alpha_{\pi_i}^{\text{prior}}, \beta_{\pi_i}^{\text{prior}} \right.\right). \tag{5.52}$$

The first term on the right-hand side of Eq. (5.52) is defined in Eq. (5.7) and the second term on the right-hand side of Eq. (5.52) is the beta distribution defined in Eq. (5.17). Hence, the posterior distribution also becomes the following beta distribution:

$$p\left(\pi_i \left| z_{(1,i)}^{(n+1)}, \alpha_{\pi_i}^{\mathrm{prior}}, \beta_{\pi_i}^{\mathrm{prior}} \right.\right) = \mathcal{B}\left(\pi_i | \alpha_{\pi_i}^{\mathrm{post}}, \beta_{\pi_i}^{\mathrm{post}}\right), \tag{5.53}$$

where

$$\alpha_{\pi_i}^{\mathrm{post}} = \alpha_{\pi_i}^{\mathrm{prior}} - z_{(1,i)}^{(n+1)} + 1 \tag{5.54}$$

$$\beta_{\pi_i}^{\mathrm{post}} = \beta_{\pi_i}^{\mathrm{prior}} + z_{(1,i)}^{(n+1)}. \tag{5.55}$$

7. The posterior distribution corresponding to $\alpha_{w_i}^{(n+1)}$ is

$$\begin{aligned} p\left(\alpha_{w_i} \left| w_i^{(n+1)}, \beta_{w_i}^{(n)}, \lambda_{\alpha_{w_i}} \right.\right) &\propto p\left(w_i^{(n+1)} \left| \alpha_{w_i}, \beta_{w_i}^{(n)} \right.\right) p\left(\alpha_{w_i} \left| \lambda_{\alpha_{w_i}} \right.\right) \\ &= \mathcal{G}\left(w_i^{(n+1)} \left| \alpha_{w_i}, \beta_{w_i}^{(n)} \right.\right) \mathcal{L}\left(\alpha_{w_i} \left| \lambda_{\alpha_{w_i}} \right.\right). \end{aligned} \tag{5.56}$$

This posterior distribution also does not belong to well-known distributions. Therefore, the Metropolis method is used to draw samples from the posterior distribution.

8. The posterior distribution for sampling β_{w_i}, $\alpha_{1/\tau_{(i,l)}}$ and $\beta_{1/\tau_{(i,l)}}$ can be obtained in the same way as the posterior distribution of α_{w_i}.

Proposed parameter estimation algorithm

The proposed parameter estimation algorithm is defined as below.

1. Initialize the following random variables: $z_{(1:N,1:K)}^{(0)}$, $\gamma^{(0)}$, $w_{(1:K+1)}^{(0)}$, $1/\tau_{(1:K,0:1)}^{(0)}$, $\pi_{1:K}^{(0)}$, $\alpha_{w_{1:K+1}}^{(0)}$, $\beta_{w_{1:K+1}}^{(0)}$, $\alpha_{1/\tau_{(1:K,0:1)}}^{(0)}$, and $\beta_{1/\tau_{(1:K,0:1)}}^{(0)}$.

2. Repeat the following steps until generated samples become independent of the initial values.

   (a) Sample $w_i^{(n+1)}$ for $i = 1, \cdots, K+1$:
   draw a candidate sample of $w_i^{(n+1)}$ from Eq. (5.42) (w^*). Sample u from a uniform distribution. Accept w^* as $w_i^{(n+1)}$ if $\min\left(1, \frac{p(w^*)}{p\left(w_i^{(n)}\right)}\right) > u$. Otherwise, the candidate sample is rejected and $w_i^{(n+1)}$ is set to $w_i^{(n)}$.

(b) Sample $\gamma^{(n+1)}$:
draw a sample from Eq. (5.45) conditioned on X, $z^{(n)}_{(1:N,1:K)}$, $w^{(n+1)}_{(1:K+1)}$, α_γ, and β_γ.

(c) Sample $z_1^{(n+1)}$:
calculate the probabilities of each possible combination of z_1 by Eq. (5.49). Calculate the cumulative probability and sample $z_1^{(n+1)}$ according to a random value drawn from uniform distribution.

(d) Sample π_i for $i = 1, \cdots, K$:
draw a sample from Eq. (5.52) conditioned on $z_1^{(n+1)}$, α_{π_i}, and β_{π_i}.

(e) Sample $z_t^{(n+1)}$ for $t = 2, \cdots, N$:
calculate the cumulative probability by Eq. (5.48) and sample z_t in the same manner as sampling z_1.

(f) Sample $\frac{1}{\tau^{(n+1)}_{(i,l)}}$ for $i = 1, \cdots, K$, $l = 0, 1$:
draw a sample from Eq. (5.50) conditioned on $z^{(n+1)}_{(1:N,1:K)}$, $\alpha_{1/\tau^{(n)}_{(1:K,0:1)}}$, and $\beta_{1/\tau^{(n)}_{(1:K,0:1)}}$.

(g) Sample $\alpha_{w_i}^{(n+1)}$ for $i = 1, \cdots, K+1$:
draw a sample from Eq. (5.56) conditioned on $w_i^{(n+1)}$, $\beta_{w_i}^{(n)}$, and $\lambda_{\alpha_{w_i}}$.

(h) Sample $\beta_{w_i}^{(n+1)}$ for $i = 1, \cdots, K+1$:
draw a sample conditioned on $w_i^{(n+1)}$, $\alpha_{w_i}^{(n+1)}$, and $\lambda_{\beta_{w_i}}$ in the same way as the sampling of $\alpha_{w_i}^{(n+1)}$.

(i) Sample $\alpha_{1/\tau_{(i,l)}}$ and $\beta_{1/\tau_{(i,l)}}$:
draw samples in the same way as sampling α_{w_i} and β_{w_i}, respectively.

5.2.5 Replica Exchange

A parallel tempering method is additionally introduced to find the globally optimal solution. This step is required because of the locality of MCMC, in which candidate random samples tend to stay close to the previous positions. This sampling strategy enables MCMC to draw critical samples from high dimensional distribution. On the other hand, in the cases when posterior distribution is multi-modal, the sampling strategy makes it difficult to explore the sample space broadly enough. Samples are often trapped to the local maxima surrounded by low-probability valleys. To overcome this difficulty, the idea of replica exchange method [64] is introduced. In the replica exchange method, multiple copies of Markov chains each having different temperature are simulated, and random samples between the copies are exchanged. It gives

samples more chance to jump out of the local space, which is similar to the idea of simulated annealing. In order to introduce a temperature coefficient, the conditional posterior distribution (β) is modified as follows:

$$p(\mathbf{x}|\beta) \propto P(\mathbf{x})^{1/\beta}. \tag{5.57}$$

The above distribution is flatter at high temperatures and it gradually becomes narrower and taller as temperature decreases. In replica exchange method, each Markov chain is independently simulated and samples at two different temperatures are exchanged with the following steps.

step 1 Generate m ($1 \leq m < M-1$) from uniform distribution. Here, M is the number of Markov chains.

step 2 Generate u ($0 \leq u < 1$) from uniform distribution. Exchange samples of the chains at temperatures β_m and β_{m+1} if $u \leq \min(1, r)$ is satisfied. Here, r is given by

$$r = \frac{P(\mathbf{x}_{m+1}|\beta_m)P(\mathbf{x}_m|\beta_{m+1})}{P(\mathbf{x}_m|\beta_m)P(\mathbf{x}_{m+1}|\beta_{m+1})}. \tag{5.58}$$

5.3 Experimental Validation

5.3.1 Preliminary Experiment Using Synthetic V_{TH} Waveform

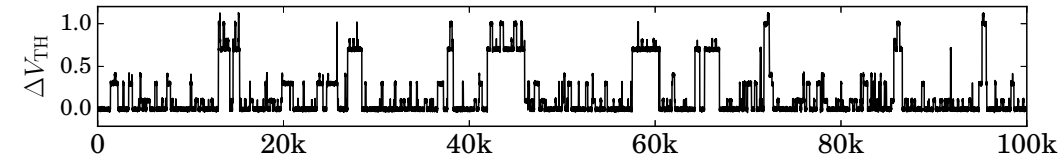
To validate the proposed estimation method, an experiment with synthetic V_{TH} waveforms is conducted. This experiment utilizes a time series signal with 100,000 steps that simulates the stair-like V_{TH}. In this particular case, the number of traps is set to three. The amplitudes and time constants are set as listed in the left half of Table 5.1 (ground truth). The parameters of the prior distributions are set to $\alpha^{\mathrm{prior}}_{w_{(1:K)}} = 10^{-1}$, $\beta_{w^{\mathrm{prior}}_{(1:K)}} = 10^{3}$, $\alpha^{\mathrm{prior}}_{\gamma} = 10^{5}$, $\beta^{\mathrm{prior}}_{\gamma} = 10^{-1}$, $\alpha^{\mathrm{prior}}_{\pi_{(1:K)}} = 1$, $\beta^{\mathrm{prior}}_{\pi_{(1:K)}} = 1$, $\alpha^{\mathrm{prior}}_{\rho_{(1:K,0:1)}} = 1$, and $\beta^{\mathrm{prior}}_{\rho_{(1:K,0:1)}} = 1$. The maximum number of defects assumed (K) is set to five.

Fig. 5.3 shows the estimated trap states and reconstructed waveform. The first 20,000 steps of the time series are plotted. From this figure, we can see that the amplitude and trap states are extracted with good accuracy from the input data.

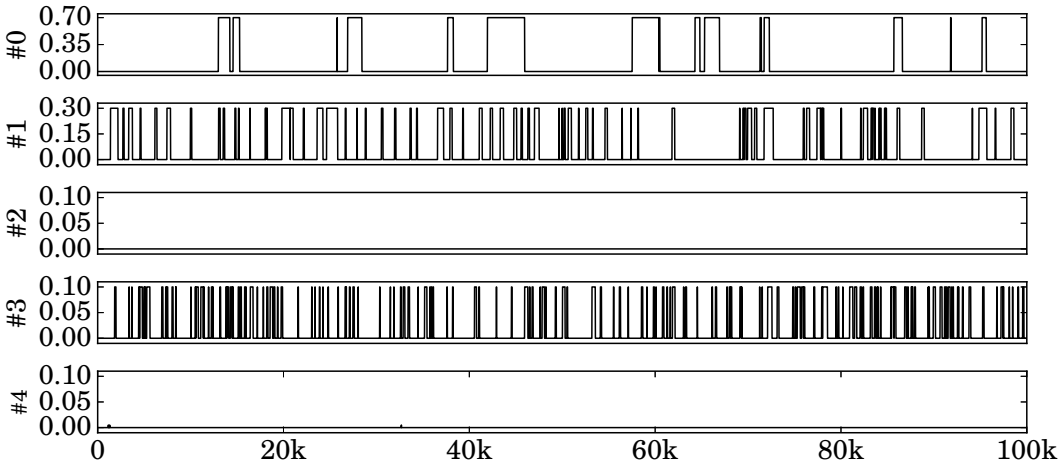
The estimated parameters are summarized in Table 5.1. We can see that the proposed method successfully estimates the magnitude of V_{TH} shift and time constants of the traps of the synthetic V_{TH} waveform. In this particular case,

Table 5.1: Parameters used for generating synthetic RTN data and its estimation result.

	Ground truth			Estimated		
Trap	#1	#2	#3	#1	#2	#3
Amplitude	0.100	0.300	0.700	0.100	0.300	0.700
τ_0	500	1000	5000	557.1	1127	5572
τ_1	100	200	1000	112.3	246.4	1106



(a) Simulated V_{TH} waveform



(b) Decomposed trap states

Figure 5.3: Estimation result on simulated RTN waveform. We can see that proposed method successfully decomposed multi-trap activity.

the estimation error of the magnitude of V_{TH} shift is 13.9% and that of the time constants is 14.2%.

The calculation time linearly increases with the number of sample points included in the V_{TH} sequence (N) and exponentially increases with the maximum number of traps (K) assumed. In the configuration considered in this section where $N = 100,000$ and $K = 5$, the calculation time was about 1 hour and 26 minutes with 16-core Xeon E5-4640 operated at 2.40 GHz.

5.3.2 Failure Analysis Using Synthetic RTN Data

For applying the proposed method to real world data, it is also important to be able to know whether a separation result is confident or not because there is no way of knowing the ground truth of the parameters, i.e. the magnitudes of V_{TH} shifts and the time constants. The posterior probability given by Eq. (5.20) perfectly works for that purpose as a quantitative measure.

To analyze the relationship between the posterior probability and the estimation error, a Monte Carlo experiment is conducted as follows. First, the synthetic V_{TH} waveforms are randomly generated and the proposed parameter extraction method is applied to extract the parameters from the synthetic V_{TH} waveforms. Then, the estimation errors are calculated by comparing the parameters extracted using the proposed method and that used for generating the V_{TH} waveforms. Finally, the estimation error and the posterior probability are plotted on the x-y plane to examine the suitability of the posterior probability as the quantitative measure.

In order to simulate the actual V_{TH} waveforms, the statistical distributions which the parameters follow are carefully selected. Here, parameters include the magnitude of V_{TH} shift corresponds to the each trap (w_i) and the time constants ($\tau_{(i,j)}$). According to [65], w_i is reported to follow a log-normal distribution. On the other hand, there seems to be no consensus to the time constants distribution. Hence, in the following, a log-normal distribution is adopted for the time constant distribution based on [66, 67].

The synthetic V_{TH} waveforms are generated as follows. First, the magnitudes of V_{TH} shift correspond to each trap ($\{w_i; i = 1, 2, \cdots, K_{\mathrm{sim}}\}$) are drawn from the following log-normal distribution:

$$w_i \sim \text{log-normal}(w_i|\mu_w, \sigma_w), \tag{5.59}$$

where μ_w and σ_w are set to -2.0 and 1.0, respectively. Here, K_{sim} is the number of traps assumed in the synthetic V_{TH} waveform and is set to two in this experiment. Similarly, time constants $\{\tau_{(i,j)}; i = 1, 2, \cdots, K_{\mathrm{sim}}; j = 0, 1\}$ are generated from the following log-normal distribution:

$$\tau_{(i,j)} \sim \text{log-normal}(\tau_{(i,j)}|\mu_\tau, \sigma_\tau), \tag{5.60}$$

where μ_τ and σ_τ are set to 7.0 and 1.0, respectively. Then, the binary trap state $z_{(t,i)}$ is simulated according to the two-state Markov model in Fig. 5.2 whose transition probabilities are given by the generated time constants. Finally, a sequence of total V_{TH} shift ($\boldsymbol{x}_{\mathrm{sim}}$) is obtained by summing up the contributions

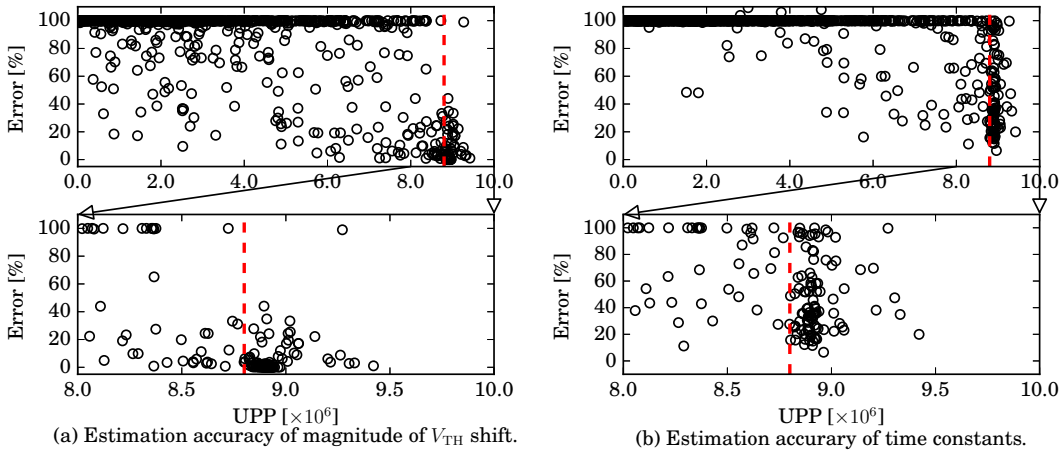


Figure 5.4: Estimation accuracy of (a) the magnitude of V_{TH} shifts and (b) the time constants.

from each trap as follows:

$$\boldsymbol{x}_{\mathrm{sim}} = \sum_{i=1}^{K_{\mathrm{sim}}} w_i z_{(1:N_{\mathrm{sim}},i)} + \boldsymbol{\xi}, \tag{5.61}$$

where N_{sim} is the total length of the simulated V_{TH} waveform and it is set to 100,000 in this experiment. $\boldsymbol{\xi} = (\xi_1, \xi_2, \cdots, \xi_{N_{\mathrm{sim}}})$ is a vector of random variables that represent the measurement error. In this experiment, each element of $\boldsymbol{\xi}$ is assumed to follow the normal distribution: $\xi_t \sim \mathcal{N}(\xi_t|\mu_\xi, \sigma_\xi)$, where μ_ξ and σ_ξ are set to 0 and 0.01, respectively. Repeating the above procedure, 1,000 synthetic V_{TH} waveforms are generated.

Fig. 5.4 shows the estimation error of the magnitude of V_{TH} shift (w_i) and time constants ($\tau_{(i,j)}$) as a function of the unnormalized posterior probability (UPP). We can see that the higher UPP is given to the cases with small estimation error. Studying Fig. 5.4, we also notice that the estimation error drops sharply around the UPP of 8.8×10^6 (indicated by red dashed lines). Hence, in the following, the estimated results with UPP of 8.8×10^6 or higher are considered to be the cases that succeeded in the parameter estimation.

Figs. 5.5(a) to (f) summarize the histogram of the UPP with respect to the estimated number of traps. Note again that, in the proposed method, the magnitude of V_{TH} shift of extra traps, i.e. traps that are unnecessary for explaining the total V_{TH} waveforms, automatically converge around zero. Hence, the number of traps is obtained by counting the number of elements in $\boldsymbol{w} = (w_1, w_2, \cdots, w_K)$ that is larger than the threshold value θ_w. Here, considering that the standard deviation of the measurement noise (σ_ξ) is 0.01, θ_w is set to

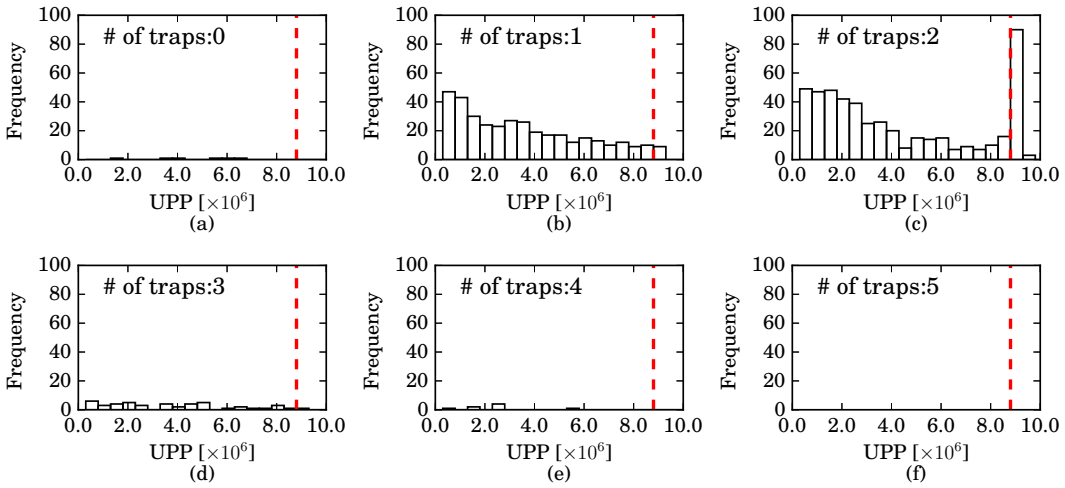


Figure 5.5: Histograms of the number of estimated traps.

0.02 that is twice as σ_ξ. In this experiment, the synthetic V_{TH} waveforms contain two traps activities and hence only the bins of Fig. 5.5(c) should have non-zero values. However, as we can see from the result, the number of traps are tend to be underestimated probably due to the inclusion of traps with small w_k that are difficult to distinguish from the measurement noise. The filtering criteria of UPP is again indicated by the red dashed lines. By taking the estimation results that lie right of the line, we can filter out the false estimations while the results that were succeeded in the estimation can be kept.

5.3.3 Experiment Using Measured V_{TH} Waveform

In order to see whether the proposed method can analyze actual V_{TH} waveforms or not, the V_{TH} waveform measured on BTIarray that is described in Chapter 3 is used. Fig. 5.6(a) is an example of V_{TH} waveform observed on pMOS transistor whose channel length and width are 360 nm and 120 nm, respectively. The sampling rate is about 125 samples per second. Observing Fig. 5.6(a), we notice the V_{TH} shift consists of two components: one is from a trap whose amplitude of V_{TH} shift is about 0.05 and the other one whose amplitude is about 0.1.

Fig. 5.6(b) shows the separation result using the proposed method. The same model parameters as those used in the previous section are used. Each column in Fig. 5.6(b) shows time series sequence of the separated five traps. Only traps #0 and #2 have non-zero amplitudes and those of the other three traps are zero. We can see that the amplitudes of traps that do not contribute for explaining the observed V_{TH} waveform have been automatically degenerated and that the proposed method successfully estimated the number of traps in

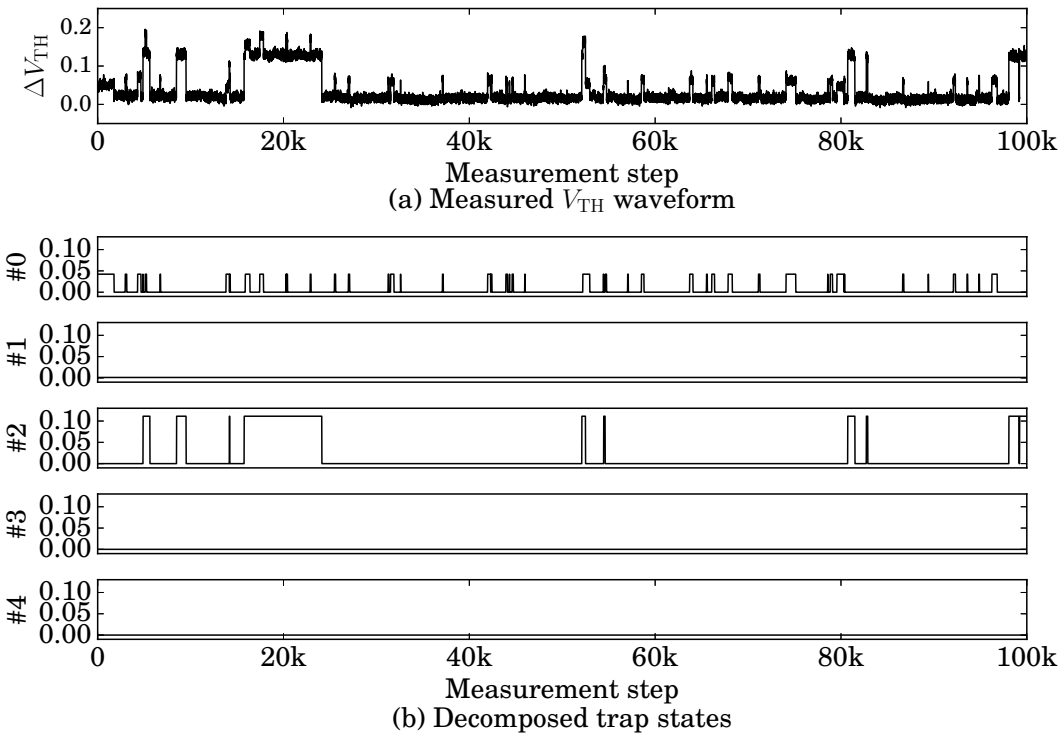


Figure 5.6: Separation result of measured RTN waveform (UPP = 1.20×10^7). Two trap components are correctly separated. Note that the traps #1, #3, and #4 have near zero amplitudes, which means the estimation of trap number is also right.

the measured V_{TH} without human intervention.

The estimated magnitudes of V_{TH} shifts and time constants of the two traps are listed in Table 5.2. Note again that τ_0 represents the average duration in which the trap is empty, i.e. the V_{TH} is in the lower state and that τ_1 represents the average duration in which the trap is occupied, i.e. the V_{TH} is in the higher state. From Table 5.2, we notice that τ_0 and τ_1 of trap #0 are smaller than those of trap #2, which means that trap #0 captures and emits electrons more actively than trap #2 does. We also notice that τ_0 of both traps are larger than τ_1. This suggests that the period at which V_{TH} is in low state is longer. These observations are perfectly consistent with the measured V_{TH} waveform in Fig. 5.6(a).

The result shown in Fig. 5.6 is from a sample that has an unnormalized posterior probability (UPP) of 1.20×10^7. In order to see the effect of the UPP-based filtering, let us take a look at the separation result with lower UPP shown in Fig. 5.7. In this case, there is only one trap that has non-zero magnitudes of

Table 5.2: Estimated amplitudes and time constants from measured RTN waveform.

trap	#0	#2
amplitude [a.u]	4.23e−2	1.11e−1
τ_0 [a.u]	1.84e+3	2.40e+4
τ_1 [a.u]	2.45e+2	3.59e+3

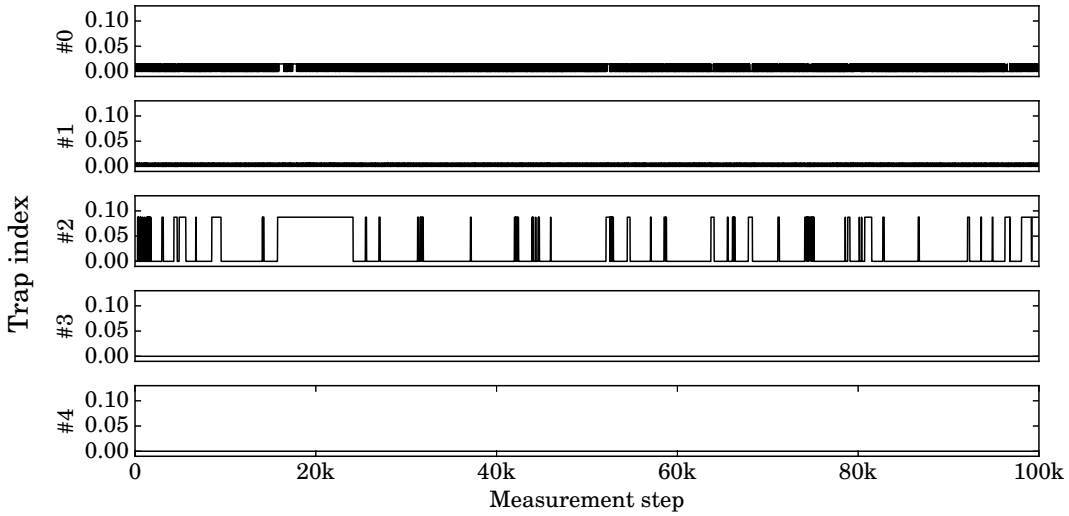


Figure 5.7: Example of unsuccessful separation result (UPP=8.72×10^6).

V_{TH} shifts and hence the proposed method underestimated the number of traps. The smaller UPP of 8.72×10^6 indicates weaker confidence of the result than that of Fig. 5.6. The UPP can be definitely a measure to judge the confidence of the estimation.

5.3.4 Comparison in Accuracy of the Extracted Parameters

Finally, the accuracy of the estimated parameters using the proposed method and hidden Markov model (HMM) [68] is examined. HMM is a popular and widely used method which is suitable to build a statistical model for time-domain sequences, such as voices. In the context of stair-like V_{TH} analysis, HMM can be used to extract transition probabilities between discrete V_{TH} states. The main difference between the proposed method and HMM is that the proposed method takes into account the generation process of the stair-like V_{TH} waveform while HMM does not. Let us consider an example situation that the two or more traps simultaneously capture carriers, resulting a large

V_{TH} fluctuation. Estimating the probability of the undesirable event, i.e. the large V_{TH} shift is invoked, is practically important because such undesirable event determines the total reliability of LSIs. We notice that the estimation of this probability is difficult because such situation that multi traps capture carriers simultaneously will hardly be observed. Hence, the proposed method complements the lack of the observation with the statistical generative model to achieve higher estimation accuracy than the conventional method such as HMM. In the following experiment, it is shown that extraction accuracy is improved over HMM by the proposed method using synthetic V_{TH} waveforms.

Experimental Setup

In this experiment, the estimation accuracy of the time constants extracted using the proposed method and that extracted using HMM is examined. In this experiment, the HMM implementation of MATLAB toolkit [69] is used.

The direct comparison of HMM with the proposed method is difficult because HMM cannot separate the trap activities. Let us take an example case where three traps are involved in the generation of the stair-like V_{TH} waveform and let $(b_1b_2b_3)$ represents the trap occupancy, where $b_i \in \{0, 1\}$. The proposed method can separately extract the time constants for each trap. Hence, for example, the transition between (011) and (111) is observed, this observation is decomposed into the series of trap activities as follows: trap #1 switched from "0" to "1" while traps #2 and #3 keep its previous states. On the other hand, HMM cannot decompose the observation into each trap activity and hence it just estimate the transition probability from (011) to (111).

In this experiment, instead of comparing the time constants, the steady state probabilities $\pi_{(b_1b_2b_3)}$, where $\pi_{(b_1b_2b_3)}$ represents the probability that V_{TH} is in state $(b_1b_2b_3)$, are compared. Further, HMM is given the ground truth of the amplitude, assuming that the amplitudes are extracted accurately using other method such as TLP in advance to the transition probability estimation using HMM.

Results and Discussion

The calculated steady state probabilities using the estimated transition probability of HMM and the proposed method are listed in Table 5.3. Note that HMM is given ground truth of amplitudes while the proposed method is not. From Table 5.3, we can see that the proposed method accurately estimates the steady state probability of state (111) which gives large V_{TH} shift with only 7.59 % error, while HMM exhibits 29.7 % error. The proposed method can utilize the prior knowledge about the physics of RTN, which contributes to the

Table 5.3: The steady state probabilities. In general, the estimation accuracy of the proposed method is higher than that of HMM.

Method	State							
	000	001	010	011	100	101	110	111
Ground Truth	5.78e−1	1.16e−1	1.16e−1	2.33e−2	1.17e−1	2.32e−2	2.32e−2	4.65e−3
HMM	5.82e−1	1.23e−1	1.23e−1	1.93e−2	1.07e−1	1.43e−2	2.91e−2	3.27e−3
Error (HMM)	0.613%	5.95%	6.07%	−17.1%	−7.88%	−38.4%	25.5%	−29.7%
Proposed	5.69e−1	1.15e−1	1.25e−1	2.52e−2	1.13e−1	2.29e−2	2.48e−2	5.01e−3
Error (Proposed)	−1.52%	−0.805%	7.60%	8.38%	−2.25%	−1.54%	6.80%	7.59%

higher estimation accuracy.

5.4 Summary

In this chapter, the statistical machine learning-based method is proposed to simultaneously estimate the amplitude and time constants of each trap from measured V_{TH} fluctuation caused by RTN. The proposed method can handle interrelated parameters of multiple traps and thereby contributes to the construction of more accurate RTN models and to the better understanding of the BTI-induced degradation. The experiments using synthetic and measurement data showed that the proposed method successfully estimated the magnitudes of the V_{TH} shift.

Chapter 6

Efficient Calculation of SRAM Yield Degradation

6.1 Introduction

Due to the increasing manufacturing variability, a circuit design has been emerged as a difficult challenge. One good example of such challenge is a bit cell design of a static random access memory (SRAM). Considering the fact that a modern microprocessor embeds tens of mega bytes of on-chip cache, extremely low failure probability is required for a single SRAM cell. A typical failure probability required for an SRAM cell is reported to be 10^{-8} to 10^{-6}, or below [70].

Estimations of such small failure probability is known to be a difficult task. A naive Monte Carlo (MC) method, which directly generates random samples in a variability space, requires millions or billions of circuit simulations to obtain only a single failure sample. Hence, it is almost impossible to accurately calculate the small failure probability and importance sampling techniques are definitely required to overcome this problem [34–36].

The variability of transistor-parameters is mostly originated in the course of manufacturing process. As the shrinkage of semiconductor manufacturing process continues, even an atomic level bump on a gate terminal or fluctuation of the number of dopant ions have a large impact on the electrical property of transistors. In addition to such "static" variability, we are currently forced to cope with an increasing impact of "dynamic" variability that originates from device degradation. Thin gate-oxide layer in highly scaled transistors poses various new problems on the reliability of LSI. Among the degradation mechanisms, NBTI attracts increasing concern as we have already seen in Chapter 4. To improve the reliability of LSI, designers must take the impact

of the NBTI induced device degradation into consideration in the design phase. Development of computer aided design (CAD) tools that accurately evaluate and countermeasure the device degradation has thus emerged as an urgent issue.

SRAM cells are considered to be one of the most vulnerable circuit components to the NBTI-induced V_{TH} shift for two reasons. First, a long stress period is frequently observed. Because of data-storage functionality as a memory, switching activity of an SRAM cell is usually lower than that of combinational circuits. Hence, a pMOS transistor in one of the coupled inverters is more likely to be exposed to constant stress, deteriorating the stability of the cell. Second, the heat produced by components surrounding an SRAM cell further complicates the problem. Although the SRAM cell itself produces only small amount of heat, components such as register files are surrounded by highly active circuits, such as instruction dispatchers, reorder buffers, etc. Heat generated by these components accelerates the NBTI-induced degradation. Therefore, circuit designers have to be extremely careful to optimize circuit structure of an SRAM cell and its placement in order to improve the reliability of LSI.

In this chapter, a novel and efficient failure probability calculation method of SRAM cells under the NBTI stress is proposed. A considerable amount of efforts are paid to accelerate the failure probability calculation of an SRAM cell [34–36]. Those methods, however, only consider the static variability such as the one caused in the manufacturing process. In order to check the change of failure probability, multiple calculations are required by changing the aging time. Even if a single failure probability calculation is accelerated using advanced sampling techniques, e.g. by using importance sampling, it still requires large simulation effort.

The proposed method solves this problem by applying the concept of the particle filter to incrementally track the time-changing characteristics of an SRAM cell. The concept of particle filter is first introduced into the CAD community in [39] to enhance the efficiency of the importance sampling (IS) based failure probability calculation. In the IS-based MC, random samples are generated from the distorted distribution, which is called "alternative distribution," instead of the original distribution. The alternative distribution is selected such that more failure samples are drawn, i.e. samples drawn from the alternative distribution is more likely to cause circuit failure. Due to the complicated shape of the alternative distribution, its analytical representation is difficult to obtain while the approximation of the distribution using a simple distribution will deteriorate the effectiveness. Hence, in [39], particles were used to represent the complex shape of the alternative distribution and drastic speed-up compared to existing importance sampling approaches was achieved. The contribution in this chapter is to extend the method proposed

in [39] so as to efficiently handle the aging effect. Due to NBTI-induced device degradation, the shape of the optimal alternative distribution changes as device age advances. Because construction of the alternative distribution from scratch is a computationally heavy task, the proposed method exploits the characteristics of the NBTI-induced device degradation. Specifically, when the change of V_{TH} is gradual, the change of the optimal alternative distribution is also gradual. Hence, in the proposed method, the temporal change of the optimal alternative distribution are tracked by using the particles moving around the variability space. This procedure substantially accelerates the total calculation time of failure probability along aging time steps by eliminating the independent explorations in variability space.

Further, a binary classifier based on a support vector machine (SVM) is integrated with a two-stage MC approach. Firstly, the binary classifier is trained using a small subset of random samples to roughly judge whether a sample causes circuit failure or not. Using the classifier, the majority of the random samples are classified as either pass or fail without executing time-consuming transistor-level simulations. Because the classifier is based on a linear model, the time required for the classifications is negligibly small. The reduction of the total calculation time is significant even though the time to train the classifier is newly introduced.

An adoption of a two-stage MC flow further reduces the calculation time while maintaining accuracy. In the first stage, a rough estimation of the optimal alternative distribution is obtained using a small number of random samples. Then, in the second stage, a failure probability is accurately calculated using the samples generated from the alternative distribution. With above, the proposed method achieves 2.3× speed-up of the failure probability calculation on a single aging time step compared to the state-of-the-art failure probability calculation method [39]. Total calculation time required to obtain the temporal change of the failure probability is also significantly reduced and the proposed method achieved 9.76× speed up compared to the conventional method [39].

This chapter is organized as follows. Backgrounds that forms the basis of the proposed method are described in Section 6.2. Then, the detail of the proposed method will be given in Section 6.3. Section 6.4 provide the numerical experiment and its result. Finally, Section 6.5 summarizes this chapter.

6.2 Background

6.2.1 Failure probability calculation

Failure probability calculation is generally formulated as

$$P_{\text{fail}} = \int I(\boldsymbol{x})P(\boldsymbol{x})d\boldsymbol{x}. \tag{6.1}$$

Here, P_{fail} is the failure probability and $\boldsymbol{x}$ is the D-dimensional random variable which corresponds to random variations of the parameters of transistor, such as V_{TH}, channel length, gate oxide thickness, etc. $P(\boldsymbol{x})$ is the probability density function (PDF) over the process variability. $I(\boldsymbol{x})$ is an indicator function that returns "1" if the given random variable $\boldsymbol{x}$ causes a malfunction of an SRAM cell, and "0" otherwise. Hereafter, the regions in which failure samples ($\boldsymbol{x}_{\text{fail}}$ such that $I(\boldsymbol{x}_{\text{fail}}) = 1$) distribute is called as "failure regions."

Because the indicator function does not have an analytical form in general, an MC approximation is adopted to evaluate Eq. (6.1). The above integral is calculated using random samples drawn from $P(\boldsymbol{x})$ as follows:

$$P_{\text{fail}} \approx \frac{1}{N}\sum_{i=1}^{N} I(\boldsymbol{x}_i), \tag{6.2}$$

where $\boldsymbol{x}_i \sim P(\boldsymbol{x})$. A naive MC method in Eq. (6.2) can not be applied to the calculation of Eq. (6.1) in low failure probability problems because very few or no samples that cause a malfunction of an SRAM cell can be generated in practical time.

In order to improve the sampling efficiency, importance sampling techniques is developed. The key idea of the importance sampling is to calculate Eq. (6.1) using samples drawn from an alternative distribution $Q(\boldsymbol{x})$. The following equation is obtained by modifying Eq. (6.1):

$$P_{\text{fail}} = \int I(\boldsymbol{x})\frac{P(\boldsymbol{x})}{Q(\boldsymbol{x})}Q(\boldsymbol{x})d\boldsymbol{x}. \tag{6.3}$$

The MC approximation of Eq. (6.3) using the samples drawn from $Q(\boldsymbol{x})$ is obtained as:

$$P_{\text{fail}} \approx \frac{1}{N}\sum_{i=1}^{N} I(\boldsymbol{x}_i)\frac{P(\boldsymbol{x}_i)}{Q(\boldsymbol{x}_i)}. \tag{6.4}$$

The optimal alternative distribution is known to be

$$Q_{\text{opt}}(\boldsymbol{x}) \propto I(\boldsymbol{x})P(\boldsymbol{x}). \tag{6.5}$$

If we can draw samples from $Q_{\mathrm{opt}}(\boldsymbol{x})$, a perfect approximation of Eq. (6.3) with zero variance can be achieved because $I(\boldsymbol{x})P(\boldsymbol{x})/Q_{\mathrm{opt}}(\boldsymbol{x})$ becomes a constant. This means that, in order to improve the efficiency, $Q(\boldsymbol{x})$ should be selected carefully so that the shape of $Q(\boldsymbol{x})$ becomes close to $Q_{\mathrm{opt}}(\boldsymbol{x})$. However, this is not a trivial task because we do not know the exact shape of the indicator function $I(\boldsymbol{x})$. In order to enable an automatic estimation of the optimal alternative distribution, a particle filter is introduced.

6.2.2 Particle filter

Particle filter is an on-line estimator of non-Gaussian distributions [9, 10]. A probabilistic density is approximated using the density of particles that move in the D-dimensional variability space. The positions of the particles are updated iteratively using the following steps as shown in Fig. 6.1.

Prediction The locations of the candidate particles in the next iteration are drawn from the proposal distribution $q(\boldsymbol{x})$. A usual choice of $q(\boldsymbol{x})$ is a mixture of normal distributions with each component centered at each position of the particles generated in the previous iteration so that the regions where the previous particles existed are more likely to be visited.

Measurement The weights that represent the goodness of fit of each candidate particle are calculated. In the context of the failure probability calculation of an SRAM cell, the weight is calculated as $I(\boldsymbol{x}) \cdot P(\boldsymbol{x})$. In case that $P(\cdot)$ is a normal distribution, large weights are assigned to the particles that are in the failure region and close to the origin of the variability space.

Resampling The particles are resampled from the candidate particles according to the probabilities proportional to the weight assigned in the **Measurement** step. Hence, the candidate particles with the larger weights attain more number of copies while the particles with the smaller weights attain smaller number of copies. Those outside the failure region are eliminated because $I(\boldsymbol{x})$ returns "0" for those samples and their weights become zero.

The population of the particles becomes gradually closer to the distribution $I(\boldsymbol{x}) \cdot P(\boldsymbol{x})$ by repeating the above procedures. The approximation of the optimal alternative distribution can be obtained as the distribution of particles.

6.2.3 Support vector machine

Although a large portion of the samples generated from the alternative distribution approximated with particles are in the failure region, a small

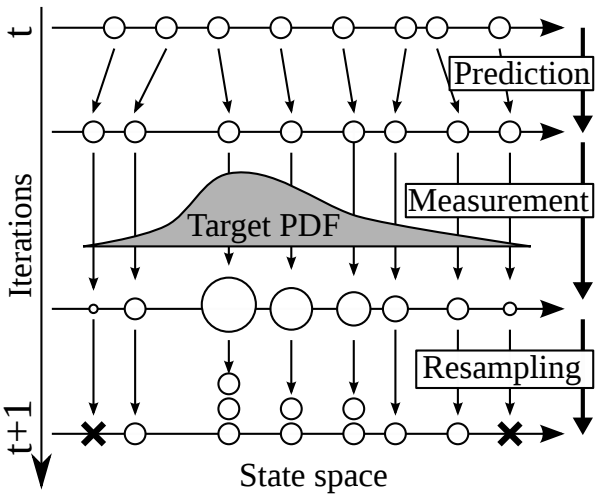


Figure 6.1: Particle filter.

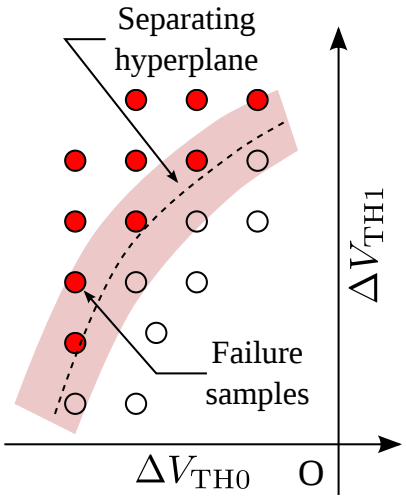


Figure 6.2: Support vector machine.

amount of pass samples are also generated due to the approximation error. If we could obtain the optimal alternative distribution, all of the random samples drawn from the distribution will be the failure samples and hence no transistor-level simulations to calculate $I(\boldsymbol{x})$ are required. Unfortunately, however, only an approximation of the optimal alternative distribution can be obtained in practice and the evaluations of the indicator function are required for all of the random samples generated.

Each time the indicator function $I(\boldsymbol{x})$ is evaluated, a transistor-level simulation is performed. This step occupies almost all of the total calculation time because the transistor-level simulation is a computationally heavy task. To accelerate the calculation of $I(\boldsymbol{x})$, a binary classifier based on a support vector machine (SVM) is introduced. SVM is a supervised training model for binary classification [71]. Given a set of training examples that consist of feature vectors and corresponding labels, SVM learns a classification model which categorizes a new feature vector into one of the two classes, i.e. pass or fail.

SVM assumes a linear classification model:

$$c = \sum_i w_i f_i. \tag{6.6}$$

Here, w_i is a coefficient of a particular feature quantity and f_i is the i-th element of a feature vector $\boldsymbol{f}$. The signature of c represents the class label of the feature vector. In other words, SVM learns a hyper plane in a feature space, which separates training examples into two classes as shown in Fig. 6.2. A good separation for a new feature vector is achieved when the distances between training examples and the hyper-plane are the largest.

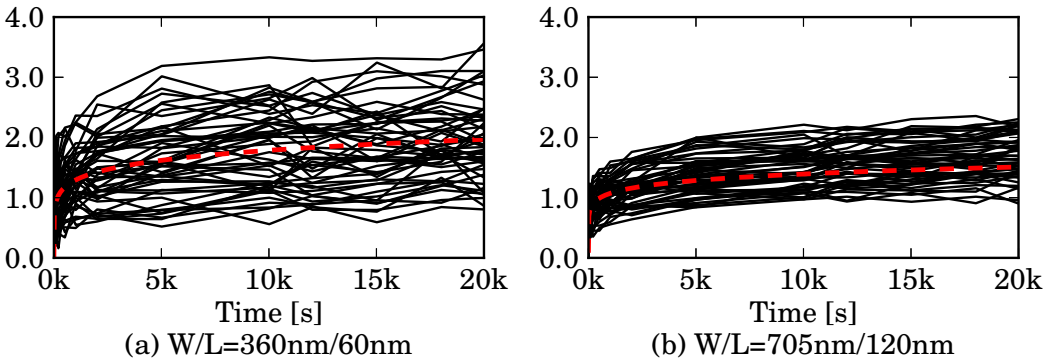


Figure 6.3: Examples of NBTI-induced V_{TH} shift observed in 50 pMOS transistors.

6.2.4 Variability on NBTI-induced V_{TH} shift

Let us again take a look at actual V_{TH} shifts acquired from a silicon measurement. Fig. 6.3 shows examples of NBTI-induced V_{TH} shifts observed in 50 pMOS transistors integrated on BTIarray which was fabricated using a commercial 65-nm CMOS process. The results of two sizes of transistors are shown. As it was investigated in Chapter 4, the V_{TH} shift varies widely among transistors just like the initial V_{TH} variation. It is clear from Figs. 6.3(a) and (b) that V_{TH} shift of smaller transistors varies widely while the overall trend of the degradations on both types of transistors are almost the same. The objective of this chapter is to propose a failure probability calculation method which can handle NBTI-induced device degradation. However, as it is clear from Fig. 6.3(a), we are also required to take the variability in device degradation into account.

In the compact model equations in Eq. (2.1) and Eq. (2.2), the model parameters that reflect the variability of the degradation are n for the RD-based model and ϕ for the TD-based model [13]. In this chapter, the RD-based model is adopted because it is simpler than the TD-based model and easy to extract the model parameters from silicon measurements. However, because the proposed method is based on an MC approach, other models such as the TD-based model can be used completely in the same way.

The red lines in Fig. 6.3 show the averaged model prediction over the 50 transistors. We can see that the model well predicts the temporal change of V_{TH}. This justifies the use of the RD-model for long-term reliability assessment. The remaining problem is to find the statistical distribution which the parameter n follows. As we have seen in Chapter 4, it is experimentally shown that n follows a log-normal distribution. According to this observation, a log-normal

distribution is employed in this chapter for the statistical distribution of n. Hence, the logarithm of n ($n_{\log}$) is assumed to follow a normal distribution:

$$n_{\log} \sim \mathcal{N}(n_{\log}|\mu_n, \sigma_n). \tag{6.7}$$

Here, $\mathcal{N}(x|\mu, \sigma)$ is the PDF of a normal distribution given by

$$\mathcal{N}(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{6.8}$$

6.3 Proposed Method

6.3.1 Variability and degradation modeling

This section first describes the variability and degradation modeling. In this dissertation, only the variabilities in V_{TH} is considered but other sources of variabilities, such as those of a channel size or a gate oxide thickness, can be easily taken into account. In [72], the authors measured 11 billion transistors and showed that the V_{TH} variation of fresh transistors follows a normal distribution. Based on this evidence, a normal distribution is adopted for the variability model of V_{TH}. Hence, V_{TH} variation of a fresh transistor is given by

$$\Delta V_{\mathrm{TH}}^{\mathrm{fresh}} = \frac{A_{\mathrm{V_{TH}}}}{\sqrt{L \cdot W}} x^{\mathrm{fresh}}. \tag{6.9}$$

Here, $A_{\mathrm{V_{TH}}}$ is a Pelgrom coefficient, and L and W are channel length and width, respectively. x^{fresh} is a random variable which is assumed to follow the standard normal distribution:

$$x^{\mathrm{fresh}} \sim \mathcal{N}\left(x^{\mathrm{fresh}} \middle| 0, 1\right). \tag{6.10}$$

The NBTI-induced V_{TH} shift is given by Eq. (2.1). Note again that n is a random variable that represents degradation variation. As stated in Section 6.2.4, n is represented using a log-normal distribution. Hence, a logarithm of n ($n_{\log}$) is given by

$$n_{\log} = x^{\mathrm{bti}} \cdot \sigma_n + \mu_n. \tag{6.11}$$

Here, x^{bti} is again a random variable following the standard normal distribution, and σ_n and μ_n are the standard deviation and mean of the distribution of $n_{\log}$.

A single MC trial proceeds as follows. A set of random samples are drawn from a probability distribution. In the failure probability calculation of an SRAM cell, there are eight random variables: six for the V_{TH} variation of six

transistors right after fabrication and other two are NBTI-induced degradation of the two pMOS transistors in the cell. Then, the corresponding V_{TH} shift of each transistor is calculated using Eq. (6.9), Eq. (2.1), and Eq. (6.11). Note that simple sum of $\Delta V_{\mathrm{TH}}^{\mathrm{fresh}}$ and $\Delta V_{\mathrm{TH}}^{\mathrm{NBTI}}$ gives a total V_{TH} shift for a pMOS transistor, because the silicon measurement of Chapter 4 showed that the V_{TH} variation of fresh transistors and their degradations are independent with each other. Then, performance of the circuit (e.g. noise margin) with the variability of the transistors is calculated using a transistor-level simulator such as SPICE. Finally, a pass or fail label, i.e. the value of the index function $I(\boldsymbol{x})$, is obtained using the calculated performance value.

6.3.2 Overview of the proposed method

A failure probability of an SRAM cell that includes the impact of NBTI can be calculated as

$$P_{\mathrm{fail}}(t_{\mathrm{age}}) = \int I(\boldsymbol{x}|t_{\mathrm{age}})P(\boldsymbol{x})d\boldsymbol{x}. \tag{6.12}$$

Here, $\boldsymbol{x}$ is a vector of random variable and t_{age} is a chip age. $I(\boldsymbol{x}|t_{\mathrm{age}})$ is an indicator function that returns "1" when the variability $\boldsymbol{x}$ causes a malfunction of the circuit whose age is given by t_{age}. From the above discussion, all the random variables now follow normal distributions. Hence, $\boldsymbol{x}$ follows a multi-dimensional standard normal distribution.

Let us see how the failure region in the variability space of $\Delta V_{\mathrm{TH}}^{\mathrm{fresh}}$ changes as the chip age increases. Sample points in Fig. 6.4 show $\Delta V_{\mathrm{TH}}^{\mathrm{fresh}}$ of failure cells. Here, the variabilities of two pMOS transistors are considered only for simplicity. The black markers show those of fresh cells while the red markers show those of 5-year-old cells. In this example, an SRAM cell who has negative read noise margin is labeled as a failure cell. We notice that there is no drastic change in $\Delta V_{\mathrm{TH}}^{\mathrm{fresh}}$ between fresh and the aged transistors. In the proposed method, the alternative distribution are "reused" by continuously modifying it among the multiple failure probability calculations along aging time steps. It eliminates the time consuming initial failure-region explorations conducted repeatedly for different chip ages.

The following steps are the overview of the proposed method. Algorithm 1 summarizes the calculation flow.

Initialization Initialize particles positions. The variability space is explored in the radial direction to find the failure regions that are close to the origin. Then, the particles are generated around the failure regions ((1) in Algorithm 1).

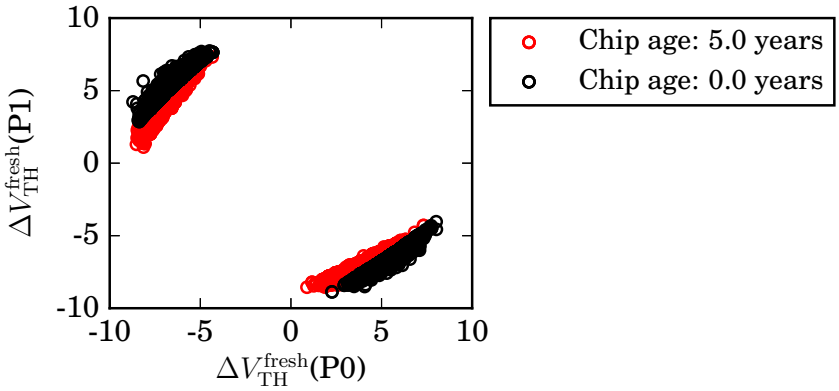


Figure 6.4: Temporal change of failure samples in the variability space of a fresh V_{TH}.

Algorithm 1 Proposed yield estimation algorithm.

1: (1) **Initial sample selection**
2: **for each** chip age **in** list of ages **do**
3: **repeat**
4: (2) **Prediction**
5: (3) **Measurement**
6: (4) **Resampling**
7: **until** Sufficient convergence of the particle density is achieved
8: (5) **Importance sampling:** construct the alternative distribution and calculate the failure probability at the current aging time step.
9: **end for**

Particle filter (first stage) The locations of the particles are then iteratively adjusted so that they best fit the density of an optimal alternative distribution ((2) to (4) in Algorithm 1).

Importance sampling (second stage) A large number of random samples is generated according to the density of the particles and the failure probability is calculated accurately ((5) in Algorithm 1).

In the failure probability calculation of the second or later aging time steps, the initialization step is skipped. Instead, the particles are copied from the previous calculation and the steps (2) to (4) are conducted so that the positions of the particles are adjusted.

6.3.3 Detailed procedures of the proposed method

(1) Initial sample selection

Random samples on the surface of a D-dimensional unit sphere are generated. The boundary of the failure region is searched using bi-section algorithm along the radial directions of the generated random samples. Candidates of initial particles $\{\boldsymbol{x}^{(0,i)}; i = 1, 2, \cdots, N\}$ are allocated near the boundary as shown in Fig. 6.5(a). Here, N is the total number of particles and $\boldsymbol{x}^{(t,i)}$ is the i-th particle at t-th iteration. Note again that the initialization step is conducted only once. In the failure probability calculations of the succeeding aging time steps, particles are copied from the previous calculations.

Steps from "prediction" to "resampling" are repeated to let the particles to follow the alternative distribution. In this experiment, five to ten times of repetitions are sufficient to achieve convergence in the estimated probability.

(2) Prediction step

The candidate particles at the next iteration $\{\widehat{\boldsymbol{x}}^{(t+1,i)}; i = 1, 2, \cdots, N\}$ are drawn from a mixture of normal distributions:

$$\widehat{\boldsymbol{x}}^{(t+1,i)} \sim \frac{1}{N} \sum_{j=1}^{N} \mathcal{N}_{\mathcal{D}}(\boldsymbol{x}^{(t+1,i)} | \boldsymbol{x}^{(t,j)}, \boldsymbol{\sigma}). \tag{6.13}$$

Here, $\mathcal{N}_{\mathcal{D}}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\sigma})$ is a D-dimensional normal distribution whose PDF is given by

$$\mathcal{N}_{\mathcal{D}}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\sigma}) = \frac{1}{\sqrt{2\pi^D |\boldsymbol{\sigma}|}} \exp\left(-\frac{1}{2} (\boldsymbol{x} - \mu)^T \boldsymbol{\sigma}^{-1} (\boldsymbol{x} - \mu) \right), \tag{6.14}$$

where $\boldsymbol{\sigma}$ is a diagonal covariance matrix assuming the "whitening" process.

(3) Measurement

For each generated candidate particle $\widehat{\boldsymbol{x}}^{(t+1,i)}$, weight $w^{(t+1,i)}$, which is a scalar value representing the fitness of the particle to the optimal alternative distribution, is calculated:

$$w^{(t+1,i)} = I(\widehat{\boldsymbol{x}}^{(t+1,i)} | t_{\text{age}}) P(\widehat{\boldsymbol{x}}^{(t+1,i)}). \tag{6.15}$$

Here, $P(\boldsymbol{x})$ is the probability that the sample $\boldsymbol{x}$ is observed. Because all of the random variables are assumed to follow the Gaussian distribution, $P(\boldsymbol{x})$ is represented by the D-dimensional standard normal distribution whose PDF is given by (6.14) with an identity covariance matrix.

For the computation of $I(\boldsymbol{x}|t_{\mathrm{age}})$, transistor-level simulations are required. N samples need to be simulated for the weight calculations of all particles. In this implementation, the number of simulations is reduced with the help of the SVM-based classifier. First, K training examples are randomly selected from N samples and give labels to them using transistor-level simulations. Then, the classifier is trained using the K training examples. Finally, the remaining $N - K$ samples are classified using the trained classifier. Hence, the number of transistor-level simulations can be reduced from N to K.

SVM-based classifier was first applied to a failure probability calculation in [73]. As we have seen, the drawback of the naive MC is that very few failure samples are drawn from the original PDF. Therefore, the authors of [73] used SVM-based classifier as a blockade so that they can skip transistor-level simulations of samples that obviously fall outside the failure region.

The difference between the proposed method and [73] is that the proposed method combines the classifier with the importance sampling. In the context of a failure probability calculation, the variability space is not equally important, i.e. $I(\boldsymbol{x}) \cdot P(\boldsymbol{x})$ represents the importance of the corresponding region. Misclassification of samples, which are rarely drawn from the alternative distribution, have almost no impact on the failure probability. Hereafter, the number of misclassification weighted by $I(\boldsymbol{x}) \cdot P(\boldsymbol{x})$ is called as "effective misclassification rate."

In this proposal, the SVM-based classifier is trained using the samples drawn from the alternative distribution. Hence, the training examples are naturally arranged around the regions that have a large impact on the effective misclassification rate. Compared to the case that the training is conducted with uniformly selected examples, the proposed approach enables to improve the classification performance with smaller number of training examples.

The computational cost required to train the classifier increases by $\mathcal{O}(n^2)$ where the number of training samples is n. To cover the failure regions of fresh and aged cells with a single classifier, a large number of training samples is required and eventually the training time of the classifier becomes unignorable. In this implementation, the old training samples are discarded and the binary classifier is newly trained for each aging time step to save the number of training samples and the time required for the training.

In order to construct a non-linear classification model, a polynomial transform of the variability vector $\boldsymbol{x}$ is used as feature quantities $\boldsymbol{f}$ in (6.6). For example, for a two-dimensional input vector $[x_1, x_2]$, the feature vector is $[1, x_1, x_2, x_1x_2, x_1^2, x_2^2]$ when degree of the polynomial transform D_{poly} is two. In this implementation, D_{poly} is set to be four.

Samples that are close to the separating hyper-plane, i.e. colored region in Fig. 6.2, may be misclassified depending on the accuracy of the classifier. Such

samples should be better classified on the basis of transistor-level simulations. However, the weights of particles do not have direct impact on the failure probability calculation. Instead, it only affects to the estimation of the optimal alternative distribution and the efficiency of the importance sampling. A rough approximation of $I(\boldsymbol{x}|t_{\mathrm{age}})$ is sufficient in this step. Hence, the transistor-level simulations can be safely skipped and all of the remaining $N-K$ samples can be classified using the trained classifier to reduce the calculation time.

Figure 6.5(b) shows particles after the prediction and measurement steps. The marker color represents the weight of each particle. We notice that the particles located closer to the origin, where the failure is more likely to occur, are assigned larger weights.

(4) Resampling

Particles at the next iteration step ($\{\boldsymbol{x}^{(t+1,i)}; i = 1, 2, \cdots, N\}$) are randomly selected from the candidates of particles $\widehat{\boldsymbol{x}}^{(t+1,i)}$ according to the probability in proportion to their weights. An example result of the resampling step is shown in Fig. 6.5(c).

While particle filters drastically speed up the estimation of the alternative distribution, a degeneration problem of particles have to be addressed. In the failure probability calculation of an SRAM cell, there are two major failure regions because of its symmetric structure. Small difference of the particle weight can make particles to concentrate on one of the two regions as the number of iterations increases. This leads to underestimation of the failure probability and thus it should be avoided. In the proposed method, multiple particle filters are utilized to track the failure regions. The resampling of particles is conducted for each particle filter respectively in order to avoid the concentration of particles into a smaller number of regions. In the example in Fig. 6.5(c), the two major failure regions are tracked by different particle filters. In this implementation, 10 particle filters are used.

(5) Importance sampling

Finally, in the second stage, the failure probability is calculated using an importance sampling. In order to optimize the alternative distribution, the outcome of the previous stage is used. Specifically, the distribution of the particles in step (4) is very close to the optimal distribution, hence it is approximated as

$$\widehat{Q}(\boldsymbol{x}) \approx \frac{1}{N} \sum_{i=1}^{N} \mathcal{N}_{\mathcal{D}}(\boldsymbol{x}|\boldsymbol{x}^{(t,i)}, \boldsymbol{\sigma}). \tag{6.16}$$

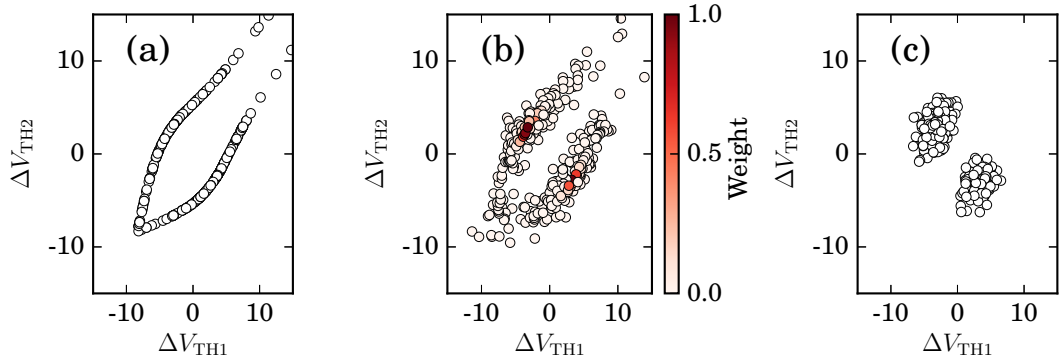


Figure 6.5: An example of particle filter based failure region tracking. (a) Particles after initialization step, (b) after prediction and weight calculation steps and (c) after resampling step.

Then, the failure probability is calculated using random samples $\{\boldsymbol{x}_{\mathrm{IS}}^k,\ k = 1, 2, \ldots, N_{\mathrm{IS}}\}$ drawn from $\widehat{Q}(\boldsymbol{x})$ as follows:

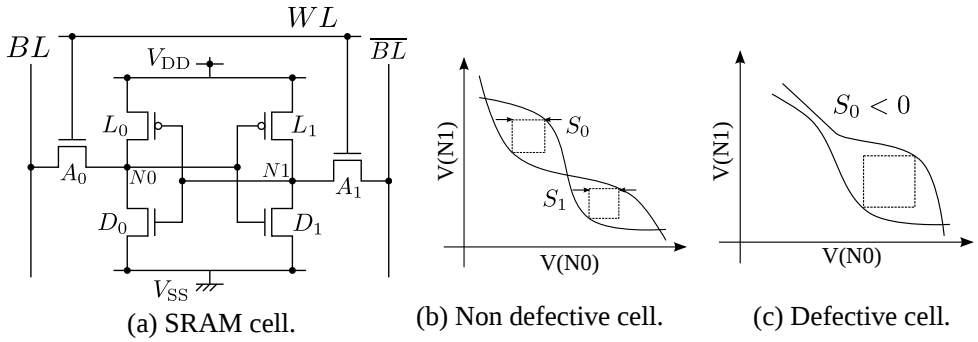
$$P_{\mathrm{fail}} \approx \frac{1}{N_{\mathrm{IS}}} \sum_{k=1}^{N_{\mathrm{IS}}} I(\boldsymbol{x}_{\mathrm{IS}}^k | t_{\mathrm{age}}) P(\boldsymbol{x}_{\mathrm{IS}}^k) / \widehat{Q}(\boldsymbol{x}_{\mathrm{IS}}^k). \tag{6.17}$$

Here, N_{IS} is the number of random samples used for the approximation. The calculation of the indicator function is again needed in the evaluation of $I(\boldsymbol{x}_{\mathrm{IS}}^k | t_{\mathrm{age}})$. In order to reduce the number of simulations, the SVM-based binary classifier is again used. Contrary to the classification in the first stage, classification accuracy in the second stage directly impacts on the accuracy of the failure probability calculation. Therefore, the samples which lie close to the separating hyper-plane go through the transistor-level simulations to obtain correct labels. The simulated samples are used to incrementally train the classifier and to increase the classification performance.

6.4 Numerical Experiment

6.4.1 Experimental setup

Figure 6.6 shows the circuit schematic of an SRAM cell. In the experiment, failure samples are defined as samples which have negative read noise margin (RNM). RNM is a stability measure of the cell [74]. It can be computed as $\min(\max(S_0), \max(S_1))$, where S_0 and S_1 are the lengths of squares embedded within the two openings of the butterfly curve as shown in Fig. 6.6(b). Two



(a) SRAM cell. (b) Non defective cell. (c) Defective cell.

Figure 6.6: (a) The schematics of the SRAM cell and (b) examples of static noise margin for a non-defective and (c) a defective cell.

examples of RNS for defective and non-defective cells are shown in Figs. 6.6(b) and (c). The mismatch of driving abilities among transistors results in negative noise margin, which causes the read failure.

In the experiment, the 16 nm high-performance model from predictive technology model (PTM) [75] is used as a transistor model. Long-term V_{TH} degradation is predicted using the following model that is slightly modified from the original RD-model in (2.1) to transform the time scale:

$$\Delta V_{\mathrm{TH}}^{\mathrm{NBTI}} = k \cdot (C_t \cdot t_{\mathrm{age}})^n \,, \tag{6.18}$$

where t_{age} is a chip age in year and C_t is a constant value to adjust the time scale. The model parameters are selected so that approximately 20 mV to 30 mV of V_{TH} shift is observed in the transistors of 5-years old. In this particular case, k and C_t are assumed to be 2×10^3 and 2×10^5, respectively. The logarithm of the power-law exponent ($n_{\log}$) is assumed to follow a normal distribution (6.7), where μ_n and σ_n are set to -1.3 and 0.1, respectively. Fig. 6.7 shows example of V_{TH} shifts of 50 pMOS transistors assumed in this experiment. Other circuit parameters such as gate length and width are summarized in Table 6.1. In order to see the temporal change of the failure probability at the early ages, in which V_{TH} rapidly increases, failure probabilities at the following aging time steps are calculated: $t_{\mathrm{age}} = 0$, 0.01, 0.03, 0.05, 0.1, 0.3, 0.5, 1.0, 2.0, 3.0, 4.0, 5.0 years.

6.4.2 Experimental results

Firstly, the proposed method is compared with one of the state-of-the-art methods proposed in [39]. Note that, in this experiment, the failure probability at a single aging time step is calculated to see the effectiveness of the two-stage MC and the SVM-based binary classifier. Fig. 6.8(a) shows the calculated failure probability of the SRAM cell and required calculation time. The calculation

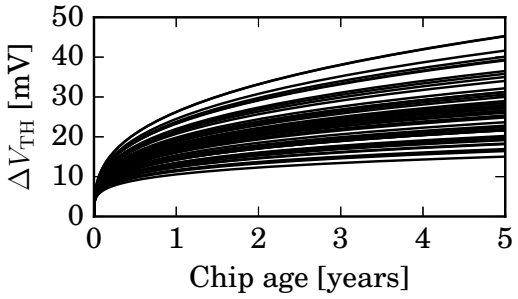


Figure 6.7: Examples of NBTI-induced V_{TH} shift assumed in this experiment.

Table 6.1: Experimental conditions

	Load (L_i)	Driver (D_i)	Access (A_i)
$A_{\mathrm{V_{TH}}}$ [mV·nm]	4×10^2		
Channel length [nm]	16		
Channel width [nm]	30	50	30
k [V/sec]	2×10^{-3}		
μ_n	-1.3		
σ_n	0.1		

time was measured on a Linux workstation with 6-core Xeon X5570 processor operated at 2.93 GHz. Among 6-cores, 4-cores are used to accelerate the calculation. In Fig. 6.8(a), the shaded regions represent the 95% confidence intervals. We can see that the proposed method converges faster than the conventional method. Fig. 6.8(b) shows relative error as a function of calculation time. The relative error is defined as the ratio of the 95% confidence interval to the calculated failure probability. In this experimental setup, the proposed method required about 657 seconds to achieve the relative error of 5% while the conventional method required 1,530 seconds to achieve the equal accuracy, which means that the proposed method achieved 2.3× speed-up. Note that the calculation time of the proposed method includes the training time of the classifier and classification. When the acceptable error is small, the difference in the calculation time becomes large. For example, the proposed method can achieve 16× speed up over the conventional method when the permissible error is set to 1%. The breakdown of the calculation time required to obtain 5% error is summarized in Tab 6.2.

Then, the temporal change of the failure probability is calculated with the proposed method. Fig. 6.9(a) summarizes the result. As a comparison, the result of the conventional method is shown in Fig. 6.9(b). The permissible error is set to 5%. We can see that the results of both methods are almost

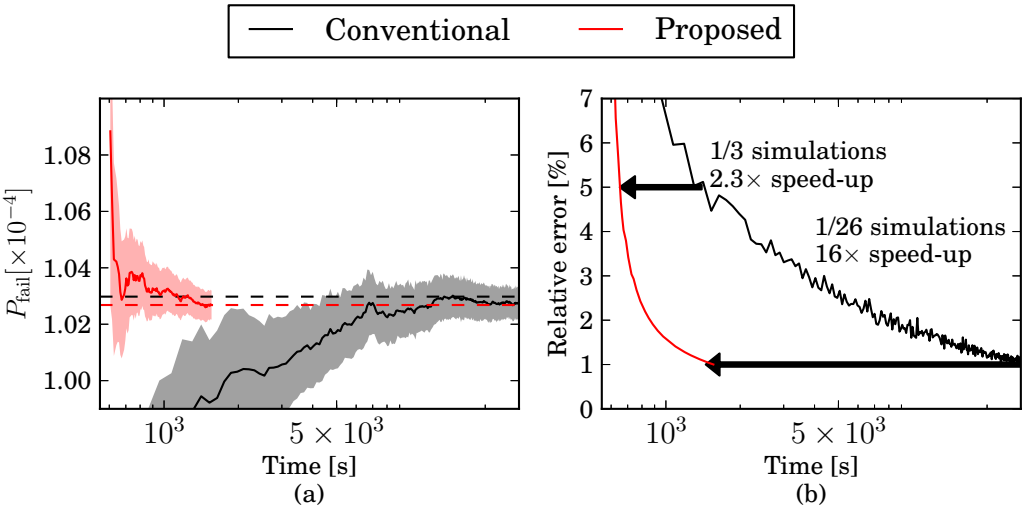


Figure 6.8: The comparison of the proposed and the conventional [39] methods. (a) The relationship between the calculated failure probability and the calculation time required. (b) The relationship between the relative error and the calculation time required.

Table 6.2: Detailed breakdown of the calculation time.

	Initial sample selection	Importance sampling
Proposed method	162 seconds	657 seconds
Conventional method [39]	162 seconds	1,530 seconds

equal, from which we can confirm the correctness of the proposed method. The total calculation time required to obtain Fig. 6.9 is about 6,980 seconds for the proposed method while 68,100 seconds for the conventional method. Hence, the proposed method achieved 9.76× speed up compared to the conventional method. The magnitude of the speed up is increased from the comparison in Fig. 6.8 because the comparison in Fig. 6.9 includes the effect of the particle reusing. In the conventional method, the alternative distribution is constructed from scratch at each aging time step while in the proposed method, the construction is conducted only once, contributing the further reduction of the total calculation time.

6.5 Conclusion

This chapter proposed a novel method to efficiently calculate the failure probability of an SRAM cell that can take the impact of NBTI-induced

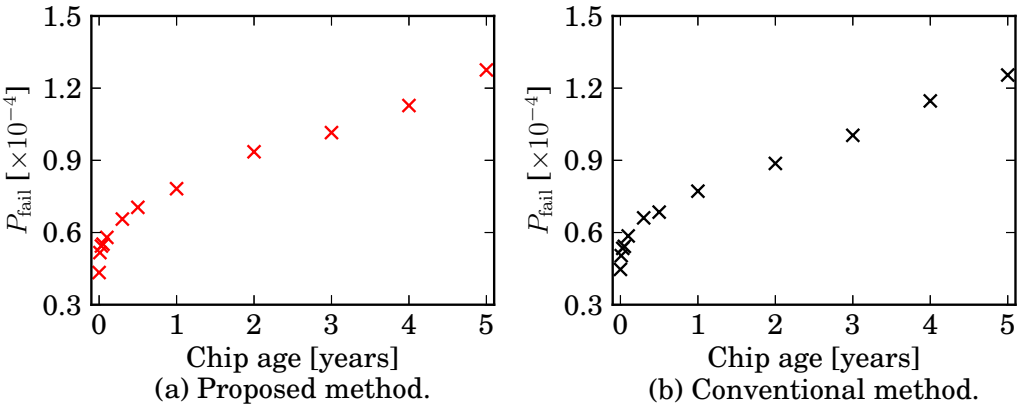


Figure 6.9: The temporal change in the failure probability.

device degradation into account. To see the temporal change of the failure probability at different device ages, multiple failure probability calculations are required by changing V_{TH} of transistor at an aging time step. Considering the gradual V_{TH} change due to aging, sequential Monte Carlo method is proposed, which utilizes particle filter to keep track the change of the near optimal alternative distribution for importance sampling. With this idea, time consuming repetitive explorations in the variability space has been eliminated. Combined with a binary classifier and two-stage MC approach to further reduce the calculation time, the proposed method achieved 9.76× speed-up over one of the state-of-the-art method [39]. With the aid of the proposed method, circuit designers can efficiently see the impact of device degradation, which finally improves reliabilities of highly scaled LSIs and contributes to broader adoption of LSIs.

Chapter 7

Conclusion

7.1 Summary of this dissertation

Since the invention of integrated circuits, the application area of the semiconductor devices has been continuously expanding. Now, our daily lives rely overwhelmingly on the semiconductor devices. Examples include mission critical systems such as self driving cars, medical equipments, or financial systems. Because failures of these system may lead to serious incidents, high-level of reliability is required for the semiconductor devices.

On the other hand, as the device dimension shrinks, ensuring circuit reliability has been emerged as an urgent issue. The thickness of the gate insulator film is now approaching few nano meter order and the increasing electric field causes various reliability problems. In particular, device degradation called bias temperature instability (BTI) attracts increasing attention.

The physical mechanism of BTI is still a controversial topic. Various models have been proposed to predict the BTI-induced V_{TH} increase but none of them fully explains the BTI-induced V_{TH} degradation. Therefore, a good experimental data is vital for the full understanding of the BTI phenomenon. Moreover, increasing manufacturing variability further complicates the problem. Due to the variability, BTI degradations are also expected to vary transistor to transistor. Hence, in order to ensure circuit reliability, the statistical characterization of BTI-induced device degradation is vitally important.

Considering these circumstances, a circuit structure suitable for the BTI measurement is proposed in Chapter 3. The difficulty for BTI measurement is that the measurement takes very long time. Hours or even days of stress period is required to invoke the noticeable V_{TH} shift, making the statistical observation of BTI-induced degradation almost impossible. Hence, the proposed circuit structure named BTIarray utilizes the stress pipelining scheme in which

the BTI measurement is overlapped among transistors to shorten measurement time.

During the measurement of BTIarray, the stair-like change in the V_{TH} is frequently observed, which indicates that capturing and emissions of carriers deeply involved in the BTI-induced device degradation. A close examination of such stair-like V_{TH} shift may lead to further understanding of BTI and random telegraph noise (RTN). The trap can be characterized using two parameters: a magnitude of V_{TH} shift and time constants. Usually, single transistor includes two or more number of traps and a summation of their activities is observed as V_{TH} shift of the transistor. Hence, in order to separately examine the characteristics of each trap activity, the activity must be individually extracted from the observed V_{TH} shift, which is an under-determined problem. In Chapter 5, this problem is addressed by using statistical machine learning methodology. The model that represent the generation process of the stair-like V_{TH} shift is developed and used as a supplemental information. The model parameters, such as the magnitudes of V_{TH} shift for each trap, are estimated so that the V_{TH} predicted by the model best fits the observed V_{TH}. The numerical experiment using synthetic V_{TH} waveforms showed that the proposed method successfully separated three trap activities from V_{TH} waveforms. In order to apply the proposed method for real world data, the confidence of the estimation plays an important role because users can filter the estimation results that are possibly failed in the parameter estimation. For this purpose, the methodology to assess the confidence of the estimation based on the posterior probability was developed. Finally, the proposed method is validated using the V_{TH} waveform obtained in the measurement of BTIarray and it is shown that the proposed method successfully separated two trap activities and that the confidence of the estimation well reflected the goodness of the estimation.

In order to increase the reliability of the semiconductor circuit, only measuring BTI on transistors or only constructing degradation models is insufficient. Circuit designers are interested in how the device degradation have impacts on their circuit reliability, such as an SRAM bit cell. To meet this demand, a method that can efficiently estimate the temporal degradation of the SRAM failure probability is proposed in Chapter 6. Modern processors embeds billions of SRAM bit cells and hence extremely high level of reliability is required for each bit cell. Estimation of such rare failure event is known to be a non-trivial task even if the device degradations are not considered. Because random samples drawn from the variability distribution rarely cause the circuit failure, thousands or millions of circuit simulations are required to obtain a single failure sample, making the application of the traditional MC almost impossible. Importance sampling, in which random samples are generated from biased distribution, also cannot be used due to temporal change

of the device characteristics caused by BTI. In order to solve this problem, this dissertation proposed an sequential importance sampling technique in which particles moving around the variability space are used to track the temporal change of the biased distribution. Numerical experiment showed that the proposed method achieved 9.76× speed up compared to conventional approach.

This dissertation especially focused on the variability of the device degradation. The efficient technique to characterize the device degradation on vast amount of transistors for constructing the statistical degradation model was proposed. On top of the silicon measurement, this dissertation further proposed the method to predict the failure probability of an SRAM bit cell. The proposed method can take into account the device degradation efficiently, which contributes to the accurate estimation of the long-term circuit reliability. With the proposed techniques, excessive design margins, which are required in the current circuit design to ensure the long-term reliability, can be safely reduced, contributing to the further cost reductions or performance increases.

7.2 Future prospects

With the proposed BTIarray, the variability of device degradation can be measured within a reasonable period of time. However, there still are remaining work to improve BTIarray. One of the possible improvements includes the integration of on-chip analog to digital converter (ADC). The degraded V_{TH} recovers so fast that it cannot be captured with the off-chip measurement equipments. Another potential improvement is to modify the switch structure of BTIarray so as to increase the variety of the stress and recovery conditions. For example, in the current implementation of BTIarray, the source and drain terminals of a DUT are shorted when the stress bias is applied to the DUT. Hence, the source and the drain voltages cannot be controlled independently, limiting the variety of measurement scenario.

Based on the measured BTI-induced degradation, BTI-aware failure probability estimation method is developed and validated using the SRAM bit cell example. However, the impact of BTI on other circuit components, such as DFF, is still difficult to estimate with the proposed method. The problem originates from the "course of dimensionality." When the number of random variables that needs to be included is large, exponentially increasing number of random samples are required to probe the variability space, limiting target of the proposed method to small circuits such as an SRAM cell. To break the simulation barrier, novel approach called subset simulation (SubSim) is developed [76, 77]. SubSim is known to be quite robust against the number of random variables. Therefore, with the aid of SubSim, it is expected to estimate

the impact of BTI on more complex circuits such as DFF.

Combining these improvements on both measurement and circuit analysis techniques, the physical mechanism of BTI and its impact on the circuit are expected to be investigated more closely. Based upon these observations, the long term circuit reliability of LSIs will further increase, contributing to extend the application area of LSIs.

Bibliography

[1] R. Degraeve, G. Groeseneken, R. Bellens, J. Ogier, M. Depas, P. Roussel, and H. Maes, "New insights in the relation between electron trap generation and the statistical properties of oxide breakdown," *IEEE Trans. Electron Devices*, vol. 45, no. 4, pp. 904–911, Apr 1998.

[2] E. Takeda and N. Suzuki, "An empirical model for device degradation due to hot-carrier injection," *IEEE Electron Device Lett.*, vol. 4, no. 4, pp. 111–113, Apr 1983.

[3] D. K. Schroder and J. A. Babcock, "Negative bias temperature instability: Road to cross in deep submicron silicon semiconductor manufacturing," *J. Appl. Physics*, vol. 94, no. 1, 2003.

[4] J. H. Stathis and S. Zafar, "The negative bias temperature instability in MOS devices: A review," *Microelectronics Rel.*, vol. 46, no. 2-4, pp. 270–286, 2006.

[5] S. Zafar, A. Kumar, E. Gusev, and E. Cartier, "Threshold voltage instabilities in high-κ gate dielectric stacks," *IEEE Trans. Device Mater. Rel.*, vol. 5, no. 1, pp. 45–64, March 2005.

[6] M. J. Uren, D. J. Day, and M. J. Kirton, "1/f and random telegraph noise in silicon metal - oxide - semiconductor field - effect transistors," *Appl. Physics Lett.*, vol. 47, no. 11, 1985.

[7] K. Aadithya, S. Venogopalan, A. Demir, and J. Roychowdhury, "MUSTARD: A coupled, stochastic/deterministic, discrete/continuous technique for predicting the impact of Random Telegraph Noise on SRAMs and DRAMs," in *Proc. of Symp. on VLSI Technology*, 2011, pp. 292–297.

[8] T. Grasser, B. Kaczer, W. Goes, H. Reisinger, T. Aichinger, P. Hehenberger, P.-J. Wagner, F. Schanovsky, J. Franco, M. Luque, and M. Nelhiebel, "The Paradigm Shift in Understanding the Bias Temperature

Instability: From Reaction-Diffusion to Switching Oxide Traps," *IEEE Trans. Electron Devices*, vol. 58, no. 11, pp. 3652–3666, Nov 2011.

[9] G. Kitagawa, "Monte Carlo Filter and Smoother for Non-Gaussian Nonlinear State Space Models," *J. Comput. and Graphical Stat.*, vol. 5, no. 1, pp. 1–25, 1996.

[10] N. Gordon, D. Salmond, and A. F. M. Smith, "Novel approach to nonlinear/non-Gaussian Bayesian state estimation," *Radar and Signal Processing, IEE Proc. F*, vol. 140, no. 2, pp. 107–113, 1993.

[11] Y. Rui and Y. Chen, "Better proposal distributions: object tracking using unscented particle filter," in *Proc. of Int. Conf. on Comput. Vision and Pattern Recognition*, vol. 2, 2001, pp. II–786–II–793 vol.2.

[12] K. O. Jeppson and C. M. Svensson, "Negative bias stress of MOS devices at high electric fields and degradation of MNOS devices," *J. of Appl. Physics*, vol. 48, no. 5, pp. 2004–2014, 1977.

[13] K. Sutaria, J. Velamala, C. Kim, T. Sato, and Y. Cao, "Aging Statistics Based on Trapping/Detrapping: Compact Modeling and Silicon Validation," *IEEE Trans. Device Mater. Rel.*, vol. 14, no. 2, pp. 607–615, 2014.

[14] T. Sato, T. Kozaki, T. Uezono, H. Tsutsui, and H. Ochi, "A device array for efficient bias-temperature instability measurements," in *Proc. of European Solid-State Device Research Conf.*, 2011, pp. 143–146.

[15] A. Ortis-Conde, F. J. García Sánchez, J. J. Liou, A. Cerdeira, M. Estrada, and Y. Yue, "A review of recent MOSFET threshold voltage extraction methods," *Microelectronics Rel.*, vol. 42, pp. 583–596, 2002.

[16] H. Reisinger, O. Blank, W. Heinrigs, A. Muhlhoff, W. Gustin, and C. Schlunder, "Analysis of NBTI degradation- and recovery-behavior based on ultra fast VT-measurements," in *Proc. of Int. Rel. Physics Symp.*, Mar. 2006, pp. 448–453.

[17] M. Denais, C. Parthasarathy, G. Ribes, Y. Rey-Tauriac, N. Revil, A. Bravaix, V. Huard, and F. Perrier, "On-the-fly characterization of NBTI in ultra-thin gate oxide PMOSFET's," in *Tech. Dig. of Int. Electron Device Meeting*, Dec. 2004, pp. 109–112.

[18] E. Kumar, V. Maheta, S. Purawat, A. Islam, C. Olsen, K. Ahmed, M. Alam, and S. Mahapatra, "Material Dependence of NBTI Physical Mechanism in Silicon Oxynitride (SiON) p-MOSFETs: A Comprehensive

Study by Ultra-Fast On-The-Fly (UF-OTF) IDLIN Technique," in *Tech. Dig. of Int. Electron Device Meeting*, Dec 2007, pp. 809–812.

[19] G. Du, D. Ang, Z. Teo, and Y. Hu, "Ultrafast measurement on NBTI," *IEEE Electron Device Lett.*, vol. 30, no. 3, pp. 275–277, 2009.

[20] T. Matsumoto, H. Makino, K. Kobayashi, and H. Onodera, "A 65 nm complementary metal-oxide-semiconductor 400 ns measurement delay negative-bias-temperature-instability recovery sensor with minimum assist circuit," *Japanese J. Appl. Physics*, vol. 50, no. 4S, 04DE06, 2011.

[21] A. Ghosh, R. M. Rao, R. B. Brown, and C.-T. Chuang, "On-chip negative bias temperature instability sensor using slew rate monitoring circuitry," in *Proc. of Int. Symp. on Low Power Electronics and Design*, 2009.

[22] J. J. Kim, R. Rao, J. Schaub, A. Ghosh, A. Bansal, K. Zhao, B. Linder, and J. Stathis, "PBTI/NBTI monitoring ring oscillator circuits with on-chip Vt characterization and high frequency AC stress capability," in *Proc. of Symp. on VLSI Technology*, 2011, pp. 224–225.

[23] T. Kim, R. Persaud, and C. Kim, "Silicon odometer: An on-chip reliability monitor for measuring frequency degradation of digital circuits," *IEEE J. Solid-State Circuits*, vol. 43, no. 4, pp. 874–880, 2008.

[24] S. Fujimoto, A. Islam, T. Matsumoto, and H. Onodera, "Inhomogeneous Ring Oscillator for Within-Die Variability and RTN Characterization," vol. 26, no. 3, pp. 296–305, 2013.

[25] T. Nagumo, K. Takeuchi, S. Yokogawa, K. Imai, and Y. Hayashi, "New analysis methods for comprehensive understanding of random telegraph noise," in *Tech. Dig. of Int. Electron Device Meeting*, Dec. 2009, pp. 1–4.

[26] T. Nagumo, K. Takeuchi, T. Hase, and Y. Hayashi, "Statistical characterization of trap position, energy, amplitude and time constants by RTN measurement of multiple individual traps," in *Tech. Dig. of Int. Electron Device Meeting*, Dec. 2010, pp. 28.3.1–28.3.4.

[27] H. Miki, M. Yamaoka, N. Tega, Z. Ren, M. Kobayashi, C. P. D'Emic, Y. Zhu, D. J. Frank, M. A. Guillorn, D. Park, W. Haensch, and K. Torii, "Understanding short-term BTI behavior through comprehensive observation of gate-voltage dependence of RTN in highly scaled high-k metal-gate pFETs," in *Dig.Tech. Papers of Symp.on VLSI Tech.*, 2011, pp. 148–149.

[28] S. Reanov and K. L. Shepard, "Random telegraph noise in 45-nm CMOS: Analysis using an on-chip test and measurement system," in *Tech. Dig. of Int. Electron Device Meeting*, Dec. 2010, pp. 624–627.

[29] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Prentice Hall, Apr. 1993.

[30] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[31] E. Maricau and G. Gielen, "Stochastic circuit reliability analysis," in *Proc. of Design, Automation Test in Europe*, 2011, pp. 1–6.

[32] E. Maricau and G. Gielen, "Efficient Variability-Aware NBTI and Hot Carrier Circuit Reliability Analysis," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 29, no. 12, pp. 1884–1893, 2010.

[33] K. Kang, H. Kufluoglu, K. Roy, and M. Alam, "Impact of Negative-Bias Temperature Instability in Nanoscale SRAM Array: Modeling and Analysis," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 26, no. 10, pp. 1770–1781, 2007.

[34] R. Kanj, R. Joshi, and S. Nassif, "Mixture importance sampling and its application to the analysis of SRAM designs in the presence of rare failure events," in *Proc. of Symp. on VLSI Technology*, 2006, pp. 69–72.

[35] J. Jaffari and M. Anis, "Adaptive sampling for efficient failure probability analysis of SRAM cells," in *Proc. of Int. Conf. on Comput.-Aided Design*, 2009, pp. 623–630.

[36] L. Dolecek, M. Qazi, D. Shah, and A. Chandrakasan, "Breaking the simulation barrier: SRAM evaluation through norm minimization," in *Proc. of Int. Conf. on Comput.-Aided Design*, 2008, pp. 322–329.

[37] M. Rana and R. Canal, "SSFB: A highly-efficient and scalable simulation reduction technique for SRAM yield analysis," in *Proc. of Design, Automation Test in Europe*, 2014, pp. 1–6.

[38] C. Dong and X. Li, "Efficient SRAM Failure Rate Prediction via Gibbs Sampling," in *Proc. of Symp. on VLSI Technology*, 2011, pp. 200–205.

[39] K. Katayama, S. Hagiwara, H. Tsutsui, H. Ochi, and T. Sato, "Sequential importance sampling for low-probability and high-dimensional SRAM yield

analysis," in *Proc. of Int. Conf. on Comput.-Aided Design*, 2010, pp. 703–708.

[40] J. H. Stathis and S. Zafar, "The negative bias temperature instability in MOS devices: A review," *Microelectronics Rel.*, vol. 46, no. 2-4, pp. 270–286, 2006.

[41] S. E. Rauch, "Review and reexamination of reliability effects related to NBTI-induced statistical variations," *IEEE Trans. Device Mater. Rel.*, vol. 7, no. 4, pp. 524–529, 2007.

[42] D. K. Schroder, "Negative bias temperature instability: What do we understand?" *Microelectronics Rel.*, vol. 47, no. 6, pp. 841–852, 2007.

[43] W. Wang, V. Reddy, A. Krishnan, R. Vattikonda, S. Krishnan, and Y. Cao, "Compact modeling and simulation of circuit reliability for 65-nm CMOS technology," *IEEE Trans. Device Mater. Rel.*, vol. 7, no. 4, pp. 509–517, 2007.

[44] J. B. Velamala, K. B. Sutaria, T. Sato, and Y. Cao, "Aging Statistics Based on Trapping/Detrapping: Silicon Evidence, Modeling and Long-Term Prediction," in *Proc. of Int. Rel. Physics Symp.*, 2012, pp. 2F.2.1–2F.2.5.

[45] J. Velamala, K. Sutaria, T. Sato, and Y. Cao, "Physics matters: Statistical aging prediction under trapping/detrapping," in *Proc. of Symp. on VLSI Technology*, 2012, pp. 139–144.

[46] J. Campbell, P. Lenahan, A. Krishnan, and S. Krishnan, "Observations of NBTI-induced atomic-scale defects," *IEEE Trans. Device Mater. Rel.*, vol. 6, no. 2, pp. 117–122, 2006.

[47] B. Kaczer, T. Grasser, P. Roussel, J. Martin-Martinez, R. O'Connor, B. O'Sullivan, and G. Groeseneken, "Ubiquitous relaxation in BTI stressing — new evaluation and insights," in *Proc. of Int. Rel. Physics Symp.*, 2008, pp. 20–27.

[48] T. Grasser, H. Reisinger, P. Wagner, F. Schanovsky, W. Goes, and B. Kaczer, "The time dependent defect spectroscopy (TDDS) for the characterization of the bias temperature instability," in *Proc. of Int. Rel. Physics Symp.*

[49] H. Reisinger, T. Grasser, W. Gustin, and C. Schlunder, "The statistical analysis of individual defects constituting NBTI and its implications for

modeling DC-and AC-stress," in *Proc. of Int. Rel. Physics Symp.*, 2010, pp. 7–15.

[50] T. Sato, H. Ueyama, N. Nakayama, and K. Masu, "Accurate array-based measurement for subthreshold-current of MOS transistors," *IEEE J. Solid-State Circuits*, vol. 44, no. 11, pp. 2977–2986, 2009.

[51] V. Huard and M. Denais, "Hole trapping effect on methodology for DC and AC negative bias temperature instability measurements in pMOS transistors," in *Proc. of Int. Rel. Physics Symp.*, 2004, pp. 40–45.

[52] G. Wirth, R. da Silva, and B. Kaczer, "Statistical Model for MOSFET Bias Temperature Instability Component Due to Charge Trapping," *IEEE Trans. Electron Devices*, vol. 58, no. 8, pp. 2743–2751, 2011.

[53] H. Reisinger, O. Blank, W. Heinrigs, W. Gustin, and C. Schlunder, "A comparison of very fast to very slow components in degradation and recovery due to NBTI and bulk hole trapping to existing physical models," *IEEE Trans. Device Mater. Rel.*, vol. 7, no. 1, pp. 119–129, 2007.

[54] T. Matsumoto, H. Makino, K. Kobayashi, and H. Onodera, "Multicore large-scale integration lifetime extension by negative bias temperature instability recovery-based self-healing," *Japanese J. of Appl. Physics*, vol. 51, no. 4S, 04DE02, 2012.

[55] T. Tsunomura, J. Nishimura, A. Kumar, A. Nishida, S. Inaba, K. Takeuchi, T. Hiramoto, and T. Mogami, "Suppression of VT variability degradation induced by NBTI with RDF control," in *Dig.Tech. Papers of Symp.on VLSI Tech.*, June 2011, pp. 150–151.

[56] K. Ito, T. Matsumoto, S. Nishizawa, H. Sunagawa, K. Kobayashi, and H. Onodera, "Modeling of Random Telegraph Noise under circuit operation – Simulation and measurement of RTN-induced delay fluctuation," in *Proc. of Int. Symp. on Quality Electronic Design*, 2011, pp. 1–6.

[57] K. Aadithya, A. Demir, S. Venugopalan, and J. Roychowdhury, "SAMURAI: An accurate method for modelling and simulating non-stationary Random Telegraph Noise in SRAMs," in *Proc. of Design, Automation Test in Europe*, 2011, pp. 1–6.

[58] C. Monzio Compagnoni, R. Gusmeroli, A. Spinelli, A. Lacaita, M. Bonanomi, and A. Visconti, "Statistical Model for Random Telegraph Noise in Flash Memories," vol. 55, no. 1, pp. 388–395, 2008.

[59] C. Andrieu, N. de Freitas, A. Doucet, and M. Jordan, "An Introduction to MCMC for Machine Learning," *Mach. Learning*, vol. 50, no. 1-2, pp. 5–43, 2003.

[60] G. Casella and E. I. George, "Explaining the Gibbs Sampler," *The American Statistician*, vol. 46, no. 3, pp. 167–174, 1992.

[61] S. Geman and D. Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 6, no. 6, pp. 721–741, Nov. 1984.

[62] S. Moussaoui, D. Brie, A. Mohammad-Djafari, and C. Carteret, "Separation of Non-Negative Mixture of Non-Negative Sources Using a Bayesian Approach and MCMC Sampling," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4133–4145, Nov. 2006.

[63] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of State Calculations by Fast Computing Machines," *The J. of Chemical Physics*, vol. 21, no. 6, pp. 1087–1092, 1953.

[64] Y. Iba, "Extended Ensemble Monte Carlo," *Int. J. of Modern Physics C*, vol. 12, pp. 623–656, 2001.

[65] S. Reanov and K. L. Shepard, "Random telegraph noise in 45-nm CMOS: Analysis using an on-chip test and measurement system," in *Tech. Dig. of Int. Electron Device Meeting*, Dec. 2010, pp. 624–627.

[66] N. Tega, H. Miki, T. Osabe, A. Kotabe, K. Otsuga, H. Kurata, S. Kamohara, K. Tokami, Y. Ikeda, and R. Yamada, "Anomalously large threshold voltage fluctuation by complex random telegraph signal in floating gate flash memory," in *Tech. Dig. of Int. Electron Device Meeting*, Dec. 2006, pp. 1–4.

[67] M. Tanizawa, S. Ohbayashi, T. Okagaki, K. Sonoda, K. Eikyu, Y. Hirano, K. Ishikawa, O. Tsuchiya, and Y. Inoue, "Application of a statistical compact model for random telegraph noise to scaled-SRAM Vmin analysis," in *Dig.Tech. Papers of Symp.on VLSI Tech.*, Jun. 2010, pp. 95–96.

[68] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[69] K. Murphy, "Hidden markov model (HMM) toolbox for Matlab," http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html.

[70] A. Bhavnagarwala, X. Tang, and J. Meindl, "The impact of intrinsic device fluctuations on CMOS SRAM cell stability," *IEEE J. Solid-State Circuits*, vol. 36, no. 4, pp. 658–665, 2001.

[71] C. Cortes and V. Vapnik, "Support-Vector Networks," *Mach. Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[72] T. Mizutani, A. Kumar, and T. Hiramoto, "Analysis of transistor characteristics in distribution tails beyond $\pm 5.4\sigma$ of 11 billion transistors," in *Tech. Dig. of Int. Electron Device Meeting*, 2013, pp. 33.3.1–33.3.4.

[73] A. Singhee and R. Rutenbar, "Statistical Blockade: Very Fast Statistical Simulation and Modeling of Rare Circuit Events and Its Application to Memory Design," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 28, no. 8, pp. 1176–1189, 2009.

[74] E. Seevinck, F. List, and J. Lohstroh, "Static-noise margin analysis of MOS SRAM cells," *IEEE J. Solid-State Circuits*, vol. 22, no. 5, pp. 748–754, 1987.

[75] Nanoscale Integration and Modeling (NIMO) Group, ASU, "Predictive Technology Model (PTM)," http://ptm.asu.edu/.

[76] S. K. Au and J. L. Beck, "Estimation of small failure probabilities in high dimensions by subset simulation," *Probabilistic Eng. Mechanics*, vol. 16, no. 4, pp. 263–277, 2001.

[77] S. K. Au and J. L. Beck, "Subset Simulation and its Application to Seismic Risk Based on Dynamic Analysis," *J. of Eng. Mechanics*, vol. 129, no. 8, pp. 901–917, 2003.

List of Publications

Journals

1. Hiromitsu Awano, Hiroshi Tsutsui, Hiroyuki Ochi: "Bayesian Estimation of Multi-trap RTN Parameters using Markov Chain Monte Carlo Method," IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, Vol. E95-A, No. 12, pp. 2272-2283, Dec. 2012.

2. Hiromitsu Awano, Masayuki Hiromoto, Takashi Sato: "BTIarray: A Time-overlapping Transistor Array for Efficient Statistical Characterization of Bias Temperature Instability," IEEE Transactions on Device and Materials Reliability, Vol. 14, No. 3, pp. 833-843, Sep. 2014.

3. Hiromitsu Awano, Masayuki Hiromoto, Takashi Sato: "Efficient Aging-Aware SRAM Failure Probability Calculation via Particle Filter based Importance Sampling," IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, (under review (conditionally accepted)).

4. Jyothi Bhaskarr Velamala, Ketul B. Sutaria, Hirofumi Shimizu, Hiromitsu Awano, Takashi Sato, Gilson Wirth, and Yu Cao: "Compact Modeling of Statistical BTI Under Trapping/detrapping," IEEE Transactions on Electron Devices, Vol. 60, No. 11, pp. 3645-3654, Nov. 2013.

5. Hirofumi Shimizu, Hiromitsu Awano, Masayuki Hiromoto, Takashi Sato: "Automation of Model Parameter Estimation for Random Telegraph Noise," IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, Vol. E97-A, No. 12, pp. 2383-2392, Dec. 2014.

6. Ikkyu Aihara, Takeshi Mizumoto, Takuma Otsuka, Hiromitsu Awano, Kohei Nagira, Hiroshi G. Okuno, Kazuyuki Aihara: "Spatio-Temporal

Dynamics in Collective Frog Choruses Examined by Mathematical Modeling and Field Observation," Scientific Reports, 4:3891 Nature Publishing Group, 27 Jan. 2014.

Peer-reviewed conference

1. Hiromitsu Awano, Tetsuya Ogata, Shun Nishide, Toru Takahashi, Kazunori Komatani, Hiroshi G. Okuno: "Human-Robot Cooperation in Arrangement of Objects Using Confidence Measure of Neuro-dynamcal Systems," in Proc. International Conference on Systems, Man, and Cybernetics (SMC), pp.2533-2538, Istanbul, Turkey, Oct. 2010.

2. Hiromitsu Awano, Shun Nishide, Hiroaki, Arie, June Tani, Hiroshi G. Okuno, Tetsuya Ogata: "Use of a Sparse Structure to Improve Learning Performance of Recurrent Neural Networks," in Proc. International Conference on Neural Information Processing (ICONIP), pp.323-331, Shanghei, China, Nov. 2011.

3. Hiromitsu Awano, Hiroshi Tsutsui, Hiroyuki Ochi, Takashi Sato: "Multi-trap RTN Parameter Extraction based on Bayesian Inference," International Symposium on Quality Electrical Design (ISQED), Santa Clara, CA, pp.597-602, Mar. 2013.

4. Hiromitsu Awano, Masayuki Hiromoto, Takashi Sato: "Statistical Observation of NBTI and PBTI Degradations," Workshop on variability modeling and characterization (VMC), San Jose, CA, Nov. 2013.

5. Hiromitsu Awano, Masayuki Hiromoto, Takashi Sato: "Variability in Device Degradations: Statistical Observation of NBTI for 3996 Transistors," European Solid-State Device Research Conference (ESSDERC), Venice, Italy, pp.218-221, Sep. 2014.

6. Hiromitsu Awano, Masayuki Hiromoto, Takashi Sato: "ECRIPSE: An Efficient Method for Calculating RTN-Induced Failure Probability of an SRAM Cell," Design, Automation and Test in Europe (DATE), Grenoble, France, Mar. 2015

7. Hiromitsu Awano, Takashi Sato: "Fast Monte Carlo for Timing Yield Estimation via Line Sampling," Variability Modeling and Characterization (VMC), Austin, TX, Nov. 2015.

8. Hiromitsu Awano, Takashi Sato: "Efficient Transistor-level Timing Yield Estimation via Line Sampling," ACM International Workshop on Timing Issues in the Specification and Synthesis of Digital Systems (TAU), Santa Rosa, CA, Mar. 2016, accepted.

9. Hiromitsu Awano, Takashi Sato: "Efficient Transistor-level Timing Yield Estimation via Line Sampling," Design Automation Conference (DAC), Austin, TX, Jun. 2016, accepted.

10. Takashi Sato, Hiromitsu Awano, Hirofumi Shimizu, Hiroshi Tsutsui, and Hiroyuki Ochi: "Statistical Observations of NBTI-Induced Threshold Voltage Shifts on Small Channel-Area Devices," International Symposium on Quality Electrical Design (ISQED), Santa Clara, CA, pp.306-311, Mar. 2012.

11. Jyothi Bhaskarr Velamala, Ketul B. Sutaria, Hirofumi Shimizu, Hiromitsu Awano, Takashi Sato, Yu Cao: "Statistical Aging Under Dynamic Voltage Scaling: a Logarithmic Model Approach," Custom Integrated Circuits Conference (CICC), San Jose, CA, pp.6.3.1-6.3.4, Sep. 2012.

12. Takashi Sato, Hiromitsu Awano, Masayuki Hiromoto: "A Scalable Device Array for Statistical Device-Aging Characterization (invited)," International Conference on Solid-State and Integrated Circuit Technology (ICSICT), Guilin, China, pp.255-258, Oct. 2014.

Domestic conference

1. Hiromitsu Awano, Tetsuya Ogata, Toru Takahashi, Kazunori Komatani, Hiroshi Okuno: "Human and Robot Cooperation for Arrangement of Objects by Prediction using Recurrent Neural Network (in Japanese)," in Proc. 72-th National Convention of IPSJ, 5V-6, Tokyo, Mar. 11, 2010.

2. Hiromitsu Awano, Tetsuya Ogata, Toru Takahashi, Kazunori Komatani, Hiroshi Okuno: "Human-Robot Cooperation in Arrangement of Objects Using Confidence Measure (in Japanese)," in Proc. 34-th Annual Conference of the RSJ, 3J1-6, Nagoya, Sep. 24, 2010.

3. Hiromitsu Awano, Tetsuya Ogata, Jun Tani, Toru Takahashi, Hiroshi Okuno: Learning Performance Improvement of Recurrent Neural Network by Sparse Structures (in Japanese), in Proc. 73-th National Convention of IPSJ, 2-131-132, 1Q-4, Tokyo, March 2, 2011.

4. Hiromitsu Awano, Hirofumi Shimizu, Hiroshi Tsutsui, Hiroyuki Ochi, Takashi Sato: "A study on parameter estimation for modeling of random-telegraph noise (in Japanese)," in IEICE Technical Report (Design Gaia 2011), VLD2011-66, pp.85-90, Miyazaki, Nov. 2011.

5. Hiromitsu Awano, Takashi Sato: "Statistical Observation of BTI Degradation using Array based Architecture (in Japanese)," in Proc. IPSJ DA symposium 2013, pp.85-90, Gifu, Aug. 2013.

6. Hiromitsu Awano, Masayuki Hiromoto, Takashi Sato: "Variability in NBTI Induced Device Degradations Observed on 3996 Transistors (in Japanese)," in Proc. IPSJ DA Symposium 2014, pp.3-8, Gifu, Aug. 2014.

7. Hiromitsu Awano, Masayuki Hiromoto, Takashi Sato: "An efficient calculation of RTN-induced SRAM failure probability (in Japanese)," in IEICE Technical Report (Design Gaia 2014), Vol.114, No.329, VLD2014-74, pp.15-20, Beppu, Nov. 2014.

8. Hiromitsu Awano, Masayuki Hiromoto, Takashi Sato: "An Efficient Calculation Method of Time Changing Circuit Failure Probability Induced by Device Aging (in Japanese)," in Proc. IPSJ DA Symposium 2015, pp.169-174, Ishikawa, Aug. 2015.

9. Hiromitsu Awano, Takashi Sato: "Fast Monte Carlo based timing yield calculation (in Japanese)," in IEICE Technical Report (Design Gaia 2015), Vol.115, No.338, VLD2015-16, pp.37-42, Nagasaki, Nov. 2015.

10. Hirofumi Shimizu, Hiromitsu Awano, Hiroshi Tsutsui, Hiroyuki Ochi, Takashi Sato: "Estimation of Model Parameters for Random Telegraph Noise Based on Information Criterion (in Japanese)," in Proc. IPSJ DA Symposium 2012, pp.49-54, Gifu, Aug. 2012.

11. Ikkyu Aihara, Hiromitsu Awano, Takeshi Mizumoto, Yoshiaki Bando, Takuma Otsuka, Kohei Nagira, Hiroshi Okuno: "Theoretical and Experimental Studies on Frog Choruses Based on the Phase Oscillator Model and Sound-Imaging Method (in Japanese)," in Proc. 39-th Meeting of Special Interest Group on AI Challenges (SIG-Challenge), pp.50-56, Kyoto, Mar. 18, 2014.

12. Motoki Yoshinaga, Hiromitsu Awano, Masayuki Hiromoto, Takashi Sato: "A Study of Chip Identification Using Random Telegraph Noise

(in Japanese)," in Proc. Society Conference of IEICE, A-7-1, p.95, Tokushima, Sept. 2014.

13. Masahiro Sato, Syoichi Izuka, Hiromitsu Awano, Masanori Hashimoto, Takao Onoye: "On Stochastic modeling of NBTI induced threshold voltage variation (in Japanese)," in Proc. IEICE General Conference, A-3-5, Shiga, Mar. 2015.

14. Motoki Yoshinaga, Hiromitsu Awano, Masayuki Hiromoto, Takashi Sato: "Physical Unclonable Function Using RTN-Induced Time-Dependent Frequency Variance in Ring Oscillator (in Japanese)," in IEICE Technical Report, Vol.114, No.476, VLD2014-174, pp.117-122, Okinawa, Mar. 2015.

Awards

1. IPSJ Student Encoragement Award, Mar. 2011.
2. IPSJ DA Symposium 2013 Excellent Student Presentation Award, Aug. 2014.
3. IEICE Design Gaia Poster Award, Nov. 2014.
4. IEEE Kansai Section Student Paper Award, Feb. 2015.
5. IPSJ DA Symposium 2014 Excellent Student Presentation Award, Aug. 2015.
6. IPSJ DA Symposium 2014 The Best Student Presentation Award, Aug. 2015.
7. IPSJ Yamashita SIG Research Award, Aug. 2015.
8. IPSJ DA Symposium 2015 Excellent Student Presentation Award, Sep. 2015.
9. IEEE CEDA All Japan Joint Chapter, Design Gaia Best Poster Award, Dec. 2015.