Accelerated Aging in Devices and Circuits

by

Devyani Patra

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved November 2017 by the
Graduate Supervisory Committee:

Yu Cao, Chair
Hugh Barnaby, Co-Chair
Jae-sun Seo

ARIZONA STATE UNIVERSITY

December 2017

ABSTRACT

The aging mechanism in devices is prone to uncertainties due to dynamic stress conditions. In AMS circuits these can lead to momentary fluctuations in circuit voltage that may be missed by a compact model and hence cause unpredictable failure. Firstly, multiple aging effects in the devices may have underlying correlations. The generation of new traps during TDDB may significantly accelerate BTI, since these traps are close to the dielectric-Si interface in scaled technology. Secondly, the prevalent reliability analysis lacks a direct validation of the lifetime of devices and circuits. The aging mechanism of BTI causes gradual degradation of the device leading to threshold voltage shift and increasing the failure rate. In the 28nm HKMG technology, contribution of BTI to NMOS degradation has become significant at high temperature as compared to Channel Hot Carrier (CHC). This requires revising the End of Lifetime (EOL) calculation based on contribution from induvial aging effects especially in feedback loops. Conventionally, aging in devices is extrapolated from a short-term measurement, but this practice results in unreliable prediction of EOL caused by variability in initial parameters and stress conditions. To mitigate the extrapolation issues and improve predictability, this work aims at providing a new approach to test the device to EOL in a fast and controllable manner. The contributions of this thesis include: (1) based on stochastic trapping/de-trapping mechanism, new compact BTI models are developed and verified with 14nm FinFET and 28nm HKMG data. Moreover, these models are implemented into circuit simulation, illustrating a significant increase in failure rate due to accelerated BTI, (2) developing a model to predict accelerated aging under special conditions like feedback loops and stacked inverters, (3) introducing a feedback loop based test methodology called Adaptive Accelerated Aging

(AAA) that can generate accurate aging data till EOL, (4) presenting simulation and experimental data for the models and providing test setup for multiple stress conditions, including those for achieving EOL in 1 hour device as well as ring oscillator (RO) circuit for validation of the proposed methodology, and (5) scaling these models for finding a guard band for VLSI design circuits that can provide realistic aging impact.

ACKNOWLEDGMENTS

First and fore-most, I would like to thank my parents Mr & Mrs N Patra, my brother, Devjyoti Patra, and sister-in-law, Madhumita Saha Patra for making my master's education possible.

I would like to sincerely thank my professor Dr. Yu Cao for giving me the opportunity to work at NIMO group and for providing invaluable guidance, teaching, and support throughout my master's studies. I am also very thankful to my committee members Prof Jae-sun Seo and Hugh Barnaby for their kind support and useful feedback. I am indebted to my colleague and friend, Shatadal Mishra and Abinash Mohanty, for their indispensable discussions throughout the course of this project. I would like to thank my friends from NIMO: Sai Rama Srivatsava Gorthy, XiaoCong Du, Zihan Xu, and Zheng Li for their help during the thesis work. I must also thank my graduate advisor Toni Mengret for her active advice regarding all the administrative procedures. Finally, I would like to thank the teams at the Boeing Research & Technology and Microsemi Cooperation as well as my collaborators at Purdue University and Peking University for the research opportunity and their enlightening inputs in furthering this work.

Last but not the least, I would like to present my heartfelt gratitude to my friends especially Gaurav Jain for his vital support.

TABLE OF CONTENTS

Page

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

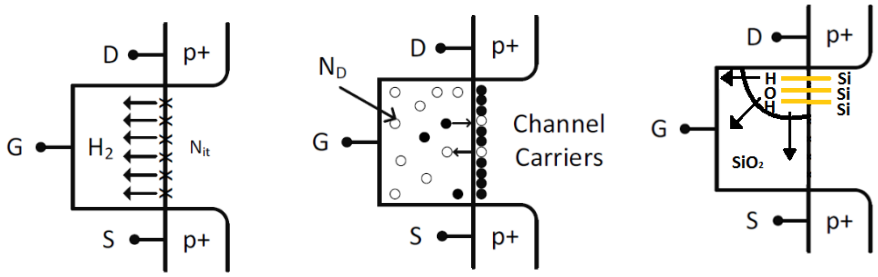
CHAPTER 1. INTRODUCTION

1.1. Overview of Current Aging Models

Aging in devices due to gradual degradation of threshold voltage (V_{th}) has been explained by multiple models. For Bias temperature inversion (BTI), there are two main models in existence that address transistor behavior degradation due to changes in the quality and physical state of the gate oxide. The earlier developed and widely accepted of these models is the Reaction-Diffusion (RD) model. The RD model assumes that applied voltage at the gate (V_g) initiates a field dependent reaction at the semiconductor-oxide interface that generates interface traps by breaking passivized Si-H bonds. During *Reaction*, the interface charges introduced by the broken bonds cause an increase in V_{th}. During *Diffusion*, the generated hydrogen species combine to form H_2 and diffuses away from Si-SiO_2 interface towards the gate, thus changing the concentration of hydrogen at the interface. When stress is lowered, the hydrogen species diffuse back and anneals the broken bonds thus causing recovery [4].

The second model known as the Trapping-Detrapping (TD) model explains the change in V_{th} through charging and discharging of trap locations in the oxide and at the interface. According to this theory, the field applied at the gate due to V_g modulates the trap energy. When the trap gains enough energy, it may capture a charge carrier thus reducing the number of available carriers in the channel. This phenomenon is called *Trapping*. Faster traps have higher probability of capturing carriers; the occupation probability increases with increase in Voltage and Temperature. When the trap releases the carrier due to decrease in its energy, it leads to *De-trapping* and hence recovery [5].

In addition to BTI, Hot Carrier Effect (HCI) impacts aging when the drain voltage (V_d) is high. The high drain-source (V_{ds}) voltage accelerates the electrons and they impact the drain causing localized generation of traps. When both V_g and V_d are high, BTI and HCI occur simultaneously and their effect on V_{th} can be seen as additive. HCI follows the t^n relation similar to BTI in RD model with n~0.45. Only thing to note is that HCI does not experience any recovery.



* (Image: Modeling and Simulation Tools for Aging Effects in Scaled CMOS Design, Ketul

Figure 1.1 Existing physical models of V_{th} change in literature. Left to right shows BTI-RD model, BTI TD model and HCI model

1.2. Issues in predicting aging in present models

The models presented above present some issues in their compact formulation. The pros and cons of each models can be listed as:

- Reaction-Diffusion model:

Pros: n^{th} order time dependence of degradation is modeled well especially for long stress time.

Cons: Short term degradation/recovery on time scales less than a milli-second are not well modeled. Moreover, this model is sensitive to statistical variations in modelling parameters. The time exponent model parameter 'n' stabilizes to the expected value of 0.16

over long stress times, but is much higher for shorter stress times. In comparison, the parameters φ, A and C of the TD model ($\boldsymbol{n(t) = \varphi(A + log(1 + Ct))}$) do not show large variations [6].

- Trapping-Detrapping model:

  Pros: Models fast recovery well, and is more reliable with log(time) prediction.

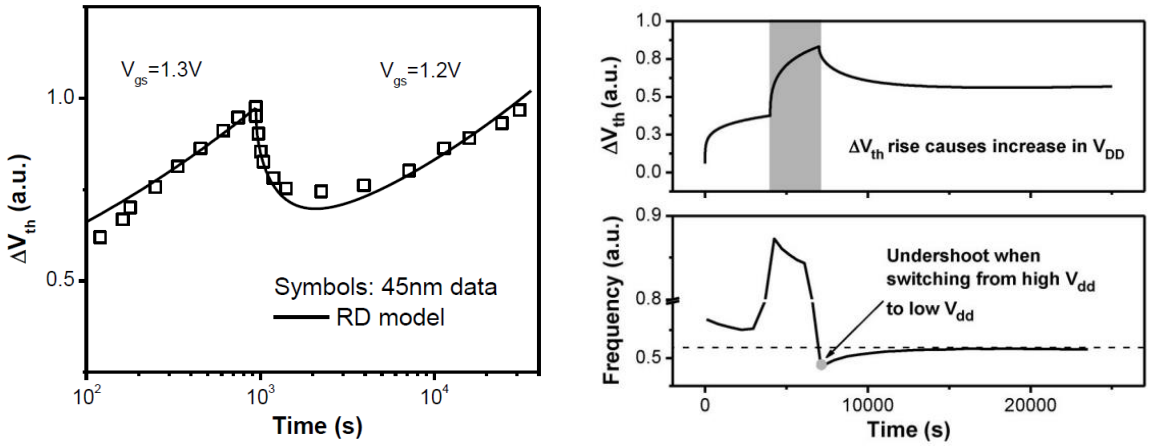
  Cons: Not reliable at longer stress times, likely due to not accounting for increase in number of available traps in time [6]. This model assumes a fixed number of initial traps then modulates the energy of the traps based on the applied stress voltage.

Owing to these observed shortcomings, attempts have been made to unify these models to be able to predict the short term and long term degradation of Vth in a better way. Previous efforts at joint modeling of RD and TD effects includes the two energy well model [7] and the structural relaxation model [8]. Built based on the RD model, the carriers in the two energy well model can be captured by traps of different energy level and recover quickly from those of low energy in order to explain the fast recovery and variation in parameters. Structural relaxation was built on the TD model and artificially introduced non-recoverable traps to explain degradation over long stress times.

In this work, I attempt a more direct unification of the RD and TD mechanisms by modeling the number of traps in a deterministic way as per the RD mechanism, while using TD to describe the stochastic process of capture and emission. This work combines the trap generation features of the RD model with the trap occupation features of the TD model to reliably predict both short- and long-term degradation and recovery due to all FEOL mechanisms. The model developed here should also be able to handle cases where Dynamic Voltage Stress (DVS) or random input stress conditions are applied.

1.3. **Motivation for this work**

Temporal degradation of devices and circuits is a key factor in failure management, causing over margining, lower performance and higher power consumption. The goal of this project is to develop efficient and accurate means of modeling and testing aging effects in the high stress regime where failure can happen in a matter of hours. The models are developed on the basis of same underlying physical principles of trap generation, population and de-population. This work not only aims at solving the problem of short term measurement error shown in Fig 1.2, but also at providing a new approach to test the device to the end of lifetime (EOL) in a fast and controllable manner adaptive to device response to mitigate the issues arising out of long term extrapolation. While HCI is still significant in analog/mixed-signal design, BTI is dominant in digital design at scaled technology nodes, for both NMOS and PMOS devices. Physical modeling and accurate prediction tools of these aging effects are essential to guarantee the quality of large-scale system design at short time scales.



* (Image: K. B. Sutaria et al," Accelerated aging in analog and digital circuit with feedback," IEEE TDMR, 2015)

Figure 1.2 (left) Mismatch in short term modelling by RD model. (Right) Mis-prediction of momentary fluctuations in AMS circuits causing failure

The prediction for reliability, such as the lifetime, presents a unique challenge to integrated circuit design community. Unlike timing and power analysis that can be directly validated by silicon measurement post fabrication, EOL can only be predicted. A common practice is to sample the device for a short time, calibrate the aging model, and rely on extrapolation methods to estimate the device lifetime. This conventional method suffers from inaccuracy in long-term prediction due to many physical and practical reasons, such as the stochastic nature of the aging process, device-to-device variations and design uncertainties. To account for these factors at the circuit level, such a method is expensive in runtime and memory resources [1-3]. As an example, Figure 1.3 presents the prediction of EOL for four 28nm NMOS devices, using the power law dependence on the stress time to extrapolate. In the presence of statistical fluctuations that are intrinsic to PBTI, even a small difference in the short-term data, which may be introduced by device variations or testing conditions, can lead to a dramatic variability in the long-term prediction [5, 13]. Starting from a similar threshold voltage (V_{th}) degradation rate, two of the devices exhibit a difference of 10X in the time needed to reach 100mV shift in V_{th} (ΔV_{th}).

Therefore in this work, analyses is done to study the impact of circuit conditions liked feedback loops, stacked inverters, level shifters, DVFS circuits etc. that may lead to accelerated V_{th} degradation. A strategy is developed to accelerate aging of a device or circuit to meet EOL requirements specified.

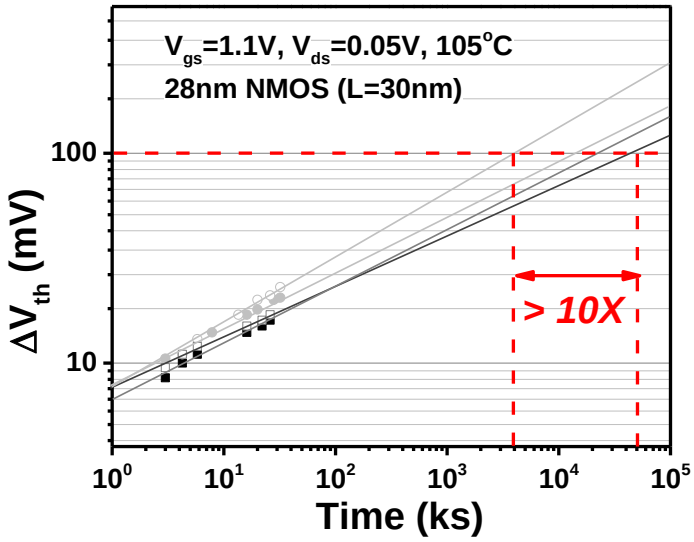


Figure 1.3 V_{th} shift based on the extrapolation from short-term measurement leads to unreliable EOL prediction under statistical variations. Though the short-term V_{th} shifts for four devices only have a small amount of variations, the difference among them is amplified during the extrapolation.

Added to these device level modelling challenges, reliability analysis at circuit level also poses problems. Simulating 28nm high-K metal-gate (HKMG) with SPICE is inefficient for large-scale digital design as shown in Fig 1.4. As the scale and the complexity of VLSI design increase, the time taken to calculate aging increases linearly. Additionally, the physics of recovery in BTI is not well described by existing models in the PDK, causing a significant overestimation of path delay degradation (ΔTd), as shown in Fig. 1.4. In design practice, such erroneous aging prediction may lead to inappropriate optimization and tradeoffs among performance, power and reliability. Thus, the framework of System Reliability Analyzer (SyRA) has been updated to SyRA-X for providing a guard-band to reliably predict circuit aging. So the motivation is to provide a fast and accurate aging prediction for VLSI circuits.

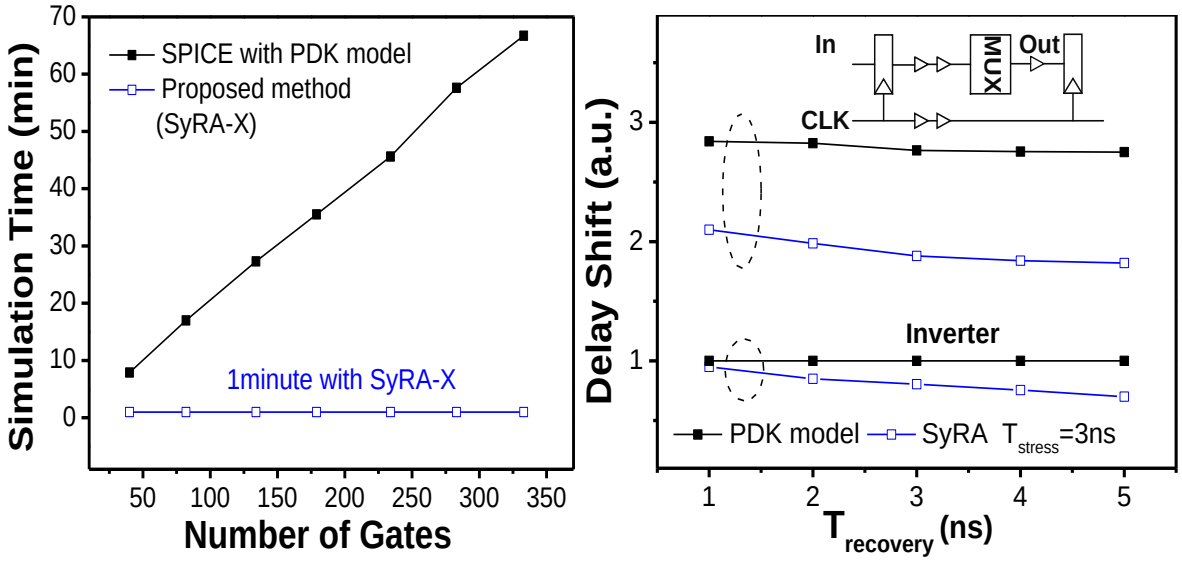


Figure 1.4 (left) Time required for aging simulation in SPICE increases with the complexity of the circuit. (right) Current PDK model at 28nm does not capture the recovery effect thus leading to overdesign

1.4. Thesis Organization

Chapter 2 discusses the development of BTI model to cater to short term as well as long term stress model. It also details the impact of TDDB aging by incorporating the additional traps and their energy into its framework. Chapter 3 discusses the analysis of positive feedback in specific circuits and implementation of methodology to accelerate aging to achieve desired EOL. Chapter 4 provides a discussion on the results of SyRA-X framework on VLSI circuits. Chapter 5 summarizes this work and sheds light on some possible future work ideas.

CHAPTER 2. BTI Model Development

2.1. Model to incorporate short term and long term effects

Bias-temperature-instability (BTI) is the dominant aging effect in scaled technology. Its underlying physics can be explained and modeled by the trapping/ Detrapping process (TD) [10]. Previous efforts on BTI modeling assume a constant trap availability and results in the logarithm model [5]. To account for the change in trap number in both gradual aging and the sudden fluctuation under TDDB (described in next section), I first introduce an increase in total trap number (N) over time as a power law on time (t^n) [2] (Table 1). The number of occupied traps in TD model is given by equation 1 [10], where N_D is the number of available traps, n(t) are occupied traps [5] and P_{01} is the occupation probability of a trap at time t.

$$\boldsymbol{n(t) = \left(\sum_{N_D=0}^{\infty} N_D \frac{N^{N_D} e^{-N}}{N_D!}\right) \int P_{01}(t, \tau_c, \tau_e)} \quad (1)$$

Expanding P_{01} in terms of capture and emission coefficients (τ_c, τ_e), and substituting for logarithmic distributions of time constants equation (2) is obtained, where p_{min} and p_{max} are the time constants of the fastest and slowest traps, $g(E_T)$ is the trap energy distribution and N is the average trap density. The assumptions made in this model are:

1. N_D is Poisson distributed, common for a discrete process of trapping-detrapping.
2. τ_c and τ_e are uniformly distributed on a logarithmic scale.
3. Distribution of trap energy is approximated as a U-shape.

$$\mathbf{n(t) = \frac{N}{\ln 10 (p_{max} - p_{min})} * \int_0^{E_{Tmax}} \frac{g(E_T) dE_T}{1 + \exp\left(-\frac{E_T - E_F}{kT}\right)} * \int_{10^{-p_{min}} t}^{10^{-p_{max}} t} \frac{e^{-u} - 1}{u} du} \quad (2)$$

In model, replace the Poisson distributed N_D by a deterministic time dependent $N(t)=At^n$ which evolves with time based on power law. Total number of traps effecting device aging is then obtained by integrating over the stress time as

$$\boldsymbol{n(t) = \varphi\left(1 - \frac{BA_1}{1+Bn-B}\right)t^n \log(A+Bt) + \frac{BA_1}{1+Bn-B}t^n - \frac{nAA_1}{1+Bn-B}t^{n-1}\log(A+Bt)} \ldots \quad (3)$$

In this equation A, B and φ are the TD model parameters discussed in [2]. In this model φ has an exponential dependence on V and T given by: $\mathbf{exp\left(\frac{\beta V_g}{T_{ox}kT}\right)exp\left(\frac{-E_0}{kT}\right)}$. The last term decays very quickly and can be ignored. This equation can be expressed in a compact form as:

$$\mathbf{n(t) = \varphi[C_1 t^n ln(1 + C_2 t) + C_3 t^n]} \quad (4)$$

where φ has an exponential dependence on stress voltage (V), temperature (T) and trap energy (E) expressed above. C_1, C_2, and C_3 are dependent on TD model coefficients which in turn depend on stress conditions and trap characteristics. These parameters will determine sensitivity to short-term and long-term stress condition and can be used to fit the model to experimental data. For dynamic stress, the first term dominates. Fig. 2.1 validates this model with 14nm FinFET data [11]. While a conventional t^n model may match the long-term data, it has a larger error in the short term (the inset in Fig. 2.1). The new model of Eq. (4) well matches the data in both the long term and the short term. The shift in threshold voltage depends on the number occupied traps as

$$\mathbf{\Delta V_{th} = \frac{q * n(t)}{C_{ox}}} \quad (5)$$

HCI does not contain a recovery step and can be explained well by the RD model [11]. Since the stress condition for BTI and HCI are different, their effects can be decoupled and the total shift in threshold voltage can be found as a linear combination of shifts produced by each mechanism (Eq. 6)

$$\Delta V_{th\_total} = \Delta V_{th\_HCI} + \Delta V_{th\_BTI} + \Delta V_{th\_BTI\_recovery} \quad (6)$$

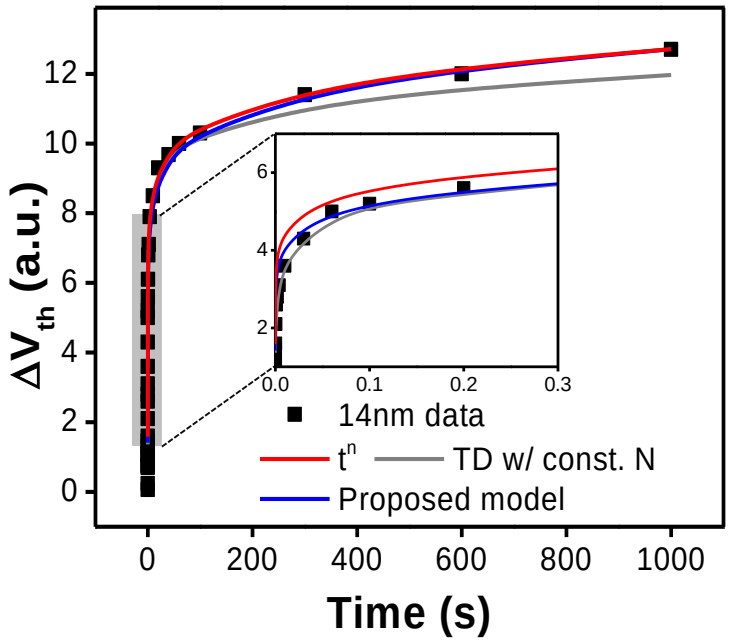


Figure 2.1 The new model improves the accuracy across a wide range of stress time [11]

2.2. Model with added TDDB impact

Usually the TD process is gradual and independent of other aging effects, e.g., time-dependent-dielectric breakdown (TDDB). Yet, with ultra-thin gate dielectric, the new traps that are generated by TDDB may lead to simultaneous increase in gate leakage (I_g) and threshold voltage (V_{th}) (Fig. 2.2). This is because the new traps from TDDB are still close to the interface between gate dielectric and the substrate and thus, are able to participate in and accelerate the capture/emission process of BTI (Fig. 2.2). Fig. 2.3 presents the study

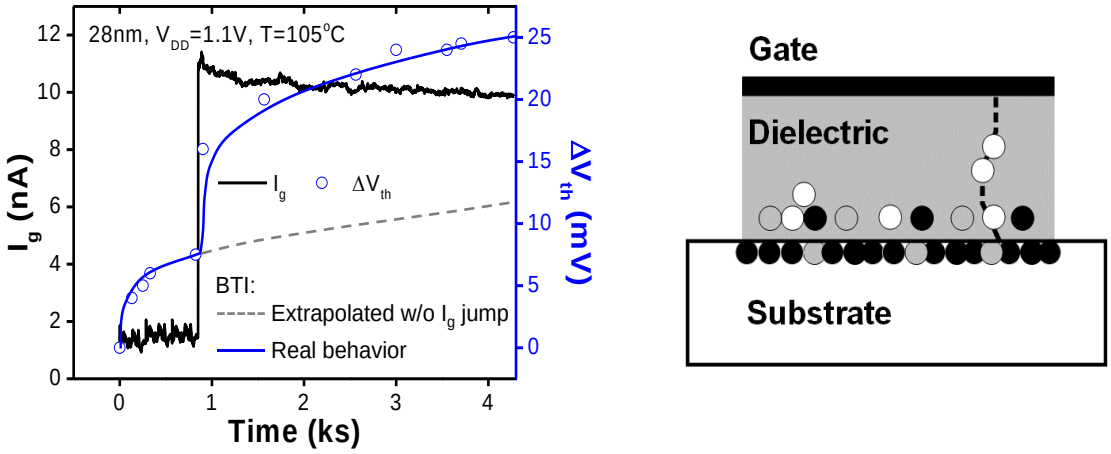


Figure 2.2 (left)The sudden jump in I_g, caused by TDDB, accelerates V_{th} change under BTI. (right)The generation of new discrete traps (shown as white dots) under TDDB lead to the final conductive path, and are also responsible to faster degradation of BTI.

flow, with the particular interest on the correlated TDDB and BTI that are induced by the same group of traps, as well as the impact on circuit lifetime.

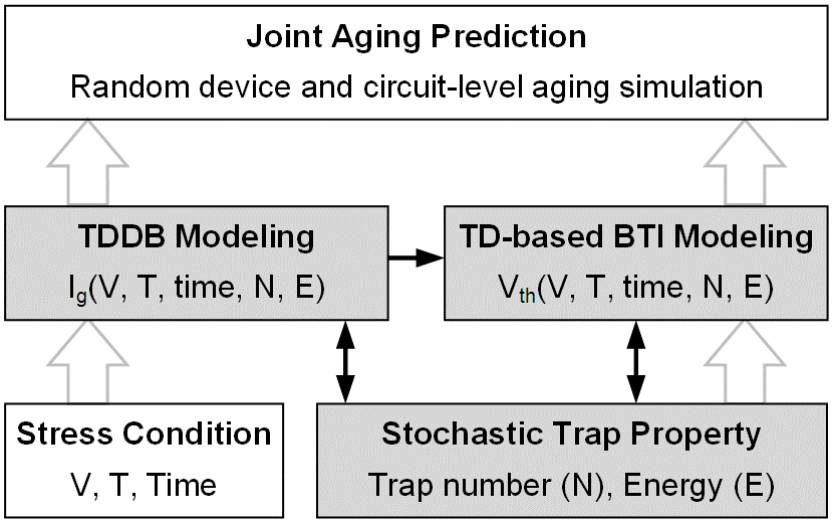


Figure 2.3 The modeling and simulation flow, with the focus on the correlation between TDDB and BTI. The stochastic properties of the trap are calibrated by joint TDDB and BTI measurement.

More importantly, the random jump in trap number due to TDDB ($\Delta N(t_1)$, where t_1 is the moment it happens) needs to be carefully considered. The occurrence of Ig jump is more probable when multiple traps are generated simultaneously leading to the percolation path [4]. These additional traps are integrated into Eq. (4), as presented in Table 1.

Effect	**Trap number**	**Compact Model**
BTI (ΔV_{th})	$At^n + \Delta N(t)$	$C_1 t^n \ln(1 + C_2 t) + C_3 t^n + \exp(E/E_0)\Delta N(t_1)\ln[1 + C_2(t - t_1)]$
TDDB (I_g)	$\Delta N(t)$	$\Delta N(t_1)\ln\left[\frac{1+\exp(\beta(E-\alpha))}{1+\exp(\beta E)}\right]$
$\Delta N(t_1)$: 2 ± 1; t_1: random; E: $1.4 \pm 0.15eV$		

Table 2.1 : Summary of BTI and TDDB models, correlated by discrete random trap generation.

Validated by Fig. 2.4, these new traps have a much more pronounced effect than the gradual aging process, due to the number of traps generated and their energy. Similar

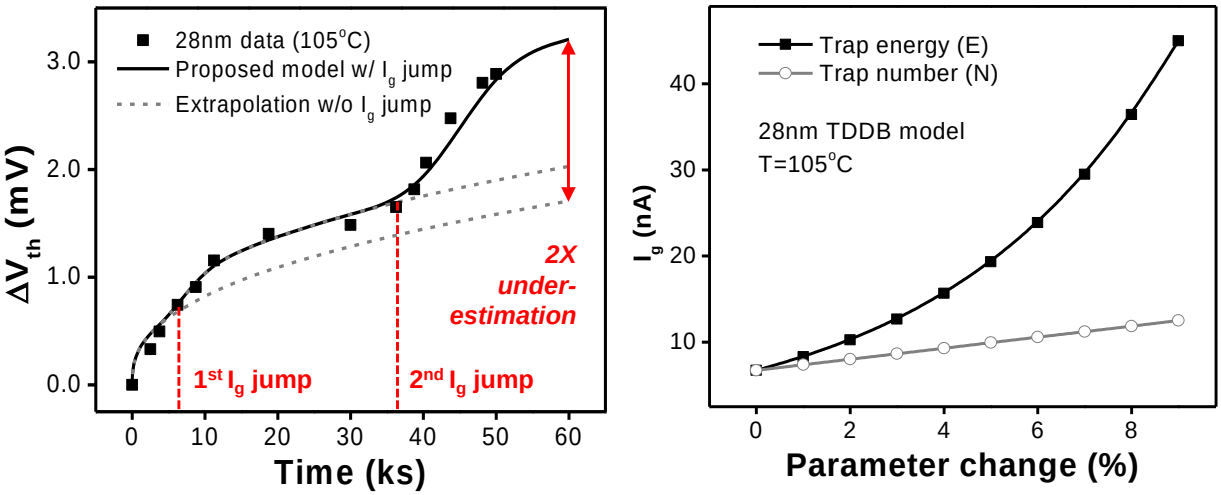


Figure 2.4 The impact of TDDB on BTI is significant in aging prediction. (right) TDDB is more sensitive to trap energy due to the exponential dependence

sensitivity to N and E is also observed in TDDB (Fig. 2.4) [9,12]. To confirm the correlation between TDDB and BTI in thin dielectric devices, multiple 28nm HKMG transistors were stressed and simultaneously monitor both I_g and threshold shift (ΔV_{th}), as shown in Fig. 2.5 (top). Note that even under severe TDDB, I_g is still much smaller than the drain current that is used to extract V_{th}.

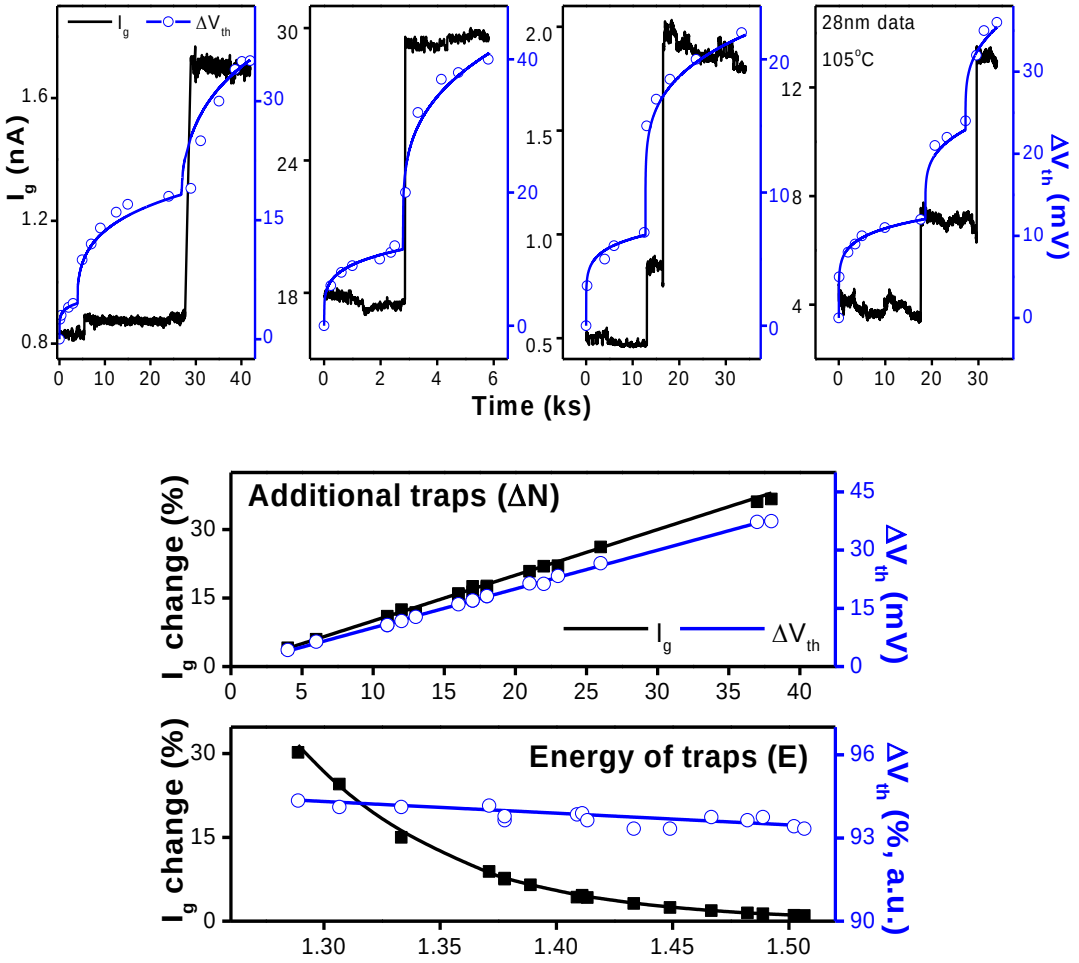


Figure 2.5 (top) The correlation of TDDB and BTI in four 28nm devices. The new BTI model well captures the acceleration that is affected by the additional traps under TDDB.(bottom) From I_g and V_{th} changes in 7 devices, the properties of new traps generated by TDDB are extracted, matching their assumption in Table 1

Therefore, ΔV_{th} is indeed due to BTI. The slope of ΔV_{th} remains the same after additional traps. From each moment where I_g jumps, the acceleration of BTI is consistently observed. Furthermore, the models of TDDB and BTI (Table 1) are applied to each moment and a pair of ΔN and E are extracted. Fig. 2.5 (bottom) summarizes the result. The range and statistics of both parameters match original assumption in deriving the models (Table 1) [10,9,12].

The new models are integrated into circuit simulation to assess the impact on timing. Fig. 2.6 presents the schematic, with both TDDB and BTI degradation affected by the same group of new traps. The moment of trap generation is a random variable in Verilog; their energy follows a normal distribution [9]. Two data-paths are statistically simulated, as shown in Figs. 2.6 and 2.7. The time to failure (TTF) is defined by timing error, i.e., when path delay increases by more than 5% of the fresh value. Indeed, the stochastic impact by TDDB on BTI significantly speeds up delay degradation. In the worst case, TTF may be reduced by more than 4X.

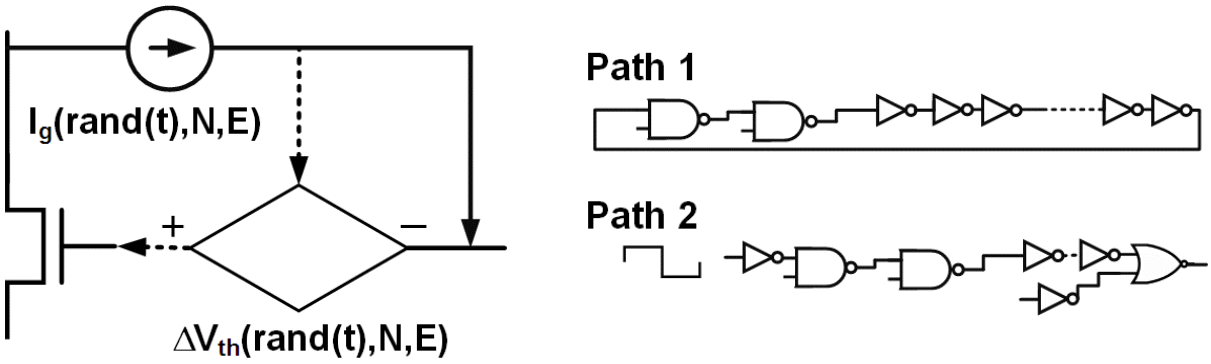


Figure 2.6 (left) Both TDDB and BTI are implemented as stochastic aging to circuits (right) the data paths analyzed.

.

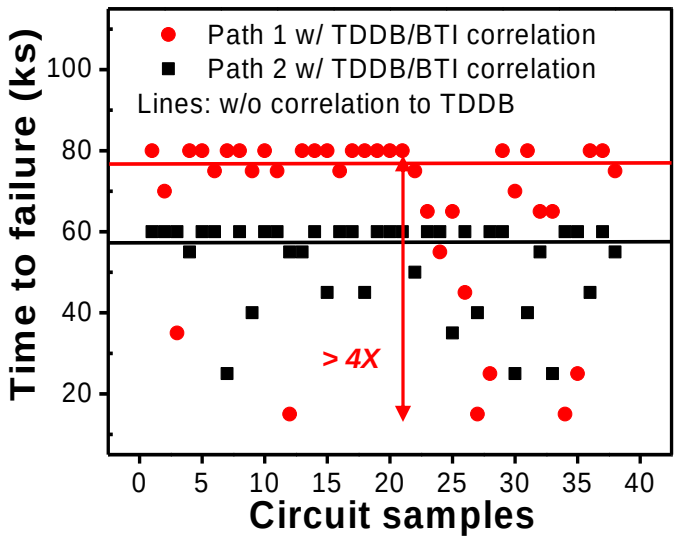


Figure 2.7 Timing failure happens much earlier due to the acceleration by TDDB, even though TDDB does not occur yet.

CHAPTER 3. Feedback Loop

3.1. Positive Feedback Loops and Accelerated Aging

Reliability modeling and analysis today lacks a robust method to directly validate the lifetime of devices and circuits. As aging mechanisms are usually gradual, i.e., a slow process, conventional aging analysis relies on the extrapolation from a short-term measurement, resulting in unreliable prediction of End of Life (EOL). Such situation is exacerbated at scaled technology nodes, where the more gradual and stochastic mechanism of Bias Temperature Instability (BTI) dominates aging, rather than Hot Carrier Injection (HCI). To improve the robustness of aging modeling and EOL prediction, this work proposes a new approach to adaptively stress the device to EOL in an accelerated and controllable manner. It enables us to monitor the entire process of degradation and validate related analysis tools. This chapters outlines the acceleration in aging caused by certain circuit configurations like diode connections, stacked inverters, level shifters etc. and leveraging this phenomenon, develop a closed-loop test methodology, Adaptive Accelerated Aging (AAA), that effectively accelerates the degradation in a device as well as circuit.

In addition to the effect of BTI, other aging mechanisms, especially HCI and time dependent dielectric breakdown (TDDB), exist together and may be correlated due to reliability physics and circuit operation conditions, making an accurate prediction even more difficult. For instance, the co-existence of PBTI and HCI in HKMG devices requires careful decomposition and physical modeling. Figure 3.1 shows the effect of HCI is significant at room temperature and increases dramatically with higher stress voltage; as

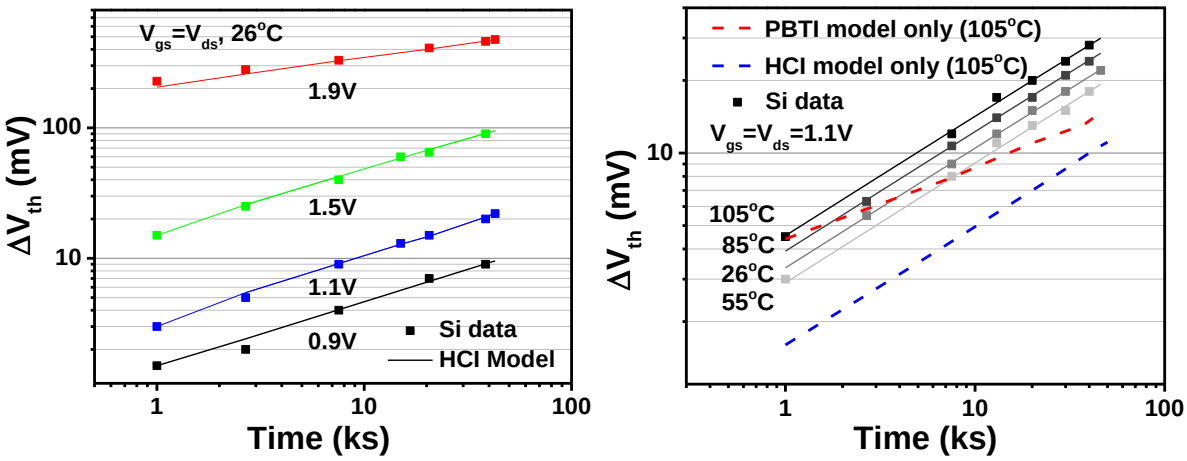


Figure 3.1 Both HCI and BTI occur when a NMOS device is biased in a diode connection (i.e., V_{gs} = V_{ds}). An appropriate temperature is selected to decompose HCI and BTI: (left) the room temperature is applied first such that the degradation is dominated by HCI; (right) then a much higher temperature is applied and thus, aging of the 28nm HKMG device is primarily induced by PBTI. The combination of these testes helps separate BTI and HCI, and calibrate corresponding models.

temperature increases, the total degradation is larger, while the contribution of HCI becomes secondary to that of PBTI. Therefore, the relative importance of PBTI and HCI depends on both the drain voltage and the stress temperature. The complexity in these physical behaviors demands an effective and controllable test method to measure the actual degradation to EOL, yet without the burden of long measurement time as in conventional methods used currently.

This capability of stressing a device to EOL will offer direct and complete validation of the individual and coupled aging mechanisms. The impact of aging also needs to be studied at circuit level to validate the device model and make it useful for large-scale designs, including both the stress and the recovery phases under dynamic system operations

[9]. A closed-loop test structure, instead of the open-loop one used in conventional aging test is configured; by adaptively tuning the gain of the feedback loop, this method is able to stress the device to EOL within one hour, offering a fast and convenient method to validate the entire aging process.

3.1.1. **Compact model for feedback loops**

Reliability analysis for complex gates and circuits presents a unique challenge while predicting the aging degradation using the existing compact models. Various configurations of devices may have stress conditions that cause accelerated aging due to presence of feedback loops. The devices in that case may degrade more than what is predicted and lead to early failure or incorrect behavior and are missed in the standard model analysis. For example, in case of stacked NMOS, with different voltage on each gate, the intermediate node's voltage decreases over time thus causing the V_{ds} drop of the transistor to increase. This causes increased HCI than predicted by a compact model assuming the initial V_{ds} as the stress condition as shown in Fig 3.8. Another type of commonly occurring structure is a diode connected device with fixed current. In this case, as the device degrades but the constant current requirement forces V_d to increase, the stress condition becomes more stringent and accelerates aging. Similar configurations exist in complex circuits like comparators, dynamic circuits, level shifters etc. with stacked NMOS with gates biased at different voltages. Other cases like digital circuits with PLL feedback where the period is adjusted to maintain the original frequency thus increasing the stress condition and increasing aging. In case of a clock buffer, as the buffers in the tree degrade, the clock needs to be made slower thus increasing the time of stress.

Thus, these cases need special treatment when it comes to aging prediction. Traditionally, analysis is performed in small time steps in multiple iterations with each iteration having updated stress condition. Though this analysis is accurate it is extremely time and computation consuming. So the development of a closed form compact model is proposed to be able to predict this accelerated aging behavior in a fast an accurate manner i.e. model bias runaway for HCI. Basically even before runaway happens, the degradation rate increases above the base line, and can model this behavior.

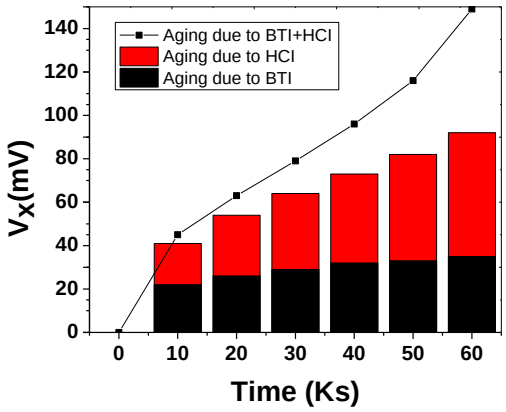


Figure 3.2 Comparison of aging due to BTI and HCI individually and combined

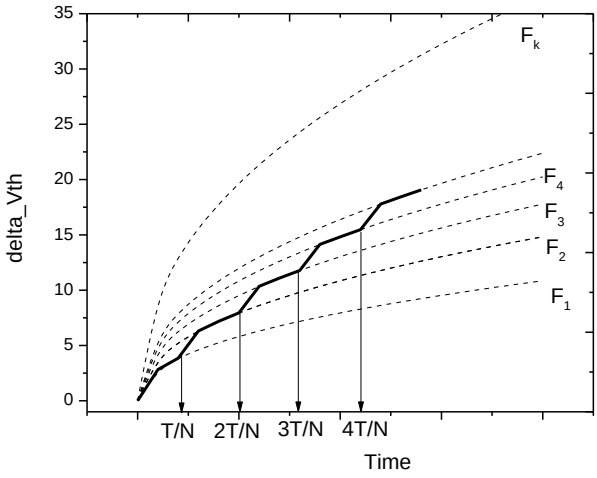


Figure 3.3 Graph showing pictorially the approach for derivation

3.1.1.1. Feedback loop model

To develop a compact model, let us look at a feedback loop case. The voltage at X in Fig 3.2 needs to be calculated as a function of initial bias voltage, time and temperature of stress. For a fixed bias current, the aging effects of HCI and BTI are:

$$BTI\colon \Delta V_{th}(t+\Delta t) = f(V_{gs}(t) - V_{th}(t), \Delta t) \tag{7}$$

$$HCI\colon \Delta V_{th}(t+\Delta t) = f(V_{gs}(t) - V_{th}(t), (V_{ds}(t), \Delta t) \tag{8}$$

To maintain constant current, $V_{gs}(t) - V_{th}(t)$ is constant so the HCI term leads to the runaway behavior. Now assuming the total stress time is divided in to N intervals. Define $F_0 = ke^{\frac{Vds}{B}}$ so that $\Delta V_{th} = F_0\left(\frac{T}{N}\right)^n$ is when a device is stressed for time a small interval of time T/N. $\Delta V_{thB} = F_0(T)^n$ is the baseline curve. Then in the next interval, Vth shift is:

$$F_0\left(\frac{T}{N}\right)^n\left(1+\alpha\frac{F_0}{B}\left(\frac{T}{N}\right)^n\right);\ F_2 = F_0\left(\frac{2T}{N}\right)^n\left(1+\alpha\frac{F_1}{B}\right) \tag{9}$$

and so on till: $F_k = F_0\left(\frac{kT}{N}\right)^n\left(1+\alpha\frac{F_{k-1}}{B}\right)$ which can also be written as:

$$F_k = F_0\left(\frac{kT}{N}\right)^n\left(1+\alpha\frac{F_0}{B}(k-1)^n\left(\frac{T}{N}\right)^n + \alpha^2\frac{F_0}{B^2}(k-1)^n F_{k-2}\left(\frac{T}{N}\right)^n\right) \tag{10}$$

Looking at Nth interval (k=N) and simplifying:

$$F_N = F_0(T)^n\left\{1+\left(\frac{N-1}{N}\right)^n CT^n + \cdots + \left(\frac{N-1}{N}\right)^n\left(\frac{N-2}{N}\right)^n\left(\frac{N-3}{N}\right)^n \ldots\left(\frac{2}{N}\right)\left(\frac{1}{N}\right)(CT^n)^N\right\} \tag{11}$$

where $C = \frac{\alpha}{B}F_0$ On taking the limit N-> inf and simplifying, CT^n <1 so it is convergent infinite sum of geometric series

$$F_N = F_0(T)^n\left[\frac{1}{1-CT^n}\right] \tag{12}$$

Therefore the shift in Vth is given by:

$$\Delta Vth = \frac{\Delta V_{th_B}}{1-\frac{\alpha}{B}\Delta V_{th_B}} = \frac{\Delta V_{th_B}}{1-g_1 g_2} \quad (13)$$

It can be rewritten in terms of loop gains as: $\frac{\alpha}{B}\Delta Vth_B = g_1 g_2$ where $g_1 = \frac{\partial Vds}{\partial \Delta Vth} = \alpha$; $g_2 = \frac{\partial \Delta Vth}{\partial Vds} = \Delta V_{th_B} f(technology\ length)$

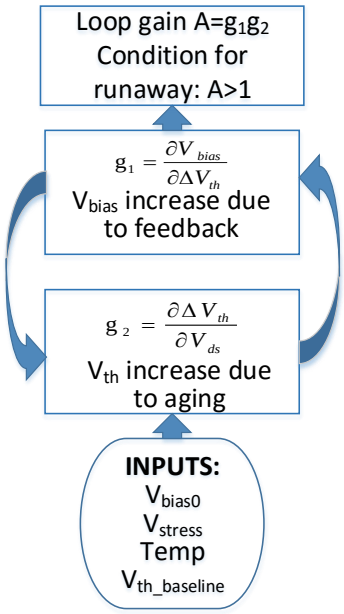


Figure 3.4 Physical formalism of bias runaway: A feedback loop accelerates the process of aging.

Diode connected	$\Delta Vth = \frac{\beta \Delta Vth_B}{1-(\alpha(g_1 g_2))}$
65nm	$\frac{\beta \Delta Vth_B}{1-(\alpha(1+\frac{1}{Vds_0})\Delta Vth_B)}$
28nm	$\frac{\beta \Delta Vth_B}{1-(\alpha(1+\frac{1}{Vds_0})\Delta Vth_B)}$
14nm	$\frac{\beta \Delta Vth_B}{1-(\alpha_1\left(e^{\frac{Vds_0}{\alpha_2}}\Delta Vth_B \ln(\alpha_3 \Delta Vth_B Vds_0)\right))}$

Table 3.1 Compact form of accelerated aging equations

3.1.1.2. Model Verification

The compact mode in Equation 13 can be used across technology nodes with proper calibration of g_1 and g_2 as seen in Table 3.1. To validate this model, measurements have been performed for 16nm finfet NMOS, 28nm HKMG NMOS and 65nm NMOS devices. The devices were first calibrated for HCI behavior as seen in Fig 3.5. In order to calibrate the devices and find g_1 and g_2, sensitivities of $V_{bias}(V_x)$ and V_{th} are plotted. For g_1, the

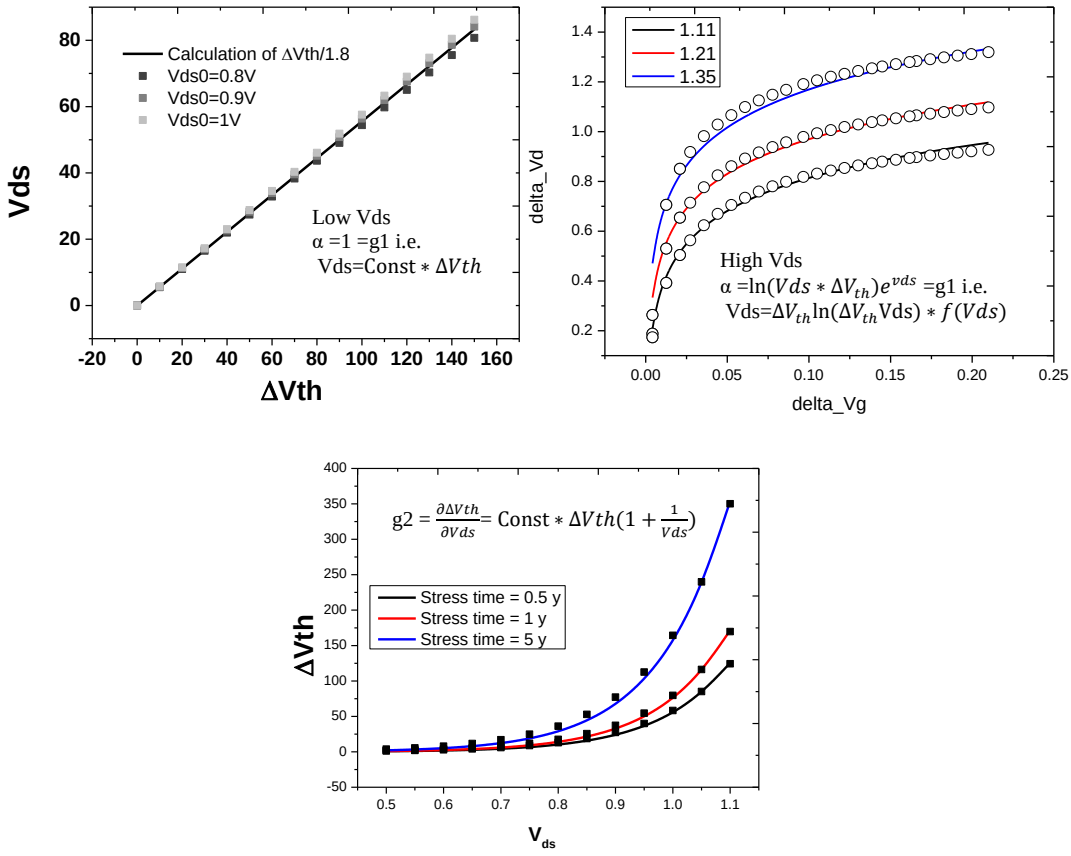


Figure 3.5 The graphs showing sensitivity of bias point for g1 and g2 calculation.

sensitivity of the bias point to ΔV_{th} Is needed. By measuring the voltage at V_{bias} while changing the V_g this can be tested. In case of a low initial bias voltage, the relation is linear. To maintain the same I_{ds}, V_{ds} increases by same amount as V_{gs}-V_{th} degrades due to aging.

When the bias voltage is high, this relation becomes non-linear as seen in Fig 3.5. Therefore calibration of g_1 should be done for a device at its operating condition for using this model. g_2 relation on the other hand is determined by the aging model used as sensitivity of ΔV_{th} to stress V_{ds}. Post calibration, for the simulation of feedback loop iterations were performed for every 2% of I_{ds} degradation. Thus the time-steps are initially small but increase with stress time.

I observe in Fig 3.7 that due to the feedback, the bias voltage increases at an accelerated rate as compared to device stressed without connecting V_g and V_d for all three

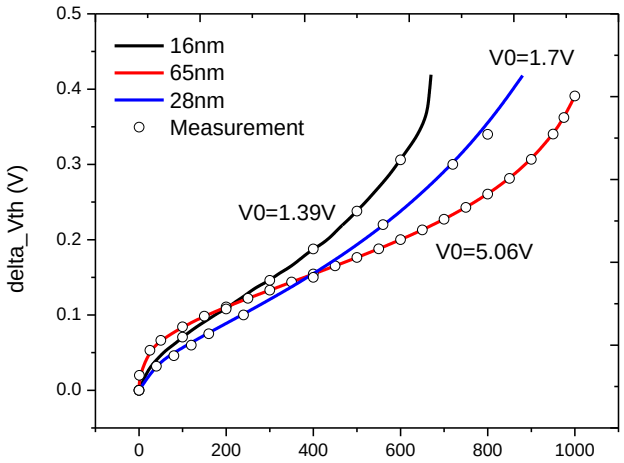


Figure 3.6 Runaway behavior correctly predicted by model

technology nodes measured. This degradation increases with initial bias V_{bias} finally leading to run-away when A>1 as seen in Fig 3.6. Once behavior of g_1 and g_2 is determined, the compact equation can predict the degradation under feedback accurately. This equation can also determine the time of runaway based on stress conditions as shown in Fig. 3.6. The nodes in a circuit that are prone to bias runaway can be identified and this model used as Verilog-A block for the effected devices in order to study its impact on circuit.

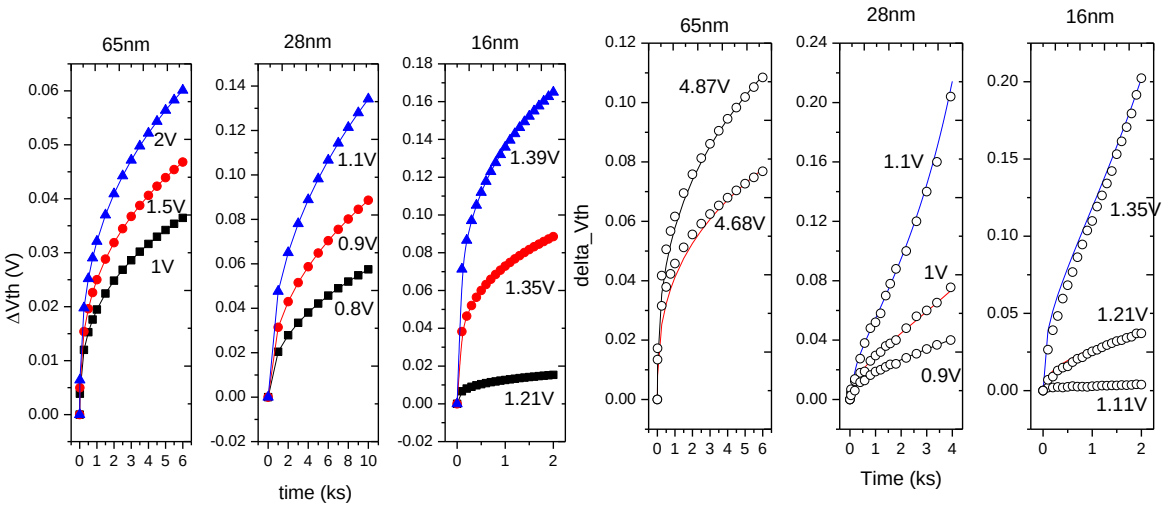


Figure 3.7 HCI calibration (left) and accelerated aging behavior (right) measured for 3 technology nodes.

3.1.2. **Accelerated aging for circuits**

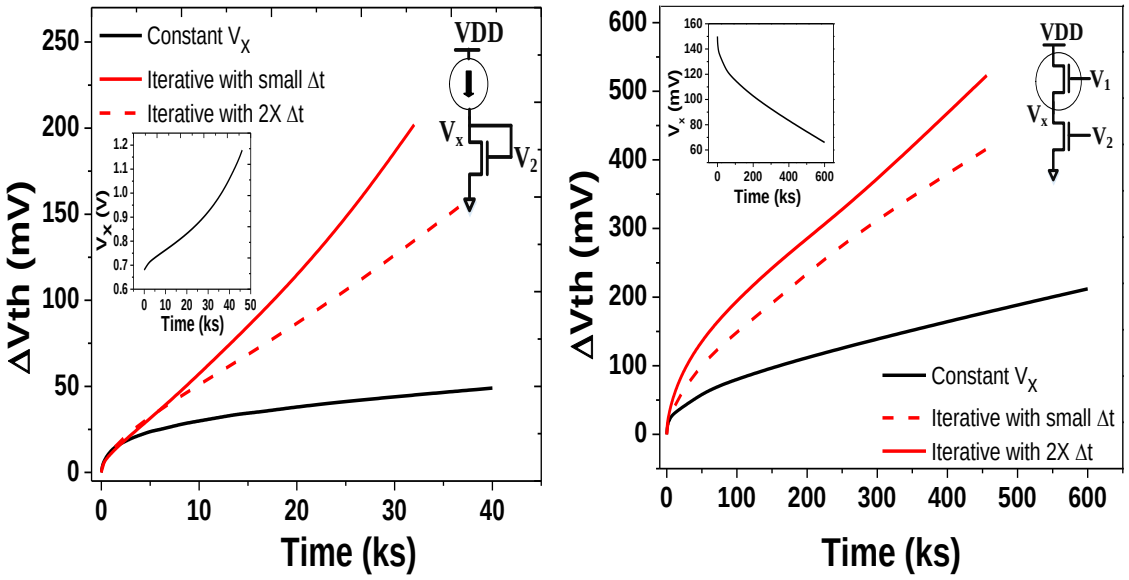


Figure 3.8 (left) Diode connected configuration showing accelerated aging. (right) stacked NMOS accelerated aging

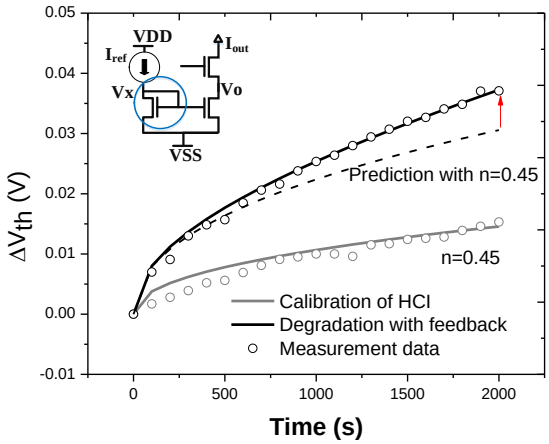


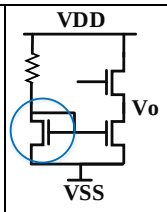
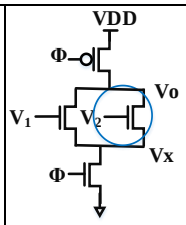
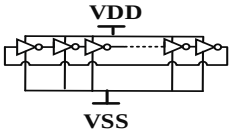
Figure 3.9 Circuit simulation showing accelerated aging in constant bias current configuration at 16nm technology node.

Simulations have been performed for various circuits [1] to determine the impact of feedback loops on their functioning. The usefulness of the compact model in terms of accuracy and speedup in computation time is validated in table 3.2. The circuits that were considered are bias generation circuits seen in AMS designs (Fig 3.8, fig 3.9), dynamic gate with stacked inverters (Fig 3.8) and ring oscillator (RO) to represent a digital circuit under DVFS feedback. Circuit I is the bias generation circuit as shown in Table II has a diode connected NMOS. This device creates the feedback loop which accelerates aging. The failure condition for this circuit is considered as 10% decrease in the bias voltage supplied at the output. Similarly in the stacked NMOS for circuit II, the intermediate node x has accelerated voltage decrease. A smaller V_x causes the V_{ds} of the transistor to increase over time thus causing increased HCI than predicted by a compact model assuming constant initial V_{ds} as the stress condition. The failure condition here is considered as 20% increase in the propagation delay. In circuit III, the frequency of RO decreases with time. The supply voltage VDD is increased to compensate for this decrease. But the increased

stress voltage causes higher degradation thus driving the RO into a feedback loop which accelerates aging. The failure condition here is considered as a 10% decrease in the initial frequency of operation.

If just traditional aging models for all the devices in these circuits is considered, the time to failure (TTF) is ~2X higher than the actual TTF seen in the iterative simulations. On the other hand performing iterative simulations for the circuits increases the computation time ~30X times. Table II shows that the compact model formulation is able to predict the TTF accurately for computation time same as traditional model as it does not require any iterations.

Table 3.2 Comparison of I. traditional, II. Iterative and III. Compact models

			
TTF(Ks)			
I	180	140	216
II	64	99	84
III	64.8	95	82
Computation time(s)			
I	1.39	3.03	2.25
II	41.7	62	67.5
III	2.12	3.22	2.96

3.2. **Adaptive Accelerated Aging: Devices**

Based on the phenomenon of accelerated aging under positive feedback, a new method of Adaptive Accelerated Aging (AAA) was devised to speed up the degradation of the devices-under-test (DUTs) in a controlled manner, in order to rapidly characterize the entire aging process until the end of their life. Similar as other works, EOL or device failure is defined by gate dielectric breakdown leading to sudden jump in gate current [9]. In addition to accelerating aging, another intention of AAA is its adaptability to the response of individual devices, i.e., a weaker device with higher level of defects will fail faster while

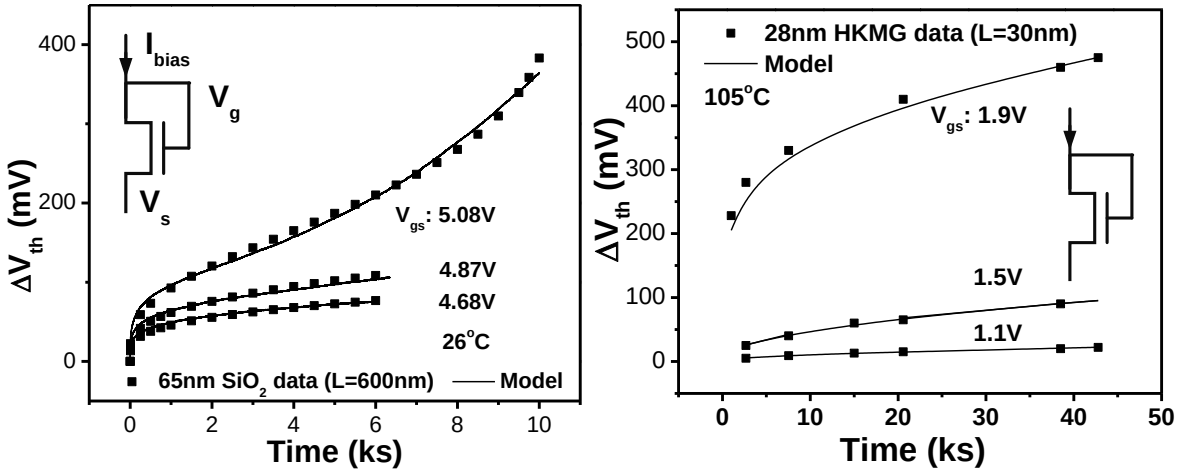


Figure 3.10 Bias runaway occurs at 65nm due to HCI, when the loop gain>1 (left); but at low temperature, it requires a very high bias voltage which reduces the controllability of the experiment. In 28nm HKMG (right), such runaway is not observed in the same diode structure, due to the dominance of gradual BTI at high temperature.

a better device will degrade less or might not even get damaged during the stress test. Detailed description and formalism of the proposed method is presented in the following sub-sections.

3.2.1. **Derivation of compact model**

From section 3.1.1.1,

$$BTI\colon \Delta V_{th}(t+\Delta t) = \Delta V_{th}(t) + \sqrt{V_{gs}(t) - V_{th}(t)}\, e^{\frac{(V_{gs}(t) - V_{th}(t))}{E_0 T_{ox}}} (A\Delta t^n)$$

Where $g_1 = \frac{\partial Vgs}{\partial \Delta Vth} = \alpha$; $g_2 = \frac{\partial \Delta Vth}{\partial Vgs} = 0$. Thus $g_1 g_2$=0 and this is a stable system. The feedback loop is modified to ensure that $V_g(t_n) = Vg(t_{n-1}) + \alpha\, \Delta V_{th}(t_{n-1})$ s.t $V_{gs}(tn) - V_{th}(tn) = (\alpha\text{-}1)\, \Delta V_{th}(t_{n-1})$ where α is a constant. This stress voltage increment is done when Id drops by x% of its present value. Now for a time duration Δt at the end of time t,

$$BTI\colon \Delta V_{th}(t+\Delta t) = \Delta V_{th}(t) + \sqrt{V_{gs}(t) - V_{th}(t)}\, e^{\frac{(V_{gs}(t) - V_{th}(t))}{E_0 T_{ox}}} (k'\Delta t^n) \quad (14)$$

Assume that the total stress time T is divided into N intervals and look at behavior of the feedback loop in each interval. Define $F_0 = k'\sqrt{V_{gs0} - V_{th0}}\, e^{\frac{(V_{gs0} - V_{th0})}{E_0 T_{ox}}}$ so that $F_1 = \Delta Vth0 = F_0 \left(\frac{T}{N}\right)^n$ when a device is initially stressed for time a small interval of time T/N. $\Delta VthB = F_0(T)^n$ is the baseline curve as before. Then in the next interval, Vth shift is:

$$F_2 = k'\sqrt{V_{gs0} + \alpha\Delta Vth0 - V_{th0} - \Delta Vth0}\, e^{\frac{(V_{gs0} + \alpha\Delta Vth0 - V_{th0} - \Delta Vth0)}{E_0 T_{ox}}} \left(\frac{2T}{N}\right)^n$$

$$= k'\sqrt{V_{gs0} - V_{th0} + (\alpha - 1)\Delta Vth0}\, e^{\frac{(V_{gs0} - V_{th0} + (\alpha-1)\Delta Vth0)}{E_0 T_{ox}}} \left(\frac{2T}{N}\right)^n \quad (15)$$

Expanding using Taylor series:

$$(x + y)^{\frac{1}{2}} = \sqrt{x} + \frac{y}{2\sqrt{x}} + \cdots \text{ (higher powers of y ignorable)}$$

$$F_2 = k'\sqrt{V_{gs0} - V_{th0}}\, e^{\frac{(V_{gs0} - V_{th0})}{E_0 T_{ox}}} \left(\frac{2T}{N}\right)^n e^{\frac{((\alpha-1)\Delta Vth0)}{E_0 T_{ox}}} + k'(\alpha - 1)\Delta Vth0 \frac{e^{\frac{(V_{gs0} - V_{th0})}{E_0 T_{ox}}}}{2\sqrt{V_{gs0} - V_{th0}}} \left(\frac{2T}{N}\right)^n e^{\frac{((\alpha-1)\Delta Vth0)}{E_0 T_{ox}}}$$

$$= \left(\frac{2T}{N}\right)^n e^{\frac{((\alpha-1)F1)}{E_0 T_{ox}}} [F_0 + k'(\alpha - 1)CF_1] \quad \text{where } C = \frac{e^{\frac{(V_{gs0} - V_{th0})}{E_0 T_{ox}}}}{2\sqrt{V_{gs0} - V_{th0}}}$$

expanding the exponent and ignoring higher powers

$$F_k = \left(\frac{kT}{N}\right)^n \left(F_0 + (\alpha - 1)\frac{F_0 F_{k-1}}{E_0 T_{ox}} + k'(\alpha - 1)CF_{k-1}\right)$$

Looking at Nth interval (k=N) and simplifying:

$$F_N = F_0(T)^n \left\{1 + \left(\frac{N-1}{N}\right)^n DT^n + \left(\frac{N-1}{N}\right)^n \left(\frac{N-2}{N}\right)^n (DT^n)^2 + \cdots +$$

$$\left(\frac{N-1}{N}\right)^n \left(\frac{N-2}{N}\right)^n \left(\frac{N-3}{N}\right)^n \ldots \left(\frac{2}{N}\right)\left(\frac{1}{N}\right)(DT^n)^N\right\} \quad (16)$$

where $D = (\alpha - 1)\left\{\frac{F_0}{E_0 T_{ox}} + k'C\right\}$. On taking the limit N-> inf $F_N = F_0(T)^n\{1 + DT^n +$

$(DT^n)^2 + (DT^n)^3 \ldots\}$ which can be simplified as DT^n <1 so it is convergent infinite sum of geometric series:

$$F_N = F_0(T)^n \left[\frac{1}{1 - DT^n}\right]$$

Therefore the shift in Vth is given by

$$\Delta Vth = \frac{\Delta Vth_B}{1 - \frac{(\alpha-1)}{F_0}\left\{\frac{F_0}{E_0 T_{ox}} + k' \frac{e^{\frac{(V_{gs0} - V_{th0})}{E_0 T_{ox}}}}{2\sqrt{V_{gs0} - V_{th0}}}\right\}\Delta Vth_B} = \frac{\Delta Vth_B}{1 - (\alpha-1)\left\{\frac{1}{E_0 T_{ox}} + \frac{1}{2(V_{gs0} - V_{th0})}\right\}\Delta Vth_B} = \frac{\Delta Vth_B}{1 - g_1 g_2} \quad (17)$$

It can be rewritten in terms of loop gains as:

$$g_1 = \frac{\partial Vgs}{\partial \Delta Vth} = \alpha\,; \quad g_2 = \frac{\partial \Delta Vth}{\partial Vgs} = \frac{\Delta Vth_B}{F_0}\left\{\frac{F_0}{E_0 T_{ox}} + k' \frac{e^{\frac{(V_{gs0} - V_{th0})}{E_0 T_{ox}}}}{2\sqrt{V_{gs0} - V_{th0}}}\right\} \quad (18)$$

Thus, $g_1 g_2 = \alpha\left\{\frac{1}{E_0 T_{ox}} + \frac{1}{2(V_{gs0} - V_{th0})}\right\}\Delta Vth_B \geq 1$ as per requirement.

In order to determine the value of x by which to reduce I_{ds} for test, consider $I_d = \frac{A}{2}(V_{gs}(t) - V_{th}(t))^2$. With aging: $I_d \Rightarrow \frac{A}{2}\left(V_{gs} - V_{th} - \Delta V_{th}\right)^2$=(1-x) $\frac{A}{2}(V_{gs} - V_{th})^2$.

Simplifying: $V_{gs} - V_{th} = \frac{1}{1 - \sqrt{1-x}}\Delta V_{th}$ and taking derivative, $\frac{\partial Vgs}{\partial \Delta Vth} = \frac{1}{1 - \sqrt{1-x}} = g_1 = \alpha$

Thus for discrete steps the decrease of I_d that prompts increase in Vgs by an amount $\alpha\Delta Vth$ is x=$\frac{2\alpha-1}{\alpha 2}$.

3.2.2. **Test Methodology**

The essential idea of AAA is to stress the device in a closed-loop fashion, and adaptively control the loop gain through the feedback. For example, a device stressed in a diode connection (i.e., gate and drain connected) with a constant bias current (I_{bias}) contains a feedback loop, as shown in Fig. 3.10 (left) [15]. The corresponding voltage at the gate node is V_{bias}. While the aging effect under constant stress voltage is usually a gradual process, the degradation rate in such a closed-loop structure is significantly elevated [9]. The DUT may even experience bias runaway, due to the positive feedback between gradual HCI degradation and constant bias control for 65nm technology. As the device degrades continuously under HCI, the threshold voltage V_{th} increases; meanwhile since the device needs to maintain a constant I_{bias}, bias voltage V_{bias} is forced to increase; the higher value of V_{bias} in turn accelerates the stress under HCI. When the initial V_{bias} is high enough, the loop gain is larger than one and thus, the device enters the positive feedback region and rapidly fails. Bias runaway that happens in this 65nm diode structure inspires the proposed AAA method to speed up the degradation. However, a direct employment of the diode structure to 28nm HKMG fails to trigger bias runaway, as observed in Fig. 3.10 (right). In a HKMG NMOS device at high temperature, PBTI is the dominant factor, not HCI (Fig. 3.1) [17]. Compared to HCI, BTI is much more gradual, especially in the long term. Therefore, the loop gain is always smaller than one in the diode connection, and the device never enters the instable region to trigger the runaway [5]. To overcome this issue, Fig.

3.11 further examines the loop gain and proposes the adaptive scheme. In the diode connection, the loop gain (A) has two components, as shown in Fig. 3.11: the first term is the gain in the aging physics (i.e., the degradation rate as a function of the bias), and the second term is the gain in the circuit topology (i.e., the change of V_{bias} to maintain I_{bias}). At 65nm, the second term is passive while the first term is significant due to the strong dependence of HCI on V_{bias}; at 28nm, however, BTI has a much weaker dependence on V_{ds} and thus, A does not exceed one in this passive connection. By recognizing such limitation of BTI, the solution of AAA is to proactively adjust the second term, the gain of V_{bias} to I_{bias} in the modified feedback loop (i.e., $\Delta V_{st}/\Delta I_{ds}$ as in Fig. 3.11), in order to force the device to enter the runaway region, even under the moderate BTI.

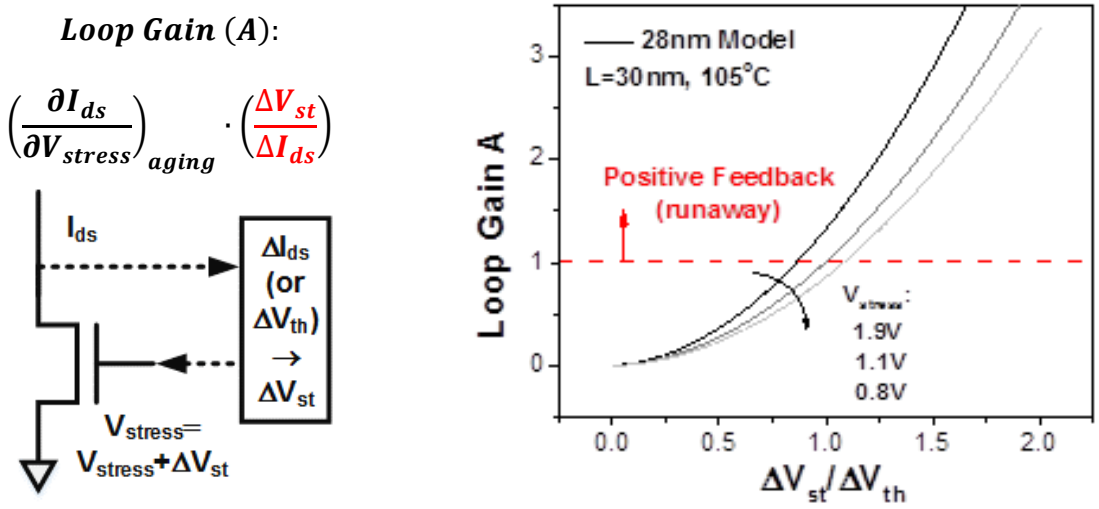


Figure 3.11 The proposed acceleration scheme actively tunes the stress voltage in order to drive the DUT into the region of the positive feedback (A>1), even under the gradual BTI effect.

In practice, continuously monitor the degradation of I_{ds} (ΔI_{ds}). When the change exceeds a threshold (η%), the stress voltage will be adaptively increased by ΔV_{st}. Note that the gain $\Delta V_{st}/\Delta I_{ds}$ here in AAA is controlled by the test scheme rather than the circuit

topology only, as in the diode connection (Fig. 3.10). By replacing the constant bias current in the diode structure with the adaptive shift and selecting appropriate $\Delta V_{st}/\Delta I_{ds}$, tune the loop gain A to be either < 1 or > 1, to set the DUT in the negative feedback or positive feedback, respectively. Figure 3.11 (right) confirms the AAA method with 28nm simulations: when $\Delta V_{st}/\Delta I_{ds}$ is large enough, this closed-loop feedback structure enables runaway and speeds up the aging process. A higher initial value of the stress voltage (V_{stress}) further helps drive the DUT into runaway with a smaller ΔVst/ΔIds.

3.2.3. **Test procedure**

Figure 3.12 summarizes the test procedure of AAA at the device level. For a discrete device, the shift of its drain current or the shift of threshold voltage (ΔI_{ds} or ΔV_{th}, respectively) is used to monitor the aging process, while the gate current (I_g) is sampled at the same time to evaluate the progress of gate dielectric breakdown. If the increase of I_g (ΔI_g) does not exceed a threshold, i.e., I_{fail}, then the stress continues; otherwise, the end of life of DUTs is reached.

During the stress period, the change of I_{ds} (or V_{th}) is the index of the degradation rate. When its amount of change is more than a given percentage, η%, the adjustment of the stress condition is activated. Different from a conventional stress test with constant conditions, the feedback in AAA is designed towards a positive loop in order to accelerate the degradation: the more degradation that happens, the higher the stress voltage is tuned to be. The exact gain at this step is controlled by the gain of $\Delta V_{st}/\Delta I_{ds}$, as shown in Fig. 3.11. Moreover, such a feedback implies an adaptation to the intrinsic quality of a DUT: a weaker DUT will degrade faster and thus, experiences more frequent increase of the stress

voltage, while a DUT with higher quality will have much fewer voltage adjustment during the test period. Therefore, for the same initial test condition and the same test period, a weaker DUT will end up with a much faster degradation. This behavior well serves the purpose to speed up the aging while protecting the strong DUT. Note in AAA, the stress voltage is used as the adaptive variable because of its faster tuning rate in the test, rather than the slower change of the stress temperature.

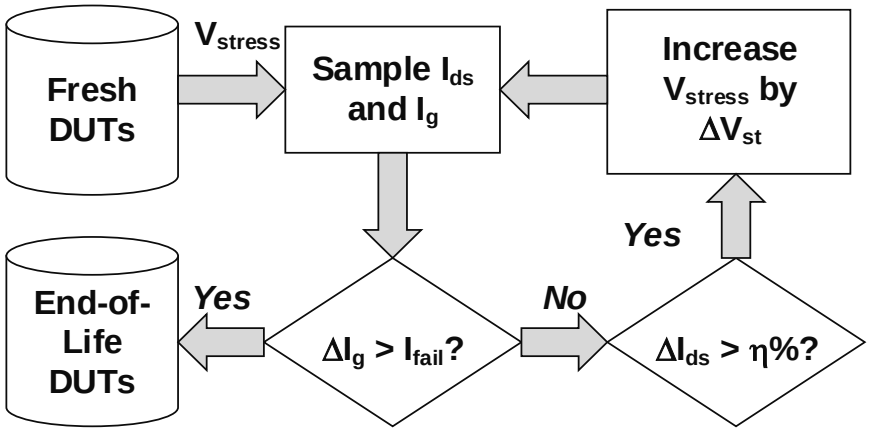


Figure 3.12 The procedure of AAA, in which the shift of I_{ds} (or V_{th}) is used as the monitor to adaptively increase V_{stress}. This iterative methodology can be performed till the end of life, such as dielectric breakdown.

Overall, the acceleration rate in AAA is controlled by the initial stress condition, the threshold of ΔIds (η%) to trigger the increase of V_{stress}, the sensitivity of V_{stress} to such change $\Delta V_{st}/\Delta I_{ds}$), and most importantly, the specific aging mechanism.

3.2.4. **Silicon Results**

The proposed AAA method is applied to 28nm HKMG devices, with the minimum gate length of 30nm. First, compact models of HCI and BTI for discrete NMOS devices

are calibrated at various stress voltages and temperatures, and decomposed as shown in Fig. 3.1 [4]. Then the AAA flow in Fig. 3.12 is conducted on multiple devices. While the end of life is defined by TDDB (e.g. I_g increased to a value I_{fail}), track both changes of I_{ds} (or V_{th}) and I_g, and explore the correlation between BTI and TDDB. When I_{ds} is η% lower than the value in previous measurement, the stress voltage, V_{stress}, will be adaptively increased by ΔV_{st}. This loop is executed till EOL is achieved, e.g. I_g is more than 100X than the fresh value. Figure 3.13 presents the test results of four NMOS devices at 28nm HKMG, all with the same width and length. They start from the same stress condition and V_ds is constant. For three of them, AAA is applied; for the fourth one, V_{stress} is kept constant at 1.1V, as the control DUT to provide a baseline of degradation.

As indicated in Fig. 3.11, the positive feedback can be triggered if $\Delta V_{st}/\Delta I_{ds}$ is carefully selected and thus, the loop gain A is more than 1, even though the gain in BTI $(\Delta I_{ds}/\Delta V_{stress})_{aging}$ is moderate. In this experiment, ΔI_{ds} is set as 3.2% (η=3.2), i.e., when I_{ds} is shifted by 3.2%, the gate voltage, V_{stress}, is added by ΔV_{st}; the device is then stressed at this new V_{stress} until I_{ds} degrades by another 3.2% or TDDB happens, as described in Fig. 3.12. For these three identical DUTs, various amounts of ΔV_{st} are tested to validate the approach and prediction by the aging model. When ΔV_{st} is small (e.g. 200mV), the device only experiences a mild elevation of the degradation rate, compared to that under constant stress. Yet when ΔV_{st} is large enough (e.g. 400mV in this case), the runaway behavior is activated, because the loop gain is significantly larger than 1 now. As a consequence of the positive feedback, the DUT rapidly degrades and eventually has a gate breakdown. Note that the moments of large I_{ds} change and I_g change are synchronized in Fig. 3.13, indicating that the discrete failures in the HK dielectric may be caused by the same trapped charges

that affect the trapping/detrapping behavior of PBTI [14]. For both gradual degradation and runaway, aging models of HCI and PBTI extracted from short-term measurement (Fig. 3.1) well predict the shift through the entire process.

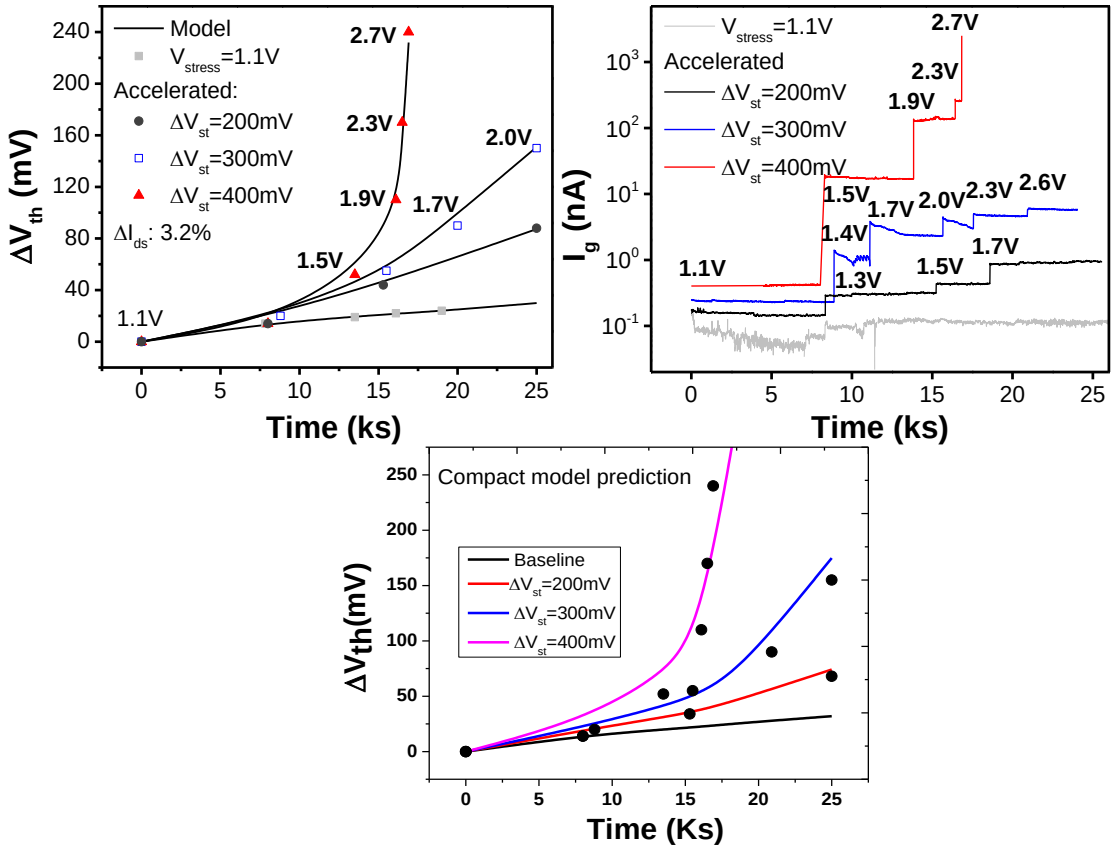


Figure 3.13 The implementation of adaptive accelerated aging at 28nm HKMG devices (L=30nm, V_{ds}=0.9V). BTI and TDDB are coupled during AAA, and match the model prediction. V_{th} (extracted from I_{ds}) and I_g are monitored simultaneously, as presented by the left and right figures, respectively. The bottom figure shows comparison with compact model.

Therefore, the AAA method provides an effective approach to calibrate aging models to EOL, using a short time period of experiments. Based on the validated model, Fig. 3.14 further explores the design space of AAA. In order to induce the runaway

behavior earlier and faster, a larger gain of ($\Delta V_{st}/\Delta I_{ds}$) is preferred, which requires a small ΔI_{ds} and a large ΔV_{st}. For instance, if the target is to

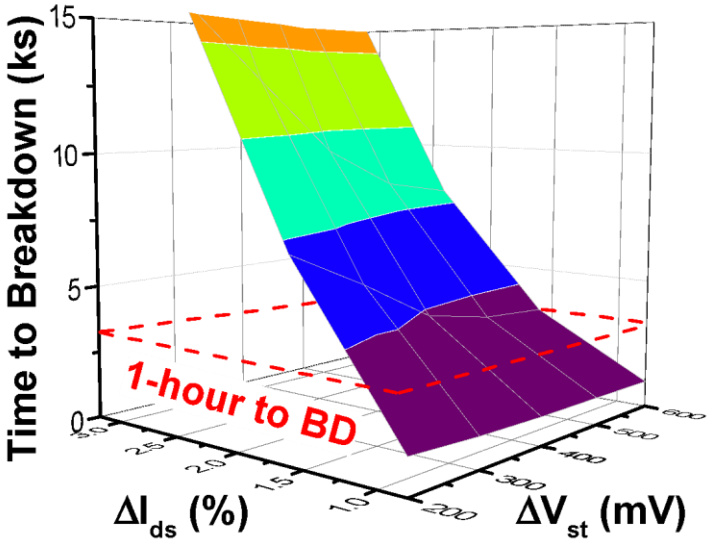


Figure 3.14 An appropriate selection of ΔV_{st} and ΔI_{ds} is able to induce failure of dielectric breakdown (BD) within one hour, driving the DUT to its EOL

drive TDDB to occur within one hour, ΔI_{ds} should be 1.6% and ΔV_{st} to be 200mV. As shown in Fig.3.13 , V_{stress} is still within 3V when TDDB happens, which is practical for the stress test. Simulation results as shown in Fig.3.14 help design the experiment, depending on the target time duration.

3.3. **Adaptive Accelerated Aging: Digital circuits**

Extensive works have been performed on physical aging models at the device level [5, 20]. These device models need to be propagated to the circuit level, helping manage system failure due to aging. However, aging analysis in digital circuits faces many additional challenges, such as process/voltage/temperature variations and dynamic workloads [19, 20]. These process and environmental uncertainties significantly affect the

aging behavior for circuits, especially the recovery phase in BTI [18]. Even though there are benchmark programs to estimate the average switching activity and long-term degradation models have been developed based on them [5], a direct and efficient measurement towards EOL is still preferred at the circuit level, such that the quality of aging analysis tools can be justified. This section extends AAA to the circuit level, using the ring oscillator (RO) as the test case for demonstration.

At 28nm and below, BTI is the dominant effect in digital circuits. BTI degradation recovers after removal of stress due to de-trapping and interface trap passivation [21]. HCI on the other hand does not undergo any noticeable recovery effect. In dynamic circuit operation, these effects work together, HCI during the transition and BTI during the constant stress in the cycle. This leads to speedy degradation. For a ring oscillator, use its oscillation frequency (F) to evaluate the performance. The shift of F (ΔF) is the index of aging, similar as ΔIds for device aging. As shown in Fig. 3.15, the mixture of BTI and HCI has different impact on RO frequency and duty cycle of the switching waveform, during static (DC) and dynamic (AC) stress. Under the static condition, duty cycle is affected asymmetrically (i.e., asymmetric aging) because different switching edge suffers from different amount of degradation. For dynamic stress, duty cycle is not affected.

To simplify the monitoring of aging, apply dynamic stress in AAA to the RO. In this case, ΔF is the single parameter to assess the degradation rate. As ΔF increases, a digital design may develop multiple paths that fail timing checks, logic correctness, etc.

which are required by design specification. For demonstration in this study, use $\Delta F > 40\%$ as an empirical value to represent end of life in RO simulation under AAA.

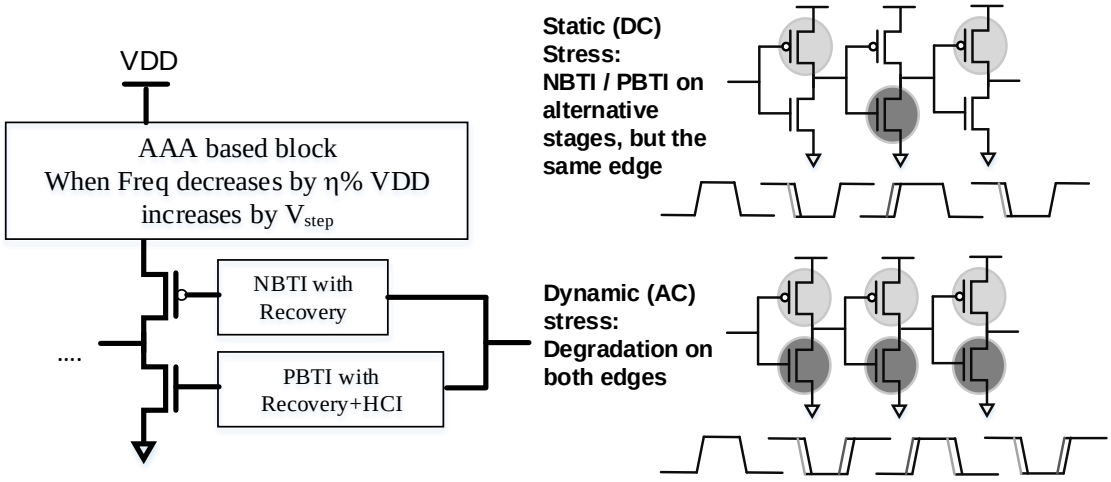


Figure 3.15 PBTI and NBTI jointly affect the degradation of a ring oscillator, while HCI is marginal. The impact on duty cycle is different between DC stress and AC stress. So Verilog-A blocks are added to model their impact.

The AAA method is demonstrated on the ROs using SPICE simulation. The aging models are introduced to each device with a Verilog-A module to model V_{th} shift under multiple aging mechanisms, such as HCI and BTI [4, 22]. The overall flow of AAA on ROs is similar as that in Fig. 3.15. The main difference include: (1) the index of aging is RO frequency, not I_{ds} or V_{th} at the device level. Once ΔF is more than a designed value, η%, the stress voltage will be increased; (2) V_{DD} is the only voltage to stress the dynamic operation of ROs. V_{DD} goes up by ΔV_{st} in each iteration loop. By choosing an appropriate ratio of ($\Delta V_{st}/\Delta F$), the goal is to trigger the positive feedback of RO, such that the degradation towards EOL can happen rapidly.

Figure 3.16 presents the simulation results of AAA on 28nm ROs. Depending on the combinations of ΔV_{st} and ΔF, the degradation of RO frequency can be gradual or runaway. A larger ΔV_{st} (i.e., the step size of stress voltage increase) and a smaller ΔF (i.e., the threshold to activate VDD increase) are preferred to induce the runaway behavior. Note that under the same ratio of ($\Delta V_{st}/\Delta F$), such as 100mV/1.6% and 200mV/3.2% in Fig. 3.10, a smaller ΔF triggers the increase of V_{DD} more frequently and thus, runaway happens faster. Finally Table 3.2 summarizes the time to reach EOL in the RO. It is feasible to apply AAA to a digital circuit and drive the circuit into rapid degradation within one hour. This is consistent with experimental results at the device level, and confirms the efficacy of AAA for fast and complete aging test.

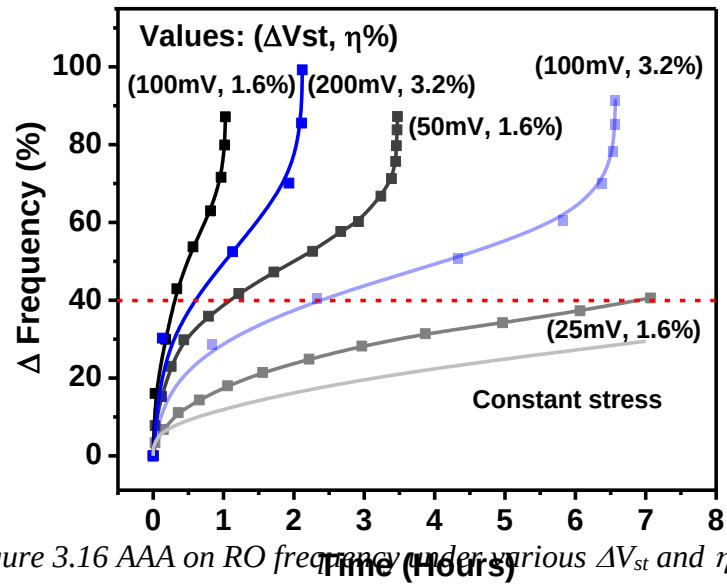


Figure 3.16 AAA on RO frequency under various ΔV_{st} and η%. The runaway behavior is possible by selecting appropriate feedback control. In these simulations, an 11-stage 28nm RO is stressed under 105°C. V_{DD} = 0.9V.

Time to EOL (seconds)		V_{st} (mV)			
		25	**50**	**100**	**200**
η% for ΔF	**1.6%**	36360	12600	3620	1300
	3.2%	172800	75600	23040	7200
	4.8%	450000	158400	57960	22320

Table 3.2 Summary of the time to EOL for an 11-stage RO at 28nm. The shadowed region represents that EOL is reached in about one hour.

CHAPTER 4. Aging in VLSI- SyRA

4.1. Framework for SyRA-X

Determining aging and End of Lifetime (EOL) for VLSI circuits can be a slow and cumbersome process. A simple extrapolation of aging results, under statistical variations and other uncertainties, may cause large errors in EOL prediction and over-margining. This section provides a fast and effective simulation solution to determine guard band against aging in a design. Here, an improved version of System Reliability Analyzer (SyRA), SyRA-X is presented, to provide: (1) circuit aging calculation which is reliable under various switching activities at each node, (2) device, gate, and path level analysis of aging with dynamic inputs, and (3) verification of SyRA-X with a minimum guard band to give prediction over 99% accuracy.

The efficiency and accuracy of path delay degradation are increasingly important as design size keeps scaling up and delay margins become tighter [23]. If the design does not account for aging within a realistic margin, it may not function correctly for its designed lifetime. Yet such needs are not sufficiently managed by existing PDK models. Fig. 1.4 illustrates the reliability analysis at 28nm high-K metal-gate (HKMG) with SPICE, which is inefficient for large-scale digital design. As the scale and the complexity of VLSI design increase, the time taken to calculate aging increases linearly. Additionally, the physics of recovery in BTI is not well described by existing models in the PDK, causing a significant overestimation of path delay degradation (ΔT_d), as shown in Fig. 1.4. In design practice, such erroneous aging prediction may lead to inappropriate optimization and tradeoffs among performance, power and reliability.

To provide an effective simulation solution, previous work, System Reliability Analyzer (SyRA), was implemented to integrate device aging models, standard gate library, large-scale static timing analysis and the management of switching activity (α), where α is defined as the average portion in a clock cycle when a signal stays 'on' (Fig. 4.1) [24]. It consists of running the design benchmark to determine α at each node, based on which ΔV_{th} of the gates are calculated. Using standard cell library at two different VDD, the delay due to change in VDD (ΔTd,VDD) is determined, which is used in Eq. 14 to generate the overall shift in delay caused by aging.

One of the shortcomings of SyRA is the estimation of α. During realistic circuit operations, α is an application-dependent variable and therefore, an empirical value, generated from specific design benchmarks, is neither accurate nor efficient. Therefore, SyRA-X is developed as an improvement over SyRA. The aforementioned model proposes to use a fixed switching activity in ΔTd prediction, instead of a dynamic α. This will reduce the runtime substantially without compromising the aging results for the design. SyRA-X method is possible because of the co-existence of PBTI plus HCI and NBTI in NMOS and PMOS, respectively, in this 28nm HKMG technology, as explained in next section. The efficacy of SyRA-X is fully validated in a 28nm design.

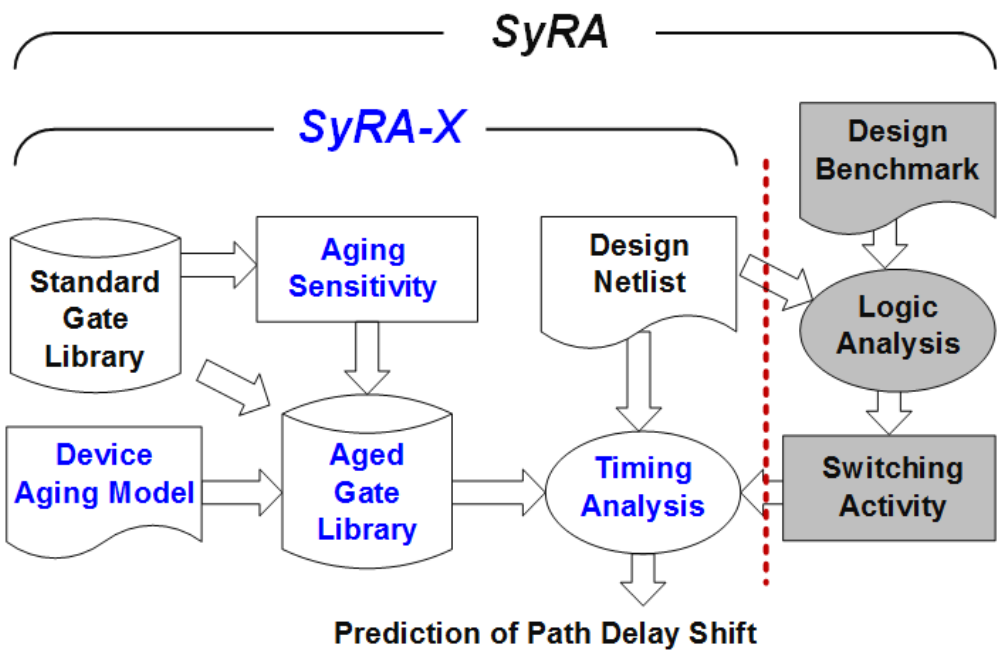


Figure 4.1 SyRA-X removes the estimation and propagation of switching activities in SyRA, which requires empirical benchmarking of the activity and logic analysis through the netlist.

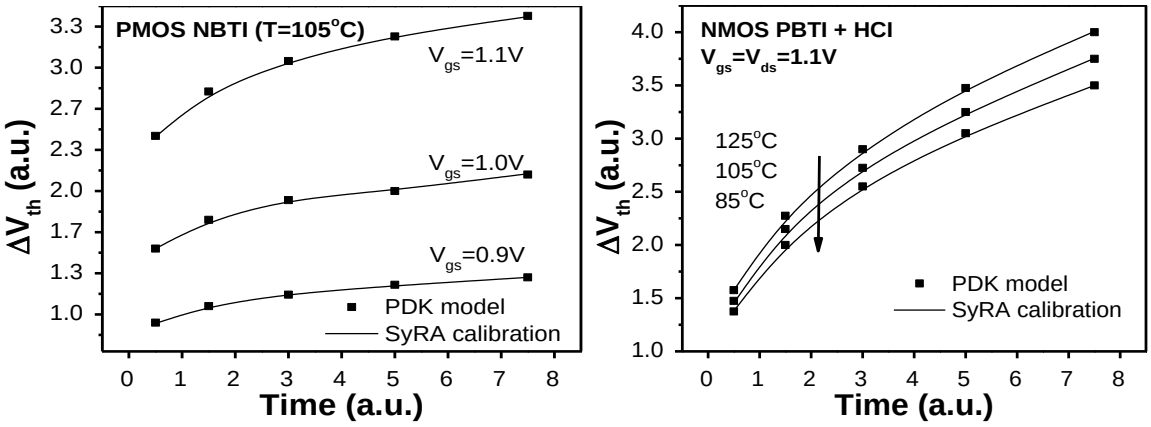


Figure 4.2 . At the device level, SyRA calibrates static models with 28nm PDK, including NBTI (left) and PBTI and HCI (right). Additional calibration is needed for dynamic aging models.

SyRA-X flow begins with Static Timing Analysis (STA) which generates a fresh timing report without aging with timing information for all gates in the design. Unlike SyRA, where α is calculated for each node, a value of α=0.5 is assigned to all nodes for dynamic stress; node voltage is assigned in case of static stress. The clock transition time is used to compute the HCI contribution to aging. ΔVth is calculated for the gates using long term NBTI, PBTI and HCI models. Furthermore, the gate delay shift ΔTd,VDD due to change in supply voltage VDD is calculated either directly from the standard cell library characterized at two closely spaced VDD values or using SPICE simulation. Eventually, the shift in delay ΔTd,Vth caused by change in ΔVth due to aging can be computed as

$$\Delta T_{d,V_{th}} = -\left(\frac{VDD}{V_{th}}\right)\left(\frac{\Delta T_{d,VDD}}{\Delta VDD}\right)\Delta V_{th} \quad (14)$$

Long term RD models are used to calculate ΔVth [24]. For complex gates, ΔVth is calculated by considering the weightage of switching gates and number of gates traversed from input to output. Also, based on the Td,VDD Vs VDD graph, slope from a relatively linear region should be chosen.

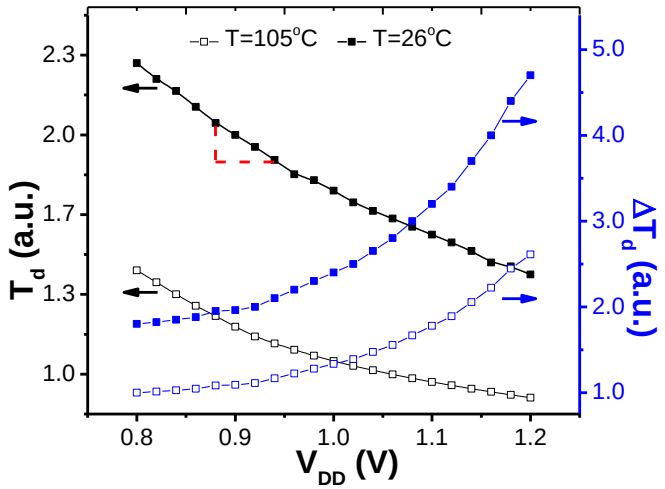


Figure 4.3 At the gate level, the sensitivity analysis converts T_d dependence on V_{DD} to V_{th} [5].

4.2. **Model Calibration and Simulation**

Fig. 4.1 presents the simulation flow with SyRA-X. The cornerstones are device aging models, an aging-aware library and the integration with static timing analyzer, while using a constant switching activity for critical path analysis. At the device level, SyRA-X leverages static aging models provided by the 28nm PDK. The calibration includes voltage, temperature and time dependence for NBTI, PBTI and HCI, as shown in Fig. 4.2. Additional parameters in the recovery phase are extracted from dynamic aging data, as the existing PDK is relatively insensitive to the recovery time (Fig. 1.4). To include aging effects into the standard gate library without resorting to expensive library characterization or silicon test, SyRA-X exploits the sensitivity analysis, based on the assumption that the shift of V_{th} (ΔV_{th}) and gate delay is still much smaller than their nominal values [24, 5]. Fig. 4.3 presents the calculation of ΔT_d as a function of V_{th} shift and the sensitivity of T_d to V_{DD}, which can be obtained from an existing library for NAND2.

The device ΔV_{th} model and standard cell library, providing T_d with changing supply voltage, readily generate two of the essential factors for predicting gate delay shift in a VLSI design. On the other hand, the dynamic switching activity complicates the calculation due to the mixture of stress and recovery [5, 23, 24]. If only NBTI exists, the impact of α on Td is monotonic, as observed in Fig. 4.4; but with contribution of PBTI, the effect of α at the input node of each gate is more uniform and can be estimated for worst case in order to get an accurate prediction [5, 24].

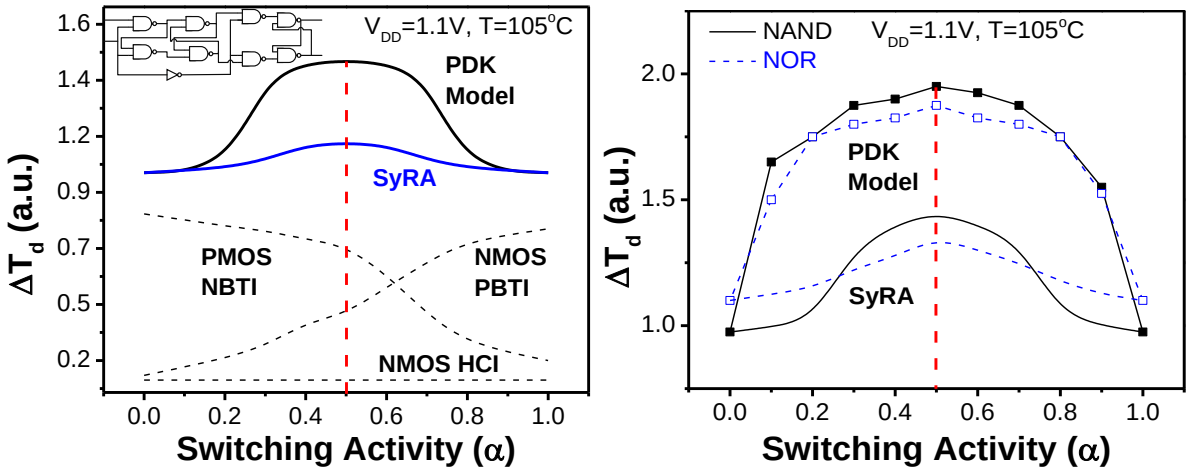


Figure 4.4 Examples of gate delay shift: D-FF (top), NAND and NOR (bottom). Due to the co-existence of NBTI and PBTI in 28nm HKMG, the peak of ΔT_d is around α=0.5

For gates like NAND and NOR, SyRA model with its account for recovery, shows much less variation in ΔTd as compared to the existing PDK model. Even for complex gate like D-flip-flop, the decomposition of NBTI and PBTI as calculated with SyRA shows an averaging effect. However with both NBTI and PBTI occurring together at 28nm HKMG, their interaction leads to the peak degradation at α~0.5 in various circuits. The effect of α becomes insignificant when this analysis is carried out for a number of timing path as shown in Fig 4.5. Note that for a specific node, only its long-term average of α is needed for aging analysis [23].

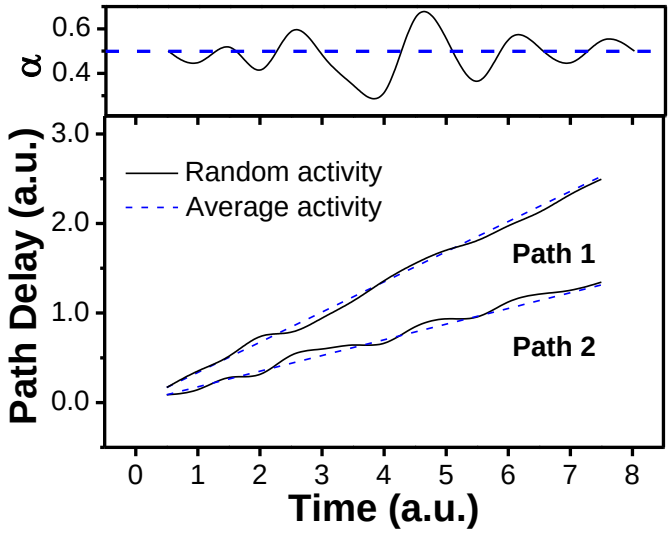
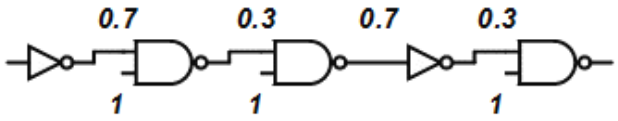
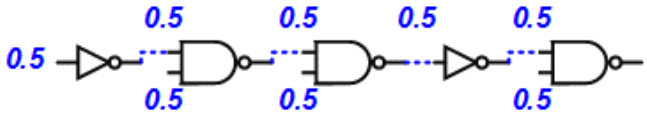


Figure 4.5 For one switching node, an average α is sufficient to predict long-term aging.

The aforementioned observation is the physical basis of SyRA-X: It is proposed to apply α=0.5 to input nodes of all gates, without acquiring α at numerous primary path inputs and then propagating them to each gate on the path as was done in SyRA. Depending on the relative strength of PBTI and NBTI, as well as the path structure, an extra guard band (η) is added for safe prediction of EOL (Fig. 4.6). This guard band can be adjusted according to the required design specification at any stage of designing order to avoid over or under optimization.



SyRA: α of each path input is estimated from design benchmarks and then propagated to each gate on the path.



SyRA-X: α of each gate at 0.5. Then a guardband (η) is added to the path delay:

$$\Delta T_d = \Delta T_d(\alpha = 0.5) \cdot 1(+\eta)$$

Figure 4.6 SyRA-X simplifies the calculation of α for each gate, using 0.5 to approximate

The full SyRA-X is implemented into a realistic 28nm design. The timing analysis is performed with a commercial tool at the gate level [25]. The guard band is empirically determined by sampling 1038 paths. Fig. 4.7 presents the result: with a small margin of 2.3%, SyRA-X achieves 99% accuracy in ΔT_d prediction, compared to SyRA with random input α's; only 4 paths have the underestimation of ΔT_d. More comprehensive evaluation is summarized in Table 4. Based on required accuracy in delay shift, a value of guard band can be chosen for the design. Indeed, since ΔTd peaks around α=0.5 at the gate level (Fig. 4.4), a small η is sufficient to improve the accuracy. Even though negative prediction error happens (Table 4), indicating an underestimation of aging, the error is still much smaller than other design margins in practice. Benefiting from the efficiency of parallel timing analysis at the gate level, SyRA-X only needs <1 minute to predict path delay shift (Fig. 1.4). As a comparison, SyRA needs more than 3 minutes to generate switching activities at each node using a benchmark.

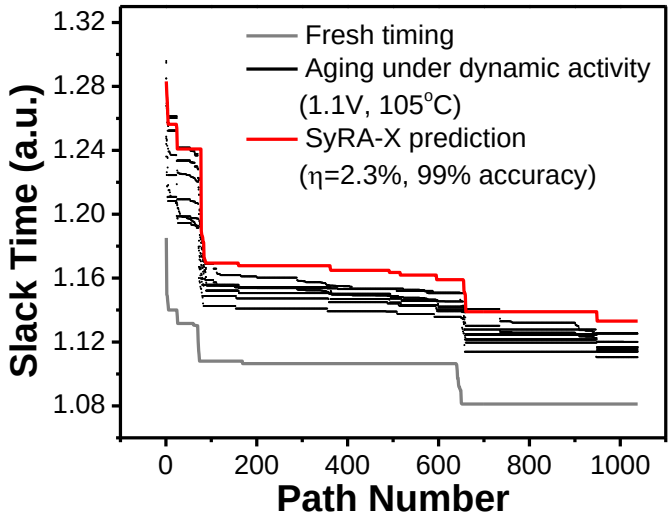


Figure 4.7 SyRA-X well predicts T_d change in 10 years. There are totally 1038 paths

ΔT_d Accuracy	90%	95%	99%	99.5%
Guardband η	0.02%	1.2%	2.3%	2.8%

28nm design	**1**	**2**	**3**	**4**	**5**	**6**	**7**
# of Gates	40	42	52	45	55	49	50
Error (%)	-0.23	-0.95	0.45	0.22	0.10	0.29	-0.74

Table 4 The evaluation of SyRA-X: (top) SyRA-X is a safe and tight prediction of ΔT_d; (bottom) the maximum error in the prediction of critical path ΔT_d compared to full SyRA with random input switching activities.

Chapter 5 Summary and Future Work

This work presents a new way to model the observed degradation of devices due to aging under multiple aging mechanisms like BTI, HCI and TDDB. It provides realistic compact models to predict aging at device as well as circuit level to help model aging and generate fast and efficient delay margins for designing large integrated circuits.

- Chapter 1 introduces the aging models and highlights the necessity and motivation behind developing this work
- Chapter 2 investigates stochastic BTI and TDDB and proposes a new model for the observed correlation. Silicon data at 28nm confirms the coincidence of I_g jump and sudden V_{th} shift. This effect may significantly reduce the lifetime of circuits, demanding more careful evaluation and management.
- Chapter 3 proposes a compact model for estimating the aging for devices in feedback loop. The method is applicable to both discrete devices and circuit-level test structure by adding a Verilog-A module with the compact model. Adaptive Accelerated Aging method is developed as an effective approach to speed up the degradation of a DUT with aging. As demonstrated with 28nm HKMG data, AAA is an efficient method to validate reliability prediction to the end of lifetime. The stress inputs can be so chosen as to decide what time frame to reach the DUT end of life for both devices and circuits.
- Chapter 4 implements SyRA for gate-level aging prediction with more robust analysis under activity uncertainty. The co-existence of both NBTI and PBTI at 28nm enables aging analysis of path delay with fixed α at each gate, rather than

pursuing more expensive activity estimation. The new tool has been demonstrated at 28nm design practice, confirming the effectiveness of SyRA-X.

To develop this work further, the model for BTI and TDDB introduced here is a stochastic model and should be validated over a large number of devices. Also, this model can be tested for modelling recovery observed in BTI. The feedback loop compact model can be extended to cover larger selection of circuit configurations. Being a generic model with dependence on loop gains, it should be able to predict the accelerated aging in various types of circuits. The analysis of SyRA-X is currently done under limited scope. It should be validated further based on silicon results.

Chapter 6 Conclusion

In conclusion, this work addresses the need of device models that can assimilate the TD model along with the gradual trap generation as well as those generated due to TDDB stress causing acceleration in aging. This model is tested for 14nm FinFet devices and 28 nm HKMG devices. Developing further, circuit configurations are studied that present the phenomenon of accelerated aging due to existence of a positive feedback loop during their operation. A compact model is developed based on loop gains that can predict the accelerated degradation of threshold voltage. This phenomenon then inspired the development of a methodology called Accelerated Adaptive Aging that leverages the positive feedback loop and based on user inputs can stress a device or circuit to desired EOL. Being able to stress a device to EOL allows us to have efficient validation of the aging models without having to perform error prone predictions. Finally to make these models usable at VLSI level, SyRA is developed into SyRA-X that can reliably predict the aging at circuit level and provide realistic margins for accommodating degradation due to aging.

REFERENCES

[1] RelXpert, Cadence.

[2] Synopsys, MOS Device Aging Analysis with HSPICE and CustomSim, 2011.

[3] Mentor Graphics, ELDO User's Manual, 2005.

[4] K. B. Sutaria, A. Mohanty, R. Wang, R. Huang and Y. Cao, "Accelerated Aging in Analog and Digital Circuits With Feedback," in IEEE Transactions on Device and Materials Reliability, vol. 15, no. 3, pp. 384-393, Sept. 2015.

[5] J. B. Velamala et al., "Compact Modeling of Statistical BTI Under Trapping/Detrapping," in IEEE Transactions on Electron Devices, vol. 60, no. 11, pp. 3645-3654, Nov. 2013.

[6] K. Sutaria, "Modeling and Simulation Tools for Aging Effects in Scaled CMOS Design," Arizona State University, 2015.

[7] S. Mahapatra et al., "A Comparative Study of Different Physics-Based NBTI Models," in IEEE Transactions on Electron Devices, vol. 60, no. 3, pp. 901-916, March 2013.

[8] D. Ielmini, M. Manigrasso, F. Gattel and G. Valentini, "A unified model for permanent and recoverable NBTI based on hole trapping and structure relaxation," 2009 IEEE International Reliability Physics Symposium, Montreal, QC, 2009, pp. 26-32.

[9] Ahmed Kamal Reza, Mohammad Khaled Hassan and Kaushik Roy, "Büttiker Probe Based Modeling of TDDB: Application to Dielectric Breakdown in MTJs and MOS Devices", et al, arXiv:1609.05919.

[10] W. Wang, V. Reddy, A. T. Krishnan, R. Vattikonda, S. Krishnan and Y. Cao, "Compact Modeling and Simulation of Circuit Reliability for 65-nm CMOS Technology," in *IEEE Transactions on Device and Materials Reliability*, vol. 7, no. 4, pp. 509-517, Dec. 2007.

[11] Z. Yu, *et al*., submitted to IEDM 2017

[12] S. Y. Kim, C. H. Ho and K. Roy, "Statistical SBD Modeling and Characterization and Its Impact on SRAM Cells," in *IEEE Transactions on Electron Devices*, vol. 61, no. 1, pp. 54-59, Jan. 2014.

[13] S. Bhardwaj, W. Wang, R. Vattikonda, Y. Cao and S. Vrudhula, "Scalable model for predicting the effect of negative bias temperature instability for reliable design," *IET Circuits, Devices & Systems*, vol. 2, no. 4, pp. 361-371, Aug. 2008.

[14] J. Yang, *et al*., "Intrinsic correlation between PBTI and TDDB degradations in nMOS HK/MG dielectrics," *International Reliability Physics Symposium*, pp. 5D.4.1-5D.4.7, 2012.

[15] K. B. Sutaria, *et al*., "Diagnosing bias runaway in analog/mixed signal circuits," *International Reliability Physics Symposium*, 2014.

[16] M. Kole, "Circuit reliability simulation based on Verilog-A," *IEEE International Behavioral Modeling and Simulation Workshop*, pp. 58-63, 2007.

[17] S. Tyaginov, M. Jech, J. Franco, P. Sharma, B. Kaczer and T. Grasser, "Understanding and modeling the temperature behavior of hot-carrier degradation in SiON nMOSFETs," *IEEE Electron Device Letters*, vol. 37, no. 1, pp. 84-87, Jan. 2016.

[18] H. Kufluoglu, C. Chancellor, M. Chen, V. Reddy, "An analysis of the benefits of NBTI recovery under circuit operating conditions," *International Reliability Physics Symposium*, pp. 15-19, April 2012.

[19] K. Sutaria, A. Ramkumar, R. Zhu, R. Rajveev, Y. Ma and Y. Cao, "BTI-induced aging under random stress waveforms: Modeling, simulation and silicon validation," *Design Automation Conference*, pp. 1-6, 2014.

[20] J. B. Velamala, K. B. Sutaria, V. S. Ravi and Y. Cao, "Failure analysis of asymmetric aging under NBTI," *IEEE Transactions on Device and Materials Reliability*, vol. 13, no. 2, pp. 340-349, June 2013.

[21] T. Grasser, *et al*., "The paradigm shift in understanding the bias temperature instability: From reaction–diffusion to switching oxide traps," *IEEE Transactions on Electron Devices*, vol. 58, no. 11, pp. 3652-3666, Nov. 2011.

[22] T. Grasser, *et al*., "Defect creation stimulated by thermally activated hole trapping as the driving force behind negative bias temperature instability in SiO_2, SiON, and high-k gate stacks," *IEEE International Integrated Reliability Workshop Final Report*, pp. 91-95, 2008.

[23] V. Huard, C. R. Parthasarathy, A. Bravaix, C. Guerin, and E. Pion, “CMOS device design-in reliability approach in advanced nodes,” in Proc. IEEE Int. Rel. Phys. Symp., Apr. 2009, pp. 624–633.

[24] Y. Cao et al., “Cross-layer modeling and simulation of circuit reliability,” IEEE Trans. Comput.-Aided Design Integr. Circuits Syst., vol. 33, no. 1, pp. 8–23, Jan. 2014.

[25] Synopsys, Inc., PrimeTime Suite, 2014.12