The die-to-die calibrated combined model of negative bias temperature instability and gate oxide breakdown from device to system

Soonyoung Cha*, Dae-Hyun Kim, Taizhi Liu, Linda S. Milor
School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA

A R T I C L E   I N F O
Article history:
Received 24 May 2015
Received in revised form 23 June 2015
Accepted 24 June 2015
Available online xxxx

Keywords:
Bias temperature instability (BTI)
Gate oxide breakdown (GOBD)
System-level reliability
System-level modeling

A B S T R A C T
In the nanoscale regime, the aggressive scaling of devices is affected by several severe reliability issues, including negative bias temperature instability (NBTI) and gate oxide breakdown (GOBD). Generally, the mathematical models of NBTI and GOBD are derived from device level test structures with accelerated tests. However, although both models are highly dependent on temperature and the gate voltage and both mechanisms are based on the probability of trap generation in the oxide layer, each model has a different impact on circuit performances. In this paper, we use a physical probability model of trap generation for both mechanisms. We first simulate the impact on circuits using process models involving threshold voltage shifts and gate oxide leakage currents for NBTI and GOBD, respectively. Then, we find a relationship between the model parameters and power/ground signal degradation. We find the stress conditions that make each of the two mechanisms dominant in the power/ground signal. We calibrate the NBTI and GOBD model parameters of each chip to experimental results. Hence, it becomes possible to identify chips that are more or less vulnerable to NBTI and GOBD.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

With the downscaling of CMOS devices, the reliability of advanced VLSI circuits becomes a key issue. Gate oxide breakdown (GOBD) and negative bias temperature instability (NBTI) are primary aging mechanisms affecting MOSFETs in state-of-the-art electronic systems. The effect of NBTI and GOBD on MOSFETs is increasing threshold voltages ($V_{th}$) and oxide leakage currents, respectively. The impact is operating speed degradation of circuits, and, in extreme cases, system failure because of timing constraints [1,2].

Recently, the Random Telegraph Noise (RTN) model is used to explain the NBTI phenomenon, which is commonly known as the charge trapping and detrapping (T–D) model. This model is considered to be the most likely explanation of the NBTI mechanism [3,4]. GOBD is also explained by the T–D model. The GOBD phenomenon has three stages: development of Trap Assisted Tunneling (TAT) current, soft breakdown (SBD), and hard breakdown (HBD). In [5], the TAT gate leakage current is explained by the T–D model and the trap configuration. The leakage current steadily increases as a function of stress time with increasing numbers of traps. Eventually a conduction path is formed in the oxide layer. After the formation of conduction paths, stress-induced leakage current (SILC) increases because more and more conduction in paths are formed by traps, commonly known as SBD. Further degradation from SBD, when the transistor fails to operate, is called HBD [6–8].

Equipped with an understanding of two wearout models with the same underlying physical model, we can apply each model in circuit simulation, through the $V_{th}$ shift by the NBTI effect and a gate-to-source resistance ($R_{GS}$) or gate-to-drain resistance ($R_{GD}$) to create SILC [9] by GOBD. This approach is used to derive gate-level simulation models as a function of time, temperature, and usage. It enables system-level lifetime prediction for timing guardband optimization or circuit adaptation analysis [10,11]. Also, because both reliability mechanisms are highly dependent on the gate voltage and temperature, it is very important to differentiate both models during parameter extraction in order to derive the combined system-level mathematical model for both mechanisms, because both wearout mechanisms occur in circuits at the same time. Therefore, in this paper, the device-level models of both reliability mechanisms are modeled with the RTN model with an appropriate probability of trap generation in the oxide layer. We compare the amount of the current and threshold voltage degradation at several stress conditions to determine stress conditions which make each mechanism dominant.

We summarize, in Section 2, the RTN model for NBTI and GOBD. The T–D model is used for trap generation. It determines the circuit impact for NBTI. A model of TAT and the percolation model (PM), combined with the quantum point contact (QPC) model, determine the circuit impact of GOBD. We apply several stress conditions (bias voltage, temperature, and stress time) in simulation to find the conditions that make each of the two mechanisms dominant in supply and ground signal degradation to enable model parameter extraction from experimental data.

In Section 3, the system level chip simulation results are generated by FastSpice (XA) [12]. We extract the signature signal and compare it
with the nominal signature signal to extract the delay and amplitude variation. Also, for NBTI, we apply shifts in \( V_{th} \) to each MOSFET to determine the amplitude and delay degradation of the signature signal as a function of time. For GOBD, we apply the oxide breakdown resistances to randomly selected devices and analyze the impact on power/ground signatures.

In Section 4, we measure the power and ground signal, extract the degraded amplitude and delay shift from the signal under different stress conditions. The shift due to degradation as a function of stress time is used to extract the model parameters. Then, using the extracted model parameters, we estimate the lifetime of each chip based on statistical timing analysis and conclude with a summary in Section 5.

2. Device-level models for NBTI and GOBD

2.1. NBTI model derivation based on RTN

RTN is the cause of charge T–D of oxide defects, and has been observed in submicron MOSFETs, as illustrated in Fig. 1(a). Because the defects capture and emit charged carriers, the charged defects lead to changes in the mobility and modulate the local \( V_{th} \) of the device [13]. Based on the T–D model, the number of defects undergoing capture and emission follows the Poisson distribution with time constants for emission \( (\tau_e) \) and capture \( (\tau_c) \) as follows [14]:

\[
\tau_c = 10^p \cdot (1 + \exp(-q))
\]

\[
\tau_e = 10^p \cdot (1 + \exp(q))
\]

where \( p \in [p_{\text{min}}, p_{\text{max}}] \), and \( q = (E_t - E_F)/k_BT \), \( E_t \geq [E_v, E_c] \), \( p \) is the signal frequency range on the log scale, \( E_t \) is the Fermi Level, \( k_B \) is the Boltzmann constant (eVK\(^-1\)), and \( T \) is temperature (K). \( E_v \) and \( E_c \) are the energy levels of the valance and conduction bands, respectively.

Under gate bias and temperature stress with the Poisson distribution, the number of defects is a function of time. A detailed explanation is available in [4,15]. The threshold voltage is proportional to the number of defects [14], and this produces the following statistics for the threshold voltage:

\[
\mu(\Delta V_{th}(t)) = \varphi(T, E_F) \cdot (A + B \cdot \log(t)),
\]

\[
\sigma(\Delta V_{th}(t)) = \varphi(T, E_F) \cdot \sqrt{A + B \cdot \log(t)},
\]

where \( A \) and \( B \) are constants, \( t > 0 \) is the stress duration, and \( \varphi(T, E_F) \) is a function that depends on the trap energy density distribution in the band-gap. Both the stress and recovery phases are considered in Eq. (3), and the Fermi level and conduction and valence bands are affected by the applied gate voltages.

Digital circuit operation has stress and recovery periods. The fraction of time under stress is called the duty cycle, \( \alpha \). Instead of modeling the shift in threshold voltage for each stress and recovery period separately, an effective Fermi level as a function of duty cycle is determined [14]:

\[
E_{F,\text{eff}}(\alpha) = \alpha \cdot E_{F,\text{on}} + (1 - \alpha) \cdot E_{F,\text{off}}.
\]
Therefore, $\phi(T, E_{\text{eff}})$ adjusts the probability of charge capture and emission in the threshold voltage according to the duty cycle. Fig. 2 presents the temperature and duty cycle dependence of $\phi(T, E_{\text{eff}})$. Because phi values are increasing with the test temperature and duty cycle, Eqs. (3) and (4) can be correctly adjusted as a function of stress conditions and the test bench. The phi value depends on the supply voltage as well, not just the temperature and the test bench [15].

### 2.2. GOBD model derivation based on RTN

For the mathematical model of GOBD, we start with the TAT gate leakage current ($\Delta I_{\text{gate}}$) using the T–D model and trap generation. Also, because of the inaccuracy of the single-trap assisted model for highly-stressed dielectrics [16], the multi-trap assisted model of a cell transistor is used to calculate $\Delta I_{\text{gate}}$ as follows [17]:

$$\Delta I_{\text{gate}} = (w \cdot L_{\text{eff}}) \cdot q \sum_i (R_{c,i} + R_{e,i}) \cdot \Delta x_{T,i},$$  \hspace{1cm} (6)

where $w, L_{\text{eff}}$ and $\Delta x_{T}$ are the effective oxide width, length, and the trap size, respectively. $R_{c,i}$ and $R_{e,i}$ are the capture and emission probabilities of the ith defect, and $q$ is the electronic charge. The contribution for each trap is summed to compute the total current. $R_{c,i}$ and $R_{e,i}$ are defined as,

$$R_{c,i} = \sum_{j=0}^{i-1} \tau_{c,i}^{-1} N_{t,j} f_{t,i} (1-f_{t,j}),$$

$$R_{e,i} = \sum_{j=i+1}^{n} \tau_{e,i}^{-1} N_{t,j} f_{t,j} (1-f_{t,i}),$$  \hspace{1cm} (7)

where $f_{t,i} = \tau_{e,i}^{-1}/(\tau_{c,i}^{-1} + \tau_{e,i}^{-1})$ is the occupancy probability of a trap, $\tau_{c,i}$ and $\tau_{e,i}$ are time constants of the capture and emission in accordance with Eqs. (1) and (2), respectively, $N_{t,j}$ is the trap density, and $n$ is the number of traps. Using the time constants, we compute the number of traps, which in turn determines the TAT gate leakage current.

As the number of traps increases with stress time, some paths are formed through the SiO$_2$, between the gate and substrate, as illustrated in Fig. 1(b). Fig. 1(b) illustrates the percolation model (PM). In the percolation model, defects are randomly placed in a 3D grid. We perform several trials, where the defects are placed randomly within the SiO$_2$ and count the number of paths that are formed for different time points. Since the defect placement is random, we determine the probability of different numbers of paths at each time point, as illustrated in Fig. 3. Hence, the number of paths is associated with a probability at each time point.

Fig. 5. (a) Inverter with the GOBD/NBTI effect. (b) The ground signature signal degradation at different time points due to NBTI. (c) The power and ground signal degradation at different time points due to GOBD.
From the number of breakdown paths, we calculate the PM-based gate leakage current as a function of time [16], which is modeled as a resistance ($R$) with the quantum point contact (QPC) model [18]:

$$R \approx V_G \left[ \frac{4e}{h\alpha} N \cdot \exp(-\alpha \cdot \Phi) \sinh\left(\frac{\alpha \cdot e(V_G - V_0)}{2}\right) \right]$$  \hspace{1cm} (8)

where $\Phi = 3 - 4$ eV, $V_0 = 0 - 0.5$ V, $\alpha = 2 - 3$ eV$^{-1}$, $h$ is Planck’s constant, $e$ is the electron charge, $V_G$ is the gate voltage, and $N$ is the number of SBD conduction paths [18].

HBD occurs when the device does not operate properly. HBD is defined when a trap density in the oxide layer reaches a critical density associated with catastrophic failure [19,20]. The voltage-dependent power-law gate oxide degradation model is used to find the HBD resistance, $R_{HBD}$, for use in simulation. HBD is also associated with a location of HBD, illustrated in Fig. 1(b), because $R_{HBD}$ is divided between the gate-to-source resistance $R_{BD \cdot G2S}$ and the gate-to-drain resistance $R_{BD \cdot G2D}$ [21]:

$$R_{BD \cdot G2S} = \frac{R_{HBD}}{x_{HBD}}, \quad R_{BD \cdot G2D} = \frac{R_{HBD}}{1-x_{HBD}}.$$  \hspace{1cm} (9)

The PM model determines the probabilities of different numbers of conduction paths as a function of stress for each device for each test bench and stress condition. The defective devices are randomly selected according to the probabilities of having different numbers of breakdown paths. Then, the appropriate breakdown resistance, $R$, is inserted in the circuit netlist for simulation of the circuit. A detailed explanation is in [22].

2.3. Adjusting stress conditions for NBTI and GOBD distinction

We apply the device-level wearout models to several circuits in simulation. However, because the NBTI and GOBD phenomena occur simultaneously, they should be differentiated for the separate extraction of parameters. We choose different stress conditions (gate voltage and temperature) for each wearout mechanism. The characteristics of the static current degradation can be used to find stress conditions that make each mechanism dominant. Fig. 4 shows the shift in the supply voltage bounce (because of static current degradation of an inverter) due to NBTI and GOBD for several stress conditions. We can distinguish NBTI and GOBD, since NBTI is dominant at low gate voltages and GOBD is dominant at high gate voltages.

3. System-level model for NBTI and GOBD

For system-level modeling through circuit simulation, several chips (a RISC microprocessor, a floating point unit (FPU), and a finite impulse response (FIR) filter) and test benches have been considered as case studies. Detailed information for each circuit is in [15].
3.1. The combined wearout model

For the analysis of the system-level NBTI and GOBD impact, our method determines the shift in delay and amplitude of degrading power supply and ground voltage bounce signatures as a function of stress time. Ground and power supply bounce signature signals are generated with an initial random set of process parameters, to mimic a true process, and at different time points, to mimic the impact of stress.

For the implementation of the degradation effect for circuit simulation, we apply Eqs. (3) and (4) for NBTI and Eq. (8) for GOBD to each process, and at different time points, to mimic the impact of stress.

In order to validate our system-level models for NBTI and GOBD, we need to check if our methodology can correctly extract device-level NBTI model parameters. For the system-level NBTI model, we check whether our methodology can correctly extract device-level NBTI model parameters in Eq. (3) using signature data. To do this, we assume values are given for coefficients A and B in Eq. (3). The signatures are computed, and Eq. (10) extracts the resulting \( \Delta V_{\text{th}} \) from the signatures. The trend in \( \Delta V_{\text{th}} \) is modeled to extract A and B, which are then compared with the original values for A and B. More details are provided in [15]. Table 1 shows the average error rate for coefficients A and B, which are extracted by several simulation runs for different circuits and test benches using full chip simulations. The overall average errors for A and B (<6%) are very small, indicating that parameter values can be computed accurately. For GOBD model validation, we also need to use the mathematical fitting approach to extract the device-level model parameters. In order to explain the relationship between trap phenomenon and stress time, many researchers have suggested a power function \( (\alpha \cdot t^n) \) to fit the number of traps based on Eq. (7). Fig. 8(a) shows the power function fitting result and the extracted coefficients.

![Table 1](image)

<table>
<thead>
<tr>
<th>Chip</th>
<th>Microprocessor FPU</th>
<th>PPU</th>
<th>FIR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test bench</td>
<td>Add</td>
<td>Cache</td>
<td>Add</td>
</tr>
<tr>
<td>( \hat{i}(A) ) (%)</td>
<td>5.48</td>
<td>4.91</td>
<td>5.12</td>
</tr>
<tr>
<td>( \hat{i}(B) ) (%)</td>
<td>3.91</td>
<td>3.37</td>
<td>3.68</td>
</tr>
</tbody>
</table>

In order to make a unified system-level NBTI/GOBD model, we used the functions below for regression analysis.

\[
\mu(\Delta V_{\text{th}}) = \xi \cdot \left( \alpha_1 \cdot \Delta t^n \right) + (1 - \xi) \cdot \left( \alpha_2 \cdot \Delta D^{\beta_1} \right) + \varepsilon_1, \tag{10}
\]

\[
\min(R_{\text{GOBD}}) = \xi \cdot \left( \alpha_1' \cdot \Delta t^{\beta_1} \right) + (1 - \xi) \cdot \left( \alpha_2' \cdot \Delta D^{\beta_2} \right) + \varepsilon_2, \tag{11}
\]

where \( \alpha_{1,2}, \beta_{1,2}, \alpha'_{1,2}, \beta'_{1,2}, \xi, \) and \( \varepsilon_{1,2} \) are fitting constants. The fitting constants are also a function of temperature and voltage.

An experiment was conducted to determine if the bounce in the signatures could be used to extract \( \Delta V_{\text{th}} \) and \( R_{\text{GOBD}} \). To do this, circuits were simulated with “true” values for \( \Delta V_{\text{th}} \) and \( R_{\text{GOBD}} \), together with random variation in process parameters. The shifts in the peaks in the signature signals were used to calculate \( \Delta V_{\text{th}} \) and \( R_{\text{GOBD}} \). The results were then compared with the “true” values. The results are shown in Fig. 7.

3.2. Validation for each system-level model

![Table 2](image)

<table>
<thead>
<tr>
<th>Wearout model</th>
<th>Temperature (°C)</th>
<th>Supply voltage (V)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BTI</td>
<td>40, 60, 80, 100, 120</td>
<td>1.4, 1.6</td>
</tr>
<tr>
<td>GOBD</td>
<td>50, 100</td>
<td>2.2, 2.7</td>
</tr>
</tbody>
</table>

In order to validate our system-level models for NBTI and GOBD, we need to check if our methodology can correctly find device-level model parameters. For the system-level NBTI model, we check whether our methodology can correctly extract device-level NBTI model parameters in Eq. (3) using signature data. To do this, we assume values are given for coefficients A and B in Eq. (3). The signatures are computed, and Eq. (10) extracts the resulting \( \Delta V_{\text{th}} \) from the signatures. The trend in \( \Delta V_{\text{th}} \) is modeled to extract A and B, which are then compared with the original values for A and B. More details are provided in [15]. Table 1 shows the average error rate for coefficients A and B, which are extracted by several simulation runs for different circuits and test benches using full chip simulations. The overall average errors for A and B (<6%) are very small, indicating that parameter values can be computed accurately. For GOBD model validation, we also need to use the mathematical fitting approach to extract the device-level model parameters. In order to explain the relationship between trap phenomenon and stress time, many researchers have suggested a power function \( (\alpha \cdot t^n) \) to fit the number of traps based on Eq. (7). Fig. 8(a) shows the power function fitting result and the extracted coefficients.

![Fig. 8: GOBD model validation. (a) Trap density vs. stress time power function \( (\alpha \cdot t^n) \) fitting result. (b) R-square value for each test bench and circuit design.](image)

![Fig. 9: Measurement of the ground signature signal degradation with the GOBD dominant stress condition (2.2 V/100 °C).](image)

Please cite this article as: S. Cha, et al., The die-to-die calibrated combined model of negative bias temperature instability and gate oxide breakdown from device to system, Microelectronics Reliability (2015), http://dx.doi.org/10.1016/j.microrel.2015.06.105
Because the number of traps generated due to the GOBD the minimum GOBD resistance has a relationship [22], we can calculate the number of traps in the oxide layer by using the extracted minimum $R_{\text{GOBD}}$ result in Fig. 7(b). Fig. 8(b) presents the model accuracy using the $R^2$ value, the coefficient of determination.

4. Measurement result

We used several microprocessor chips to apply our methodology. The MOSFET gate lengths range from 90 nm to 500 nm, and a wide gate width is used in the designs, up to 5 $\mu$m. The nominal $V_{\text{th}}$ was around 0.4 V at 25 °C. To measure the degradation, we have used a digital sampling oscilloscope (DSO) with “average” and “bandpass filter” functions to remove noise and jitter. The time point resolution was 0.2 ps and the amplitude resolution was 360 $\mu$V, which is enough to detect the delay and amplitude degradation.

4.1. Computation of model parameters

The degraded ground and power voltage signature signals were recorded with the DSO for nine microprocessor chips that were stressed for several conditions, as shown in Table 2. Under the stress conditions, we captured several ground and supply voltage bounce signals at different stress times. Fig. 9 shows an example ground bounce signal which is degraded as a function of stress time under the GOBD dominant stress condition.

Then, we extract the voltage and delay shift from the degraded signals with error bars from three of the microprocessors in Fig. 10. The

Fig. 10. Degradation of amplitude and delay by GOBD, extracted from the ground signature signal tested with 2.2 V and 50 °C/100 °C.

Fig. 11. (a) Threshold voltage degradation due to different supply voltages (1.4 V and 1.8 V) for the microprocessors, derived by Eq. (10). (b) Extracted degraded oxide resistances for PMOS and NMOS devices in a microprocessor for several stress conditions (50 °C/100 °C @ 2.2 V).

Please cite this article as: S. Cha, et al., The die-to-die calibrated combined model of negative bias temperature instability and gate oxide breakdown from device to system, Microelectronics Reliability (2015), http://dx.doi.org/10.1016/j.microrel.2015.06.105
extracted data shows that the amplitude and delay degradation increases as a function of stress time. We show that the harsher stress conditions make GOBD dominant not only in Figs. 4 and 6, but also in Fig. 10, because the amplitude shift increases as a function of stress time. Also, it is evident that the shift is larger for higher temperatures. The ground signature is affected by NMOS degradation, while PMOS degradation causes shifts in the power supply signature. Therefore, we use both the ground and power supply signature signals to extract both the NMOS and PMOS oxide degradation resistances.

Results for extraction of NBTI and GOBD parameters are shown in Fig. 11. For GOBD, we calculated the degraded oxide resistance as a function of stress time in Fig. 11(b). The solid lines indicate the average. The results show that NMOS degradation is less than PMOS degradation, and higher temperatures cause MOSFETs to degrade more quickly. Our results match previous experimental results, which compare degradation in NMOS and PMOS devices [23], and experimentally show the temperature dependence of GOBD [24]. Moreover, variation increases as stress time increases.

Using the result in Fig. 11, we calibrate the model parameter coefficients for each chip. In order to further study the joint impact offrontend wearout on the microprocessor, the critical paths have been extracted by the statistical static timing analysis method in [25]. Because the models we have developed depend on the test conditions, NBTI and GOBD are scaled to use conditions by adjusting $v(T,E_f)$ and the coefficient of the power function ($\alpha, n$), respectively.

Because the lifetime is the time until the system suffers from timing violations and frontend reliability impacts these timing margins, lifetime is a function of frequency, since higher operating frequencies have smaller timing margins. Fig. 12 shows the estimated lifetime of each chip as a function of operating frequency.

Even though we have developed the system-level model and have obtained silicon measurement results from a SiO$_2$-based CMOS technology, the device-level reliability model is from the T–D model. The trap configuration probability in the T–D model mainly depends on the energy level and temperature (Eqs. (1) and (2)). The model coefficients depend on the manufacturing process. Because the T–D phenomenon occurs also in high-k dielectrics and FinFETs [26], we can apply our system-level models to state-of-the-art manufacturing processes as well.

**Acknowledgment**

The authors thank the Defense Advanced Research Projects Agency under grant HR0011–11–1–0011 for financial support. The authors would also like to thank the National Science Foundation under Award #116786 for financial support.

**References**


