AVERT: An elaborate model for simulating variable retention time in DRAMs

Dae-Hyun Kim *, Soonyoung Cha, Linda S. Milor

School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

1. Introduction

As the main memory in computer systems, dynamic random access memory (DRAM), consisting of one transistor and one capacitor, is widely used. A DRAM cell stores data on a capacitor in the form of charge. Because of leakage from a storage node, a DRAM cell loses data without periodic refreshes. The memory controller operates refresh periodically based on the length of time that a DRAM cell can retain data, referred to as the “retention time.” To meet the standard refresh rate, manufacturers require accurate characterization of the DRAM retention time. However, finding a precise profile of the retention time becomes a challenge because of random fluctuations in retention time, also known as variable retention time (VRT). Since DRAM retention time randomly varies for each test, it is hard for vendors to screen all retention errors due to VRT during testing. Therefore, to enhance DRAM reliability under VRT, predicting its influence with device-level models is essential.

This paper proposes AVERT, an elaborate device model for simulating VRT in DRAM circuits. To understand the configuration of traps, we adopt the charge trapping and detrapping (TD) model. Based on trap configurations, we generate random telegraph noise (RTN) in trap-assisted gate-induced drain leakage (TA-GIDL) and trap-assisted gate leakage (TA-GL), known as the origin of VRT [1][3]. Finally, we simulate DRAM arrays with VRT by introducing gate-level models in the netlist and generating control voltage inputs converted from RTN leakage for circuit simulation using a Perl script. Employing experimental data and our model, we propose a method for determining the appropriate test time per DRAM cell and for reducing the total test time of DRAMs considering the VRT phenomenon. Our paper makes the following contributions: (1) Unlike prior work [4,5], we model the RTN in leakage currents as the origin of VRT, not on-current RTN; (2) we deploy RTN in TA-GL to achieve a more accurate VRT model; (3) we exploit the TD model to obtain the trap density and calculate the influence of the trap density on RTN leakage currents; and (4) we propose a strategy for testing VRT based on our model.

2. Background

2.1. Variable retention time in DRAMs

DRAM retention time is inversely proportional to total leakage current [7]. A DRAM cell has various leakage sources: leakage from a storage node to a plate poly, a gate, a source (sub-threshold current), a junction, another transistor (isolation leakage), and a body (drain leakage or GIDL). Since GIDL current empirically shows random telegraph noise among various current sources, trap-assisted GIDL is known as the origin of variable retention time [3,7]. In addition to TA-GIDL, gate leakage also experimentally exhibits RTN, also explained using traps, so that trap-assisted gate leakage can be another cause of VRT [2,8]. A trap can be occupied randomly and increases leakage current. As a result, a cell leaks faster and exhibits lower retention time. However, when the trap becomes empty again, leakage current reduces, resulting in a higher retention time. Such random variations in leakage current resulting from different trap conditions cause VRT in DRAMs.
2.2. Trap-assisted gate-induced drain leakage

Fig. 1(a) shows the device structure of a cell transistor and the mechanism of GIDL current. In the case of DRAM standby or precharge mode, with data ‘1’ written on a cell, a drain (a storage node) is charged up to a high voltage level by a bitline (VDD), and a gate (a wordline) has negative bias for reduction of the sub-threshold voltage. Since the bias between the drain and the gate is high enough to deplete the n+ drain region under the gate and cause high-field effects, such as avalanche multiplication and band-to-band tunneling, increased electrical field and band bending generate electron–hole pairs in the depletion region. While electrons that flow to the capacitor increase the GIDL current, holes that drift to the substrate contribute to an increase of substrate current [6]. Fig. 1(b) shows the trap-assisted tunneling mechanisms [7]: (1) an electron moves from the valence band to a trap by thermal emission, and tunnels from the trap to the conduction band; (2) an electron tunnels from the valence band to a trap and then tunnels from the trap to the conduction band like a stepping stone; and (3) an electron tunnels from the valence band to a trap and thermal emission helps the electron move up to the conduction band from the trap.

2.3. Trap-assisted gate leakage

The random fluctuation in DRAM retention time results from not only trap-assisted GIDL but also trap-assisted gate leakage current [8]. During DRAM standby mode, as explained in Section 2.2, Fig. 2(a) and (b) depicts the trap-assisted tunneling and two mechanisms of gate leakage: (1) direct tunneling and (2) trap-assisted tunneling. Although the energy level of the SiO2 is higher than that of the gate or the drain, the tunneling effect enables gate leakage to occur. Trap-assisted tunneling facilitates gate leakage [9,10]. Because the occupancy of traps randomly varies, gate leakage current resulting from trap-assisted tunneling also randomly fluctuates. To explain multi-state gate leakage fluctuation, we exploit multi-trap-assisted tunneling from recent research rather than conventional single-trap considerations [11].

3. RTN model for VRT

3.1. Leveraging the trapping and detrapping model

As shown in Fig. 3, RTN is explained with the TD model of silicon dioxide (SiO2) defects, which randomly capture and emit charge in submicron FETs [12]. Since captured and emitted charge changes the energy of a trap, the variation in the leakage current of a transistor, especially TA-GIDL and TA-GL, depends on the number of captured defects. Based on the TD model, the number of defects undergoing

![Fig. 1. (a) GIDL during precharge mode in DRAMs ([6]) and (b) mechanism of trap-assisted GIDL (source [7]).](image)

![Fig. 2. (a) Trap-assisted gate leakage current during precharge mode in DRAMs and (b) mechanism of gate leakage.](image)

![Fig. 3. The charge trapping and detrapping model.](image)

![Fig. 4. Flowchart of AVERT.](image)
capture and emission follows a Poisson distribution with a wide range of time constants for emission ($\tau_e$) and capture ($\tau_c$) [13]. Such constants are random variables that depend on temperature, bias, and trap location [1,12]:

$$\tau_c = 10^9 \cdot (1 + \exp(-q)),$$

(1)

$$\tau_e = 10^9 \cdot (1 + \exp(q)),$$

(2)

where $p \in [p_{\text{min}}, p_{\text{max}}]$ and $q = (E_t - E_p)/k_B T$ [3], $p$ is the signal frequency range on the log scale, $E_p$ is the Fermi level, $k_B$ is the Boltzmann constant ($eV^{-1}$), $T$ is temperature ($K$), and $E_t$ is the energy drop at the trap in SiO$_2$, which is a function of [1]:

$$E_t = \frac{T_{\text{ox}} - \Delta x}{\Delta x} q V_{\text{ox}}.$$  

(3)

$$V_{\text{ox}} = -V_{\text{DC}} - V_{\text{FB}} - \frac{q N_D T_{\text{ox}}^2 e}{\varepsilon_{\text{ox}}^2} + \frac{\sqrt{-2(V_{\text{DC}} - V_{\text{FB}})^2 q N_D T_{\text{ox}}^2 e}}{\varepsilon_{\text{ox}}^2} - (V_{\text{DC}} - V_{\text{FB}})^2.$$  

(5)

Algorithm: Circuit Simulations with non-stationary RTN in AVERT

Input: Circuit simulation end time, $t_{\text{max}}$ (length*time), width (w), thickness (t), Mesh resolution ($N_{\text{mesh}}$, $N_{\text{mesh}}$, $N_{\text{mesh}}$), total number of cell transistors ($N_{\text{tr}}$), simulation time step ($t_s$)

Output: Time-variant VRT signals for all cell transistors

//Initialize trap configuration in a transistor
//Define 3D device structure $N_{\text{tr}}$ (number of mesh)
//Build the probability table of capture and emission for each mesh site

while ($i < N_{\text{tr}}$) do

if (uniRand(seed) $<$ $N_{\text{traps}}$) then

$t_i$, $t_{\text{tr}}(x,y,z)$ = $\frac{i}{N_{\text{traps}}}$, $t_{\text{te}}(x,y,z)$ = $\text{exprand}_{\text{te}}(y,T,V_{\text{gate}})$ //capture time update

else

$t_i$, $t_{\text{tr}}(x,y,z)$ = $\frac{i}{N_{\text{traps}}}$, $t_{\text{te}}(x,y,z)$ = $\text{exprand}_{\text{te}}(y,T,V_{\text{gate}})$ //emission time update

end

$x_i$, $y_i$, $z_i$++;

end

//Generate non-stationary leakage of a transistor up to $t_{\text{max}}$

while ($i < t_{\text{max}}$) do

if ($i < N_{\text{traps}}$) then

$t_i$, $t_{\text{tr}}(x,y,z)$ = $\text{exprand}_{\text{tr}}(y,T,V_{\text{gate}})$, $t_{\text{te}}(x,y,z)$ = $\text{exprand}_{\text{te}}(y,T,V_{\text{gate}})$;

else

$t_i$, $t_{\text{tr}}(x,y,z)$ = $\text{exprand}_{\text{tr}}(y,T,V_{\text{gate}})$, $t_{\text{te}}(x,y,z)$ = $\text{exprand}_{\text{te}}(y,T,V_{\text{gate}})$;

end

$x_i$, $y_i$, $z_i$++;

end

//Calculate leakage currents resulting from all traps at a given time step

$\Delta J_{\text{GIDL}}(i) = \frac{A \cdot \exp(-B_{\text{eff}}/F_I)}{\exp(-B_{\text{eff}}/F_I)} (6)$

$B_{\text{eff}} = \frac{4}{h} \cdot \left(2m_t \right)^{1/2} \cdot \left(\frac{E_j - E_D}{3q} \right). (7)$

$\Delta J_{\text{GIDL}} = \frac{A \cdot \Delta x}{\sum_i \exp(-B_{\text{eff}}/F_I) \cdot \Delta x}, \text{ for } (8)$

where $A$ depends on the interface trap density, $F_I$ is the total field in the deep depletion region, $h$ is Planck’s constant, $m_t$ is the effective mass of an electron, $E_j$ is the energy of the conduction band, $\Delta x$ is the trap location, $E_D$ is the energy drop at a trap in SiO$_2$, $q$ is the electronic charge, and $\Delta x$ is the effective action range of electrical field. As for gate leakage, instead of the single-trap model that is less accurate in the high-stressed dielectrics [11], we utilize the multi-trap model for calculating the
variation in trap-assisted gate leakage current ($\Delta I_{\text{gate}}$) of a cell transistor. We calculate $\Delta I_{\text{gate}}$ as follows [15]:

$$\Delta I_{\text{gate}} = \frac{w}{L_{\text{eff}}/C_0/C_1 \frac{\text{effective area}}{C_1} q X_i R_c},$$

where $w$ and $L_{\text{eff}}$ are the width and the length of the effective oxide area, $q$ is the electronic charge, $R_c$ and $R_e$ are the capture and emission rate in a trap, and $\Delta x_T$ is the trap location. We calculate $R_c$ and $R_e$ with:

$$R_c = \sum_{i=1}^{N_k} \left( \frac{\tau_c}{\sum_{j=1}^{N_k} \tau_c \cdot f_i \left( 1 - f_i \right) \cdot \sum_{j=1}^{N_k} \frac{\tau_c}{\sum_{j=1}^{N_k} \tau_c \cdot f_i \left( 1 - f_i \right)} \right),$$

$$R_e = \sum_{i=1}^{N_k} \left( \frac{\tau_e}{\sum_{j=1}^{N_k} \tau_c \cdot f_i \left( 1 - f_i \right) \cdot \sum_{j=1}^{N_k} \frac{\tau_e}{\sum_{j=1}^{N_k} \tau_c \cdot f_i \left( 1 - f_i \right)} \right),$$

where $N_k$ and $f_i = \frac{\tau_c}{\tau_c + \tau_e}$ are the density and the occupancy probability of a trap, respectively.

4. The algorithm and the circuit simulation methodology of AVERT

4.1. The device model in AVERT

A flowchart in Fig. 4 shows the algorithm of AVERT. The pseudo-code shown in Fig. 5 explains the device model to generate VRT signals in DRAMs. For generating RTN in leakage current, AVERT defines trap configurations with a 3D-structure and then initializes traps by determining if traps are filled or not, based on the times of capture/emission following exponential random distributions. After initialization, AVERT generates a time-variant RTN signal for a transistor until the target time ($t_{\text{target}}$) with a resolution of the time step ($t_s$). For each trap, AVERT checks if it is filled or not first. If it is filled, AVERT statistically determines whether a trap will be evicted or not based on the trap emission probability, which follows a Poisson distribution with a shape parameter of $\tau_e$. After eviction, the capture time is updated based on the exponential random distribution. On the other hand, if a trap is not filled, AVERT determines whether a trap will be filled or not based on the trap capture probability of a Poisson distribution with $\tau_c$. After updating the trap configuration, AVERT calculates the current variation of both gate leakage and TA-GIDL. In the end, to get RTN in the leakage current of each DRAM cell in an array, AVERT repeats the generation of the time-varying VRT signal for each cell with a different seed, which changes the random distribution for the trap configuration in every iteration.

4.2. Circuit simulation methodology with VRT

To deploy time-varying RTN in the leakage current from the VRT device model in circuits, we propose a gate model. Since a time-varying resistor, as depicted in Fig. 6(a), is not applicable to circuit simulation, we exploit a voltage-controlled resistor shown in Fig. 6(b). We simply define a new voltage source (e.g., $V_{\text{gate,off}}$) and change the voltage level using a piecewise linear signal that allows the resistor to have a time-
varying resistance based on the relationship between the voltage and the resistance defined in the G-element. Using a resistor ($R_G$), RTN in gate leakage is modeled with the time-varying resistance with a control voltage of $V_{gate,offset}$. Similarly, RTN in GIDL is modeled using $R_{GIDL,offset}$, a resistance determined by a user-defined voltage $V_{GIDL,offset}$. Using a Perl script, we modify the netlist so that every cell transistor has a user-defined voltage source and voltage-controlled resistors. Using the script, we also generate input stimuli for circuit simulations.

5. Results

Fig. 7 depicts changes in captured trap counts and leakage current fluctuations corresponding to the number of filled traps over time. Despite continuous leakage current levels, we quantize the possible leakage current levels of a transistor for comparing results with those from experiments in prior work. To optimize the simulator in terms of accuracy and performance, we conduct sensitivity simulations with various numbers of possible leakage states as shown in Fig. 8. With fewer leakage levels ($N$), we have more quantization errors and a faster execution time. Based on the error-runtime product with the $N > 10^3$ case as a baseline, we found $N = 8$ to be optimum and use it for the rest of the simulations.

To validate our model, we compare our simulation results with prior empirical results, such as measurements of the bias/temperature dependence, results from the Poisson distribution for high and low retention states, and data on the power spectral density. As shown in Fig. 9, the bias/temperature dependence on the leakage current generated by AVERT shows more frequent transitions with higher bias and higher temperature, which corresponds well to empirical observations in [1]. The distribution of retention states ($\tau_{\text{high}}$ and $\tau_{\text{low}}$) follows a Poisson distribution, shown in Fig. 10(a) and (b), and the power spectral density

Fig. 9. Simulated dependence of (a) temperature with 1.2 V $V_{DG}$ and (b) bias ($V_{DG}$) on leakage current at 85 °C.

Fig. 10. Distribution of (a) $\tau_{\text{high}}$ and (b) $\tau_{\text{low}}$ using AVERT and (c) power spectral density of fluctuation.
follows a Lorentzian-type spectrum, shown in Fig. 10(c). Such results correspond to results from prior experiments [3]. Moreover, such properties match those of the channel RTN of MOSFETs and the GIDL RTN of non-stressed devices [3]. Therefore, AVERT successfully models VRT in leakage currents.

We use SPICE with IBM 90 nm PDK and the BSIM model to obtain the circuit simulation result for DRAM retention time shown in Fig. 11(a). The upper part of the graph depicts the waveform of a piecewise linear signal to control the resistance of the voltage-controlled resistor for gate leakage and GIDL in a cell transistor. The graph at the bottom shows that charge leaks over time after the storage node is written as leakage and GIDL in a cell transistor. The graph at the bottom shows that signal to control the resistance of the voltage-controlled resistor for gate leakage cause VRT of a DRAM cell, an accurate profile of retention time requires prohibitive efforts for DRAM manufacturers to screen VRT bits during testing. To predict the impact of VRT on devices and support circuit simulations that are aware of VRT, we have proposed AVERT, an elaborate device model of RTN in leakage current, and a circuit simulation methodology for variable retention time in DRAMs. AVERT adopts the TD model for better temporal and spatial configurations of

Fig. 11(b) shows the results of a case study of retention time with VRT. Using AVERT, we obtain the degree of variability in leakage current of DRAM arrays under VRT. Assuming one part per million (ppm) retention errors, a lognormal distribution of retention time with a mean of 10 seconds, and no variability in cell capacitance whose nominal value is 20 fF, we add the variability of VRT to the original retention distribution to obtain the cumulative probability of retention time with VRT. As a result, the number of retention errors increases from one ppm to 20 ppm because of VRT.

6. A strategy for testing VRT

Based on the analysis from AVERT, we propose to optimize the total test time of a DRAM in the presence of VRT through (1) the optimization of test repetitions and (2) the reduction of the number of bits to be tested. Fig. 12 shows the test coverage with different test repetitions, in the presence of VRT, using the same assumptions as in Fig. 11(b). As the number of tests increases, which also increases the test time with a given time step for testing, the number of retention errors due to VRT in DRAM arrays decreases and saturates, as shown in Fig. 12(b). Such a trend is useful for optimizing the test time per cell and determines the test coverage. With the optimal test time per cell, we can obtain the degree of variability in retention times of DRAM arrays in the presence of VRT using AVERT, which indicates how frequently a VRT bit can transition from one state to another. By exploiting the degree of the variation, we can reduce the total number of cells that must be re-tested because every VRT cell does not cause a retention error. Only transitions from a higher retention time to a lower retention time below the standard refresh rate (64 ms) such as type (d), depicted in Fig. 13(a), result in retention errors [17]. Therefore, we can confine the re-testing of retention time in the presence of VRT to only portions of the DRAM, illustrated in Fig. 13(b). The decreased number of bits for testing VRT eventually reduces the overall test time. Assuming that maximum variation of retention times due to VRT is two seconds, the test interval is 64 ms, and the other assumptions in Fig. 11(b), the cumulative error rate becomes 6.78%. Therefore, we require only 6.78% of the total conventional test time with our proposed testing methods and the reduced number of cells.

7. Conclusion

Since random fluctuations in trap-assisted GIDL and trap-assisted gate leakage cause VRT of a DRAM cell, an accurate profile of retention time requires prohibitive efforts for DRAM manufacturers to screen VRT bits during testing. To predict the impact of VRT on devices and support circuit simulations that are aware of VRT, we have proposed AVERT, an elaborate device model of RTN in leakage current, and a circuit simulation methodology for variable retention time in DRAMs. AVERT adopts the TD model for better temporal and spatial configurations of
traps and calculates leakage current variations of TA-GIDL and TA-GL considering the influence of multiple traps. Our results have shown that AVERT can generate stochastic RTN leakage current signals, fed to circuit simulations with ease after modification of a netlist so that every transistor has voltage-controlled resistors, use of the gate-level model proposed with VRT, and the conversion of RTN current leakage into a piecewise linear voltage signal as user-defined voltage sources for the resistors. Simulations based on our RTN model have shown reasonable results that match well with prior empirical studies. In addition, results from circuit simulation have demonstrated that AVERT enables circuit simulations of DRAM arrays under VRT for characterizing retention times. AVERT is useful for optimizing the test time per cell and predicting the variability in retention times. Based on the degree of variability, AVERT can contribute to reducing the overall or average test time per cell and the maximum retention time transition after model calibration.

Fig. 13. (a) Different types of VRT cells: only type (d) causes a retention error due to the VRT phenomenon and (b) definition of the portion of DRAM cells that are of interest based on the maximum retention time transition after model calibration.

Prior publications [18] have investigated the trapping/detrapping phenomenon of charge in high-k dielectrics, and the TD model appears to be valid for high-k dielectrics. Moreover, because prior experimental studies [19] have shown that GIDL of high-k dielectric MOSFETs also exhibits RTN, the method for RTN current calculations based on the trap configuration is also viable with different energy levels for high-k dielectrics. Therefore, although our model was developed based on SiO2, AVERT is also applicable to high-k materials.

Acknowledgments

The authors would like to thank the National Science Foundation under Award #116786 for financial support.

References