This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Modeling of Retention Time for High-Speed Embedded Dynamic Random Access Memories

Swaroop Ghosh, *Senior Member, IEEE*

***Abstract*—Embedded dynamic random access memory (eDRAM) is becoming a popular choice for large cache applications due to its density, speed, and power benefits. One of the crucial challenges in eDRAM design is meeting the retention time specification. Due to implementation in logic process, usually eDRAM suffers from poor retention time compared to commodity DRAM. The retention time of eDRAM designed in scaled technologies not only depends on bitcell leakage but also on effects such as reference voltage variations, frequency-dependent writeback voltage, and various pattern-dependent coupling noise. Under the strict frequency and power budgets, these second-order mechanisms start playing a major role in determining the array retention time. Designing eDRAM array for certain retention time requires detailed modeling and understanding of the noise sources and variations. This paper investigates these components and provides a model of eDRAM retention time. Our results in 22 nm predictive technology shows that retention time can be impacted by as much as 10–16× if the noise and variations are not contained in the design.**

***Index Terms*—Embedded dynamic random access memories (eDRAM), low-power memory, retention modeling.**

I. Introduction

EMBEDDED dynamic random access memories (eDRAM) [1] are promising candidates for last level caches or replacing the off-chip DRAM to meet the high bandwidth requirement of graphics processors. eDRAM is attractive because it can be integrated in the logic process, eliminating the need for commodity DRAM and achieving better throughput as well as lower input/output (I/O) power [2]. Due to its throughput and power benefits eDRAM has found wide applications (e.g., playstation-2 [3], Power7 [4], [5]). eDRAM and commodity DRAM are conceptually same but totally different from implementation standpoint and both technologies face their own unique challenges. Few differences are as follows [6]: a) commodity DRAM contains 2–3 metal layers whereas eDRAM may contain 8–9 metal layers due to integration with logic process. Availability of upper layers allow DRAM to achieve superior cell capacitance (>25 fF) whereas eDRAM suffers from poor capacitances; b) access transistor of commodity DRAM can be made extremely low leakage by employing recessed long channel devices whereas eDRAM suffers from leaky access transistor due to its performance constraints; and c) some of the noise sources can be contained in DRAM (e.g., WL-WL coupling by embedding the poly in the substrate [7]) whereas this flexibility may not be present in eDRAM.

Retention time (i.e., the time interval between which the bitcell must be read and restored in order to preserve its value under bitcell leakage) plays a key role in determining the idle power dissipation of large eDRAM caches. Commodity 1T1C DRAM achieves very high retention (~ms) due to low leakage process, high cell cap, and special recessed channel access transistor, as discussed before. eDRAM suffers from poor retention not only due to first-order effects like low cell cap and high transistor leakage but also due to second-order effects, e.g., reference voltage variations, frequency-dependent writeback voltage, and various pattern-dependent coupling noise.

In DRAM or eDRAM, most of the design issues eventually boil down to sense margin and retention time. Therefore, modeling and understanding of factors affecting the retention time is crucial for designers. A perspective on the major factors for low retention would help them adjust the key circuit design parameters ahead of time and maintain high yield. Retention time behavior for commodity DRAM has been studied in detail in the past [8]–[14]. Traditionally, bitcell leakage has been pointed out as the critical factor. In [8], the authors have investigated subthreshold leakage, junction leakage, gate-induced drain leakage (GIDL), capacitor leakage, dielectric leakage, etc., and have concluded that the retention time of tail bits are determined by junction leakage. A test element group based method of determining retention time has been developed in [9] by considering various leakage currents. In [10], generation current via Shockley–Read–Hall process and GIDL has been modeled to study the retention time behavior. Retention time variation due to thermal stress has been studied in [11], where the GIDL is found to be the key component behind retention time variation. A detailed treatment of various factors influencing the DRAM yield has been proposed in [12] where the authors have considered the impact of bitcell leakage as well as few noise components, e.g., bitline-bitline (BL-BL) and wordline-bitline (WL-BL) and offset voltage of senseamps on sense margin. However, the eDRAM-specific coupling noises and high-frequency effects have not been studied. IBM's 32 nm eDRAM retention time has been analyzed in [13], and subthreshold leakage has been pointed as the main component determining the retention time of the 32 MB array. A testing methodology for eDRAM arrays has been developed in [14] where the authors have proposed specific test patterns to accentuate the bitcell leakage and WL-BL and WL-bit coupling noises. A retention time optimization methodology for Intel eDRAM has been presented in [20]. Although the authors report loss in retention time due to coupling effects, the retention model is not described.

Manuscript received August 21, 2013; revised December 17, 2013 and January 23, 2014; accepted February 22, 2014. This paper is based on work supported by the Semiconductor Research Corporation under Grant 2442.001. This paper was recommended by Associate Editor T. S. Gotarredona.

The author is with Department of Computer Science and Engineering, University of South Florida, Tampa, FL 33647 USA (e-mail: sghosh@cse.usf.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSI.2014.2312481

1549-8328 © 2014 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

To the best of our knowledge, this is the first attempt toward modeling of eDRAM retention time considering the leakage, various noise sources, high-frequency effects, and variations. In summary, we make following contributions in this paper.

a) We present leakage sources, variations, high-frequency effects, and pattern-dependent coupling noise sources in eDRAM design. We also model their impact on retention time for both storage polarities.
b) We propose a new coupling noise model called BL-WL-BL coupling that plays a major role in determining the array retention.
c) We present a retention time model for both storage polarities. Our results using this model indicate that eDRAM retention time in scaled technologies can be severely limited by coupling noise reducing the yield significantly.

The rest of the paper is organized as follows. In Section II, we describe the basic eDRAM operation, timing waveforms and bitcell retention. We discuss the effects contributing toward retention time in Section III. The retention time model and results have been presented in Section IV and techniques to improve the retention are introduced in Section V. Conclusions are drawn in Section VI.

II. Basic eDRAM Operation

In this section, first we describe the column circuit of eDRAM. Next, we explain the timings and bitcell retention.

A. Structure of Column I/O

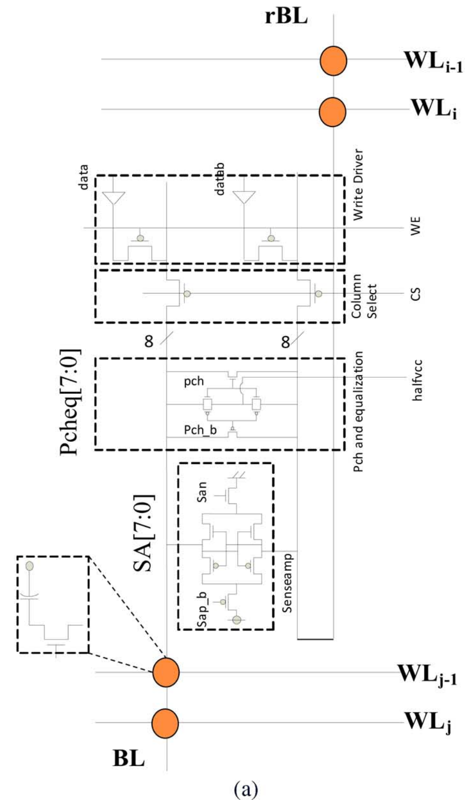
Fig. 1(a) shows the simplified column structure of open bitline architecture [15]. It shows the bitline (BL), reference bitline (rBL), 1T1C, and peripheral circuits such as senseamp, precharge, column select, write driver, and halfvcc generator. Note that the column circuit is different from conventional SRAM design in following ways: a) senseamp is placed on per-column basis (instead of per-global column) in order to restore the bits associated with unselected columns during the read/write operation; b) senseamp is fired by enabling both the header and footer transistors (instead of footer transistor firing) in order to prevent static current due to halfvcc bitline precharge; c) the precharge and equalization circuit consists of full CMOS gates due to halfvcc precharge (instead of NMOS only pass transistor); and d) the wordline is boosted to $V_{PP}(\sim V_{cc} + V_{tn})$ in order to write a full "1" through the NMOS access transistor and under-driven to V_{BB} to reduce subthreshold leakage during idle mode.

The column select is PMOS type in order to enable writing a good "1" through the write driver. Writing of "0" is accomplished by first writing a good "1" on the reference bitline (rBL) while the bitline (BL) stays close to V_{tp} (threshold voltage of pmos transistor) and then firing the senseamp. Once the senseamp turns on, it pulls the BL all the way to ground and a good "0" is written to the bitcell.

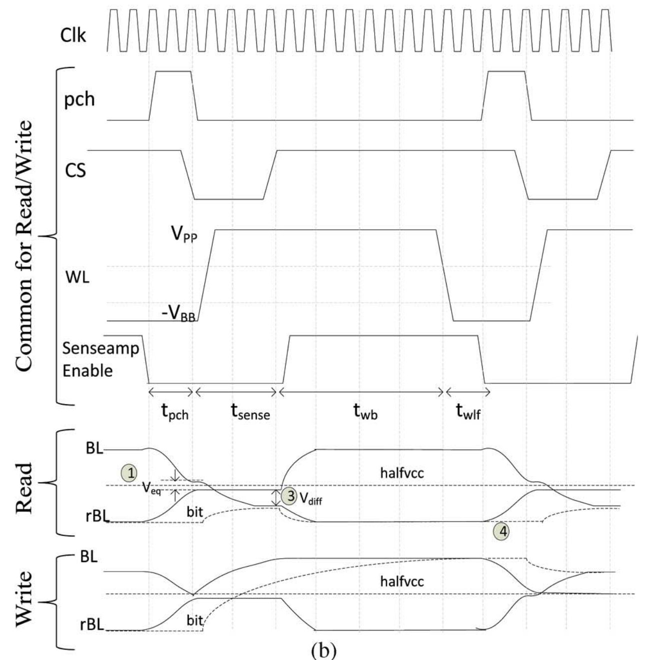
B. Circuit Timing

Fig. 1(b) shows the timing of eDRAM operation. The entire access can be divided into four categories as described below.

a) *Precharge*: The BL and rBL are precharged to halfvcc to prepare them for the next access. Note that there may be some amount of un-equalization between the bitlines (V_{eq}) due to high-frequency operation and higher threshold voltages of PMOS and NMOS under process variations.
b) *Sense*: When wordline (WL) turns on, the access transistor starts conducting and the bitline starts to develop differential by sharing charge with the bitcell (during read operation or refresh of unselected columns during write operation). This operation is known as "sense."
c) *Writeback*: After the development of differential, the senseamp is fired and the bitlines resolve to their re-



(a)



(b)

Fig. 1. (a) eDRAM column I/O circuit. (b) Example timing diagram of a high-frequency eDRAM macro.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

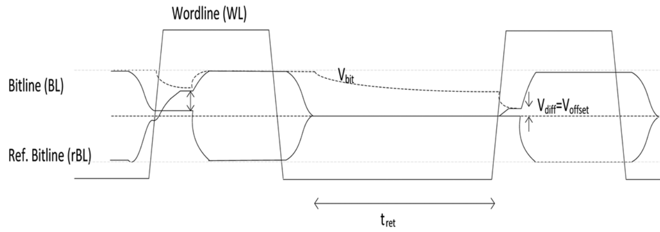


Fig. 2. Retention time of the bitcell. The bitcell voltage is V_{bit} after writeback but the charge leaks out during idle period and the next access results in just enough sense margin to meet the offset voltage of the sense amplifier (V_{offset}).

spective values. Since the WL is ON, the bitcell starts to regain the charge that was lost during sensing period. This operation is known as "writeback."

d) *WL fall*: Typically, turning ON/OFF the WL takes some extra amount of time because of the large voltage swing (V_{BB} to V_{PP} and vice versa). During sensing, the WL front edge partly eats up from differential margin. However, the WL back edge needs extra time to discharge during which bitlines should be held by the senseamp. We call this "WL fall" operation.

The number of cycles assigned for each operation shown in Fig. 1(b) is just an example. In reality, the timing of each of these operations is determined by extensive statistical simulation for a given clock frequency. The total number of clock cycles allotted for a single access guides the architectural decisions such as number of independent banks that the system can access back-to-back (to maintain certain throughput). Therefore, clock cycle assignment for each of the above operation is extremely important. However this work is mainly focused on retention modeling with fixed timing assumptions and detailed timing can be included in the model for better accuracy.

C. Bitcell Retention Time

Retention is defined as the amount of time before which the cell can be read correctly (and restored). In 1T1C cell, the capacitor acts as storage element, and therefore, the cell value leaks over time. Once the bitcell voltages goes above/below a certain level (determined by the reference voltage), the access results in reading of wrong value. The retention phenomenon is further explained in Fig. 2 for a bitcell storing "1." During t_{ret} the bitcell loses its value due to leakage and subsequent read results in a differential voltage which is just enough to meet the offset voltage of the senseamp [12]. Similar mechanism holds true for store "0" where the bitcell can gain charge through leakage and corrupt the stored value. Typical eDRAM macro requires periodic refresh in order to maintain the functional correctness of the read/write operation. Ideally, the retention time is desired to be purely guided by leakage through the access transistor and the senseamp offset. However, eDRAM in scaled technologies suffers not only from access transistor leakage but also from differential loss due to capacitor leakage, reference voltage variations as well as coupling noise (BL-BL, BL-rBL, WL-BL, BL-WL-BL). The details of these retention limiting factors have been discussed in the next section.

III. Retention Limiting Factors

In previous section, we discussed the basic eDRAM circuit, timings and retention time. In this section, we present the details of various factors responsible for reducing the retention time.

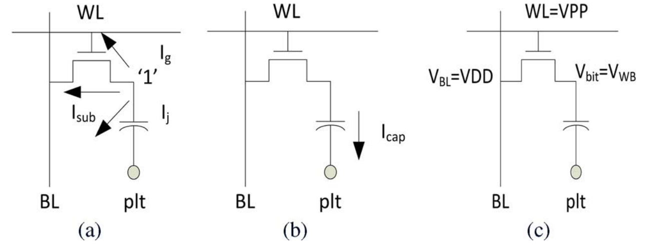


Fig. 3. (a) Various sources of bitcell leakage. (b) Cell capacitor leakage. (c) Incomplete writeback that results in $V_{bit} = V_{WB} < V_{CC}$.

A. Access Transistor Leakage

Access transistor leakage is one of the well-known leakage mechanisms in conventional DRAM. There are three major sources of access transistor leakage, namely subthreshold leakage (I_{sub}), junction leakage (I_j), and gate leakage (I_g) as shown in Fig. 3(a). The total leakage through the access transistor is given by

$$I_{leak} = I_{sub} + I_g + I_j.$$

The detailed expressions of the above leakage components can be found in [16] and have been omitted here for brevity. If the transistor leakage is known then the voltage loss for store-0 and store-1 can be expressed as

$$\Delta V_{tran}(st0) = \frac{I_{leak(st0)} t_{ret}}{C_{cell}}$$
$$\Delta V_{tran}(st1) = \frac{I_{leak(st1)} t_{ret}}{C_{cell}} \quad (1)$$

where

$I_{leak(st0)} =$ total leakage for store $-0 = I_{sub}$

$I_{leak(st1)} =$ total leakage for store $-1 = I_{sub} + I_g + I_j$

$t_{ret} =$ retention time

Although leakage current under process variation can be modeled, we have used Monte Carlo to estimate the impact of bitcell retention due to leakage in order to improve the accuracy of results (Section IV-C).

B. Capacitor leakage

Ideally, leakage through the storage capacitor [as shown in Fig. 3(b)] should be zero. However, the storage capacitor in scaled technologies experiences significant leakage, depending on voltage across the plate and operating temperature [17]. If the cell capacitor leakage is assumed to be constant (for simplicity) with magnitude $I_{cap(st-0)}$ and $I_{cap(st-1)}$ for store-0 and store-1 respectively, then the voltage loss of the bitcell is given by

$$\Delta V_{cap}(st0) = \frac{I_{cap(st0)} t_{ret}}{C_{cell}}$$
$$\Delta V_{cap}(st1) = \frac{I_{cap(st1)} t_{ret}}{C_{cell}}. \quad (2)$$

In this work we have assumed 0.1 pA/cell of capacitance leakage [17].

C. Insufficient Writeback

Writing of the last tens of millivolts on the storage capacitor is extremely challenging and time consuming due to poor V_{ds} in that operating region. This is especially true for store "1" when

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

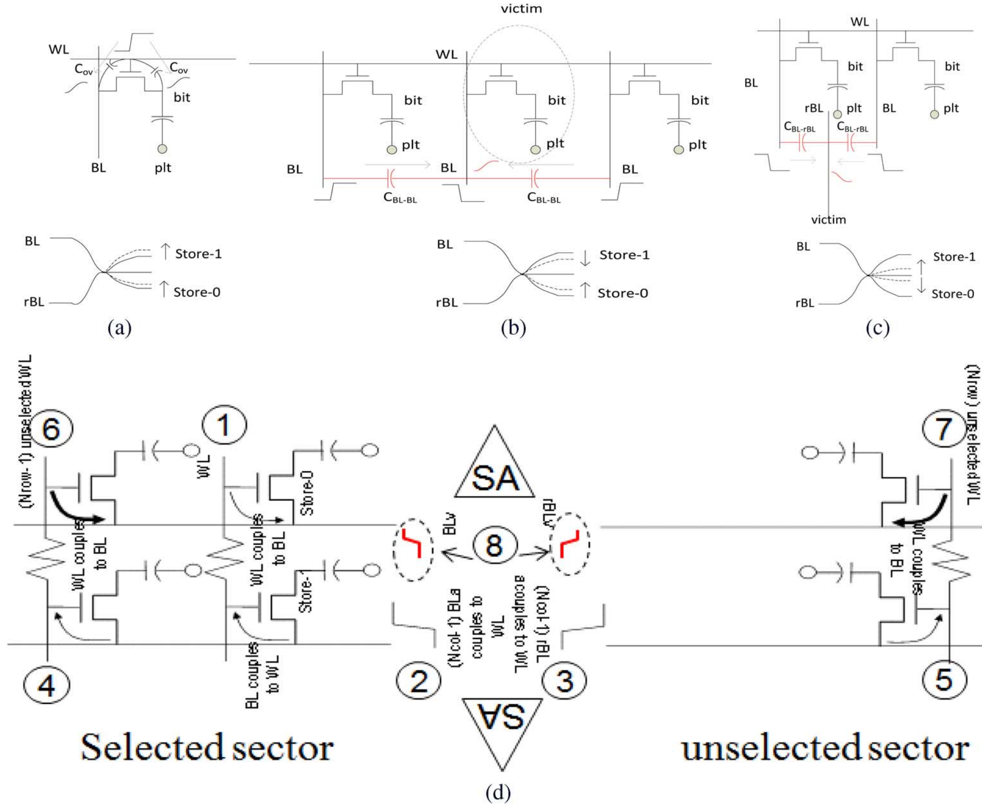


Fig. 4. Impact of coupling noise on the sense margin. (a) WL-bit/WL-BL coupling (store-1 sense margin increases whereas store-0 sense margin decreases). (b) BL-BL coupling (store-1 and store-0 margins are negatively impacted). (c) BL-rBL coupling (store-1 and store-0 margins are negatively impacted). (d) BL-WL-BL coupling mechanism. The sequence of events leading to coupling is marked.

$V_{ds} \sim 10$ mV and $V_{gs} \sim V_{tn}$ where transistor drivability is very poor. Writeback of store "0" is relatively faster because of NMOS access transistor with large $V_{gs} (\sim V_{PP})$. Commodity DRAM avoids the loss of writeback voltage by employing very relaxed access time (~50 ns). eDRAM access, on the other hand, is very tightly designed (with random access time in the order of few nanoseconds); therefore the writeback time is limited. This is further accentuated due to high V_{tn} access transistor (to control transistor subthreshold leakage) that eventually limits the write speed of last tens of millivolts. Incomplete writeback reduces the retention time because a lower voltage on the storage node lowers the amount of stored charge. Writeback voltage [V_{WB} as shown in Fig. 3(c)] depends on frequency of operation (F), number of cycles allocated for writeback (N_{WB}), V_{tn} of access transistor, and wordline voltage (V_{PP}). Assuming linear region of operation (due to very poor V_{ds}), V_{WB} can be found by solving

$$\frac{\mu C_{ox} W N_{WB}}{LF}(V_{pp} - V_{tn} - V_{WB})(V_{WB} - V_{BL}) = C_{cell}(V_{WB} - V_{halfvcc} - V_{diff})$$

where C_{ox} is oxide capacitance, μ is mobility, W is width, and L is channel length of the access transistor. If the writeback voltage loss is $\Delta_{WB(st-0)}$ and $\Delta_{WB(st-1)}$ for store-0 and store-1 respectively (estimated from above equation) then the impact of incomplete writeback on bitcell voltage is given by

$$\Delta V_{WB}(st0) = \Delta_{WB}(st0)$$
$$\Delta V_{WB}(st1) = \Delta_{WB}(st1). \tag{3}$$

Note that writeback voltage is temperature dependent. Cold temperature is worse for writeback as the threshold voltage of the access transistor increases and degrades gate overdrive. Therefore, actual estimate of writeback voltage loss should be performed at cold temperatures.

D. Insufficient Bitline Equalization

Variation in transistor threshold voltage due to process fluctuations affects the bitline equalization (during the precharge). High threshold voltage of the equalization transistors prevents full equalization and results in residual voltage between BL and rBL. This residual voltage reduces the differential margin when opposite values are read back-to-back. This is shown in Fig. 1(b) where previous read "1" leaves un-equalization voltage V_{eq} that eventually reduces the next read "0" sense margin. The insufficient equalization gets worse at lower temperatures (due to higher V_{th}) and higher operating frequencies. If the amount of un-equalization voltage between BL and rBL is Δ_{eq} then the impact on sense margin (for store-0 and store-1) is given by

$$\Delta SM_{eq}(st0) = \Delta SM_{eq}(st1) = \Delta_{eq}. \tag{4}$$

E. Reference Voltage Variation

Reference voltage plays a major role in determining the sense margin during read/refresh operation. In halfvcc sensing scheme the bitlines are precharged to halfvcc (that is generated internally). If the halfvcc level moves down, sense "1" is favored compared to sense "0" and vice versa. Under DC conditions,

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

$\mathrm{BL} = \mathrm{rBL} = \mathrm{V_{halfvcc}}$; however, during high-frequency operation bitlines in high state discharge quite a bit of charge on halfvcc network injecting noise. If the halfvcc generator is strong and the halfvcc network resistance is small then the injected noise vanishes quickly otherwise the noise persists. This is in addition to the halfvcc movement due to process fluctuations in halfvcc generator. If the bitcell capacitance is $\mathrm{C_{cell}}$, bitline capacitance is $\mathrm{C_{BL}}$, and change in halfvcc level is Δ_h, then the impact on sense margin is given by

$$\Delta SM_h(st1) = \Delta SM_h(st1) = \Delta_h \left(\frac{C_{BL}}{C_{cell} + C_{BL}}\right). \quad (5)$$

Note that equalization voltage is temperature dependent. Cold temperature is worse as the threshold voltage of the equalization transistor increases and degrades the gate overdrive. Therefore, actual estimate of equalization voltage loss should be performed at cold temperatures.

F. Coupling-Induced Noise

Coupling-induced noise from the bitlines, wordlines, and senseamps lowers the sense margin, thereby reducing the retention time. In the following paragraphs, we will describe several important coupling noise sources that are prominent in determining the eDRAM retention time.

i) WL-bit/WL-BL coupling (WL rise): Wordline couples to the bit and BL due to gate-source and gate-drain overlap capacitance and also due to fringing capacitance between poly and source/drain contacts. This capacitance component is becoming prominent in finfet type devices that are currently being used in 22 nm technologies [18]. Store-1 and store-0 are affected differently due to this coupling noise. The differential margin in store-0 reduces because the bit as well as BL moves up as the WL ramps from $\mathrm{V_{BB}}$ to $\mathrm{V_{PP}}$ (the rBL remains precharged to halfvcc). On the other hand, store-1 gets benefitted due to improved sense margin. This is shown in Fig. 4(a). If the bitcell capacitance is $\mathrm{C_{cell}}$, bitline capacitance is $\mathrm{C_{BL}}$, and coupling capacitance is $\mathrm{C_{ov}}$, then the coupling noise is given by

$$\begin{aligned}\Delta SM_{WLr-bit/WLf-BL}(st0) \\ = \Delta SM_{WLr-bit/WLf-BL}(st1) \\ = (V_{PP} + V_{BB})\left(\frac{2C_{ov}}{C_{cell} + C_{BL} + 2C_{ov}}\right)\end{aligned} \quad (6)$$

where $\mathrm{V_{PP}}$ is WL high and $-\mathrm{V_{BB}}$ is WL low voltage. Note that the coupling capacitance units will cancel out leaving out voltage on RHS (which is the unit of sense margin).

ii) WL-bit coupling (WL fall): Wordline couples to the bitcell during fall time (which is similar to the coupling described above) reducing the bitcell voltage. Note that WL fall coupling only affects the store-1 retention negatively as it reduces the writeback voltage. The amount of coupling is given by

$$\begin{aligned}\Delta V_{WLf-bit}(st0) &= \Delta V_{WLf-bit}(st1) \\ &= (V_{PP} + V_{BB})\left(\frac{C_{ov}}{C_{bit} + 2C_{ov}}\right).\end{aligned} \quad (7)$$

iii) BL-BL coupling: Bitline couples to the neighboring bitlines due to parasitic capacitance in the bitcell and tightly drawn bitline pitches in the column I/O area. This coupling reduces the sense margin when the neighboring columns sense opposite data [Fig. 4(b)]. If the bitline capacitance is $\mathrm{C_{BL}}$, coupling capacitance is $\mathrm{C_{BL-BL}}$, and differential voltages of the aggressor bitlines are $\mathrm{V_{diff}}$, then the coupling noise is given by

$$\begin{aligned}\Delta SM_{BL-BL}(st0) &= \Delta SM_{BL-BL}(st1) \\ &= V_{diff}\left(\frac{2C_{BL-BL}}{C_{BL} + 2C_{BL-BL}}\right).\end{aligned} \quad (8)$$

iv) BL-rBL coupling: BL couples to the rBL mostly due to tightly drawn bitline pitches in the column I/O area. Majority of the coupling also originates from senseamp gate to drain overlap capacitance. This coupling reduces both store-1 and store-0 sense margins because rBL moves in the same direction as BL [Fig. 4(c)]

$$\begin{aligned}\Delta SM_{BL-rBL}(st0) &= SM_{BL-rBL}(st1) \\ &= V_{diff}\left(\frac{2C_{BL-rBL}}{C_{BL} + 2C_{BL-rBL}}\right).\end{aligned} \quad (9)$$

v) BL-WL-BL coupling: This is one of the interesting and non-intuitive coupling phenomena where firing of senseamps creates coupling noise on the selected as well as unselected wordlines from aggressor bitlines (through access transistor gate overlap capacitance). This noise, in turn, couples back to the victim bitlines through access transistor reducing the sense margin of victim bitlines that are slowly being resolved by a weak senseamp.

This coupling mechanism is shown in Fig. 4(d). For the sake of simplicity the unselected wordlines in selected/unselected sector are lumped together. Similarly the selected and unselected bitines are lumped together. The coupling event starts with selection of a WL (1), firing of senseamp and resolution of bitline voltages in both selected and unselected sectors (2 and 3). Since both selected and unselected bitlines resolve together, the bitline voltage couples to both selected and unselected WL through access transistor gate overlap capacitance (4 and 5). This coupling noise in turn couples back to victim (or selected) bitline through the gate-overlap capacitance (6 and 7). The resulting noise on bitline and reference bitline move in opposite direction and compromises the sense margin (8).

This is again a pattern-dependent noise and a store-0 (store-1) in a sea of store-1 (store-0) creates the worst case coupling. The interesting fact about this coupling is that it lowers the sense margin through coupling from both BL and rBL side. If $\mathrm{N_{row}}$ and $\mathrm{N_{col}}$ is the number of rows and columns respectively (per sector), WL capacitance is $\mathrm{C_{WL}}$, and the differential of aggressors are $\mathrm{V_{diff}}$, then the noise on BL is given by

$$\begin{aligned}&\Delta SM_{BL-WL-BL}(st0) \\ &= \Delta SM_{BL-WL-BL}(st1) \\ &= \underbrace{|V_{halfvcc} - V_{diff}|}_{a} \underbrace{\left(\frac{N_{row}N_{col}C_{ov}}{N_{row}N_{col}C_{ov} + C_{WL}}\right)}_{b} \\ &\times \underbrace{\left(\frac{N_{row}C_{ov}}{N_{row}C_{ov} + C_{BL}}\right)}_{c}.\end{aligned} \quad (10)$$

In the above equation, term (a) represents the aggressor voltage responsible for coupling. Term (b) is the ratio of total overlap coupling capacitance $(\mathrm{N_{row}N_{col}C_{ov}})$ that couples to all WLs

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

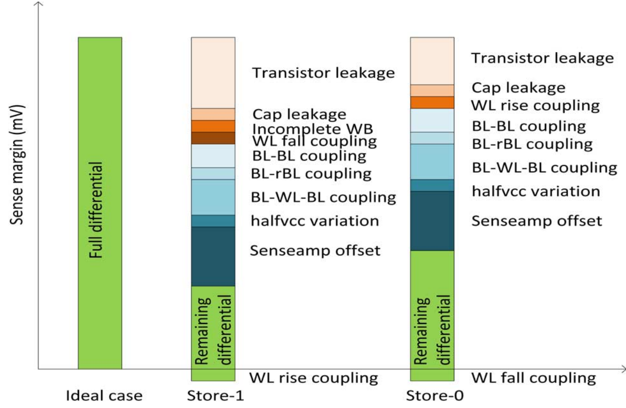


Fig. 5. Impact of leakage and coupling noise on the store-0 and store-1 sense margin. The ideal sense margin is also shown for reference.

TABLE I
PARAMETERS USED FOR RETENTION TIME ESTIMATION

Param	values	Param	values	Param	values
N_{row}	64	C_{cell}	20fF	V_{diff}	0.25V
N_{col}	1024	C_{BL}	20fF	V_{offset}	100mV
V_{cc}	1V	$C_{BL\text{-}BL}$	0.5fF	Δ_{eq}	15mV
V_{PP}	1.8V	$C_{BL\text{-}rBL}$	1fF	I_{cap}	0.1pA
V_{BB}	-0.15V	C_{WL}	200fF		
$V_{halfvcc}$	0.5V	Δ_{WB}	50mV		
C_{ov}	0.1fF	Δ_h	50mV		

TABLE II
MODELING OF RETENTION TIMES FOR (14) AND (15)

t_{ret}	Model
ideal	$\frac{C_{cell}(V_{WB}-V_{halfvcc})}{I_{leak}}$
V_{offset}	$\left(\frac{C_{cell}}{I_{leak}}\right)\left(\frac{V_{offset}}{CR}\right)$
BL-WL-BL	$\left(\frac{C_{cell}}{I_{leak}}\right)\left(\frac{\Delta SM_{BL-WL-BL}}{CR}\right)$
rBL-WL-rBL	$\left(\frac{C_{cell}}{I_{leak}}\right)\left(\frac{\Delta SM_{rBL-WL-rBL}}{CR}\right)$
Writeback	$\left(\frac{C_{cell}}{I_{leak}}\right)\Delta V_{WB}$
Halfvcc	$\left(\frac{C_{cell}}{I_{leak}}\right)\left(\frac{\Delta SM_{h}}{CR}\right)$
Equalization	$\left(\frac{C_{cell}}{I_{leak}}\right)\left(\frac{\Delta SM_{eq}}{CR}\right)$
BL-rBL	$\left(\frac{C_{cell}}{I_{leak}}\right)\left(\frac{\Delta SM_{BL-rBL}}{CR}\right)$
BL-BL	$\left(\frac{C_{cell}}{I_{leak}}\right)\left(\frac{\Delta SM_{BL-BL}}{CR}\right)$
WL_f-bit	$\left(\frac{C_{cell}}{I_{leak}}\right)\Delta V_{WLf-bit}$
Cell Leakage	$C_{cell}\left(\frac{1}{I_{leak}}-\frac{1}{I_{leak}+I_{cap}}\right)(V_{WB}-V_{halfvcc})$
WL_r-bit/WL_r-BL	$\left(\frac{C_{cell}}{I_{leak}}\right)\left(\frac{\Delta SM_{WL_r-bit/WL_r-BL}}{CR}\right)$

and the total capacitance of WLs. This term determines the amount of coupling that will propagate to the WLs. Term (c) is the ratio of total overlap coupling capacitance ($N_{row}C_{ov}$) that couples to victim BL and the total capacitance of victim BL. This term determines the amount of coupling that will propagate from WLs to the victim BL.

The coupling noise on rBL is similarly given by

$$\begin{aligned}\Delta SM_{rBL-WL-rBL}&(st0)\\ &=\Delta SM_{rBL-WL-rBL}(st1)\\ &=V_{halfvcc}\left(\frac{N_{row}N_{col}C_{ov}}{N_{row}N_{col}C_{ov}+C_{WL}}\right)\left(\frac{N_{row}C_{ov}}{N_{row}C_{ov}+C_{BL}}\right).\end{aligned}\quad(11)$$

The BL-WL-BL coupling is sensitive to gate to drain overlap capacitance of access transistor and relative timing of senseamp firing. If the victim senseamp is fired late, the noise will fully couple to BL and rBL reducing the sense margin to almost zero. However if the victim and aggressor senseamps are fired simultaneously then the coupling noise will be partially neutralized by the victim senseamps (which have gained sufficient strength). To account for (a) relative timings of aggressor and victim senseamp firing, (b) noise shielding due to WL resistance, and (c) WL driver sinking the noise, we have scaled the noise magnitude by 50% in simulations.

IV. RETENTION MODEL

For the given bitcell and bitline capacitance, the ideal sense margin is given by

$$SM=(V_{bit}-V_{halfvcc})CR\ where\ CR=\left(\frac{C_{cell}}{C_{BL}+C_{cell}}\right).\quad(12)$$

In absence of any coupling noise the ideal retention time would be given by

$$\begin{aligned}SM&=V_{offset}\\ &=\left(V_{WB}-\frac{I_{leak}t_{ret}}{C_{cell}}-V_{halfvcc}\right)CR\\ t_{ret}&=\frac{\left(V_{WB}-\frac{V_{offset}}{CR}-V_{halfvcc}\right)}{I_{leak}}\\ &=\left(\frac{C_{cell}}{I_{leak}}\right)(V_{WB}-V_{halfvcc})-\left(\frac{C_{cell}}{I_{leak}}\right)\left(\frac{V_{offset}}{CR}\right)\\ &=t_{ret}(ideal)-t_{ret}(V_{offset}).\end{aligned}\quad(13)$$

If we assume $I_{leak}=10$ pA, $V_{halfvcc}=0.5$ V, $V_{WB}=1$ V, $V_{offset}=0.05$ V, $C_{cell}=20$ fF [5], and $C_{BL}=20$ fF, then approximate value of retention time would be 800 μs. Note that the total retention time in (14) has been broken into two components: a) ideal retention time (t_{ret} [ideal)] and b) loss in retention time [$t_{ret}(V_{offset})$] due to senseamp offset. In this example, ideal retention time would be 1 ms; however, the senseamp offset reduces the retention by 0.2 ms, resulting in 800 μs of final retention. Similarly, various leakage mechanisms (as discussed in previous sections) reduce the sense margin that in turn lowers the retention time. Therefore the eDRAM macro is refreshed frequently, which increases the refresh power. Fig. 5 pictorially illustrates the impact of various factors affecting sense margin for store-1 and store-0 in high-speed eDRAM arrays. It can be noted that store-0 sense margin is expected to be better compared to store-1 value due to less transistor leakage and usage of NMOS access transistor that helps in fast writeback. In the following paragraphs, first we will present the store-0 and store-1 retention time model. Next, we will outline the simulation results.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

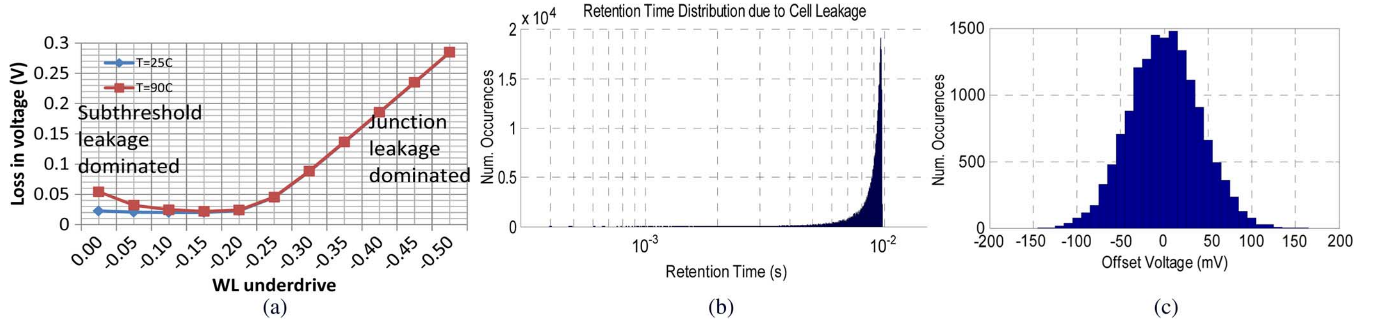


Fig. 6. (a) Optimization of wordline underdrive for leakage reduction. (b) Impact of process variation on store-1 retention time. (c) Impact of process variation on offset voltage of senseamp.

A. Store-1 Retention Model

The sense margin for store-1 can be written as

$$\begin{aligned} SM(st1) =& V_{offset} \\ =& (V'_{bit} - V_{halfvcc})\,CR - \Delta SM_{eq} - \Delta SM_h \\ &+ \Delta SM_{WL-bit/WL-BL}(st1) \\ &- \Delta SM_{BL-BL}(st1) - \Delta SM_{BL-rBL}(st1) \\ &- \Delta SM_{BL-WL-BL}(st1) \end{aligned}$$

where $V'_{bit} = V_{bit} = \Delta V_{tran}(st1) - \Delta V_{cap}(st1) - \Delta V_{WB}(st1) \Delta V_{WL-bit}(st1)$.

Putting the values of ΔV and ΔSM from (1)–(11) and solving for t_{ret} results in

$$\begin{aligned} t_{ret} =& t_{ret}(ideal) - t_{ret}(V_{offset}) - t_{ret}(BL - WL - BL) \\ &- t_{ret}(rBL - WL - rBL) - t_{ret}(WB) \\ &- t_{ret}(halfvcc) - t_{ret}(BL - rBL) - t_{ret}(eq) \\ &- t_{ret}(BL - BL) - t_{ret}(WL_f - bit) - t_{ret}(I_{cap}) \\ &- t_{ret}(WL_r - bit/WL_r - BL) \end{aligned} \tag{14}$$

where the individual components are defined in Table I.

B. Store-0 Retention Model

The retention time for store-0 can be written as

$$\begin{aligned} t_{ret} =& t_{ret}(ideal) - t_{ret}(V_{offset}) - t_{ret}(BL - WL - BL) \\ &- t_{ret}(rBL - WL - rBL) - t_{ret}(WB) \\ &- t_{ret}(halfvcc) - t_{ret}(BL - rBL) - t_{ret}(eq) \\ &- t_{ret}(BL - BL) + t_{ret}(WL_f - bit) - t_{ret}(I_{cap}) \\ &- t_{ret}(WL_r - bit/WL_r - BL) \end{aligned} \tag{15}$$

where $t_{ret}(^*)$ are retention times using store-0 terms in Table II.

C. Simulation Results

For the estimation of retention time, we have assumed the technology parameters based on 22 nm predictive technology [19]. The values of bitcell cap, bitline cap, WL cap, overlap cap, and leakage numbers have been estimated from PTM as well as from 32 nm eDRAM design [5]. The values of these parameters for our study have been shown in Table I. The WL sleep voltage during retention (V_{BB}) is determined by measuring leakage-induced voltage loss at each WL underdrive value at 25C and 90C. From Fig. 6(a) we find that $V_{BB} = -0.15$ V is optimal in terms of leakage, and this value was used for our retention simulation. The process variation has been modeled by threshold voltage variation with $(\mu, \sigma) = (0, 30\ \mathrm{mV})$ and channel length variation with $(\mu, \sigma) = (0, 3\ \mathrm{nm})$. Fig. 6(b) shows the impact of process variations on retention time of store-1 bitcell (at 90C) whereas Fig. 6(c) shows the variation in offset voltage of senseamp. We have used the 4.5-sigma values of store-0 and store-1 retention time and 1.5-sigma value of senseamp offset to model variation in a 1 Mb array. The row and column size of the eDRAM macro is determined by following IBM eDRAM design [5]. V_{pp} is chosen conservatively considering charge pump variation (assumed to be 20%). Typically, $V_{pp} = V_{cc} + V_{tn}$, which is ~1.5 V; however charge pump output may change to 1.8 V under variations. Halfvcc is equal to 0.5 V for $V_{cc} = 1$ V. The halfvcc variation is assumed to be 10% (50 mV) which is possible due to process variation. V_{diff} is 250 mV due to charge sharing between bitcell capacitance and bitline capacitance. Writeback voltage is assumed to be 50 mV because writing the last 50 mV becomes extremely slow. Equalization voltage is assumed to be 5–10% of V_{diff} which amounts to 12.5 to 25 mV.

The impact of each component (leakage and noise) on retention time has been computed using (15) and shown in Fig. 7(a), (b) for store-0 and store-1. Note that these plots should not be confused with the sense margin trend that is qualitatively illustrated in Fig. 5. The ideal retention time for store-0 is store-1 is 1000 and 667 μs, respectively for these parameters. It can be observed that BL-WL-BL coupling noise plays the most crucial role in reducing the retention time (for both store-0 and store-1) even after assuming only 50% noise impact. This coupling noise must be contained by adjusting the senseamp firing timings. The negative retention components in Fig. 7(a), (b) indicate that they improve the retention time. We have also plotted how the retention time changes once we start accounting for each noise/leakage components [Fig. 7(c), (d)]. It can be clearly seen from these plots that even though the ideal retention time (leakage dominated) is high, noise, high-frequency operations, and variations can reduce the retention by as much as 10–16×. Furthermore, store-1 retention dominates the array retention. These results strongly indicate that understanding and detailed modeling of eDRAM retention time is crucial. Moreover, it can be concluded that the success of eDRAM in future technology nodes would largely depend on containing these noise and variation sources.

D. Model Validation

Retention time of commodity DRAM is in the order of milliseconds (ms) and is typically limited by bitcell leakage (the

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

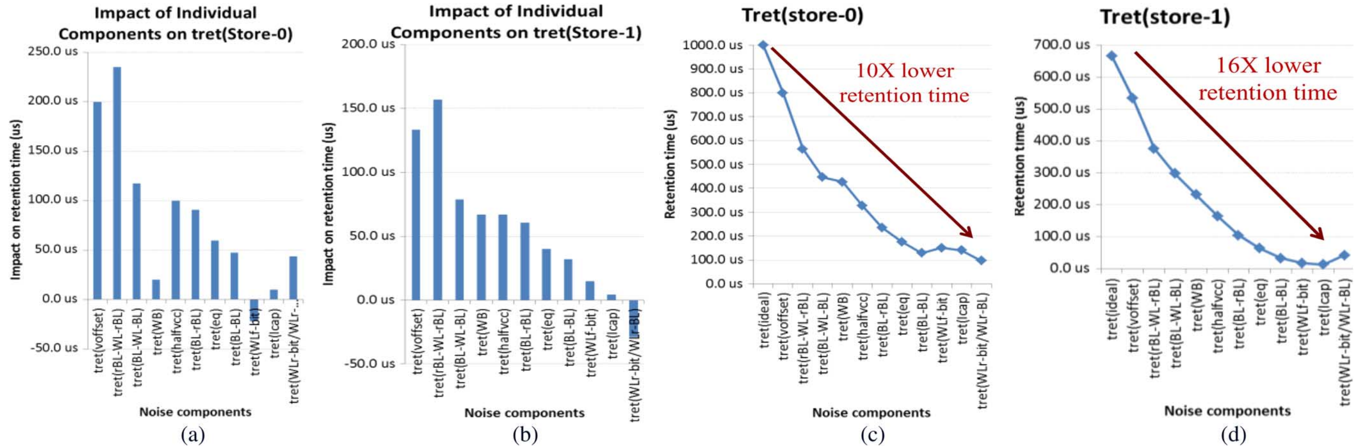


Fig. 7. Impact of leakage and coupling noise on retention time. (a) Store-0. (b) Store-1. (c) Retention time for store-0. (d) Retention time for store-1.

contribution of noise is minimal). Hence the current literature on DRAM does not describe Silicon data on noise-induced retention time loss. eDRAM retention time is in the order of microsecond (μs). Therefore factors such as noise and high-speed operation are more noticeable. However the individual retention limiting cases depend on bitcell architecture. The recently published Intel eDRAM paper [20] show orders of magnitude increase in bit failure rate due to coupling noise. This outcome is analogous to the results obtained through our proposed model. We believe that the proposed model is generic in nature and will provide a tool to the designer in estimating the retention time by plugging the right parameters according to their eDRAM specifications.

V. Improving Retention Time

In previous section, we described the retention model and the impact of various noise sources. This section presents possible techniques to improve the retention time.

A. Lowering the Coupling Noise

The coupling noise can be lowered by reducing the coupling capacitance which requires either process or design modifications. For example, metal-metal coupling capacitance can be reduced by shortening the metal height (at the cost of resistance) or by using low-k dielectric. Similarly, gate-drain/source coupling capacitance can be reduced by manipulating the gate overlap area. The design solution to reduce the coupling capacitance would involve shielding whenever possible however this may not be feasible due to tight column pitch.

B. Increasing Storage Capacitance

Another way to regain the lost retention time is to increase the storage capacitance. This is a costly approach in terms of additional process steps. In order to maintain a retention time of 200 μs after the coupling events the storage capacitance would have to be increased by 15% whereas restoring the retention time back to ideal would require doubling of storage capacitance.

VI. Conclusions

Designing eDRAM array to meet certain retention time specification at high operating frequency requires detailed understanding of the impact of leakage, noise, and variations. Commodity DRAM solves the retention issue by employing extremely low leakage access transistor and implementing high C storage capacitance (both of which are not feasible in eDRAM technology). The leaky access transistor along with low cell capacitance leaves eDRAM designers no choice but to trade various design parameters in order to achieve reasonable retention time. Therefore accurate modeling of retention time during design time is extremely important for making right decision in terms of determining the transistor parameters (e.g., V_{th}, parasitic caps) as well as for adjusting other circuit knobs (e.g., timings, voltage levels). This paper describes the factors affecting the retention time of high-speed eDRAM array design. We have presented the often overlooked factors that are becoming prominent in scaled technologies and need careful attention for high-speed and low-power eDRAM array design.

References

[1] P. W. Diodato, “Embedded DRAM: More than just a memory,” *IEEE Commun. Mag.*, 2000.

[2] R. E. Matick and S. E. Schuster, “Logic-based eDRAM: Origins and rationale for use,” *IBM J. Res. Devel.*, vol. 49, no. 1, pp. 145–165, 2005.

[3] A. K. Khan, H. Magoshi, T. Matsumoto, J. Fujita, M. Furuhashi, M. Imai, and Y. Kurose *et al.*, “A 150-MHz graphics rendering processor with 256-Mb embedded DRAM,” *IEEE J. Solid-State Circuits*, vol. 36, no. 11, pp. 1775–1784, Nov. 2001.

[4] Kalla, Ron, B. Sinharoy, W. J. Starke, and M. Floyd, “Power7: IBM’s next-generation server processor,” *IEEE Micro*, vol. 30, no. 2, pp. 7–15, 2010.

[5] Barth, John, D. Plass, E. Nelson, C. Hwang, G. Fredeman, M. Sperling, and A. Mathews *et al.*, “A 45 nm SOI embedded DRAM macro for the power processor 32 MByte on-chip L3 cache,” *IEEE J. Solid-State Circuits*, vol. 46, no. 1, pp. 64–75, Jan. 2011.

[6] Suzuki and Satoshi, The Development of Embedded DRAM Statistical Quality Models at Test and Use Conditions Portland State Univ.. Portland, OR, USA, 2010.

[7] K. Kim, “Technology for sub-50 nm DRAM and NAND flash manufacturing,” *IEDM*, 2005.

[8] Hamamoto, Takeshi, S. Sugiura, and S. Sawada, “On the retention time distribution of dynamic random access memory (DRAM),” *IEEE Trans. Electron Devices*, vol. 45, no. 6, pp. 1300–1309, 1998.

[9] Y. Mori, R.-I. Yamada, S. Kamohara, M. Moniwa, K. Ohyu, and O. Yamanaka, IEEE, “A new method for predicting distribution of DRAM retention time,” in *Proc. 39th Annu. IEEE Int. Reliability Physics Symp.*, 2001, pp. 7–11.

[10] Jin, Seonghoon, J.-H. Yi, J. H. Choi, D. G. Kang, Y. J. Park, and H. S. Min, “Prediction of data retention time distribution of DRAM by physics-based statistical simulation,” *IEEE Trans. Electron Devices*, vol. 52, no. 11, pp. 2422–2429, Nov. 2005.

[11] M. H. Cho, Y. I. Kim, D. S. Woo, S. W. Kim, M. S. Shim, Y. J. Park, W. S. Lee, and B. I. Ryu, IEEE, “Analysis of Thermal Variation of DRAM Retention Time,” in *Proc. 44th Annu. IEEE Int. Reliability Physics Symp.* , 2006, pp. 433–436.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

[12] Li, Yan, H. Schneider, F. Schnabel, R. Thewes, and D. Schmitt-Landsiedel, "DRAM yield analysis and optimization by a statistical design approach," *IEEE Trans. Circuits Syst. I: Reg. Papers*, vol. 58, no. 12, pp. 2906–2918, Dec. 2011.
[13] Kong, Wei, P. C. Parries, G. Wang, and S. S. Iyer, IEEE, "Analysis of retention time distribution of embedded DRAM-A new method to characterize across-chip threshold voltage variation," in *Proc. IEEE Int. Test Conf.*, 2008, pp. 1–7.
[14] Yang, Hao-Yu, C.-M. Chang, M. C. Chao, R.-F. Huang, and S.-C. Lin, "Testing methodology of embedded DRAMs," *IEEE Trans. Very Large Scale Integration (VLSI) Syst.*, vol. 20, no. 9, pp. 1715–1728, Sep. 2012.
[15] Jacob, Bruce, S. Ng, and D. Wang, *Memory Systems: Cache, DRAM, Disk*. San Mateo, CA, USA: Morgan Kaufmann, 2010.
[16] Y. Taur and T. H. Ning, *Fundamentals of Modem VLSl Devices*. New York, NY, USA: Cambridge Univ. Press, 1998.
[17] Park, Donggun, H. Ban, S. Jung, H. Yang, and W. Lee, IEEE, "Stress-induced leakage current comparison of giga-bit scale DRAM capacitors with OCS (one-cylinder-storage) node," in *IEEE Int. Integrated Reliability Workshop Final Report*, 2000, pp. 116–119.
[18] C. Auth, C. Allen, A. Blattner, D. Bergstrom, M. Brazier, M. Bost, and M. Buehler *et al.*, IEEE, "A 22 nm high performance and low-power CMOS technology featuring fully-depleted tri-gate transistors, self-aligned contacts and high density MIM capacitors," in *Proc. Symp. VLSI Technology*, 2012, pp. 131–132.
[19] Predictive Technology Model ASU [Online]. Available: http://www.asu.edu/~ptm
[20] Wang, Yih, U. Arslan, N. Bisnik, R. Brain, S. Ghosh, F. Hamzaoglu, N. Lindert, M. Meterelliyoz, J. Park, S. Tomishima, and K. Zhang, "Retention Time Optimization for eDRAM in 22 nm Tri-Gate CMOS Technology," *IEEE Int. Electron Device Meeting*, 2013.

Swaroop Ghosh (S'04–SM'13) received the B.E. (Hons.) degree from the Indian Institute of Technology, Roorkee, India in 2000, the M.S. degree from University of Cincinnati, Cincinnati, OH, USA, in 2004, and the Ph.D. degree from Purdue University, West Lafayette, IN, USA, in 2008.

He joined the faculty of the University of South Florida in Fall 2012. He was a Senior Research and Development Engineer in Advanced Design, Intel Corp. from 2008 to 2012. At Intel, his research was focused on low-power and robust embedded memory design in scaled technologies. He has three U.S. patents, published over 40 papers, and authored a book chapter. His research interests include energy-efficient, secure and robust circuit/system design, and digital testing for nanometer technologies.

Dr. Ghosh has served in the technical program committees of ISLPED, Nanoarch, VLSI Design, ISQED, ASQED, and VLSI-SOC.