Energy versus Data Integrity Trade-Offs in Embedded High-Density Logic Compatible Dynamic Memories

Adam Teman*, Georgios Karakonstantis*, Robert Giterman†, Pascal Meinerzhagen‡, Andreas Burg*
*Telecommunications Circuits Lab (TCL), École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
†Faculty of Engineering, Bar-Ilan University, Ramat Gan, Israel
‡Circuit Research, Intel Labs, JF3-334, 2111 NE 25th Ave, Hillsboro, OR 97124
Email: {adam.teman, georgios.karakonstantis, andreas.burg}@epfl.ch

***Abstract*—Current variation aware design methodologies, tuned for worst-case scenarios, are becoming increasingly pessimistic from the perspective of power and performance. A good example of such pessimism is setting the refresh rate of DRAMs according to the worst-case access statistics, thereby resulting in very frequent refresh cycles, which are responsible for the majority of the standby power consumption of these memories. However, such a high refresh rate may not be required, either due to extremely low probability of the actual occurrence of such a worst-case, or due to the inherent error resilient nature of many applications that can tolerate a certain number of potential failures. In this paper, we exploit and quantify the possibilities that exist in dynamic memory design by shifting to the so-called approximate computing paradigm in order to save power and enhance yield at no cost. The statistical characteristics of the retention time in dynamic memories were revealed by studying a fabricated 2 kb CMOS compatible embedded DRAM (eDRAM) memory array based on gain-cells. Measurements show that up to 73% of the retention power can be saved by altering the refresh time and setting it such that a small number of failures is allowed. We show that these savings can be further increased by utilizing known circuit techniques, such as body biasing, which can help, not only in extending, but also in preferably shaping the retention time distribution. Our approach is one of the first attempts to access the data integrity and energy trade-offs achieved in eDRAMs for utilizing them in error resilient applications and can prove helpful in the anticipated shift to approximate computing.**

***Index Terms*—Embedded Memories, DRAM, Refresh Power, Data Integrity, Energy Efficiency, Error Resilience**

I. Introduction

The large amount of data that needs to be handled by today's systems has increased the memory requirements, which already often occupy more than 50% of silicon real-estate and power in embedded systems [1]. This has led to a rise in popularity of embedded dynamic-random-access-memory (eDRAMs) due to their high-density and low retention power, as compared to static random access memory (SRAM) [2]–[5]. However, eDRAMs still require periodic, power-hungry refresh cycles to retain the stored data. Traditional design approaches dictate that the frequency of these refresh cycles is determined by the worst-case retention time of the most leaky cell. While such an approach guarantees error-free storage, it results in high refresh power consumption, most of which is wasted due to the extremely rare occurrence of pessimistically assumed worst-case conditions [5]. The design margins and resulting wasted power is expected to further increase as silicon predictability reduces with technology scaling, putting the feasibility of such a design approach in doubt [6].

This reality has led to the quest for alternative design strategies and to the promising *approximate computing* paradigm [6]–[8], in which the error resilient nature of many applications is exploited to relax the design constraints and save power. The approximate computing paradigm includes the development of processors and software that may not always produce 100% precise results, but their output fidelity is acceptable for human consumption at a significantly reduced power [6]–[8]. Error resilient applications also open up possibilities in dynamic memory design, providing an opportunity to potentially reduce the memory refresh rate, without caring about a certain extent of resulting failures [9], [10]. However, this raises questions regarding the yet to be quantified energy vs. reliability trade offs in eDRAM, and if such a paradigm can be realized and lead to more efficient operation.

In this paper, we try to address these questions by investigating the viability of the idea for a paradigm shift in gain-cell based embedded DRAM (GC-eDRAM). These memories have been gaining attention in the research community, due to features such as small cell size, low leakage, and logic compatibility [2], [3], [5], [11], [12]. As a basis for our analysis, we use the retention time of a fabricated 2T gain-cell array, which has been characterized across several chips, allowing us to extract the failure probability as a function of the refresh rate. Our measurements not only reveal the large spread of retention time in GC-eDRAM, but also its graceful degradation, which is a preferred characteristic for relaxing the refresh rate and promoting the proposed approximate storage idea. In particular, this means that only very few cells have short retention times and the error probability increases monotonically and slowly as the refresh rate is decreased. This allows to save a significant amount of refresh power with only a small number of failures that can be limited to the tolerance threshold of a particular error resilient application.

Contributions: Our study advances recent works [9], [13] that have attempted to exploit the error resilient nature of various applications for implementing approximate DRAM storage by revealing the actual DRAM characteristics. Such early works were not able to reveal the actual statistical characteristics of the retention time since they were based mainly on high-level models of DRAM cells and focused on their utilization within simulators at the microarchitecture and software layer. Our contributions can be summarized as:

- Characterization of the retention time of a silicon fab-

This work was partially supported by the EU OPEN-FET SCoRPiO grant (no. 323872), the EU Marie-Curie DARE grant (no. 304186) and the ICYSoC RTD project (no. 20NA21 150939) funded by Nano-Tera.ch.

978-3-9815370-4-8/DATE15/©2015 EDAA

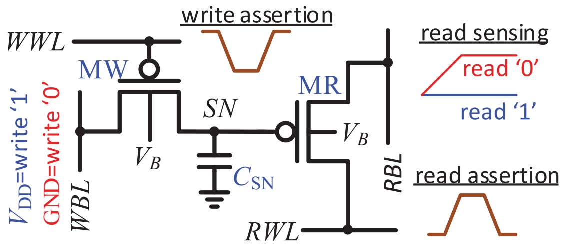


Fig. 1. (a) Schematic of all-PMOS 2T Gain Cell. (b) Operation waveforms.

ricated 2 kb GC-eDRAM array, revealing a large spread both within a single array and across different chips.

- Analysis of the wasted power, assuming that the traditional worst-case cell criterion is used to determine the refresh rate.
- Discussion of an alternative criterion for determining the refresh rate, and analysis of the potential power savings by applying it. We show that the failure rate increases gracefully with reduction of the refresh rate, thereby obtaining large power savings, while keeping the number of failures under application tolerable levels.
- Utilization of other circuit level techniques, such as the application of body bias, to shift the retention time distribution in order to increase the potential power savings and favorably shape it to better exploit the previous methods.
- Analysis of yield improvement through error tolerance, assuming a fixed refresh rate and power requirement.

The rest of the paper is organized as follows: Section II describes the operation of GC-eDRAM and analyzes the retention time of a fabricated array. Section III discusses the wasted power due to traditional worst-case design. The proposed approach and trade-offs between error probability and power savings are presented in Section IV, including yield improvement through error tolerance. Section V discusses circuit level techniques for achieving further power savings and improve the efficacy of the proposed idea. Finally, Section VI concludes the paper.

II. Gain Cell based eDRAM and Data Retention

GC-eDRAM is a low-cost, high-density alternative to SRAM for the implementation of on-chip, logic-compatible embedded memories, without the need for additional process steps. Several topologies of GC-eDRAM have been proposed, consisting of two to four transistors [2], [3], [5], [12], [14], providing a smaller unit-cell footprint than comparable 6T or 8T SRAM. However, as opposed to SRAM that uses a statically driven active feedback for data retention, GC-eDRAM solutions store data as temporary charge on in-cell (parasitic) capacitors. This charge leaks away over time, and therefore, these cells require periodic refresh cycles for data retention. As with conventional DRAM, for the majority of the GC-eDRAM solutions, this refresh power comprises the vast majority of their standby power consumption [2], [3], [5], [11].

In this paper, we will focus on the popular all-PMOS 2T

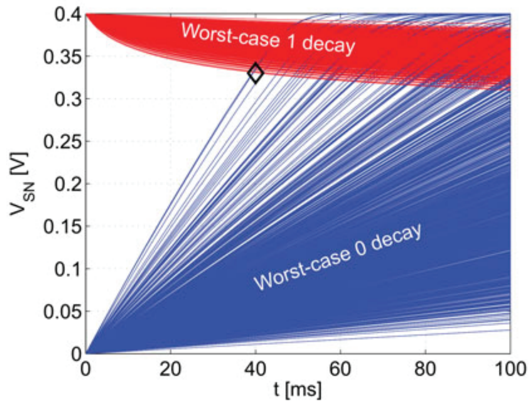


Fig. 2. Level degradation of an all-PMOS 2T GC-eDRAM bitcell over 1 k MC samples in a CMOS 0.18 µm process.

GC-eDRAM[1], shown in Fig. 1 [2], [5], [11], [12]. This cell comprises a write transistor (MW), a read transistor (MR), and a storage node (SN) made up of the parasitic capacitance of the devices and their connecting wires. Typical operational waveforms are illustrated in Fig. 1 for write and read access cycles for both '0' and '1' data. Write operations are initiated by biasing the write bitline (WBL) at V_{DD} ('1') or GND ('0'), and subsequently pulsing the write word line (WWL) to a negative voltage, thereby charging or discharging the SN. Subsequently, a read operation can be initiated by discharging the read bitline (RBL) and pulsing the read word line (RWL) to V_{DD}. Depending on the voltage level stored in the cell, MR is either cutoff (in case of a '1') or conducting (in case of a '0'). Therefore, RBL remains discharged or is charged, respectively, and amplified to a digital output level.

As with all dynamic circuits, the level stored on the internal capacitance of a GC-eDRAM cell degrades over time. In the case of the all-PMOS 2T cell, the dominant leakage current has been shown to be the subthreshold (sub-V_{T}) current through MW [5], [11], given by:

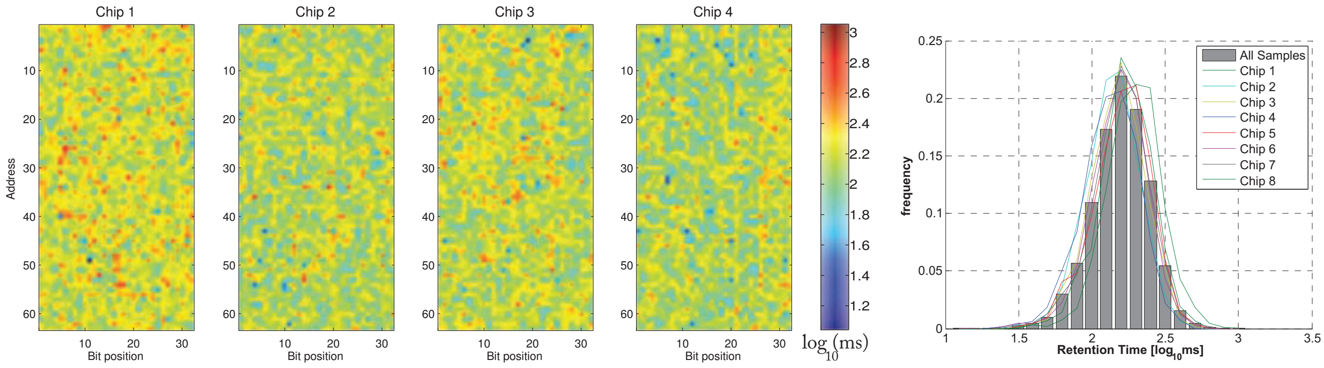
$$I_{\mathrm{MW}} = I_0 e^{\frac{V_{\mathrm{SG,MW}} - |V_{\mathrm{T,MW}}|}{n\phi_t}} (1 - e^{\frac{V_{\mathrm{SD,MW}}}{n\phi_t}}) e^{\frac{\eta V_{\mathrm{SD,MW}}}{n\phi_t}}, \quad (1)$$

where $V_{\mathrm{SG,MW}}$, $V_{\mathrm{SD,MW}}$, and $V_{\mathrm{T,MW}}$ are the source-to-gate, source-to-drain, and threshold voltages, respectively, of MW; ϕ_t is the thermal voltage; n is the sub-V_{T} slope; and η is the drain-induced barrier lowering (DIBL) coefficient.

The SN level degradation is simulated in Fig. 2 for 1000 Monte Carlo (MC) samples of initially stored '0' and '1' levels with WBL biased at a worst-case opposite level. It is clear from this figure that following a write operation, data integrity can only be ensured for a certain time period, during which a stored '0' and '1' can still be unequivocally differentiated. This time period is known as the data retention time (DRT), and it is exponentially dependent on $V_{\mathrm{SG,MW}}$, $V_{\mathrm{SD,MW}}$, and $V_{\mathrm{T,MW}}$, as shown in (1).

The exponential dependence of I_{MW} on $V_{\mathrm{T,MW}}$ is of particular importance, as assuming WWL is off (V_{DD}) and

[1] The 2T GC-eDRAM cell is used to demonstrate these concepts; however, the majority of the proposed ideas are valid for other eDRAM topologies and for conventional DRAM, as well.



(a) Per-bit retention maps of four chips

(b) PDF of retention time of all bits (bars) with PDFs of single chips overlayed

Fig. 3. Distribution of retention time over all chips.

WBL is held high, $V_{\mathrm{SG,MW}}$=0, and any fluctuation in $V_{\mathrm{T,MW}}$ causes a significant change in the DRT of the bitcell. Since $V_{\mathrm{T,MW}}$ is susceptible to random doping fluctuations (RDF) and other intra-die and inter-die variations, the DRT of a bitcell for given operating conditions is a random variable with extreme differences, even between two adjacent bitcells within the same array. This randomness is clearly depicted in Fig. 3(a), showing the per-bit mapping of DRT for four 2 kb GC-eDRAM arrays, fabricated in a commercial CMOS 0.18 µm process. The distribution of the DRT in each of the 8 measured chips and the probability density function (PDF) of all measured samples is further shown on a semi-logarithmic scale in Fig. 3(b). These figures reveal that the DRT approximately follows a log-normal distribution, with up to two orders-of-magnitude difference in DRT between cells.

III. Power waste due to worst-case design

The consequences of the observed variation of the retention times are severe. Even though, the majority of the data written to one of these measured cells will remain intact for hundreds of milliseconds, the statistical low-end outliers assumption require a refresh operation according to the DRT of the worst cell under worst-case biasing. Worst-case biasing occurs when the WBL of a cell storing a '0' is driven to V_{DD}, thereby incurring the highest possible sub-V_{T} leakage to the cell's SN. The actual probability that such a worst-case biasing will be present is extremely low, as it requires a constant write '1' operation applied to cells on the same column as the worst cell [5], [11]. Furthermore, in most cases, the refresh rate will be much higher than those dictated by the worst cell in a given chip, as it is set according to the worst cell in an entire lot of chips or according to pre-fabrication statistical simulations. This is emphasized in Fig. 4 by plotting the retention time of the worst cell of each of the 8 measured chips, showing an almost 3× difference between the best and worst chip.

During data retention, the power consumption of a GC-eDRAM array, comparable to the static power of an SRAM, is the combination of the leakage and the refresh power, and is often called *retention power*. Assuming a refresh operation is applied exactly at DRT, this power can be calculated as:

$$P_{\mathrm{ret}} = \frac{1}{T_{\mathrm{ref}}} E_{\mathrm{ref}} + P_{\mathrm{leak}} \approx \frac{E_{\mathrm{ref}}}{T_{\mathrm{ref}}} \qquad (2)$$

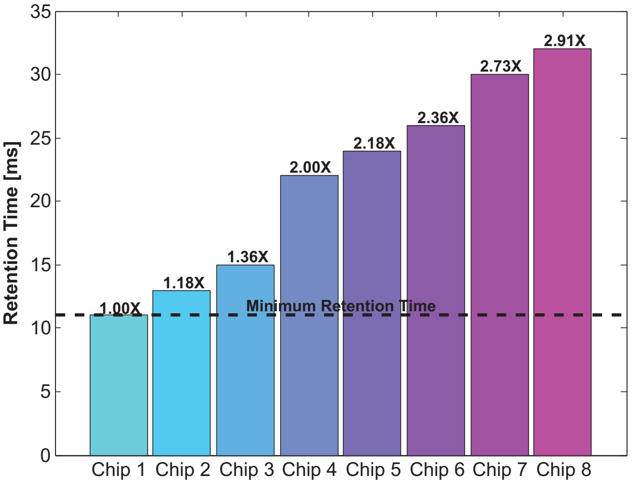


Fig. 4. DRT of 8 measured chips. Text above each bar shows the power wasted by each chip due to setting T_{ref} according to the minimum DRT (in this case, Chip 1) of the entire lot.

with T_{ref} representing the refresh period; E_{ref} representing the energy required to refresh the entire array; and P_{leak} representing the array leakage power, which is negligible, as compared to the refresh power [11]. In other words, the retention power of a GC-eDRAM array is inversely proportional to T_{ref}. Therefore, as shown in Fig. 4, by setting the refresh rate according to the worst cell in the entire lot of chips, a large potential for power savings is wasted. In the example of the eight measured chips, the T_{ref} would be set to 11 ms, according to the DRT of chip 1. However, if the DRT of each chip was measured separately and T_{ref} was configured independently, as much as 65% of P_{ret} could be saved (chip 8) with an average of almost 50% across the entire lot.

IV. Exploiting Retention Time Statistics for Power Savings

A. Operation with non-zero error probability

The experimental results discussed in the previous section show that by setting the refresh rate according to an overall worst-case DRT, a large potential for power savings is wasted. However, it may not be feasible in all cases to find the DRT of the worst cell of each fabricated array, and accordingly

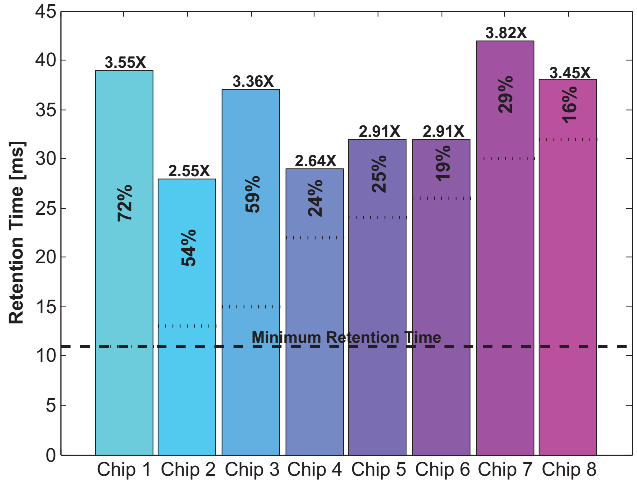
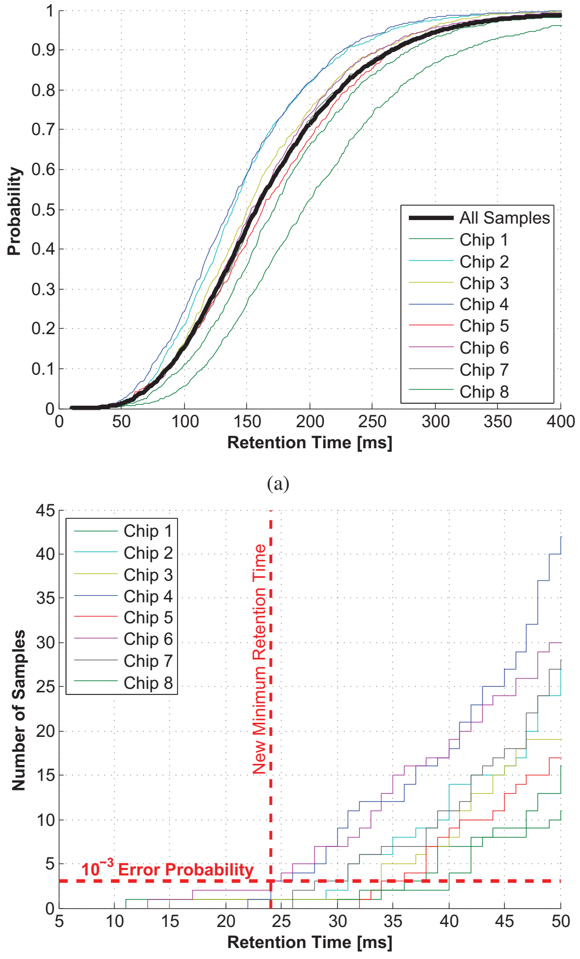


Fig. 5. DRT of 8 measured chips, assuming five errors can be tolerated. Text above each bar shows the power savings vs. setting T_{ref} according to the minimum DRT of the entire lot. The dotted line on each bar shows the zero-error DRT, and the rotated text shows the additional percentage of power savings for each chip by accepting up to five errors.

configure its T_{ref}, as this requires additional hardware and especially expensive test time. An alternative approach would be to use a different criterion for setting T_{ref}, in order to take advantage of the potential trade-offs between data integrity and refresh power enabled by the approximate computing paradigm. Of course, this will not be desirable in safety critical applications and memories used for storing control, which require 100% reliable data; however, it could be exploited in many applications that are known to be inherently error resilient, such as communications, multimedia, and data mining [7], [8]. In fact, several recent studies have shown that in such applications, memory failures in data buffers up to a certain number or percentage of failing cells will not lead to catastrophic failures of the overall system [15]–[17]. Therefore, we propose to depart from the traditional worst-case design paradigm, and instead exploit the error resilience of these applications. Rather than limiting ourselves to a particular application, we try to show the power benefits that can be achieved by extending the retention time, at the expense of an increased error rate. In particular, for an array of N cells, we set T_{ref}, such that it leads to a maximum number of bit failures k, which coincides with a maximum error rate of $P_{err} \leq k/N$. By doing so, a designer would be able to select the preferred T_{ref} that is suitable for the targeted worst-case application quality and power budget.

The motivation for the proposed method can be observed in Fig. 6(a), plotting the cumulative density function (CDF) of the DRT of the 8 measured test chips. It is clear from this figure that only a small percentage of bitcells suffer from low retention times, resulting in a relatively flat curve at the left part of the CDF. Therefore, by allowing a small number of potential failures, the DRT could be significantly increased, leading to considerable power savings. To emphasize this point, Fig. 6(b) provides an enlarged perspective of the per-chip CDF plots of Fig. 6(a). While a number of cells fail at refresh rates starting from 11 ms, no more than two cells per 2 kb array fail with T_{ref} as high as 24 ms. This is equivalent to an error probability of approximately 10^{-3}, which can be tolerated by various kernels, such as those used in communication systems, as reported in [15], [16], [18]. Operating with this maximum error probability results in less than half the retention power consumption during standby periods. This point is further elaborated upon in Fig. 7, showing the potential power savings, assuming a predefined maximum error probability across the entire lot of measured samples. For example, with an error tolerance of 10^{-3}, 55% of the power can be saved, or savings of 75% can be approached if an error rate of 10^{-2} can be tolerated. These above numbers are only an example to showcase the potential benefits, which will vary according to the target application.



(a)

(b)

Fig. 6. (a) Cumulative Density Function of DRT across measured chips. (b) Zoomed in view of CDF.

B. Data distribution across several banks

To further extend this concept, for a given memory comprising several banks or sub-arrays with different CDFs, data can be distributed between memory banks to exploit the power vs. error probability trade-off demonstrated above. A similar concept has been shown in recent studies for approximate off-

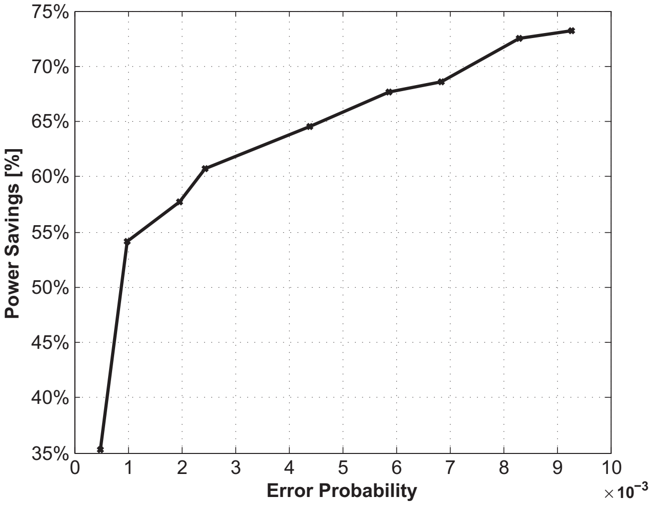


Fig. 7. Potential power savings, assuming a pre-defined tolerable maximum error rate (based on 8 measured chips).

chip DRAM [9], by mapping data to separate banks according to their error-tolerance.

Assuming a memory made up of three 2 kb sub-arrays with the DRT distributions of chips 1, 5, and 8, shown in Fig. 6, the T_{ref} of the sub-arrays could be independently set to 11, 24, and 32 milliseconds, respectively, for 100% error-free operation. This would already provide a 40% power reduction over traditional worst-case T_{ref} design; however, error-tolerant data offers much more power savings. For example, if the software was to distribute data to the three banks with 0.1%, 1%, and 0% error tolerances to the three sub-arrays, respectively, their T_{ref} could be extended to 32, 35, and 48 ms. Accordingly, this would result in a power reduction of 70%, as compared to the traditional, worst-case design, which would use a constant T_{ref} of 11 ms for all sub-arrays.

C. Yield-improvement through error tolerance

The previous sub-sections introduced the idea of exploiting the retention time statistics for lowering the refresh rate, leading to substantial retention power savings. A reciprocal approach is to use error tolerance in order to increase the fabrication yield, assuming the refresh rate is predetermined to meet a power specification.

The conventional, 100% defect-free requirement for array operation requires that a refresh operation is applied before the minimum DRT lapses. Assuming an independent distribution of the DRTs of all cells, the yield can be estimated as [18]:

$$Y(\text{zero error}) = \left(1 - \mathrm{P}(\mathrm{t_{ret}} < \mathrm{T_{ref}})\right)^{\mathrm{N}}, \quad (3)$$

with T_{ref} the predetermined refresh period, $P(t_{\text{ret}} < T_{\text{ref}})$ the probability that the refresh time is smaller than the refresh period, and N the number of bitcells in the array. Taking, as an example, the PDF of the measured cells (Fig. 3(b)) as the distribution function of all fabricated bitcells, and budgeting 1.2 nW of retention power per 2 kb sub-array, a fabrication yield of only 53% would be achieved. For tolerating up to K errors, the yield equation is expanded to:

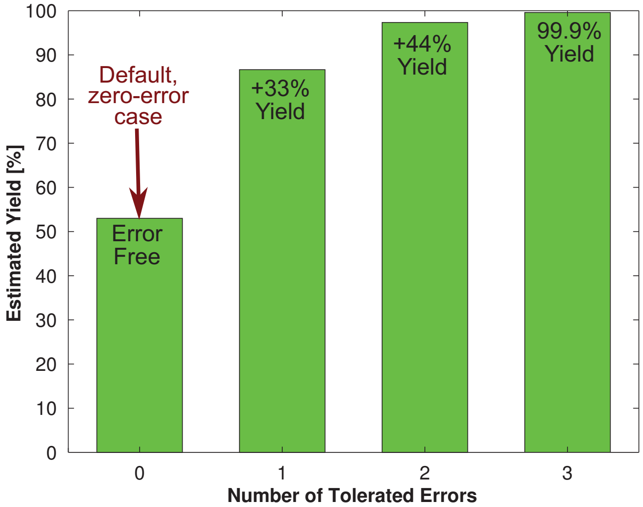


Fig. 8. Yield estimation of 2 kb sub-arrays with increasing error tolerances.

$$Y(K) = \sum_{k=0}^{K} \binom{N}{k} P(t_{\text{ret}} < T_{\text{ref}})^k \left(1 - P(t_{\text{ret}} < T_{\text{ref}})\right)^{N-k}. \quad (4)$$

Allowing only a single error per sub-array, the yield estimation increases to 86%, and a 99.99% yield can be achieved by tolerating up to six errors. This yield increase is illustrated in Fig. 8, for an error tolerance of up to six errors or a maximum error rate of 0.3%.

V. Curve Shaping for Additional Power Savings

The DRT distributions, shown in Fig. 6, show that the shape of the CDF is an important factor in achieving efficient power/error probability trade-offs, when following the methods presented in Section IV. In this section, we present one circuit level and one application level technique for improving this efficiency by preferably shaping these distributions.

A. Body biasing

The DRT of a GC-eDRAM bitcell is exponentially dependent on $V_{\text{T,MW}}$, as shown in (1). One well-known circuit technique for post-fabrication modification of this parameter is body biasing. The application of a reverse body bias raises the threshold voltage of an MOS transistor, thereby decreasing the sub-V_{T} leakage and extending the retention time. Of course, this will come at the expense of increased access time; however, this can often be acceptable during long retention periods [12].

Fig. 10 plots the measured retention time of a 2 kb GC-eDRAM array under various body bias conditions. As expected, the application of a higher the body voltage V_B (resulting in a higher RBB for the PMOS write transistor), results in increased retention time. However, in addition to this, the CDFs of Fig. 10 show that body biasing also flattens the curves, thereby allowing a much larger extension of T_{ref} for the same maximum number of errors. For example, with V_B=500 mV, allowing a $5{\cdot}10^{-3}$ error probability enables a 6 ms lower refresh rate over the zero-error constraint, whereas with V_B=750 mV, a 52 ms extension can be added.

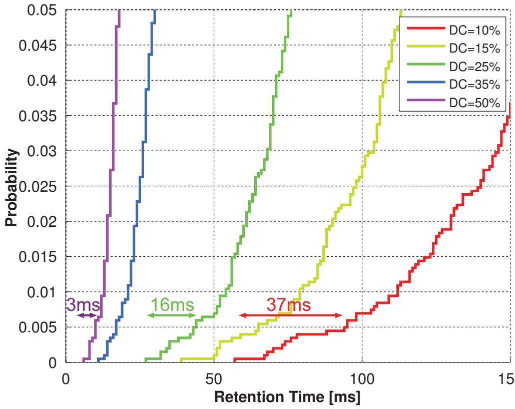


Fig. 9. Retention time CDF of the worst cell in a 2 kb array as a function of the access (write) statistics.

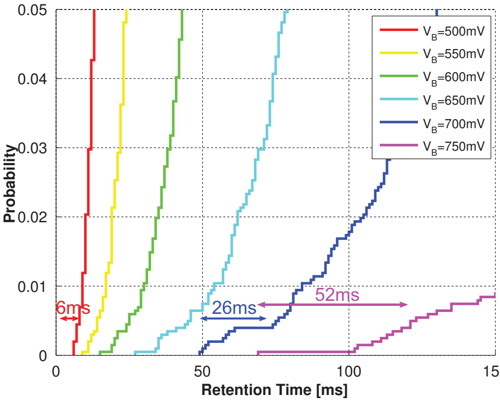


Fig. 10. Retention time CDF of the worst cell in a 2 kb array as a function of the body biasing level.

B. Write access statistics

Traditional, worst-case DRT design assumes that the WBL is constantly biased at V_{DD} to cause maximal sub-V_{T} leakage from the SNs when the weaker logic '0' level is stored. However, previous works [11] have proposed to minimize this level degradation by biasing WBL at GND during all standby cycles. In this case, the worst-case DRT is actually set by the access statistics, as the level degradation only occurs during a write '1' operation to a column with other bitcells storing a '0'. Therefore, taking the write access statistics into account at a software level, the refresh rate can be lowered for power savings.

The impact of the write access duty cycle (DC) on the DRT is plotted in Fig. 9 for a 2 kb GC-eDRAM array. As expected, when write accesses are initiated during a high percentage (*e.g.,* 50%) of the system operation, the retention time is much lower than when only few write operations (*e.g.,* 10%) are applied. However, similar to the case of body biasing, shown in Fig. 10, a lower write DC also results in curve flattening of the CDFs. For example, with 50% write cycles, allowing a $5{\cdot}10^{-3}$ error probability enables the increase of the refresh rate by 3 ms. But when the write DC is only 10%, a similar error tolerance allows as much as a 37 ms refresh rate relaxation.

VI. Conclusions

In this paper, we propose and demonstrate several techniques for exploiting the statistical characteristics of dynamic memories for achieving power savings and/or yield enhancement within the increasingly popular approximate computing paradigm. Our study was based on measurements of fabricated Gain Cell embedded DRAM chips, providing true retention time statistics, rather than estimated, high-level models. Our observations revealed the large spread in retention time distribution within an array and across chips, resulting in large refresh power overheads, when adhering to traditional worst-case design. We propose extending the refresh rate according to the retention time distribution to the error tolerance limits of error resilient applications, thereby achieving significant power savings while retaining sufficient data integrity, as required by such applications. Alternatively, we show that the gradual retention time degradation of these memories can be used to achieve higher yields, given a specified power and budget and maximum error rate. Finally, we propose utilizing circuit techniques and/or software level knowledge of access statistics, not only to save power through extension of worst-case data retention, but also to re-shape the underlying distributions to increase the efficiency of the proposed error tolerant methods.

References

[1] "ITRS - 2013 edition," 2013. [Online]. Available: http://www.itrs.net

[2] D. Somasekhar *et al.*, "2 GHz 2 Mb 2T gain cell memory macro with 128 GBytes/sec bandwidth in a 65 nm logic process technology," *JSSC*, vol. 44, no. 1, pp. 174–185, 2009.

[3] K. Chun *et al.*, "A 667 MHz logic-compatible embedded DRAM featuring an asymmetric 2T gain cell for high speed on-die caches," *JSSC*, 2012.

[4] Y. S. Park *et al.*, "Low-power high-throughput LDPC decoder using non-refresh embedded DRAM," *JSSC*, vol. 49, no. 3, pp. 783–794, 2014.

[5] A. Teman *et al.*, "Replica technique for adaptive refresh timing of gain-cell-embedded DRAM," *IEEE TCAS-II: Express Briefs*, vol. 61, no. 4, pp. 259–263, April 2014.

[6] P. Gupta *et al.*, "Underdesigned and opportunistic computing in presence of hardware variability," *IEEE TCAD*, vol. 32, no. 1, pp. 8–23, 2013.

[7] J. Henkel *et al.*, "Multi-layer dependability: From microarchitecture to application level," in *ACM-DAC '14*, 2014, pp. 1–6.

[8] S. Venkataramani *et al.*, "Quality programmable vector processors for approximate computing," in *IEEE/ACM ISM 2013*, 2013, pp. 1–12.

[9] S. Liu *et al.*, "Flikker: Saving DRAM refresh-power through critical data partitioning," *SIGPLAN Not.*, vol. 46, no. 3, pp. 213–224, Mar. 2011.

[10] A. Sampson *et al.*, "Approximate storage in solid-state memories," in *Proceedings IEEE/ACM ISM '13*, 2013, pp. 25–36.

[11] Y. Lee *et al.*, "A 5.42nW/kB retention power logic-compatible embedded DRAM with 2T dual-VT gain cell for low power sensing applicaions," in *Proc. IEEE A-SSCC*, 2010.

[12] P. Meinerzhagen *et al.*, "Impact of body biasing on the retention time of gain-cell memories," *IET JoE*, vol. 1, no. 1, 2013.

[13] L. Tran *et al.*, "Heterogeneous memory management for 3D-DRAM and external DRAM with QoS," in *ASP-DAC '13*, 2013, pp. 663–668.

[14] R. Giterman *et al.*, "4T gain-cell with internal-feedback for ultra-low retention power at scaled CMOS nodes," in *Proc. IEEE ISCAS 2014*.

[15] M. S. Khairy *et al.*, "Algorithms and architectures of energy-efficient error-resilient MIMO detectors for memory-dominated wireless communication systems," *IEEE TCAS-I*, vol. 61-I, no. 7, 2014.

[16] H. Cho *et al.*, "ERSA: error resilient system architecture for probabilistic applications," *IEEE TCAD*, vol. 31, no. 4, pp. 546–558, 2012.

[17] V. K. Chippa *et al.*, "Analysis and characterization of inherent application resilience for approximate computing," in *ACM-DAC '13*, 2013.

[18] G. Karakonstantis *et al.*, "On the exploitation of the inherent error resilience of wireless systems under unreliable silicon," in *ACM-DAC '12*, 2012, pp. 510–515.