Replica Technique for Adaptive Refresh Timing of Gain Cell embedded DRAM

Adam Teman, *Student Member, IEEE,* Pascal Meinerzhagen, *Student Member, IEEE,* Robert Giterman, Alexander Fish, *Member, IEEE,* and Andreas Burg *Member, IEEE*

***Abstract*—Gain cells have recently been shown to be a viable alternative to SRAM in low-power applications due to their low leakage currents and high density. The primary component of power consumption in these arrays is the dynamic power consumed during periodic refresh operations. Refresh timing is traditionally set according to a worst-case evaluation of retention time, under extreme process variations and worst-case access statistics, leading to frequent, power hungry refresh cycles. In this paper, we present a replica technique for automatically tracking the retention time of a gain cell embedded DRAM macrocell according to process variations and operating statistics, thereby reducing the data retention power of the array. A 2 kb array was designed and fabricated in a mature 0.18µm CMOS process, appropriate for integration in ultra-low power applications, such as biomedical sensors. Measurements show efficient retention time tracking across a range of supply voltages and access statistics, lowering the refresh frequency by more than 5×, as compared to traditional worst-case design.**

I. Introduction

Recent publications have shown a renewed interest in gain cell based embedded DRAM (GC-eDRAM) as a viable high-density alternative to SRAM [1]–[7]. The majority of the target applications for such memories have been large caches that benefit from the reduced transistor count of gain cells (GCs) [1], [2], [8]. However, several groups have shown the potential for integration of GC-eDRAM as a candidate for the replacement of SRAM in ultra-low power (ULP) applications [3]–[7], [9]. These studies have shown that the standby power of a GC-eDRAM array can be comparable to, or even lower than that of a similar capacity SRAM, while providing additional benefits, such as smaller area, robust low-voltage operation, and zero overhead concurrent read and write access [5]. While the standby power of an SRAM array is entirely comprised of leakage currents, GC-eDRAMs require periodic, power consuming refresh cycles to retain data. This refresh power tends to overtake the inherently low leakage power of GC-eDRAMs, such that the ability to achieve low power operation is directly proportional to the data retention time (DRT) of the GC array [5].

Manuscript submitted April 11, 2013. Manuscript revised Dec. 27, 2013. Manuscript accepted Jan. 12, 2014. This work was supported by the Swiss National Science Foundation under project number PP002-119057. A. Teman is supported by a Swiss Government Excellence Scholarship. P. Meinerzhagen is supported by an Intel Ph.D fellowship.

A. Teman, P. Meinerzhagen, and A.Burg are with the Telecommunications Circuits Laboratory (TCL) of the Institute of Electrical Engineering, EPFL, Lausanne, VD, 1015 Switzerland (tel.: +41-21-69-31027; Fax: +41-21-69-32687; e-mail: adam.teman@epfl.ch, pascal.meinerzhagen@epfl.ch, andreas.burg@epfl.ch)

A. Teman and R. Giterman are with the VLSI Systems Center, Ben-Gurion University of the Negev, P.O. Box 653, Beer Sheva, 84105.

A. Fish is with the Faculty of Engineering, Bar-Ilan University, Ramat Gan (e-mail: alexander.fish@biu.ac.il).

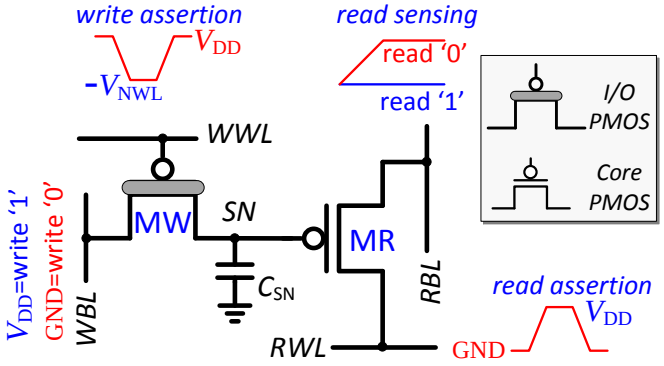

Fig. 1. Schematic of the all-PMOS 2T gain cell with I/O write transistor (MW), including biasing levels for access operations.

In contrast to 1-Transistor 1-Capacitor (1T-1C) eDRAM solutions that require special process steps to implement large in-cell storage capacitors, GC-eDRAM is logic compatible and relies on parasitic (device) capacitances for charge storage. The authors of [3] and [4] have shown that the use of a high-threshold (HVT) or I/O transistor is the best choice for the implementation of a GC's write transistor. Retention times of several milliseconds have been shown in mature, 0.18µm CMOS technologies, appropriate for the fabrication of ULP applications, such as biomedical sensors [3], [7]. However, these DRTs are generally measured for extreme worst-case conditions, not only in terms of Process-Voltage-Temperature (PVT) variations, but also in terms of operational behavior. For example, in the case of the all-PMOS 2T bitcell, illustrated in Fig. 1, data '0' exhibits a significantly reduced retention time, as compared to data '1', and the degradation of the data '0' level is expedited by driving the write bitline (WBL) to the supply voltage (V_{DD}). Therefore, DRT is simulated under these conditions, without taking into account the actual levels of WBL during operation. As a result, when operating on real data, refresh cycles will be initiated well before the actual stored data approaches hazardous levels.

In this paper, we propose a replica technique for tracking both the PVT variations of a GC-eDRAM macrocell, as well as the acute characteristics of the data access to adaptively time refresh operations. To demonstrate the compatibility of this technique with ultra-low power applications, such as biomedical sensors where small storage macros are often needed, the proposed technique was integrated with a 2 kb

GC-eDRAM array, and fabricated in a mature 0.18µm CMOS technology. The replica technique was found to efficiently track the DRT of the array, providing more than a 5× reduction in the frequency of refresh cycles. Section II of this paper presents the concept of the replica technique and its integration into a functional test chip, along with a brief description of the control architecture and logic implemented for post-silicon measurements. A subset of measurement results are presented in Section II-D for proof of concept, and Section IV concludes the paper.

II. Replica Technique for Auto-Refresh Timing

A. Retention Time of a 2T Gain Cell

The all-PMOS 2T GC circuit (Fig. 1) comprises a write transistor (MW), a read transistor (MR), and a storage capacitor (C_{SN}), made up of the parasitic capacitances of the connected devices and interconnect. Data is written to the cell by applying an underdrive voltage (V_{NWL}) to the write wordline (WWL) that transfers the biasing level of WBL to the storage node (SN). This level can be read out by pre-discharging the read bitline (RBL) and subsequently raising the read wordline (RWL), conditionally charging RBL if the voltage level stored on SN is low. The circuit's leakage power, shown to be dominated by subthreshold conduction at sub-micron process technologies [4], is extremely low, since during standby and write, the drain-to-source voltage (V_{DS}) of MR is zero, and the subthreshold leakage through MW is limited to (dis)charging the storage capacity of SN. The obvious issue is that any leakage to or from SN results in a degradation of the stored data level, requiring periodic refresh cycles. Therefore, the standby, or *retention* power of a GC-eDRAM is given by (1):

$$P_{\mathrm{retention}} = P_{\mathrm{leakage}} + P_{\mathrm{refresh}} = V_{\mathrm{DD}} I_{\mathrm{leak}} + \frac{E_{\mathrm{refresh}}}{t_{\mathrm{refresh}}} \quad (1)$$

where I_{leak} is the standby leakage current, E_{refresh} is the energy required to refresh the entire array, and t_{refresh} is the time between refresh operations. Clearly, in order to minimize the retention power, t_{refresh} must be maximized; however, in order to ensure data integrity, this parameter must be set lower than the estimated DRT. Therefore, an accurate estimation of DRT is required to achieve low power operation.

Various metrics have been used for simulating the DRT of a bitcell [3], [5], [6], but the unequivocal definition of this important parameter is the time at which the voltage written to C_{SN} degrades to the point where it results in an incorrect readout. This time is set by four primary factors: the initial level stored on C_{SN} following a write, the size of C_{SN}, the leakage currents to and from SN, and the readout mechanism. All of these factors are significantly affected by both environmental and manufacturing variations, as demonstrated in measurements by [3]. This results in a large spread of DRT distribution [2], [7], and as with any memory array, necessitates design for the worst cell. However, in addition to the effects of PVT variations, SN leakage currents are highly sensitive to the biasing level of WBL. For a stored '1', the highest discharge leakage occurs when WBL is low, while the worst case for a stored '0' occurs when WBL is high. As shown in [3]–[5], the worst-case biasing of a stored '0' exhibits an orders-of-magnitude lower DRT than that of a '1' for an all-PMOS cell. Consequently, DRT is calculated assuming that WBL is constantly held high. However, this situation would only occur if a *write* '1' operation was executed on a given column during *every clock cycle*, leading to early, power consuming refresh operations in any typical scenario.

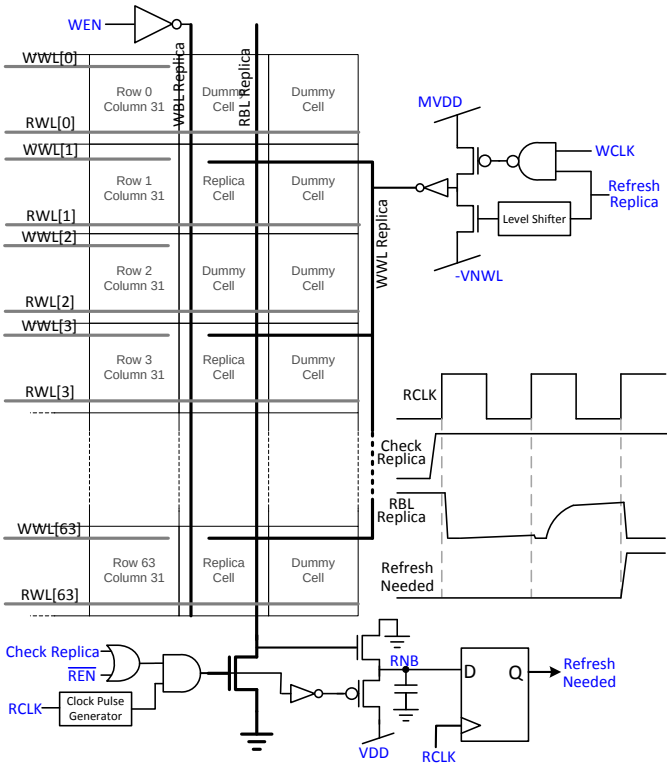

Fig. 2. Schematic illustration of the read and write circuitry for operation and control of the proposed replica technique, including timing diagrams.

B. Replica Technique Concept

The design for worst-case conditions, coupled with the wide spread of DRT due to PVT variations and access statistics, almost always results in the initiation of refresh cycles when the stored data is still at strong levels. By implementing a replica technique to track the global parametric variations, environmental conditions, and acute operating characteristics, a significant amount of this overhead power can be saved. An additional post-silicon calibration step is added to adjust the tracking mechanism for each die to handle local variations.

The foundation of the proposed technique relies on the superiority of the DRT of the data '1' state of the all-PMOS bitcell. This superiority is due to a number of factors, starting with the PMOS write-transistor that easily passes a high-level to SN, as opposed to a low-level, which requires WWL underdrive to completely discharge C_{SN} in a reasonable amount of time. Subsequently, both charge injection from MW and the coupling capacitance between WWL and SN drive charge onto C_{SN} during the rising edge of WWL, causing a slight voltage rise on SN, resulting in a degraded initial '0' state and an overcharged initial '1' state. Furthermore, the subthreshold leakage to/from C_{SN} is exponentially dependent on the overdrive of MW, which is constant for the worst-case

of ‘0’ storage (WBL=V_{DD}) but self limiting in the case of ‘1’ degradation (WBL=0, $V_{\mathrm{SG}} = V_{\mathrm{SN}} - V_{\mathrm{DD}}$). A full description of these processes is presented in [4].

Two primary mechanisms are incorporated to simultaneously extend the DRT of the entire array, while maintaining data stability. First, during all non-write cycles, WBL is driven low, thereby enhancing the level of a stored ‘0’ bit while minimally affecting the level of a stored ‘1’. Second, a column of replica cells are integrated with the GC-eDRAM array and are periodically read out to analyze the state of the array’s data retention. These replica cells are standard all-PMOS 2T bitcells, designed with slightly reduced C_{SN} (less metal stacking above the bitcells). With proper post-silicon calibration of the write-frequency to these replica cells, they can be adjusted to fail before the worst data cell in the array, while tracking the PVT variations of the fabricated array. In addition, the replica column is designed to track the access statistics of the array, rather than assuming unlikely worst-case conditions (i.e., write operations during every clock cycle). Immediately prior to an array refresh, data ‘0’ is written to all of the replica cells, and during read and standby cycles, the WBL of the replica column is driven low, as is applied to the WBL of the data columns. Significant data level degradation only occurs when the WBL is high, which can only happen to a cell storing a ‘0’ when a ‘1’ is written to a cell on the same column. Therefore, during write cycles, the WBL of the replica column is driven high, thereby applying worst-case conditions only when they can actually occur. In this way, the retention time of the replica cells is always slightly worse than a data cell on a column with cells that were repeatedly written as ‘1’ over the retention period. However, instead of assuming an extreme worst case of *tying* WBL to ‘1’ (which would only occur if the array was written to during every operating cycle), this setup tracks the actual frequency of write operations. Therefore, the replica cells track the access statistics of the array (i.e., the relationship between non-write and write operations), while still ensuring that the replica cells will fail before the real data is lost.

As mentioned above, the DRT of a 2T Gain-Cell suffers from significant variations, both global and local, as has been well documented and demonstrated in previous publications [2]–[4], including recent measurements by the authors [7]. However, the replica mechanisms described above are only designed to track the global variations and access statistics of the array. Therefore, several guardbands have been incorporated into the memory macro in-order to further combat the local variations that may impede the overall array DRT. First, while the layout of the replica cells is identical to the core array bitcells, a layer of metal stack has been removed in order to reduce their C_{SN} and ensure a slightly lower DRT than the data bits. Second, a post-silicon calibration technique has been incorporated into the refresh controller to skew the DRT below the measured worst-case DRT of the array. This is done by employing periodic *pseudo-write* cycles to the replica column. During these operations, the WBL of the replica column is charged, causing the replica cells to degrade at a slightly higher rate than dictated by the write statistics, thereby ensuring the initiation of an array refresh prior to data-loss in the worst cell of the array.[1]

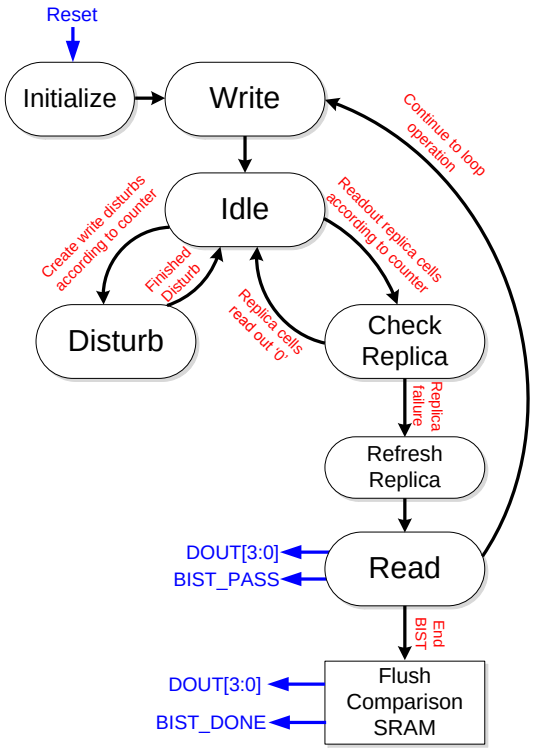

Fig. 3. State machine of the test controller.

C. Replica Technique Integration into Gain Cell Array

The proposed replica technique was integrated into a 2 kb all-PMOS GC array in an 0.18µm CMOS technology according to the schematic illustration in Fig. 2. A total of 32 replica cells were placed in an additional column to deal with the large distribution of local variations [2]. These cells are driven low upon the assertion of the external *RefreshReplica* signal within a single cycle, independent of the operation of the rest of the array. The same write mechanism as used for the data bit WWL drivers is incorporated for driving the negative write voltage to the replica cells. In order to track the write statistics of the array, the WBL of the replica column is tied to the write enable (WEN) signal.

Readout of the replica cells is achieved through a mechanism similar to the readout of the data cells with the addition of a designated *CheckReplica* signal. As the replica cells were designed to fail due to the deterioration of a stored ‘0’ level, reading out a ‘1’ from the replica column indicates the need for a refresh cycle. Therefore, the readout level is propagated to the control blocks as the *RefreshNeeded* signal.

[1]We note that in the extremely unlikely case of a continuous write ‘1’ operation applied to a column with a bitcell with a DRT that is worse than the worst replica cell, this calibration would be insufficient. However, this scenario would hardly ever occur in any application known to the authors. If such a case is relevant, it is possible to impose a write access policy to the array, which, for example, permits writing to the array only every second clock cycle. In addition, to avoid such a write access policy, it is possible to limit the WBL pulse time for the storage array, while using a pulse width that is equal to a full clock cycle for the replica columns.

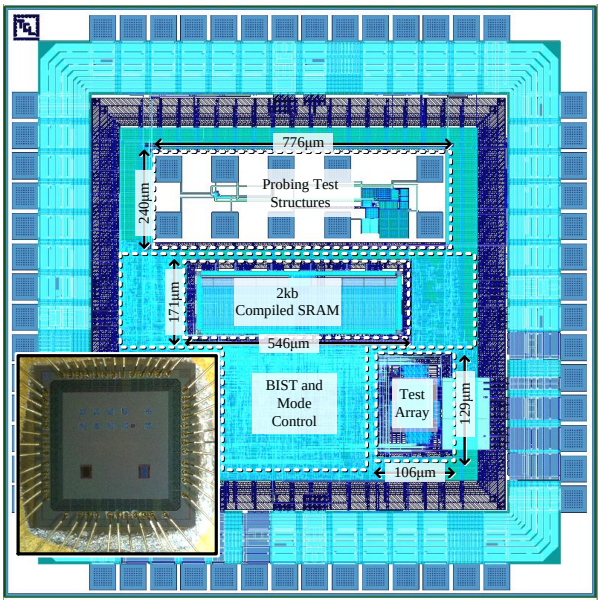

Fig. 4. Full layout of the GC-eDRAM test chip with major components.

D. Testing and Characterization Procedure

Testing and characterization of the replica technique was implemented with an on-chip controller, incorporating the finite-state machine (FSM) illustrated in Fig. 3. This controller initially writes data to the entire array, and subsequently proceeds into an *Idle* (standby) state for a configurable time period. In *Idle*, the controller performs one of two operations. To measure tracking of write statistics, the controller initiates periodic *Disturb* cycles, during which a row of '1's (0xFFFFFFFF) is written to a predetermined "victim" address[2]. This operation drives the WBL of all columns high, thereby causing deterioration of stored '0' bits in the entire array. A similar mechanism is incorporated through a post-silicon calibration to further deteriorate the replica cells in order to account for local variations that may further skew the worst cell in the array below the replica cells.

The second operation is the *CheckReplica* state, during which the 32 replica cells are serially read out to determine the onset of a refresh operation. If the *RefreshNeeded* signal is asserted (i.e., the data in one of the replica cells reads out erroneously), the controller proceeds to refreshing the replica cells, before refreshing the actual storage array and looping back to the *Idle* state for an additional retention period.

The *Read* state (part of the array refresh sequence) of the test controller provides important measurement data for analysis. The read-out data is compared with the originally written data to ensure equality and the per-bit comparison results are stored in an on-chip 2 kb SRAM. Concurrently, the one-bit comparison result of the current row data is driven off-chip via the *BIST_PASS* signal, and the four MSB bits are propagated to the external *DOUT*[3:0] pads to enable further observation. An external interrupt signal can break the refresh loop, sending the controller into its termination state, during which the comparison data can be analyzed. In this state, the

[2]The victim address will always store 0xFFFFFFFF, and therefore is not considered for comparison with expected responses.

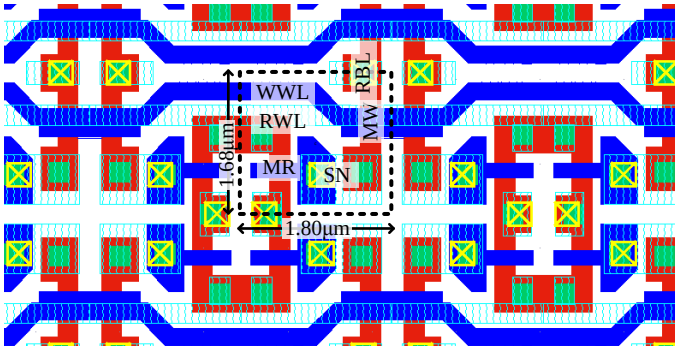

Fig. 5. Small section of the GC-eDRAM array layout showing the dimensions of the unit cell.

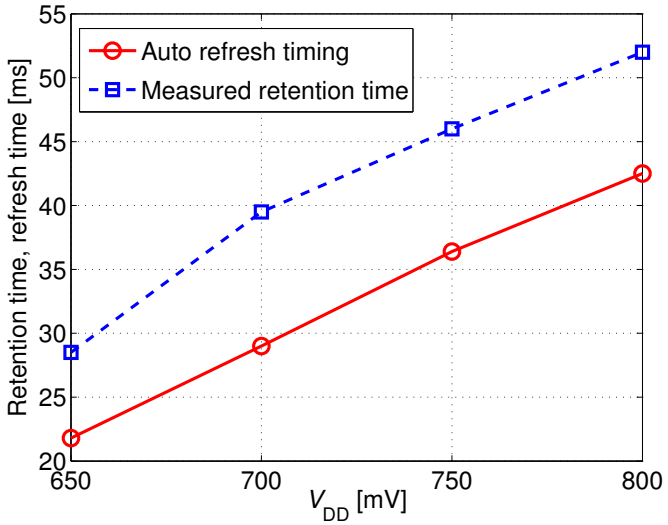

Fig. 6. Automatic refresh timing vs. measured DRT for a range of supply voltages.

BIST_DONE signal is raised, and subsequently, the full, per-bit comparison data that was stored in the SRAM is flushed out to the *DOUT*[3:0] pads by means of scan chains.

This control scheme enables at-speed testing of the GC-eDRAM array, including the ability to observe the functionality of the replica technique under various write disturb statistics. However, additional bypass schemes were implemented to enable further measurement control, as discussed in Section II-D.

III. Implementation and Measurement Results

A 2 kb (64×32) GC-eDRAM array with integrated replica technique was designed and fabricated in a commercial 0.18µm CMOS technology, as part of the test chip shown in Fig. 4. In addition to the array, the test chip included the on-chip test controller, the 2 kb SRAM for data comparison, and several other test components. The test chip was designed to enable three primary test modes: full, at-speed, controller testing; array operation through scan chain configuration; and external direct access to the array. A combination of these three modes was used to test the functionality of the array and produce the measurement data shown below.

The GC-eDRAM bitcell was laid out in a compact array with mirrored rows and columns, as shown in Fig. 5,

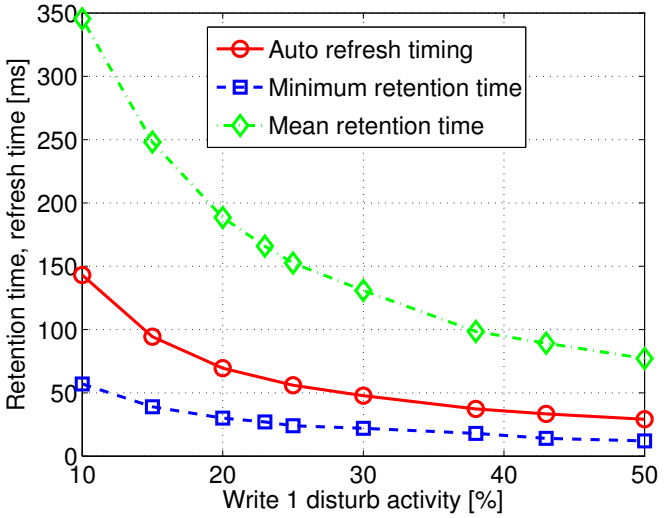

Fig. 7. Automatic refresh timing vs. measured DRT for a varying degree of write disturbs.

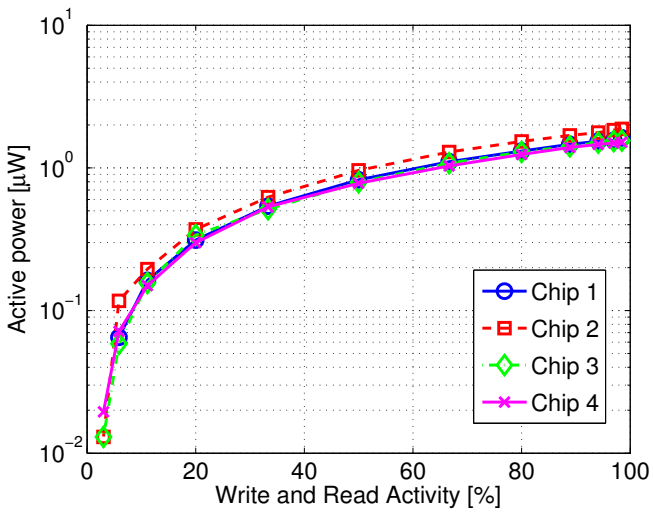

Fig. 8. Dynamic power consumption of 2 kb GC-eDRAM array as a function of the write and read activity factor for several measured chips.

with a unit cell size of 3.024 µm^2 (1.8 µm×1.68 µm). The array, including peripheral circuits, occupies 0.013 mm^2 (106 µm×129 µm) and is biased by a separate, low-voltage supply (MVDD) different than the supply (VDD) of the BIST and the other digital peripheral circuits of the test chip. In addition, an external negative voltage (VNWL) is supplied for write underdrive. The addition of the replica columns led to an overall area overhead of only 6.25%, without significantly affecting the read and write access times. All measurements were carried out at room temperature, which is appropriate for most ultra-low power applications.

Figure 6 illustrates the ability of the replica technique to automatically track the DRT of the array. The figure shows the automatically triggered refresh period for various supply voltages, as compared to the minimum DRT measured at this voltage, following a post-silicon adjustment in the write disturb frequency to account for local variations. Refresh is consistently initiated just prior to the array's minimum DRT for a range of supply voltages.

Tracking of the write statistics is shown in Fig. 7. This figure plots the automatic refresh timing, as compared to the measured DRT of the array with a given frequency of write operations. The figure shows both the mean and minimum DRT of the array. For the shown die, the minimum DRT is lower than the uncalibrated automatic refresh timing; however, the write activity tracking mechanism is shown to work correctly, such that the post-silicon calibration can easily skew the refresh time below the worst-case failure. This plot emphasizes the efficiency of integrating the replica technique. Traditional, worst-case design assumes 100% write activity, resulting in a retention period of well below 10 ms, even for this typical die. Application of the replica technique adapts this period, refreshing at a more the 5× lower frequency with 10% write activity.

Fig. 8 plots the dynamic power consumption of the array, as a function of the write and read activity. For a retention period of 20 ms, the active refresh power of the array is 635 fW/bit, which is comparable with similar low-power GC-eDRAM publications [3].

IV. Conclusion

In this work, we proposed a replica technique for tracking the PVT variations and operating statistics of a GC-eDRAM array for efficient DRT extension. The technique was implemented on a 2 kb all-PMOS 2T array in a commercial 0.18µm CMOS process along with an advanced control scheme for extensive testing and measurement. The replica technique was shown to effectively track the retention time of the array across various supply voltages and write activity frequencies, enabling as much as a 5× improvement in DRT, thereby significantly reducing the frequency of power hungry refresh operations.

References

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

[2] K. C. Chun *et al.*, "A 2T1C embedded DRAM macro with no boosted supplies featuring a 7T SRAM based repair and a cell storage monitor," *IEEE JSSC*, vol. 47, no. 10, pp. 2517–2526, 2012.

[3] 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.

[4] P. Meinerzhagen, A. Teman, R. Giterman, A. Burg, and A. Fish, "Exploration of sub-VT and near-VT 2T gain-cell memories for ultra-low power applications under technology scaling," *MDPI Journal of Low Power Electronics and Applications*, vol. 3, no. 2, pp. 54–72, 2013.

[5] A. Teman, P. Meinerzhagen, A. Burg, and A. Fish, "Review and classification of gain cell eDRAM implementations," in *Proc. IEEE IEEEI*, 2012.

[6] P. Meinerzhagen, A. Teman, A. Mordakhay, A. Burg, and A. Fish, "A sub-VT 2T gain-cell memory for biomedical applications," in *Proc. IEEE Sub-VT Microelectronics Conference*, 2012.

[7] P. Meinerzhagen, A. Teman, A. Fish, and A. Burg, "Impact of body biasing on the retention time of gain-cell memories," *The IET Journal of Engineering*, vol. 1, no. 1, 2013.

[8] X. Liang *et al.*, "Replacing 6T SRAMs with 3T1D DRAMs in the L1 data cache to combat process variability," *Micro, IEEE*, vol. 28, no. 1, pp. 60–68, 2008.

[9] K. C. Chun *et al.*, "A 0.9 V, 65nm logic-compatible embedded DRAM with 1ms data retention time and 53% less static power than a power-gated SRAM," in *Proc. ACM/IEEE ISLPED*. ACM, 2009, pp. 119–120.