Aging-Compromised Computing-In-Memory Dot-Product Calculation Technique Through DVFS

[1]Yu-Guang Chen, [1]Chi-Hsu Wang, , [2]Ing-Chao Lin
[1]Dept. of EE, National Central University, Taoyuan, Taiwan
[2]Dept. of CS, National Cheng Kung University, Tainan, Taiwan
andyygchen@ee.ncu.edu.tw, wch861997@gmail.com, iclin@mail.ncku.edu.tw

ABSTRACT

Von Neumann Architecture which separates the computing logic and the storage area has been considered as the fundamental architecture of nearly all digital computers nowadays. The data-intensive applications such as image recognition or cryptography may transfer large amount of data between memory and the computing cores, which causes a well-known von Neumann bottleneck due to the limitation of communication bandwidth. Computing In-Memory (CIM), which directly performs in-situ operations at memory, has been considered as one of the promising solutions to overcome von Neumann bottleneck. Previous researchers have proposed an 8T-SRAM-based CIM architecture to perform multi-bit dot product computations by analog charging/discharging operations. However, such operations are very sensitive to variations as well as aging effects such as Bias Temperature Instability (BTI) and/or Hot Carrier Injection (HCI). To provide a reliable CIM multi-bit dot product engine, in this paper we propose an aging-aware computing in-memory framework which consists of an aging detection method and an aging tolerance technique. Specifically, we apply Dynamic Voltage Frequency Scaling (DVFS) on CIM structure to compensate the current drop due to variations and aging effects. Experimental results show that we can double the lifetime of CIM structure with 1.185x extra power consumption in average.

Keywords—Computing-In-Memory, PBTI, HCI, DVFS, 8T SRAM

1. INTRODUCTION

Von Neumann Architecture (VNA) has been considered as the fundamental of nearly all digital computers nowadays. With the concept of VNA, the processing unit and storage unit are separated in stored-program system which can provide the flexibility to change the control sequence by simply modifying the program. However, the major limitation of VNA, also known as von Neumann bottleneck, comes from the speed limitation of retrieving instructions and data from storage unit to processing unit. The limitation becomes non-neglectable at present since various data-intensive applications such as motion detection and pattern recognition have been widely adopted in our daily life. Moreover, with the rapid developing of machine learning related algorithms, the huge data transferring between processing unit and storage unit will not only slowdown the system performance but also consume more power. As shown in [1], 62.7% of energy is consumed on data movement.

To overcome the von Neumann bottleneck, researchers have studied for over a decade to seek more efficient computing architecture and techniques. Computing In-Memory (CIM), which processes operations at memory cell in-situ, has been considered as one of the promising solutions with very limited area and power overhead. Various proposals for CIM architecture with various technologies have been proposed. Amount them, SRAM-based CIM architecture is considered as a practical solution with respected to CMOS technology. Therefore, in this paper, we focus on 8T SRAM CIM structure for Dot Product (DP) operation proposed by [2] and is shown in Fig. 1(a). A 8T SRAM cell, which serves as a basic component in the structure, is shown in Fig. 1(b). By separating write bit-line and read bit-line, the read disturbance problem [3] can be avoided and the read/write operations can be performed more reliable. We use Fig. 1(c) to demonstrate how the DP operation A • B is performed where A is a 1x2 matrix with binary values and B is a 2x1 matrix. The two elements of A are stored in cell 1 and cell 2. Then the two elements of B is appropriately converted into analog signals V_1 and V_2 respectively. When the A • B is performed, RWL_1 (Read Word Line) and RWL_2 are triggered and the current will flow from V_i to RBL (Read Bit Line) if the $Cell_i$ stores logic 1 where i={1, 2}. Then the accumulated current on RBL, I_{RBL}, is sent to a sensing circuit follows by a Analog to Digital Converter (ADC) to obtain the computing result. More details will be provided in section 2.1. With such a structure, DP operations can be accomplished just inside the memory and the overhead of data movement can be significantly reduced.

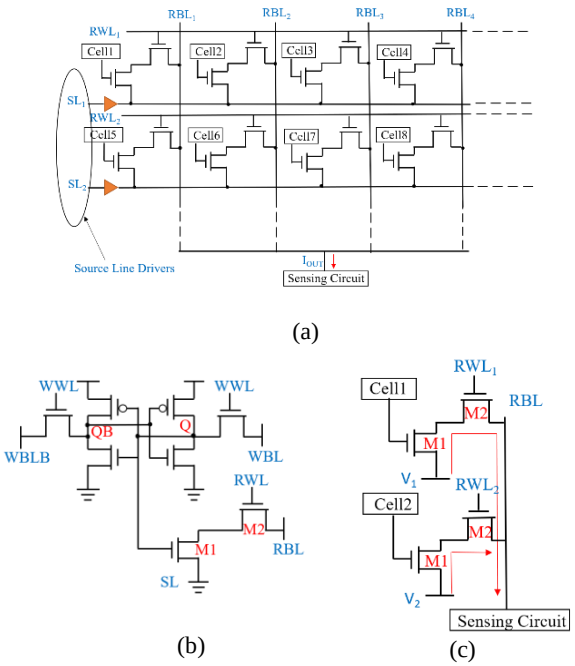

Fig. 1 (a) A 8T SRAM multi-bit DP engine; (b) A 8T SRAM cell; and (c) An example of DP operation with single column and two rows.

With the advancement of CMOS technologies, reliability is also an non-negligible concern. Aging effects, especially Bias Temperature Instability (BTI) and Hot Carrier Injection (HCI), are both the most serious threats on nMOS in advanced CMOS technologies. The two aging effects increase the threshold voltage of nMOS (in absolute value) with time when transistors are stressed. The increased threshold voltage on M1 in Fig. 1(c) will degrade the I_{RBL} during the CIM DP operation and leads to unreliable results. Therefore, a reliable 8T SRAM CIM operation technique is in need.

In this paper, we propose an aging detection and aging tolerance framework to strike the aging effects on 8T SRAM CIM engine. The major idea is to adjust the operating voltage during the computation so the degraded I_{RBL} due to the aging effects can be compromised and the result of ADC could be accurate. Specifically, we address the aging effect on M1 in Fig. 1 (b) since the stored data may remain the same for a long time and lead to long period stress. When M1 is aged, the operating voltage is increased by Dynamic Voltage Frequency Scaling (DVFS) technique so I_{RBL} can maintain the integrity as before aging. Although the idea of applying DVFS on aging tolerance is not an novel idea, it is challenging when we apply DVFS on 8T SRAM CIM structure. Firstly, an aging detection method is in demand to identify the aged row(s) in the CIM computing unit. Second, we need an accurate voltage regulator to perform fine-grained DVFS without causing troubles on other peripheral circuits. Third, we need to minimize the extra power consumption overhead while maintaining

similar performance with the proposed aging tolerance technique. In this paper, we attempt to overcome the above challenges and propose an aging-aware 8T SRAM CIM technique.

The major contributions of this paper are as follows:

1. In this paper we propose an aging-aware technique for 8T SRAM CIM, which consists of an aging detection method and aging tolerance CIM operation method.
2. In our aging detection method, we identify rows with aged nMOS(s) during the testing period and recorded the aged-rows.
3. In our aging tolerance CIM operation method, we firstly identify whether the rows stored the operands are aged rows. If the operands are in aged rows, we then apply DVFS during the computation period to compensate I_{RBL} degradation so the ADC output can be accurate.
4. To the best of the authors' knowledge, this is the first work to address the DVFS on SRAM CIM structure for aging tolerance. Experimental results also show that our framework can not only successfully overcome the aging-induced threats but also can double the system lifetime with only 1.18x power consumption in average.

The rest of this paper is organized as follows: Section 2 details the 8T SRAM CIM operations and potential threats from PBTI and HCI. Section 3 elaborates the problem formulation and Section 4 details the proposed framework. Section 5 provides the experimental results, and the conclusions are given in Section 6.

2. PRELIMINARIES

In this section, we first review the 8T SRAM CIM structure and DP operations we target, and then discuss aging effects and aging models we used in this paper. Finally, we recall the aging estimation method under different operating voltages.

2.1. 8T SRAM Computing In-Memory

Computing in-memory is one of widely used approaches to avoid large amount of data transfer between memory and computing units. As shown in [4], the standard 6T-SRAM structure can perform computing in-memory well. However, the 6T-SRAM structure will suffer from the read-disturb problem due to its coupling read-write paths and may lead to inaccurate results. Therefore, in this paper we target the 8T SRAM CIM structure as shown in Fig. 1(a) which adds two extra transistors as shown in Fig. 1(b) to separate RBL and WBL (Write Bit Line) [2].

The 8T SRAM provides two basic operations, read and write. By adding peripheral circuits, the DP computations can be performed just like read operation. The DP operations consists of multiplications and an addition to sum the result of the multiplications. In CIM structure, each multiplicand is stored in the memory cell with the same column and each multiplier is translated into analog voltage through DAC (Digital to Analog Converter) and is sent through Source Line (SL) as V_i shown in Fig 1(c) respectively. When the computation is performed, the control signal RWL_i of all corresponding rows are turned on, and transistor M2 of that row is turned on. If the stored value is logic "1", transistor M1 will also turn on and form a current path through V_i to RBL. The I_{RBL} is proportional to $V_i \cdot g$, where g is the series conductance of the transistors and it can be seen as a constant value. Otherwise, transistor M1 will be turned off and almost no current will pass through. Then all the current go through V_i to RBL is considered as "partial sum" of each multiplication. Since each column shares the same RBL, all the current will accumulate on RBL to perform "summation" result. Finally, a sensing circuit as shown in Fig. 2 is connected to RBL, which is used to readout the I_{RBL} and generate an appropriate output voltage V_o. The V_o is then sent to an ADC to decide the logic value of the DP computation. In the 8T SRAM CIM structure proposed in [2], each computing unit is form by four columns and can support 4-bit by 4-bit DP operations. Note that the transistor size of each column is well adjust to reflect the significance of each bit stored in the memory.

One thing we need to take care is the relations between V_i and I_{RBL}. Figure 2(b) shows the I_{RBL} value under different V_i with only 8T CIM cell shown in Figure 2(a) where V_{pos} is set to 0.1V. From the figure we can find that a linear relation of V_i and I_{RBL} is held in the linear region otherwise the I_{RBL} will increase slower than the V_i. It would be easier to design the ADC for converting final digital results if V_i and I_{RBL} are in linear relationship. Therefore, the range of V_i should be carefully set for different multipliers.

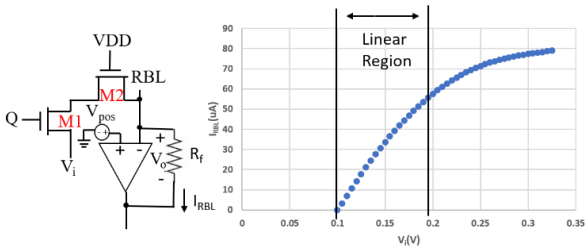

Fig. 2 (a) The sensing circuit structure; and (b)The I_{RBL} versus V_i

2.2. Aging Effect and Aging Model

Recently, Positive-Bias Temperature Instability (PBTI) and Hot Carrier Injection (HCI) effect are considered as huge threats on nMOS and attracts lots of attentions due to the technology scaling. The PBTI effect is caused by charge trapping. Defects generated during and after the gate oxide formation process acts as interface states. The defects might trap the carriers and the interface states will raise the threshold voltage of the transistor. The interface traps catch the electron when the positive bias voltages is applied to the gate of a nMOS transistor [5][6]. On the other hand, HCI is formed when carriers are accelerated by the lateral electric field along the channel. The accelerated carriers attain an ability to undergo impact ionization close to the drain side. Impact ionization leads to the formation of electron-hole pairs. If electrons do not have the sufficient energy for tunneling, they get trapped in the oxide and create interface states. Accumulation of trapped electrons raises the transistor threshold voltage [6]. Since both PBTI and HCI effects will increase the threshold voltage (in absolute value) of nMOS, decrease I_{RBL}, and may cause failure when we perform DP operation on CIM structure, it is important to analyze the influences of these aging effects and provide efficient tolerance and/or mitigation methods.

To analyze the impact of the PBTI and HCI effect on threshold voltage shifting, we need an aging model. In this paper we use HSPICE MOSFET Model Reliability Analysis (MOSRA) model [7] which takes both PBTI and HCI into consideration to perform aging simulation. Fig. 3 shows a long-term PBTI and HCI induced threshold voltage shifting on a nMOS with 90nm technology with 0.75V, 20˚C, and 100% stress period (i.e., the nMOS is always turned on) in grey line obtained from HSPICE MOSRA simulation. In this paper, we use HSPICE MOSRA to get threshold volatge shifting of the nMOS, and apply it to 8T SRAM structure to observe the current degradation.

2.3. Aging Estimation Under Different Operating Volatges

As mentioned above, in this paper we will adjust the operating voltage to tolerate the aging effects. However, aging model curve is related to the operating voltage. Therefore, for more precise and real situation, we need to estimate the aging model curves of different operating voltages. We apply the aging estimation method proposed in [8] which will adjust the aging period between different operating voltage. Fig. 3 shows the aging curves with different operating voltages in red line and black line. In our simulation, black line (V'') is the highest operating voltage, and the rise of threshold voltage is also the most serious. Through this method, we can make our aging model more practical.

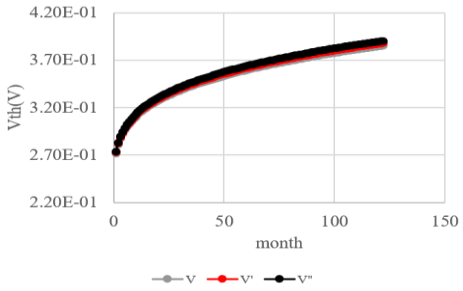

Fig. 3 Vth vs. time under aging effects

3. PROBLEM FORMULATION

In this section, we first list the assumptions we apply in this paper, and describe the aging-aware 8T SRAM CIM operation problem.

In this paper we only concentrate on the PBTI/HCI-induced reliability problems on 8T SRAM CIM structure. From HSPICE MOSRA simulation, we can assume that the size of each nMOS will not influence the stressed situation in the 8T SRAM structure. In this paper, the lifetime of the system we definite is that the any row of the CIM result is incorrect.

Here we formulate the aging-aware 8T SRAM CIM operation problem as follows: Given 8T SRAM with 256 rows, we want to identify the aged rows and the DP operations can be performed correctly after the circuit is aged so that the system lifetime can be maximized. Moreover, the extra energy consumption should be minimized.

4. FRAMEWORK

In this section, our aging-aware operation method is proposed to detail how we operate DP operation while considering aging effects. Besides, we describe the proposed aging detection method and aging tolerance method with DVFS to make the proposed method more practical.

4.1. Technique Overview

Because aging effects will impact the accuracy of CIM, if we still perform DP operation without any detection and tolerance method, it might lead us to get a wrong answer. Therefore, we propose an aging-aware operation method to address the problem. Fig. 4 shows the overview flowchart of our aging-aware 8T SRAM CIM operation method. Our operation method consists of testing phase and computing phase. During testing phase, we will use our aging detection method which will be discussed carefully in section 4.2 to distinguish whether the row is aged or not. During computing phase, the healthy row can directly perform the CIM to get the answer. On the other hand, the aged row needs to use our aging tolerance method first, then perform the CIM to get the answer. Through the proposed structure, we can distinguish the aged row and address it properly. Through our aging-aware operation method shown in 4.3, we can perform CIM operating with aging tolerance technique and maximize the lifetime and accuracy of the CIM structure.

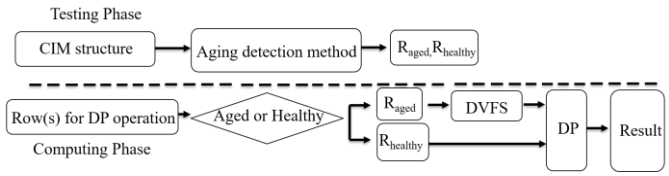

Fig. 4 Overview Flowchart of Framework

4.2. Aging Detection Method

To detect aged rows, we propose an aging detection method as shown in Fig. 5. During testing process, we firstly move the data to the datakeeper, and assert value 1 to the cells of the row. The testing row will then perform DP operation with only this row and V_i voltage is set to digital 15 because this value is impacted by aging effects most seriously. If this value will not affected by aging effects, we can assume other value will not be impacted, too. V_{ref} is got by V_i voltage is set to digital 15 and I_{RBL} is healthy situation. Then check the output value of the OPA Comparator. The principle of OPA Comparator is that V_{out} will generate a positive value when the positive node is larger than the negative node, as shown in Fig 6. On the other hand, if positive node is smaller than negative node, V_{out} will generate a negative value. Because V_o is related to I_{RBL} and a constant resistance as shown in Fig. 2(a). I_{RBL} will decrease due to the aging effects, and therefore V_o will decrease, too. Therefore, when the row is healthy, V_o is larger than V_{ref}. And V_{out} will generate a negative value. If the row is aged, V_o is smaller than V_{ref} due to the decreasing I_{RBL}. And V_{out} will generate a positive value. Through the detection method, we can know the health situation of each row. After we find out which row is aged, we need to record the aged row and move the data back to the cell.

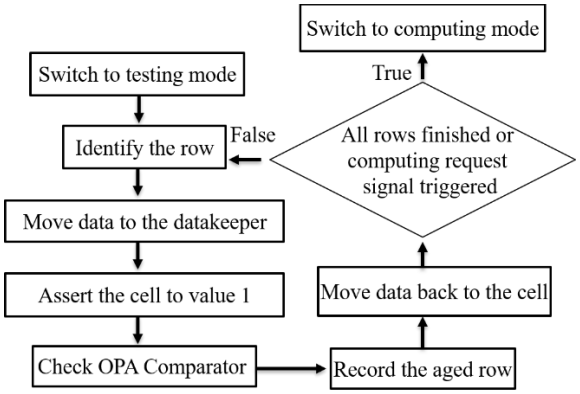

Fig. 5 Flowchart of Aging Detection Method

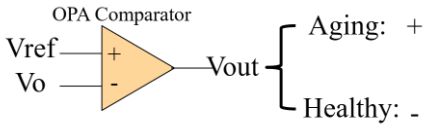

Fig. 6 The principle of OPA Comparator

4.3. Aging Tolerance Method with DVFS

From equation (1), we know the well known formula of nMOS current in linear region. From the equation we can know the parameter which will influence the value of the current. After the nMOS suffers from aging effects, the threshold voltage of the nMOS will increase. It will cause the I_{RBL} decreasing and the value read out at RBL incorrectly. As equation (1) shows, after we increase the operating voltage, V_{gs} will increase, too. The nMOS current in linear region will increase and the I_{RBL} will recover. Before we tune the operating voltage, we also learn from [9] that 8T SRAM is indeed able to operate in different operating voltage normally. Therefore, tuning operating voltage can tolerate the aging effects and recover the decreased I_{RBL}. Besides, we also reference the voltage regulator from [10][11], it can make our voltage regulator more practical. Now we cooperate our aging detection method and our aging tolerance method. As Fig. 4 shows, when the row is under computing phase, the DVFS controller will check the healthy situation of the row. If the row is aged, the DVFS controller will increase the VDD, then perform the DP to get the answer. On the other hand, if the row is healthy, the row can directly perform the DP to get the answer. The principle of increasing the VDD is based on experimental result. For example, the row is aged one month, and we execute the experiment to find out the operating voltage which can compensate the current. After increasing the operating voltage, the aged row can behave as the healthy row and get the correct answer. However, if the row is aged two months, the structure might get a wrong answer under the same operating voltage. Therefore, increasing the operating voltage again is in demand. From the experimental simulation, we can know the correlation between the operating voltage and the aged months. Thus, the DVFS controller can decide the operating voltage correctly.

$$I_D=K[(V_{gs}-V_t)V_{ds}- V_{ds}^2/2] \quad (1)$$

5. EXPERIMENTAL RESULTS

We realize the spice netlist of the aging-aware 8T SRAM CIM structure with DVFS, and perform HSPICE simulation with an industrial 90nm library. We construct an SRAM platform with 256 cells per column and 64 cells per row. The basic operating voltage is 0.75V. In order to estimate aging degradations, we use the threshold voltage shifting and PTM model cards with different aging conditions [12]. We also perform Monte-Carlo simulations for 1000 times with different geometric and statistical models that were built in our industrial library to analyze the PVT variation.

Fig. 7 shows the result of Monte-Carlo variation simulation with 5% variation of the width of transistors in HSPICE, and the figure shows the I_{RBL} current curve with digital input growing from 1 to 15. The Monte-Carlo variation analysis has proved the robustness of the structure. From Fig. 7 we can find the current maintains a good linear relationship between input and I_{RBL}. Through the linear relationship, we can get the correct result of each operation by the value of the I_{RBL} current.

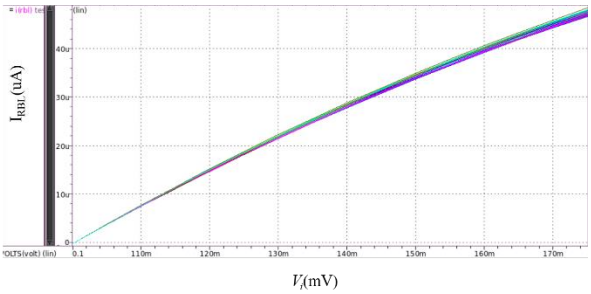

Fig. 7 Monte-Carlo variation simulation in HSPICE

Fig. 8 shows the comparison between the healthy I_{RBL}, the aged I_{RBL} and the aged_DVFS I_{RBL}. Aged_DVFS I_{RBL} adopts our proposed method. It is obvious that once the 8T SRAM cells suffer from aging effects, the I_{RBL} will decrease. Due to the decreasing I_{RBL}, from the figure we can see that the same I_{RBL} will overlap two different points in the horizontal line. In this condition, it will lead to a wrong answer because the circuit will not know which point is the correct one. Once we adopt our DVFS tolerance method, from the figure we can know that I_{RBL} will recover and match with the healthy I_{RBL}. Therefore, our method indeed tolerate the problem caused by aging effects and promise the operation result.

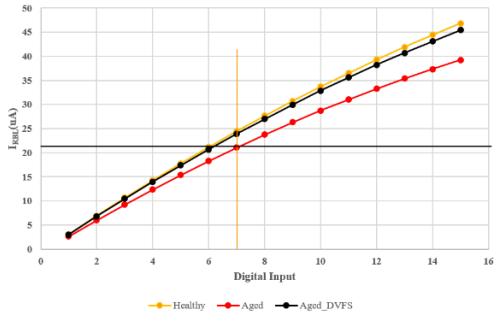

Fig. 8 Comparison between the healthy I_{RBL}, the aged I_{RBL} and the aged_DVFS I_{RBL}

However, once we increase the operating voltage, we will face another problem is power consumption. There is a tradeoff between lifetime and power consumption and in Table I we show the power consumption and the target lifetime with different cases. In Table I we also list average power and worst power to make the result more practical. From Table I we can know that if we do not increase the operating voltage, the circuit will fail. If we adopt our tolerance method and target at 1x_lifetime, in average we need 1.142x power consumption and 1.172x in worst power consumption. If we want to extend the lifetime and target at 2x_lifetime, in average we need 1.185x power consumption and 1.229x in worst power consumption. Therefore, the tradeoff between power consumption and lifetime depends on our purpose.

Table I The correlation between lifetime and power consumption overhead

Techniques		[2]	Ours	
Nor. Lifetime		1	1	2
Nor. Power	Avg	1x	1.142x	1.185x
	Worst	1.118x	1.172x	1.229x

6. CONCLUSIONS

This paper proposes aging-aware 8T SRAM CIM operation method with Memory DVFS and aging detection method to compensate aging-induced problem. This paper also considers both PBTI and HCI effects, and estimates the aging curves of different operating voltages to make our work more practical. Our method can extend the lifetime while maintaining high precision of DP operation.

7. REFERENCES

[1] A. Boroumand, et al., “Google Workloads for Consumer Devices: Mitigating Data Movement Bottlenecks” in Proc. of International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), pp. 316- 331, 2018

[2] A. Jaiswal, I. Chakraborty, A. Agrawal and K. Roy, "8T SRAM Cell as a Multibit Dot-Product Engine for Beyond Von Neumann Computing," in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 27, no. 11, pp. 2556- 2567, Nov. 2019

[3] H. Kim, T. Kim, S. Manisankar and Y. Chung, “Read disturb-free SRAM bit-cell for subthreshold memory applications,” in Proc. of International Conference on Electron Devices and Solid-State Circuits (EDSSC), pp. 1-2, 2017

[4] A. Agrawal, A. Jaiswal, C. Lee, and K. Roy, “X-SRAM: Enabling In-Memory Boolean Computations in CMOS Static Random Access Memories,” IEEE Transactions on Circuits and Systems (TCAS-I), vol.65, issue 12, pp. 4219-4232, 2018

[5] W. Chang, Y. -G. Chen, P. -Y. Huang and J. -F. Li, "An Aging-Aware CMOS SRAM Structure Design for Boolean Logic In-Memory Computing," *2021 IEEE International Symposium on Defect and Fault Tolerance in VLSI and Nanotechnology Systems (DFT)*, 2021, pp. 1-4, doi: 10.1109/DFT52944.2021.9568343.

[6] E. Afacan, M. Berke Yelten and G. Dündar, "Review: Analog design methodologies for reliability in nanoscale CMOS circuits," *2017 14th International Conference on Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design (SMACD)*, 2017, pp. 1-4, doi: 10.1109/SMACD.2017.7981608.

[7] M. Karimi, N. Rohbani and S. Miremadi, “A Low Area Overhead NBTI/PBTI Sensor for SRAM Memories,” in Proc. of IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 25, no. 11, pp. 3138-3151, 2017

[8] Y. -G. Chen, I. -C. Lin and J. -T. Ke, "ROAD: Improving Reliability of Multi-core System via Asymmetric Aging," *2019 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2019, pp. 1-8, doi: 10.1109/ICCAD45719.2019.8942178.

[9] Y. Morita *et al.*, "An Area-Conscious Low-Voltage-Oriented 8T-SRAM Design under DVS Environment," *2007 IEEE Symposium on VLSI Circuits*, 2007, pp. 256-257, doi: 10.1109/VLSIC.2007.4342741.

[10] R. Jain and S. Sanders, "A 200mA switched capacitor voltage regulator on 32nm CMOS and regulation schemes to enable DVFS," *Proceedings of the 2011 14th European Conference on Power Electronics and Applications*, 2011, pp. 1-10.

[11] P. -C. Wu, Y. -P. Kuo, C. -S. Wu, C. -T. Chuang, Y. -H. Chu and W. Hwang, "PVT-aware digital controlled voltage regulator design for ultra-low-power (ULP) DVFS systems," *2014 27th IEEE International System-on-Chip Conference (SOCC)*, 2014, pp. 136-139, doi: 10.1109/SOCC.2014.6948914.

[12] Predictive Technology Model (PTM). [Online]. Available http://www.ptm.asu.edu