EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Page 138 – 143

Proceedings of Emerging Tech Conference:
Edge Intelligence 2023

Aging alleviation technique for 8T IMC SRAMs

Dounavi Helen-Maria[1] and Tsiatouhas Yiorgos[1]
[1] *Dept. of Computer Science and Engineering, VCAS Lab, University of Ioannina, Greece*
edounavi@cse.uoi.gr, tsiatouhas@cse.uoi.gr

Abstract

CMOS Static Random Access Memories (SRAMs) are widely used for In-Memory Computing (IMC) in modern systems to achieve fast and efficient logic and arithmetic computations. However, aging, such as BTI, is a serious threat to the reliability of the SRAMs operations, affecting also significantly the results of the IMC operations in these memories. Hence, it is crucial to develop aging mitigation strategies to maintain the reliability of the memory. The current work proposes an aging alleviation technique for 8T CMOS SRAMs that are frequently used for IMC, by adopting a special purpose and commonly exploited in conventional 8T SRAMs source line, with proper voltage bias on it during non-active periods, as an effective solution to mitigate aging.

1 Introduction

The Von-Neumann architecture, which isolates the memory from the processing unit, is regularly employed in systems dedicated for computing. Yet, data-centric applications such as artificial intelligence, have serious performance limitations due to the requirement for data movements between the memory and the compute cores (Zhiting, L. et al, 2022), (Si, X. et al, 2019). IMC, which realizes data processing within the memory, is an efficient method for reducing the costs related to data transfer. SRAM memory on the other hand, is extensively used in cutting-edge technologies since it is fast, less power-demanding and reliable. As a result, SRAM-based IMC approaches give great potential for speeding up a wide range of applications, which have a significant impact on the future of the computing industry. In SRAM-IMC the complete execution of logical and arithmetic operations is realized within the SRAM. A variety of SRAM-IMC designs in literature propose computations to be carried out directly on the bitlines for the acceleration of the IMC logical operations (Yin, S. et al., 2020), (Jaiswal, A. et al., 2019).

Nevertheless, Bias Temperature Instability (BTI) poses a significant risk to the accuracy of SRAM IMC computations. Timing failures related to transition delays are caused by BTI; an aging phenomenon which causes the absolute threshold voltage value of a transistor under stress to rise over time. nMOS transistors are affected by Positive BTI (PBTI), whereas pMOS transistors are affected by Negative BTI (NBTI). The reliability of the SRAM-IMC is seriously degraded, and under the influence of strong BTI, logic operations carried out within the SRAM can produce false results. The intermediate output value in SRAM-IMC for logic operations is a voltage level, digitized afterwards to the final output. The aging of IMC cell's transistors causes the response voltage levels to change, producing incorrect output results. To provide sustainable

in-memory computations, it is vital to build an aging alleviation strategy. For SRAM-IMC, a variety of memory architectures have been proposed. Numerous works (Jaiswal, A. et al., 2019), (Chang, W. et al., 2021), (Agrawal, A. et al., 2018) have employed the typical 8T SRAM cells for logic IMC operations instead of the common 6T SRAM cells due to read write disturb issues in them. The 8T SRAM IMC structure suggested in (Chang, W. et al., 2021) is used as a case study in the present work. The main concept is the layout modification of the typical 8T SRAM by separating a transistor source line from the ground (Gnd) and properly biasing it, to alleviate the impact of BTI on the transistors within the memory's computation path.

2 Preliminaries

The typical 8T SRAM architecture commonly utilized in contemporary IMC technologies is the main focus of this study. As depicted in Figure 1a, two additional to the typical 6T cell nMOS transistors (MN5 and MN6), create the read path through the extra Read bitline (RBL), dividing the read from the write operation which is performed through the traditional Write bitline (WBL). This modification does not alter the write operation while, for the read operation, the memory cells are accessed through the additional Read wordline (RWL) (set to logic '1'), maintaining the Write wordline (WWL) to low (logic '0'). The RBL is initially pre-charged to Vdd and during the read operation the value stored in the memory cell determines the output RBL voltage level (V_{RBL}). In read mode, when the value stored in the cell is logic 1 ($Q = '1'$) transistors MN5 and MN6 are activated and the VRBL drops through the activated reading path. When the cell's stored value is logic '0' ($Q = '0'$) the RBL remains charged to V_{dd}. Reading can be executed with a substantial voltage swing on the RBL when using the 8T cell's decoupled read port, eliminating completely any read disturb failures (Jaiswal, A. et al., 2019). A part of the memory array with two 8T SRAM cells attached on a RBL is depicted in Figure 1b, for the demonstration of the read paths formed in a column, through the corresponding MN5 and MN6 transistors of the cells. This architecture, exploited by (Chang, W. et al., 2021) for IMC operations, is hereby explored as a case study of an 8T SRAM-IMC scheme (Figure 1b). Initially, a pair of RWLs (such as RWL1 and RWL2) are enabled to perform an IMC logic operation, and then the combined stored value of Cell1 and Cell2 determines the VRBL, within a specified time duration. Then, to carry out various logic operations (OR, AND, NOR, NAND), the read out V_{RBL} is transferred to properly skewed inverters for digitization. In Figure 1c, the VRBL voltage levels are presented for the different stored values stored in Cell1 and Cell2 accordingly (cases 00, 01, 10 and 11 are given). Each properly skewed inverter, treats the V_{RBL} as a logic '1' or '0' respectively, delivering the response of the corresponding logic operation.

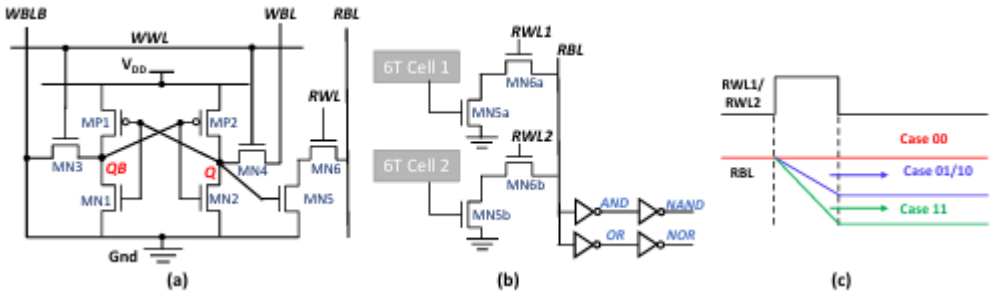

Figure 1: (a) The typical 8T SRAM cell structure, (b) an 8T SRAM-IMC column, with skewed inverters to sense logic operations, (c) the voltage levels of RBL under different cases of stored values in the cells

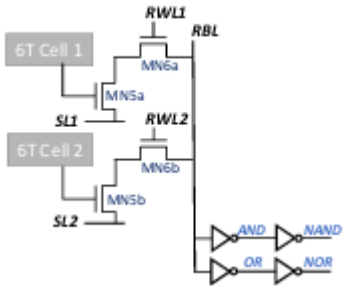

Figure 2: The proposed 8T SRAM aging alleviation modifications

However, SRAM memory cells are prone to transistor aging, such as BTI, raising the likelihood of false IMC responses, which is a serious reliability issue in contemporary technologies. Extreme stress circumstances, such as high temperatures and gate-to-source voltage levels, hasten BTI aging even more. Numerous solutions have been proposed to mitigate the effects of NBTI, nevertheless, with technology evolution, PBTI also poses a serious threat to SRAM-IMC (Jaiswal, A. et al., 2019), (Chen, Y.-G. et al., 2022). To properly address the PBTI-induced issue, (Chang, W. et al., 2021) suggests a modified 8T SRAM IMC architecture with supplemental transistors and the use of an aging detection algorithm to determine the health state of each memory array row. Additional area overhead is required for the implementation of the aging detection scheme. More recently, the authors in (Chen, Y.-G. et al., 2022) suggests adjusting the operating supply voltage during computation to counteract the deteriorating RBL discharge current brought on by the effects of aging. This solution needs both an aging monitoring method and a dynamic voltage frequency scaling mechanism, while it presents an important increase in the circuit's average power consumption.

With a minimal area overhead and a small power consumption penalty, the current work seeks to alleviate PBTI-induced reliability degradation on the nMOS network of the reading/computing path of the 8T SRAM IMC structure presented in Figure 1. The typical 8T SRAM cell is modified with the separation of the MN5 transistor source line from the ground (Gnd), and the application of an appropriate voltage level to it, as explained below. It should be noted that any potential aging of the 6T Cells does not affect the IMC reliability and thus, the aging of MN5 and MN6 transistors is examined.

3 Proposed technique

An 8T CMOS SRAM aging alleviation approach is proposed to achieve accurate IMC results and maintain the memory's reliable operation. Initially, note that the reading operation corresponds only to a brief time interval of the total SRAM's operating time. During this time, the gate of the MN6 transistors (Figure 1) are set to high (logic '1') and the transistors only enter the stress state during this brief phase. Then, the gate is set back to low (logic '0') and the transistors return to the relaxed state for the rest operation of the memory. Hence, aging does not have a significant impact on MN6 transistors. On the contrary, when a logic '1' value is the one stored in a memory cell ($Q = '1'$), the accompanying MN5 transistor constantly experiences extreme DC stress since VGS=Vdd for as long as the cell's stored value is kept the same. Because AC BTI stress involves recovery cycles that minimize the influence on Vt, DC stress is the main reason for transistor's Vt degradation.

In the proposed 8T SRAM architecture, depicted in Figure 2, the source terminals of MN5 transistors are separated from Gnd and act as independent source lines (SL). To reduce the PBTI on MN5, a positive

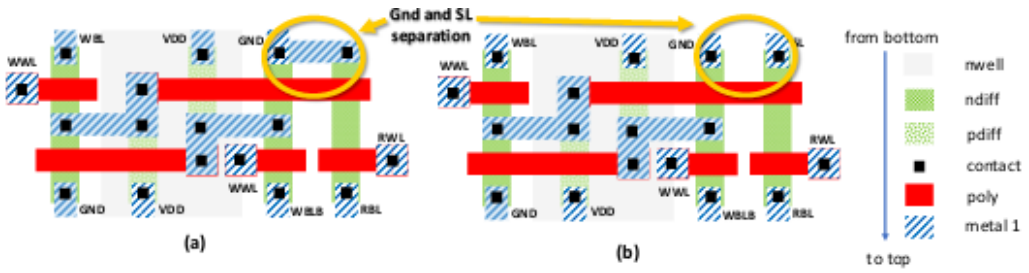

Figure 3. (a) Layout of the standard 8T SRAM cell
(b) Layout for the 8T SRAM cell with separated Gnd and SL

voltage $V_{\mathrm{SL}} > 0$ (eg. V_{dd}) is by default applied on the SL lines in all memory rows when they are not activated for any operation (inactive state). As these periods correspond the vast majority of the memory's operational time, it follows that for MN5 transistors $V_{\mathrm{GS}} < V_{\mathrm{dd}}$ for most of their lifetime. Thus, MN5 transistors experience extreme BTI stress only when the related cell is activated for reading or computation, remaining to a decreased BTI stress or even to a fully relaxed state (when our selection is $V_{\mathrm{SL}} = V_{\mathrm{dd}}$) during the rest of the cell's prolonged inactive period, reducing the impact of PBTI on it. Note that the stress on MN5 will be negligible as that on MN6, when VSL is equal to or close to V_{dd}

To maintain the memory operation, an SL is turned to Gnd when a reading/computing operation is initiated and the pertinent RWL is set to high. Since the reading path through MN5 leads to Gnd when Q='1', the RBL will be discharged, whereas when Q='0', MN5 will be inactive and the cell will not contribute to the RBL's voltage drop. To examine the additional area cost of the suggested solution, Figure 3a depicts the typical 8T SRAM cell layout and Figure 3b the modified 8T SRAM cell layout of the suggested method, where the SL and Gnd nodes are separated. Note that the SL line is shared by all MN5 transistors of every two consecutive rows. As shown, there is no additional silicon area cost at the memory cell level. However, due to the routing requirements of higher-level power supply metal lines, the exploitation of the SL line as a separate line will result in a 14% area overhead penalty at memory array level. Nevertheless, according to the literature, the adoption of a separate source line, with the same area cost, is a common practice in order to support other special functionality in 8T SRAMs (Jaiswal, A. et al., 2019).

4 Simulation Results

Proper simulations on a memory array structure were conducted for the study of the aging effect on the 8T SRAM IMC. An 8T SRAM array with 256 rows and 64 cells per row, according to the topology of Figure 1a, was designed (with Virtuoso of CADENCE) and simulated (with Spectre) exploiting the 28nm CMOS technology of UMC ($V_{\mathrm{dd}} = 0.9\mathrm{V}$). Figure 4 illustrates the outcomes of IMC operations for the case where two 'fresh' cells are activated with 01/10 as stored values (Figure 4a) or 11 (Figure 4b) accordingly. Once the RBL voltage level falls below a selected threshold value, the skewed inverters (Figure 1) output the results of the relevant logic operation. Under severe PBTI stress, MN5 transistors' current will not develop an adequate VRBL level, within the IMC time duration, lower than the skewed inverters' transition threshold, resulting in a false result. To assess the impact of the PBTI on the IMC results, different threshold voltage shift (ΔV_{t}) levels were applied to MN5 transistors, by inserting a DC voltage source of proper polarity at their gate terminals. When only one cell has stored logic '1' (01/10 scenario), Figure 5 shows the transition of the RBL and its final voltage level at IMC mode for the "fresh" and the aged case (MN5 transistor with $\Delta V_{\mathrm{t}} = 50\mathrm{mV}$or 100mV). It also depicts the skewed inverter's transition threshold. For the aged cases, the inverter's transition threshold is violated and a failure is produced as expected.

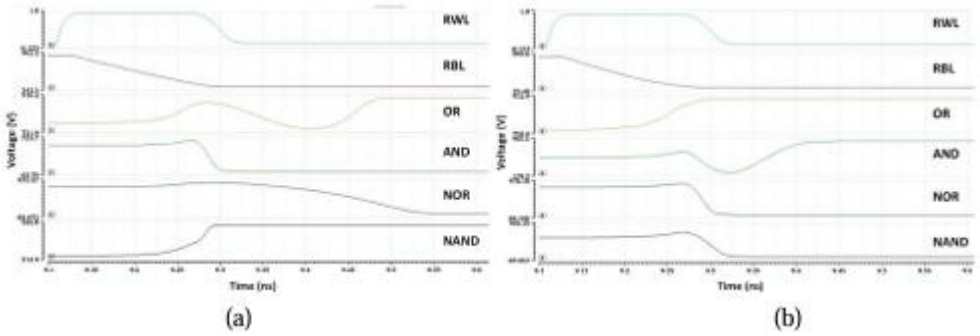

Figure 4. (a) The result of IMC when 01/10 are stored and **(b)** when 11 is stored in the memory cells.

To further accentuate BTI influence, Table 1 presents the required delay time for the generation of the logic result and the V_{RBL} after the NOR and NAND operations are performed, for both MN5 transistors' "fresh" and aged (at $\Delta V_{\mathrm{t}} = 50\mathrm{mV}$) state. Since case 00 will not result to an important RBL voltage change in either a fresh or an aged state, it has not been listed in the Table. It is clear that PBTI has a significant impact on IMC results. Biasing the SL at appropriate voltage levels above zero and near V_{dd}, the voltage stress on MN5 transistors is drastically reduced or even eliminated (when $V_{\mathrm{SL}} = V_{\mathrm{dd}}$) so that the BTI effect is significantly alleviated maintaining the reliability of the memory operation.

The suggested technique guarantees the SRAM's reliable operation with a small but acceptable area overhead penalty. The simulations also revealed that the proposed method will result in a 8.52μW increase of the overall dynamic power consumption in the memory array, over 13.73μW of the original design, during the read/compute operations when $V_{\mathrm{SL}} = V_{\mathrm{dd}}$, while it is smaller for lower voltage levels (e.g. 2.32μW for VSL=Vdd/2). Note that there is not any additional delay penalty, since the discharging/charging of the SL is achieved during the charging/discharging of the pertinent wordline.

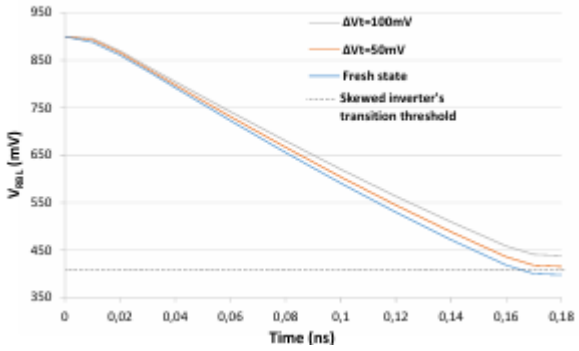

Figure 5: RBL voltage level for different ΔVt (01/10 scenario)

	Operation	Delay (ps)	V_{RBL} (mV)
Fresh state	NOR (Case 01/10)	400	400
	NOR (Case 11)	170	116
	NAND (Case 11)	200	116
ΔVt= 50mV	NOR (Case 01/10)	Failed	416
	NOR (Case 11)	180	126
	NAND (Case 11)	Failed	126

Table 1: Delay time and remaining RBL voltage level for (a) the fresh state and (b) under 50mV of aging

5 Conclusion

In this work, a scheme and a method for the alleviation of PBTI-induced aging in the reading/computing path of 8T IMC SRAMs is proposed. It involves the layout modification of the typical 8T SRAM cell for the insertion of an additional properly biased source line. When the SRAM is active, the source lines of the cells that are accessed are set to the predefined voltage level that is suitable for the corresponding operation. However, during the SRAM's non-active periods, the source line is set to a selected positive voltage level in order to alleviate aging.

The proposed technique does not induce any performance degradation, inserts a small, but acceptable according to the common practice, area overhead and increases the power consumption depending on the used bias voltage. On the other side, it offers an easy to implement solution against the influence of BTI aging on read/computing performance, that ensures the reliable memory operation throughout its lifetime, a crucial issue in high reliability systems. Previously presented techniques, mitigate aging by periodically monitoring the memory and properly adjusting its operation. These approaches affect performance and may not be effective in case of overaged transistors.

6 Acknowledgments

We acknowledge support of this work by the project "Dioni: Computing Infrastructure for Big-Data Processing and Analysis." (MIS No. 5047222) which is implemented under the Action "Reinforcement of the Research and Innovation Infrastructure", funded by the Operational Program "Competitiveness, Entrepreneurship and Innovation" (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund).

7 References

[1] Agrawal, A. et al. (2018). X-SRAM: Enabling In-Memory Boolean Computations in CMOS Static Random Access Memories. IEEE Trans. Circ. Syst. I (TCAS-I), vol. 65, no. 12, pp. 4219-4232.

[2] Chang, W. et al. (2021). An Aging-Aware CMOS SRAM Structure Design for Boolean Logic In Memory Computing. IEEE Intern. Symp. on Defect and Fault Toler. in VLSI and Nanotech. Syst. (DFT), (pp. 1-4). Athens.

[3] Chen, Y.-G. et al. (2022). Aging-Compromised Computing-In-Memory Dot-Product Calculation Technique Through DVFS. 24th Workshop on Synth. and Syst. Integr. of Mixed Inf. Tech. (SASIMI), (pp. 44-47). Hirosaki.

[4] Jaiswal, A. et al. (2019). 8T SRAM Cell as a Multibit Dot-Product Engine for Beyond Von Neumann Computing. IEEE Trans. on VLSI Syst., vol. 27, no. 11, pp. 2556-2567.

[5] Si, X. et al. (2019). A Dual-Split 6T SRAM-Based Computing-in-Memory Unit-Macro With Fully Parallel Product-Sum Operation for Binarized DNN Edge Processors. IEEE Trans. on Circ. and Syst. I: Regular Papers, vol. 66, no. 11, pp. 4172-4185.

[6] Yin, S. et al. (2020). XNOR-SRAM: In-Memory Computing SRAM Macro for Binary/Ternary Deep Neural Networks. IEEE J. of Solid-State Circ., vol. 55, no. 6, pp. 1733-1743.

[7] Zhiting, L. et al. (2022). A review on SRAM-based computing in-memory: Circuits, functions, and applications. J. Semicond., vol. 43, no. 3, pp. 173–210.