Exploration Methodology for BTI-Induced Failures on RRAM-Based Edge AI Systems

Alexandre Levisse, Marco Rios, Miguel Peón-Quirós and David Atienza
Embedded Systems Laboratory (ESL), EPFL, Switzerland

***Abstract*—Resistive switching memory technologies (RRAM) are seen by most of the scientific community as an enabler for Edge-level applications such as embedded deep Learning, AI or signal processing of audio and video signals. However, going beyond a "simple" replacement of eFlash in micro-controller and introducing RRAM inside the memory hierarchy is not a straightforward move. Indeed, integrating a RRAM technology inside the cache hierarchy requires higher endurance requirement than for eFlash replacement, and thus necessitates relaxed programming conditions. By doing so, the reliability bottleneck is moved from programming to the read operations (i.e., read margin is reduced and the risk of read failure is increased). Based on this observation, in this work, we propose to explore how Edge-level applications running on a RRAM-based Edge device could fail because of Bias Temperature Instability (BTI). BTI causes threshold voltage (Vt) degradation on the transistors along the memory WordLines (WL), leading to a reduction of the read margin along regularly used WLs. We thereby propose a 3-steps methodology consisting in (i) characterizing the RRAM bitcell and identifying beyond which Vt shift the read operation is going to fail. (ii) characterizing applications and extracting the memory traces. And (iii) running a long term BTI simulation to extract the actual Vt shift of the bitcells sharing the same array WordLine. Based on this, we show that for a 1T1R bitcell featuring a 250% High/Low Resistance State (HRS/LRS) ratio, read failures tend to happen after less than a month in the case of a constantly running convolution kernel. These simulations highlight the fact that transistor-level reliability can be critical for embedded RRAM and that specific workload aware simulation frameworks are required to assess their effects.**

I. Introduction

With the arrival in the consumer market of edge AI systems running more and more complex applications such as deep neural networks, video/audio encoding and filtering or cryptography, the requirements in terms of energy efficiency and leakage power have evolved. Hence, emerging resistive non volatile memories came in as a promising solution as they enable zero-leakage, low-cost integration and relatively high performances. In regard with these specifications, they are usually considered as a potential game changer and, beyond the replacement of eFlash in sub-28nm nodes [1], the scientific community considers their direct integration inside the memory hierarchy (Caches replacement or enhancement for e.g. [2]). In that context, while memory activity is relatively high, programming conditions have to be relaxed to enable higher endurance [3], thereby shifting the reliability bottleneck of the memory from the write operation to the read operations as the read window becomes smaller. Consequently, it has been shown that some technologies (Magnetic Memories [4] for e.g.) or programming conditions (low programming current for filamentary RRAM [5], [6] or Phase Change Memories [7] - PCM) enable higher endurance at the cost of smaller read margins (i.e. more complex read operations [8]). Thereby, while the reliability constraints are still high during the programming operations, the bottleneck tends to move towards the read operation which becomes extremely sensitive. While most of reliability studies explore the failure mechanisms of the RRAM device by itself [6], reliability of its selection transistor is usually not considered in reported publications. In this work, we propose to open the question of the selection transistor reliability and we chose to focus on the impact of Bias Temperature Instability [9], [10] (BTI)-induced threshold voltage (Vt) shift on the selection transistors inside 1Transistor-1RRAM memory arrays while running Edge-level applications. Along the memory array WordLines (WL), all the transistors are stressed in the same way, thereby inducing workload dependent Vt shift.

The main contributions of this work are the following:

- A bitcell-level analysis of the impact of BTI-induced Vt shift on the RRAM read margin.
- A WL-level memory traces framework simulating the evolution of BTI-induced Vt shift with Edge-level application patterns.
- An exploration of the effect of BTI-induced failures in memory. We show that for a 250% HRS/LRS ratio, while running a convolution kernel, first failures tend to happen from a few days to a few months of constant operation.

The remainder of the paper is organized as follows. Section II presents the background of the paper, presenting RRAM memory array organizations and BTI-induced Vt shift. Section III presents the simulation conditions and the simulation framework for Edge level applications considered in this work. Section IV presents the simulations results and discusses some potential solutions. Finally, Section V concludes the paper with a summary of the main findings of this work.

II. Background

A. Resistive Memory Architectures

Resistive Random Access memories (RRAM) have been gaining popularity along the last 10 years as eFlash technologies became hard to co-integrate in advanced technology process [1]. In that context, 2-terminal RRAM like filamentary RRAM (ReRAM) [5], Phase Change Memories (PCM) [7] or Spin Transfer Torque Magnetic RRAM (STT-MRAM) [4] are now considered as serious industrially viable solutions.

In a nutshell, these technologies consist in the non-volatile resistance variation of an insulating material sandwiched between two metallic electrodes. While each technology is based on drastically different physical phenomenons and exhibit different behaviors, from a functional point of view, Single Level Cells (SLC) RRAM can be modeled as a two resistance states device, assuming that the programming operation is managed adequately. In that context, the two resistance states can be assumed to be a normal distribution with a given dispersion which depends on the technology characteristics and the programming conditions.
From the architectural point of view, while crosspoint architecture is gaining a lot of interest in the last years, it introduces deep technology, integration, physical design, circuit and reliability constraints [11], [12] targeting it towards standalone memory chips [13]. Thereby, most of the viable solutions reported in the last years are integrating RRAM as part of a 1T1R array architecture [14], [3], [15], [16]. In that context, reliability of the access transistor has to be analysed, modelled and optimized by the designers and the foundry.

B. BTI Induced Aging

Among the various sources of failure, soft reliability issues, such as Bias Temperature Instability (BTI), are under high interest in the community [10] as they tend to exhibit highly application dependent degradation [17]. BTI occurs when a transistor gate is forward biased: Positive BTI (PBTI) for n-type and negative (NBTI) for p-type transistors. BTI occurs at the oxide-channel interface, and is caused by the creation of charge trapping defects. Such traps can then catch or release charges following a stochastic process. Over time, two kind of situations are possible: (i) The transistor is stressed, inducing the creation of traps, thereby causing the transistor Vt to shift toward higher values. (ii) The transistor is unbiased, inducing a release of trapped charges, thereby causing the transistor Vt to shift towards lower values. We use the defect-centric model proposed in [9] to determine the evolution of the Vt along time with regards to the considered workload.

III. Simulation Framework

In this work, we propose an innovative simulation framework that can give insights on the effect of BTI-induced threshold voltage (Vt) drift on RRAM memory architectures when complex edge AI applications are executed. We considered three simulation steps which we iterate among them. First, for a given bitcell and RRAM technology, we extract the read margin versus the Vt shift. Then, we characterize Edge level applications and extract memory traces. Finally, we identify the most used memory array WordLine (WL) and we calculate the Vt shift occurring on the access transistors based on memory access patterns from step 2. Based on the failure characterization from step 1 we can determine the potential lifetime of the architecture running a given application.

A. Bitcell Level Vt Shift

In order to assess the read margin, we consider a 1T1R bitcell and run monte-carlo simulations considering RRAM and CMOS variability. Then, we introduce Vt shift by progressively reducing the WL voltage (up to 50mV).
In this work, we consider a low HRS-LRS window (5k-12.5k - corresponding to a 250% ratio) case which corresponds to STT-MRAM regular windows or low programming current PCM or ReRAM technologies [4], [7], [5], [6]. We then introduce a 10% (5% respectively) variability on the HRS (LRS respectively). We consider a 28nm bulk CMOS industrial PDK and run simulations at $20\,^{\circ}\mathrm{C}$ (morever an increase in temperature would make the situation worse [3]).

B. Memory Traces Extraction

Application characterization is performed through the extraction of memory traces from real edge level application such as signal processing (filtering, compression, convolutional Neural Networks, machine learning as we did in [18]). We thereby apply a methodology analogous to [18] enabling the extraction of memory traces from different edge level applications. For the sake of concept demonstration and in order to open the discussion and highlight potential issues, in this work we focus on a single convolution kernel (widely used in recent machine learning algorithms [19], [20]) involving 3x3 and 30x30 random matrix.

C. Application Level BTI Extraction

We then translate the memory traces extracted in section III-B into signals understandable by the BTI simulator [9]. In that sense, we consider only the addresses corresponding to the most used WL and then translate each read access into a 50ns 0.8V pulse, each write access into a 50ns 1.5V pulse. When no operation is performed, on that WL, we consider a relaxation period equivalent to the amount of operations performed on the other WLs. This way, assuming that each memory access is following the previous, we can extract an accurate application-aware behavior of the access transistors BTI-induced Vt shift.

IV. Application Aware Aging

A. BTI impact on Read Margin

The first step of the proposed methodology is to assess the effect of BTI-induced Vt shift at the bitcell level. Hence, we simulate 1T1R bitcells and extract the corresponding read current (assuming a current-based read sense amplifier). Figure 1 shows the read current distributions in LRS and in HRS of the considered bitcell at $20\,^{\circ}\mathrm{C}$. When introducing BTI-induced Vt shift, the distributions shift toward lower read current. It can be noted that the LRS read current tends to shift more than the HRS read current, leading to a point where distribution tails overlap. This unbalanced shift is due to the fact that in LRS, the access transistor Vds is higher, thus, leading to more variations in the Ids current. Such distribution tails overlap will lead to read failures. Read margin with no Vt shift is shown in black (5uA) and it becomes negative (in red) when Vt shift is introduced. Thereby, we define the read margin as the difference of current between the LRS and HRS

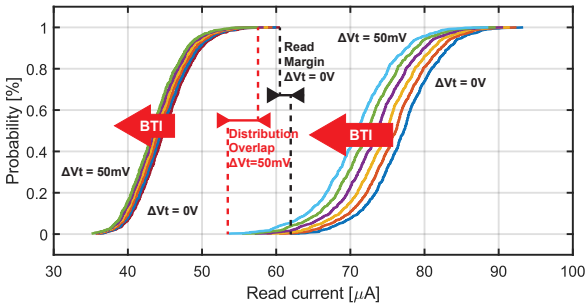

Fig. 1. Read current distributions for a 250% HRS/LRS ratio considering RRAM and CMOS variability at 20 °C. Distribution shift due to BTI-induced Vt degradation are shown and reduction of the read margin is highlighted (from black to red).

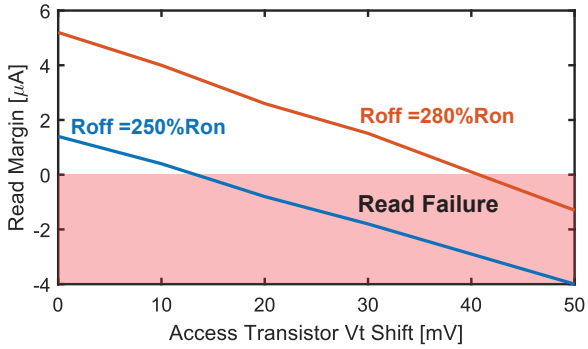

Fig. 2. Evolution of the read margin versus the BTI-induced Vt shift. A 250% ratio exhibits a 15mV maximum Vt shift before failure while a 280% ratio fails with 40mV of shift at 20 °C.

current tails and extract it for 250% and 280% HRS/LRS ratio at 20 °C. As shown in figure 2, for a 250% ratio, a 15mV Vt shift is enough to cause read failures while a 280% ratio requires 40mV before exhibiting read failures at 0.8V.

B. BTI impact on Application

In this section we consider the application traces from Section III-C and explore their effect at the application level. Following up the methodology described in Section III, we run the selected convolution kernel and observe its effect over time. An interesting feature of the simulation framework we propose, and that has not been considered in any of the previously reported works, is the support for several voltages over the device lifetime. As a mater of fact, we consider a relaxation voltage (0V), a read voltage (0.8V), a programming voltage (1.5V) and stress the transistor accordingly following the application pattern. In that sense, as Figure 3 shows, over time the Vt shift tends to reach 3 different states: (i) a relaxed state in which the charges are de-trapped; (ii) a read stress state in which only a portion of previously created traps can be charged; and (iii) a write stress state in which all the available traps are getting occupied. Consequently, the value of each of these states is controlled by: the application activity factor, the read voltage value, the write voltage value and the temperature. Overall, as it can be seen in figure III, a higher temperature

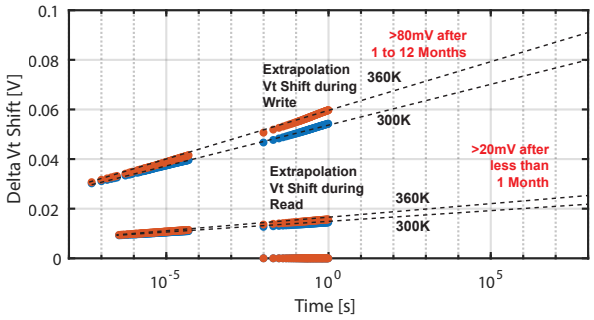

Fig. 3. Workload aware evolution of the Vt shift for the most used WL in the considered memory array at 20 and 80 °C. Linear extrapolation shows read failures happening after less than a month for a 250% HRS/LRS ratio. On the other hand, during write, more than 80mV Vt shift can be exhibited after less than a year.

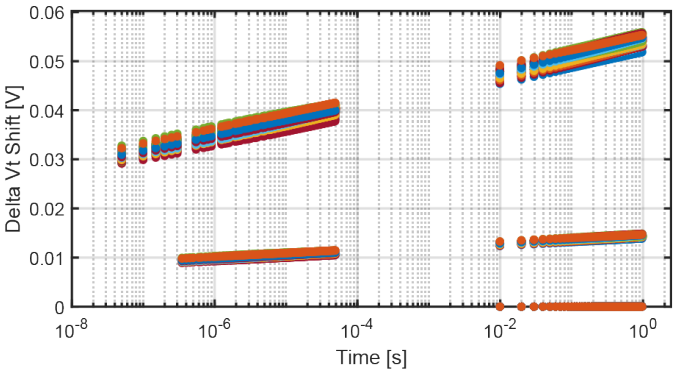

Fig. 4. Workload aware evolution of the BTI-induced Vt shift considering stochastic nature of the defect-centric BTI model [9]. Random creation/release of traps introduces variability in the aging process.

eases trap creation, in turn leading to a higher Vt shift.

As a BTI-induced Vt shift can lead to potentially extremely long simulations (i.e., a second of BTI simulation using the model from [9] takes a few tens of minutes in a high end Intel server), we consider a linear (and pessimistic) approximation for the extrapolation of BTI-induced Vt shift. Based on this extrapolation, we show that during read operation, a 20mV shift can appear in less than a month of stress. On the other hand, during write operation, a 80mV Vt shift can be reached from 1 (at 80 °C) to 12months (at 20 °C) after the initial stress. Such shift will induce read failures, as it was introduced Figure 2 and potential write failures (out of the scope of this paper but briefly discussed at the end of this section). It could be concluded that such effect might make a edge level completely unusable within a month.

Beyond temperature corners, it must be noted that trap creation, is a completely random process that must be considered from a statistical perspective [10], [9]. Figure 4 shows the results of 100 runs of a 1second BTI-induced Vt shift running the previously introduced convolution kernel on the most used WL of the RRAM array. As traps might get created completely randomly among the gate oxide, it is mandatory to consider such exploration when looking into BTI-induced stress.

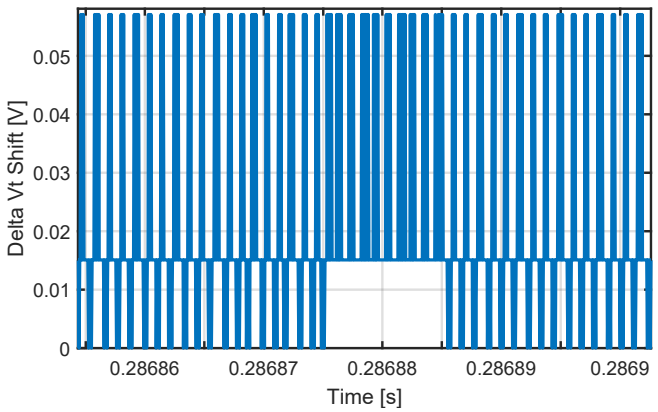

Fig. 5. Detailed view of 60us of the convolution kernel running. This view shows how the BTI-induced Vt shift varies over time, switching between relaxation periods, read and write phases.

Figure 5 presents the detailed evolution of the considered BTI-induced Vt shift over time. As previously discussed, the application actually alternates between relaxation, read and write phases. Hence, this example shows a typical edge AI context in which the memory is running at 20Mhz during 60us. While we only discussed the failures due to BTI-induced Vt shift during read, we did not discuss in this paper any of the other sources of failures induced by BTI during write operations. However, two additional families of failures can occur, namely: (i) Write failures during write: this failure is prone to happen during a reset operation. A too high Vt shift may not enable the access transistor to drive enough current to trigger a programming operation. In that context, the probability of failure is increased, leading to stuck-at fault in the memory array. (ii) Read failure due to degraded distribution. A lower programming current may induce in some technologies (filamentary RRAM or Phase Change Memories) a distribution spread closing the HRS/LRS window [16]. In that context, the read failure is induced by a degraded write condition.

Although in this work, we do not consider RRAM memory aging itself. It is widely reported that such effect closes the window [21], exacerbating the aforementioned effects. As a countermeasure, introducing a relaxation period could be considered until the trapped charges are released and the transistor reaches a lower Vt shift value. On the other hand, address relocation could be considered, at one point, for WLs that tend to be used a lot. Finally, native high read margin area expensive solutions relying on differential read, such as, 2Transistors-2RRAM (2T2R) [15] could enable better reliability to BTI-induced Vt shift.

V. Conclusion

In this work we have proposed an innovative BTI-induced failure exploration framework for 1T1R arrays of RRAM memories. This framework contains a bitcell characterization phase that identifies for a given CMOS and RRAM technology the maximum admissible Vt shift. Then, an application characterization providing memory traces for a given WordLine. Finally, a long term exploration using a defect-centric BTI prediction model and a fast extrapolation methodology to identify the lifetime of a given RRAM architecture running a given application. In that context, we show with our framework that a 250% HRS/LRS ratio RRAM integrated with a regular 28nm industrial CMOS transistor starts to exhibit read failure events after less than a month of operation while running a convolution kernel. Overall, in this work, we highlight (i) the need for workload aware simulations when considering RRAM technologies and (ii) the fact that CMOS reliability has to be considered when designing 1T1R bitcells arrays. Such considerations have to be taken into account when dealing with low-window RRAM technologies in the context of next-generation edge AI systems.

Acknowledgment

This work has been supported by the ERC Consolidator Grant COMPUSAPIEN (GA No. 725657), the EC H2020 WiPLASH (GA No. 863337) project, and the EC H2020 RECIPE (GA No. 801137) project.

References

[1] F. Arnaud et al. Truly innovative 28nm fdsoi technology for automotive micro-controller applications embedding 16mb phase change memory. In *IEEE IEDM*, 2018.
[2] H. K. Ahn et al. Evaluation of stt-mram l3 cache in 7nm finfet process. In *IEEE ICEIC*, 2018.
[3] Q. Dong et al. A 1mb 28nm stt-mram with 2.8ns read access time at 1.2v vdd using single-cap offset-cancelled sense amplifier and in-situ self-write-termination. In *IEEE ISSCC*, 2018.
[4] D. Apalkov et al. Magnetoresistive random access memory. 2016.
[5] H-S Philip Wong et al. Metal–oxide rram. 2012.
[6] C. Nail et al. Understanding rram endurance, retention and window margin trade-off using experimental results and simulations. In *IEEE IEDM*, 2016.
[7] H-S Philip Wong et al. Phase change memory. 2010.
[8] L. Wei et al. 13.3 a 7mb stt-mram in 22ffl finfet technology with 4ns read sensing time at 0.9v using write-verify-write scheme and offset-cancellation sensing technique. In *IEEE ISSCC*, 2019.
[9] D. Rodopoulos et al. Understanding timing impact of bti/rtn with massively threaded atomistic transient simulations. In *IEEE ICICDT*, 2014.
[10] J. H. Stathis. The physics of nbti: What do we really know? In *IEEE IRPS*, 2018.
[11] A. Levisse et al. Architecture, design and technology guidelines for crosspoint memories. In *IEEE/ACM NANOARCH*, 2017.
[12] A. Levisse et al. Sneakpath compensation circuit for programming and read operations in rram-based crosspoint architectures. In *NVMTS*, 2015.
[13] F. T. Hady et al. Platform storage performance with 3d xpoint technology. *Proceedings of the IEEE*, 2017.
[14] G.-W. Burr et al. Neuromorphic computing using non-volatile memory. *Advances in Physics: X*, 2017.
[15] M. Bocquet et al. In-memory and error-immune differential rram implementation of binarized deep neural networks. In *IEEE IEDM*, 2018.
[16] M. Alayan et al. Switching event detection and self-termination programming circuit for energy efficient reram memory arrays. In *IEEE TCASII*, 2019.
[17] D. Stamoulis et al. Capturing true workload dependency of bti-induced degradation in cpu components. In *ACM GLSVLSI*, 2016.
[18] S. Tuli et al. Rram-vac: A variability-aware controller for rram-based memory architectures. In *IEEE/ACM ASP-DAC*, 2020.
[19] A. Vasudevan et al. Parallel multi channel convolution using general matrix multiplication. In *IEEE ASAP*, 2017.
[20] A. Krizhevsky et al. Imagenet classification with deep convolutional neural networks. In *Advs. in Neural Information Process. Syst. 25*, 2012.
[21] G. Sassine et al. Sub-pj consumption and short latency time in rram arrays for high endurance applications. In *IEEE IRPS*, 2018.