Switching event detection and self-termination programming circuit for energy efficient ReRAM memory arrays

M. Alayan[1], E.Muhr[1], A. Levisse[2], M. Bocquet[1], M. Moreau[1], E. Nowak[3], G. Molas[3], E. Vianello[3], J.M. Portal[1]

***Abstract*— Energy efficiency remains a challenge for the design of non-volatile resistive memories (ReRAMs) arrays. This memory technology suffers from intrinsic variability in switching speed, programming voltages and resistance levels. The programming conditions of memory elements (e.g. pulse widths and amplitudes) must cover the tail bits to avoid programming failures. Switching time of ReRAMs shows wide distributions. Therefore, fast cells are subjects for electrical stress after their switching and energy waste since programming currents are typically large for this technology (tens of µA). In this paper, we present a Write Termination (WT) circuit to stop the programming operation when the switching event occurs in the selected memory element. The proposed design is sensitive to current variations that take place when the memory element switches between two different resistance states (LRS and HRS). This WT scheme reduces the power consumption by 97+%, 93+% and 65+% during Forming, RESET and SET operations respectively. Our estimations show that area efficiency of 70% for a memory array is achievable when the presented WT circuit is integrated in near-memory peripheries. The demonstrated WT circuit is suitable for different ReRAM technologies with small overhead penalty and shows robustness against CMOS and ReRAM variabilities.**

***Keywords*—ReRAM; Write termination (WT); Switching time; Energy efficiency; Area overhead.**

I. INTRODUCTION

With ever-increasing requirements for energy efficiency, conventional memories have become a major bottleneck for the development of ultra-low power electronic systems [1,2]. Many research institutions and manufacturers have actively been involved in research and development for "emerging nonvolatile memories" [3]. Among different emerging memory technologies, resistive RAMs (ReRAMs) are considered as leading candidates for embedded nonvolatile memory (NVM) and storage class memory (SCM) applications [4-7]. ReRAMs are characterized by their fast switching speed, low operating voltages compared to charge storage solutions, good scalability and compatibility with BEOL CMOS process [8,9]. Thanks to recent integration architectures and device engineering, Mb-scale capacities for embedded applications in advanced nodes have been demonstrated [10]. However, ReRAMs suffer from intrinsic cell-to-cell and cycle-to-cycle variabilities in resistance levels and switching parameters (e.g. speed and voltages) [11-13]. This variability stands as one of the main challenges for ReRAM technology development. Programming times ($T_{Forming/SET}$ and T_{RESET}) of ReRAM devices have shown wide distributions [14]. During programming phases, the time width of the applied voltage pulses must cover the worst cases corresponding to the tail bits in the distribution. This leads to energy waste and over programming stress for the fast switching cells.

M. Alayan, E.Muhr, M. Bocquet, M. Moreau and J.M. Portal are with Institut Matériaux Microélectronique et Nanosciences de Provence (Im2np), France (e-mail: jean-michel.portal@univ-amu.fr).

A. Levisse is with École Polytechnique Fédérale de Lausanne (EPFL), Switzerland (e-mail: alexandre.levisse@epfl.ch).

E. Nowak, G. Molas and E. Vianello are with Commissariat à l'énergie atomique et aux énergies alternatives (CEA), France (e-mail: etienne.nowak@cea.fr).

Therefore, Write Termination (WT) is mandatory for energy and latency efficiencies in ReRAM arrays. Typical WT circuits consist of two stages:

1. Detection of the switching event during either Forming/SET or RESET.
2. Feedback loop to stop the programming operations when the switching event occurs at $T_{Forming/SET}$ or T_{RESET}

In previous works, different WT circuits have been proposed. In [15], the demonstrated current-mode WT circuit consists of verdict module, write bias module, Self-Adaptive Write Mode (SAWM) module and polarity selector. One can note that this circuit exhibits large area overhead. In other works [16,17], voltage-mode WT circuits are proposed. These circuits are based on the detection of voltage variations that take place on bit-lines when resistive switching occurs. Such designs might have impacts on programming operation-biasing conditions. Ideally, isolating sensing path from programming path is important to reduce the impact of WT measurements on programming operations. Therefore, a WT scheme with small area overhead, fast response time, and robustness against process and ReRAM variabilities is required. In this paper, we propose a novel current mode WT circuit for ReRAM technologies. This WT circuit minimizes power consumption while keeping low area overhead. The demonstrated approach releases on the detection of predefined threshold current during programming operations. When threshold current is detected, the programming operation is interrupted by an implemented logic function. Our WT design is functional with different programming conditions (e.g. Compliance Currents (CC) and programming voltages) with strong immunity to variability. Consequently, such design could be adopted for different ReRAM technologies. Simulated in a 130nm CMOS process and accounting for ReRAM and CMOS variabilities, we demonstrated a reduction of 97+%, 93+% and 65+% in power consumption during Forming, RESET and SET operations respectively. In addition, we strongly believe that our WT strategy reduces ReRAM variability since the same programming current and electrical stress are fixed over all the devices. Variability reduction with WT solutions was demonstrated in previous work [16]. Estimations at array level show that area efficiency of 70% is reachable with the presence of the proposed WT in the memory array overhead.

The rest of this paper is organized as follows. Section II introduces the general background of this paper. Section III presents the proposed WT circuit with its functionality and energy estimations. Section IV introduces the estimation methodology of area efficiency for a memory array with the proposed WT circuit in its near-periphery. Finally, section V concludes the paper.

II. BACKGROUND

This section introduces OxRAM technology adopted in this work with the associated physics-based model.

A. Oxide-based ReRAM (OxRAM) Technology

OxRAM devices have shown good performances among different candidate technologies [18]. Typically, an OxRAM device consists of a binary oxide material (e.g. HfO_x, TaO_x) sandwiched between Top (TE) and Bottom (BE) metallic Electrodes in so-called metal-insulator-metal (MIM)

structure. OxRAM working principle is based on the formation and rupture of a Conductive Filament (CF) through the active oxide layer [19]. The resistance of the memory element can be modulated between two different states: 1) Low Resistance State (LRS) that corresponds to the presence of CF and 2) High Resistance State (HRS) corresponding to the absence of a complete CF. During Forming/SET operations, the CF is created and OxRAM device switches to LRS. Forming operation is required only once during OxRAM lifetime to initialize the device. The Forming and SET operations are similar. However, the Forming requires higher voltage. Over RESET operation, partial dissolution of CF takes place and OxRAM device switches to HRS. The majority of OxRAM devices are bipolar. Therefore, applied voltages with opposite polarities are used to perform the Forming/SET and RESET operations. On one hand, HRS level is controlled by the applied voltage amplitude during RESET operation. On the other hand, LRS level depends on the CC used during Forming/SET operations [9]. Typically, a CMOS transistor is integrated in series with the memory element in 1T1R structure to access the selected cell and to control the CC.

B. *OxRAM Model*

In this work, we have used the OxRAM model presented in [20] to perform our electrical simulations. In this physics-based compact model, resistance modulation is associated to CF radius variation during OxRAM operations. The model was confronted with experimental characteristics and shows excellent agreement with measured data. Thus, this model is appropriate to study the functionality of the proposed WT circuit. Fig.1-a shows the Scanning Electron Microscopy (SEM) cross section of OxRAM technology at which the model was based [21]. The electrical characteristics with simulations for switching times $t_{Forming}$, t_{SET} and t_{RESET} as a function of applied voltages are shown in Fig.1-b. Simulation results reproduce very accurately the experimental data including corners simulations.

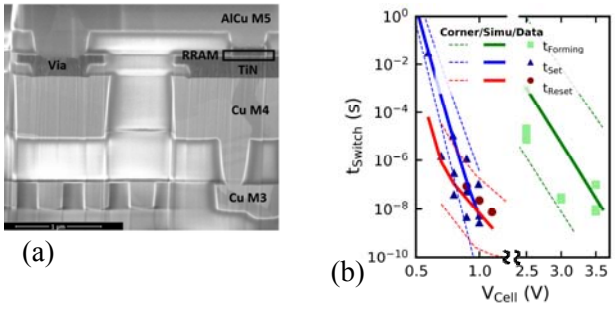

Fig. 1. (a) TEM cross section of TiN/HfO_2/Ti/TiN OxRAM [21]. (b) Measured switching times $t_{Forming}$, t_{Set} and t_{Reset} versus applied programming voltage V_{Cell} (symbols), and the corresponding simulation results with corners (lines). Experimental data was measured on the adopted OxRAM technology in [9,22].

III. PROPOSED WT CIRCUIT AND SIMULATION RESULTS

In this section, we present our WT design with electrical simulation results as well as Monte Carlo (MC) simulations to estimate the energy consumption of the circuit. OxRAM switching times (i.e. $t_{Forming}$, t_{SET} and t_{RESET}) show wide distributions [23]. To guarantee successful programming of all memory cells in OxRAM array, one pulse programming approach is used. The programming pulse width covers worst-case switching time. Fig.2 shows an example of voltage and current waveforms during SET operation of two OxRAM devices with variability in SET time (T_{SET}). SET voltage is continuously applied on cells with T_{SET} smaller than $T_{Worst\text{-}Case\text{-}SET}$ even after resistance switching. With high dc current in LRS, such programming approach wastes a large amount of energy and time in the memory array.

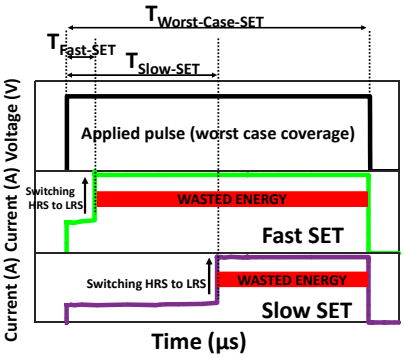

Fig. 2. Current and voltage waveforms during SET operation with different switching times ($T_{Fast\text{-}SET}$ and $T_{Slow\text{-}SET}$). Programming pulse width covers the worst-case SET time ($T_{Worst\text{-}Case\text{-}SET}$)

We proposed a current-mode WT circuit to detect the switching event and to stop dynamically the programming operations. Fig.3 shows the schematic diagram of the designed WT circuit. The circuit consists of: i) two current mirrors (M3/M4 and M6/M7) that copy the current flowing through the selected OxRAM cell in the array, ii) two switches (M1,M2) to select the programming operation either Forming/SET or RESET, iii) two pull-up transistors (M8,M5) with variable gate voltages (V_{bias_RESET} and V_{bias_SET}) to deal with variable current levels and iv) two logic blocks for Forming/SET and RESET termination. During either stand-by mode or Read operation with external sense amplifier, WT circuit is power gated to avoid static leakage.

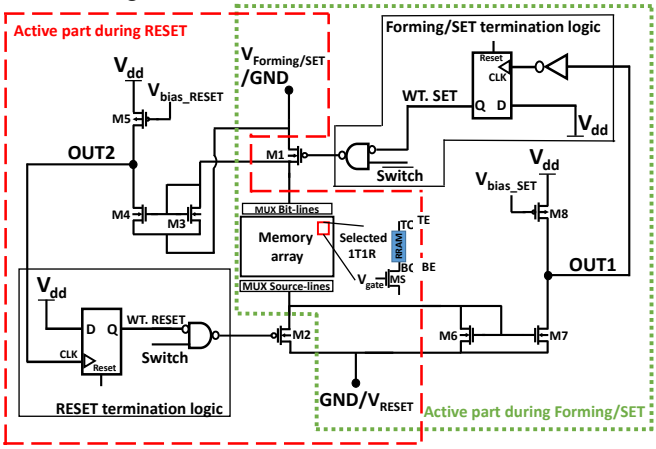

Fig. 3. Schematic diagram of the proposed WT circuit with 1T1R memory array. Active parts of the circuit with respect to programming operation are highlighted using dotted peripheries with different colors (Red for RESET and green for Forming/SET).

In the following, the functionality of the proposed circuit for all possible ReRAM operations (Forming/SET and RESET) is demonstrated. Simulation conditions (i.e. pulse widths, voltage values and CC) were chosen to agree with typical programming conditions for the considered OxRAM technology [24]. The conditions were set as following: Forming ($V_{Forming}$=5V, Pulse width=10µs, CC~120µA), RESET (V_{RESET}=3V, Pulse width=6µs) and SET (V_{SET}=2.6V, Pulse width=100ns, CC~120µA). Programming voltages of typical ReRAM technologies exceed the voltages limits of low voltage CMOS logic.

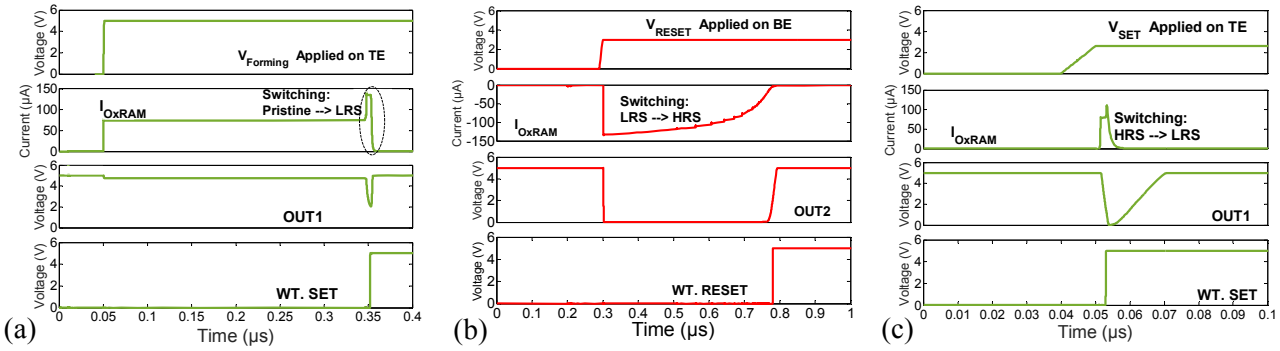

Fig. 4. Simulations waveforms of: applied voltage ($V_{Forming}$, V_{SET} and V_{RESET}), current through OxRAM (I_{OxRAM}), voltage on OUT1 and OUT2 nodes, WT.SET and WT.RESET signals during (a) Forming, (b) RESET and (c) SET operations.

Therefore, to avoid any use of level shifter circuits, all the logic gates of the WT circuit are based on thick oxide CMOS with a V_{dd} of 5V (CMOS node 130nm with middle voltage options) to guarantee the highest gates speed. The different active parts of the circuit during Forming/SET and RESET operations are highlighted in Fig.3 by using dotted lines peripheries.

A. *Forming/SET Operations*

An initialization phase is adopted to reset the flip-flops outputs (WT.RESET=0 and WT.SET=0) before any operation. To perform the Forming/SET operations, programming pulse must be applied on OxRAM TE that corresponds to the source of M1. Figs.4-a,c show the waveforms during Forming/SET operations of: Applied voltage on TE ($V_{Forming}$ and V_{SET}), current through OxRAM (I_{OxRAM}), voltage variation on OUT1 node and WT.SET signal. In this phase, 'Switch' signal is low to select M1 as active switch. The selection transistor of 1T1R structure (MS) controls the CC. When the OxRAM switches from HRS to LRS, the current increases abruptly up to the CC limit, in our case the CC is set to ~120µA. The increase of current is copied with M6/M7 mirror then OUT1 voltage decreases until a positive CLK edge for the active flip-flop is generated. WT.SET signal switches to high leading M1 to cut the current and stop the Forming/SET operation.

B. *RESET Operation*

For the RESET operation the voltage must be applied on OxRAM cell BE node that corresponds to M2 switch. Fig.4-b shows the waveforms of: Applied RESET voltage, current through OxRAM, voltage variation on OUT2 node and WT.RESET signal during RESET operation. Over RESET, the OxRAM switches gradually from LRS to HRS. Thus, the current through the OxRAM decreases gradually to achieve low levels depending on the final resistance value of the memory element. In this phase, there is no need for any current limitation. Therefore, the imposed voltage on MS gate (V_{gate}) is higher compared to V_{gate} during Forming/SET. The current variation during RESET is copied by M3/M4 mirror. As shown in Fig.4-b, initially, I_{OxRAM} current corresponds to the CC achieved during Forming. Thus, OUT2 switches to low. When the current decreases below a given level, OUT2 switches again to high and generates a positive CLK edge for the active flip-flop. WT.RESET signal switches to high and M2 stops the RESET operation. Generally, RESET is a self-termination process. However, with low HRS levels after RESET, large dc current through OxRAM cells still exists. Thus, WT circuit suppresses completely this residual dc current and guarantees zero energy waste in OxRAM cells after the RESET.

C. *Energy Efficiency*

To investigate the energy efficiency with the proposed WT solution, we performed MC simulations and we estimated the energy consumption with enabled and disabled WT circuit. 2000 statistical runs were performed with the programming conditions defined previously in this section. The MC simulations account for CMOS variability (global and local) and OxRAM variabilities in terms of programming voltages and switching time. Memory variability is reproduced very accurately in the compact model adopted in this work [20]. Fig.5 shows the energy distributions resulting from MC simulations during Forming, RESET and SET operations. As expected, the WT solution strongly reduces the energy consumption during Forming and SET operations where the operating currents are considerably high for this technology (tens to hundreds of µA). At 50% of the distributions, energy consumption is reduced by 97% and 65% during Forming and SET operations, respectively (Figs.5-a,c). We demonstrated an energy reduction of 93% at 50% of the distribution during RESET operation for HRS resistance between 70kΩ and 1MΩ (Fig.5-b). This HRS level corresponds to the OxRAM technology considered in this work [19]. Note that for ReRAM technologies with higher HRS levels (e.g tens to hundreds of MΩ) the dc current after RESET operation is self-limited by the memory element. Therefore, the proposed WT circuit will not show significant effects on energy consumption during RESET operation.

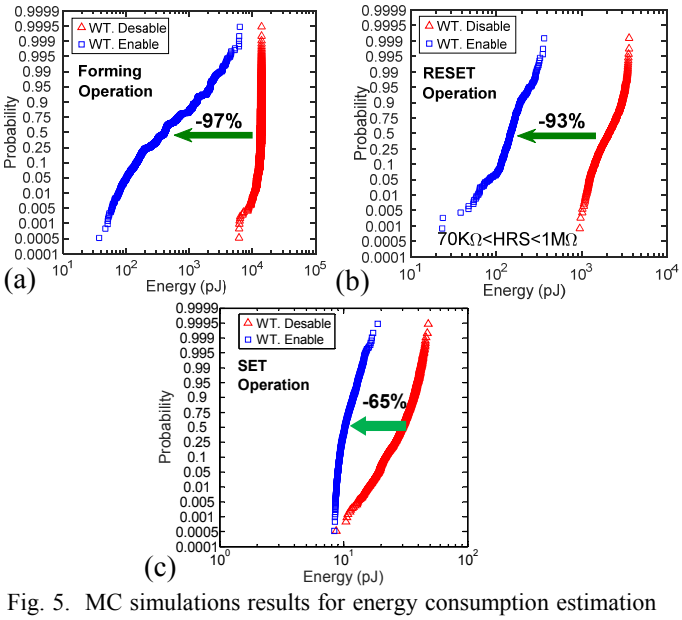

Fig. 5. MC simulations results for energy consumption estimation with enabled and disable WT circuit during (a) Forming, (b) RESET and (c) SET operations.

IV. Near Memory Periphery Exploration

In this section, we explore the effect of such WT circuit at the near-memory periphery level. In that case, we adapted the area estimation of near-array periphery (transistor count and size) methodology from [25] for 1T1R arrays. Fig.6 shows the schematic of subarray architecture used for area estimation. In this design, the array overhead consists of: 1) Word Line (WL) decoder (YDEC) with an area related to WLs number, 2) WL driver, 3) Bit lines (BL) multiplexors (XMUX) (the area of XMUX depends on the programming current (I_{prog}) level. This latter determines the size of XMUX transistors and consequently the area of XMUX), 4) BL decoders (XDEC) and 5) the proposed WT circuit. We defined a *Multiplexing Factor* (MF) as the number of XMUX inputs that correspond to the number of multiplexed BLs. In other words, the MF defines the number of parallel access bits in the array. Thus, the number of WT circuits and XDECs needed to cover the memory array depends on the adopted MF. The output of such estimation methodology is the *memory array efficiency* defined as the ratio between memory array area and its overhead area. In the following paragraph, MUX-N stands for a multiplexing factor of N.

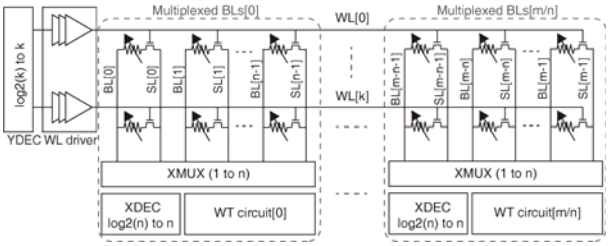

Fig. 6. Block schematic of the proposed subarray architecture with different near-memory periphery blocks

Fig.7-a shows the layout view of the WT active part during Forming/SET operations (highlighted on schematic Fig.3). The total area of the WT circuit is twice to the one shown in Fig.7-a. Therefore, WT calculated circuit area is 425µm². Estimations show that the area efficiency is strongly related to the adopted multiplexing factor. Fig.7-b presents the area efficiency evolution with WT circuit area for various multiplexing factors (4 to 32). In this example, 128x128 memory array in 130nm CMOS technology is considered. With the proposed WT circuit area, more than 70% of area efficiency could be obtained in MUX-32 configuration (4bits in parallel access) whereas a MUX-4 configuration (32bits in parallel access) would lead to ~55% of area efficiency.

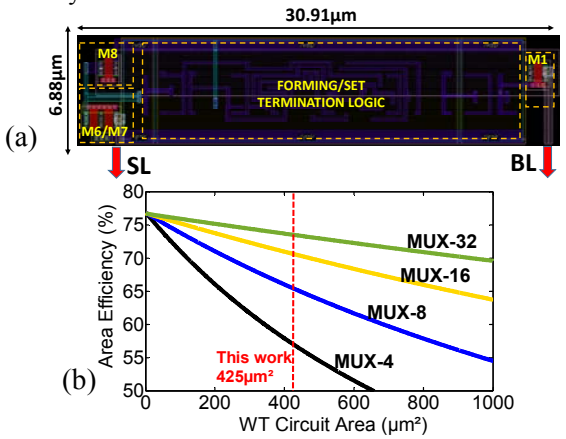

Fig. 7. (a) WT circuit layout of the active part during Forming/SET operations, (b) Array area efficiency versus WT circuit area for different multiplexing factors.

These considerations were extended to the full design space of memory sizes (from 16 to 512 WLs and 32 to 512 BLs). Fig.8 shows the achievable write throughput per memory array that corresponds to the number of parallel access bits (i.e. the ratio between the number of BLs and the MF). To be compliant with standard embedded memories in which 50% to 70% of area efficiency is expected, we considered 70% of area efficiency as target. For each BL/WL couple, array efficiency is determined, and the MF is minimized (maximizing the write throughput) until the efficiency reaches 70%. Two zones were identified: (i) the purple zone corresponds to the memory sizes for which the array efficiency cannot reach 70% and (ii) the red zone corresponds to the zone for which the throughput is limited by the instantaneous current consumption at the beginning of the programming operation. Based on the programming conditions defined in section II (CC~120µA), the corresponding current for 32bits parallel access is 3.8mA per array. In between the two zones, various configurations providing various aspect ratios and write throughout are possible while providing a 70% area efficiency. On the one hand, 1bit-write per array could be achieved in 256x32 array with MUX-256 configuration or 64x256 array with MUX-64 configuration. On the other hand, higher bandwidth requires bigger array size to satisfy the area efficiency criterion of 70% (512x32 array with MUX-16 or 128x416 array with MUX-4 for 32bits parallel access).

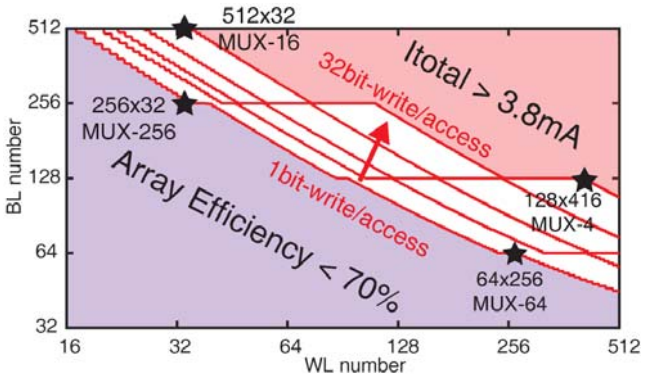

Fig. 8. Write throughput of ReRAM subarray considering the proposed WT circuit and featuring 70% array area efficiency.

V. CONCLUSION

In this work we proposed an innovative Write Termination (WT) circuit for energy and time efficient ReRAM based memory arrays. We demonstrated the WT circuit functionality and reliability through electrical simulations. We performed MC simulations considering CMOS and ReRAM variability for energy estimations during ReRAM operations. We showed that the proposed WT enables 97+%, 93+% and 65+% of energy reduction during Forming, RESET and SET operations respectively. In addition, we extracted the proposed WT circuit area and we estimated the optimal array sizes to maximize the write throughput while achieving 70% array area efficiency. With such smart adaptive programming strategies, energy consumption drastically reduces in ReRAM arrays while keeping high memory density. Finally, we aim to improve the cycling endurance of memory devices with the proposed WT circuit since the devices would experience soft programming conditions. The WT circuit effects on ReRAM's performances and reliability will be investigated on a test chip in a future work.

VI. REFERENCES

[1] G P.-F. Chiu *et al.*, "Low store energy, low VDDmin, 8T2R nonvolatile latch and SRAM with vertical-stacked resistive memory (memristor) devices for low power mobile applications," IEEE Journal of Solid-State Circuits, vol. 47, no. 6, pp. 1483–1496, 2012.

[2] H. Hidaka, "Evolution of embedded flash memory technology for MCU," IEEE International Conference on IC Design & Technology (ICICDT), pp. 1–4, 2011.

[3] T. Endoh *et al.*, "An Overview of Nonvolatile Emerging Memories — Spintronics for Working Memories," IEEE journal on emerging and selected topics in circuits and systems, vol. 6, no. 2, pp. 109–119, 2016.

[4] C. H. Lam, "Storage class memory," IEEE International Conference on Solid-State and Integrated Circuit Technology (ICSICT), pp. 1080–1083, 2010.

[5] H.-S. Philip Wong *et al.*, "Metal–Oxide RRAM," Proceedings of the IEEE, vol. 100, no. 6, pp. 1951–1970, 2012.

[6] H. Wu *et al.*, "Resistive Random Access Memory for Future Information Processing System," Proceedings of the IEEE, vol. 105, no. 9, pp. 1770–1789, 2017.

[7] D. Ielmini, "Resistive switching memories based on metal oxides: mechanisms, reliability and scaling," Semiconductor Science and Technology, vol. 31, no. 6, pp. 1–25, 2016.

[8] B.Govoreanu *et al.*, "10×10nm2 Hf/HfOx crossbar resistive RAM with excellent performance, reliability and low-energy operation," IEEE International Electron Devices Meeting (IEDM), pp. 729–732, 2011.

[9] E. Vianello *et al.*, "Resistive Memories for Ultra-Low-Power embedded computing design," IEEE International Electron Devices Meeting (IEDM), pp. 144-147, 2014.

[10] H. S.-S. Sheu *et al*., "A 4Mb Embedded SLC Resistive-RAM Macro with 7.2ns Read-Write Random-Access Time and 160ns MLC-Access Capability," IEEE International Solid-State Circuits Conference (ISSCC), 2011.

[11] N. Raghavan *et al*., "Stochastic variability of vacancy filament configuration in ultra-thin dielectric RRAM and its impact on OFF state reliability," IEEE International Electron Devices Meeting (IEDM), pp. 554-557, 2013.

[12] S. Ambrogio *et al.*, "Statistical Fluctuations in HfOx Resistive-Switching Memory: Part I - Set/Reset Variability," IEEE Transactions on Electron Devices, vol. 61, no. 8, pp. 2912-2919, 2014.

[13] A. Fantini *et al.*, "Intrinsic Switching Variability in HfO2 RRAM," IEEE International Memory Workshop (IMW), pp. 30-33, 2013.

[14] A. Lee *et al*., "A ReRAM-Based Nonvolatile Flip-Flop With Self-Write-Termination Scheme for Frequent-OFF Fast-Wake-Up Nonvolatile Processors," accepted for inclusion in a future issue of journal of solid-state circuits.

[15] X. Xue *et al*., "A 0.13 μm 8 Mb Logic-Based Cu Si O ReRAM With Self-Adaptive Operation for Yield Enhancement and Power Reduction," IEEE Journal of Solid-State Circuits, vol. 48, no. 5, pp. 1315–1322, 2013.

[16] W.-H. Chen *et al.*, "A 16Mb Dual-Mode ReRAM Macro with Sub-14ns Computing-In-Memory and Memory Functions Enabled by Self-Write Termination Scheme," IEEE International Electron Devices Meeting (IEDM), pp.657-660, 2017.

[17] M.-F. Chang *et al.*, "Low VDDmin Swing-Sample-and-Couple Sense Amplifier and Energy-Efficient Self-Boost-Write-Termination Scheme for Embedded ReRAM Macros Against Resistance and Switch-Time Variations," IEEE Journal of Solid-State Circuits, vol. 50, no. 11, pp. 1–10, 2015.

[18] M. Azzaz *et al*., "Endurance/retention trade off in HfOx and TaOx based RRAM," IEEE International Memory Workshop (IMW) (IMW), 2016.

[19] B. Traore, *et al*., "Microscopic understanding of the low resistance Ztate retention in HfO2 and HfAlO based RRAM," IEEE International Electron Devices Meeting (IEDM), pp. 546-549, 2014.

[20] M. Bocquet *et al*., "Robust Compact Model for Bipolar Oxide-Based Resistive Switching Memories," IEEE transactions on electron devices, vol. 61, no. 3, pp. 674-681, 2014.

[21] D. R. B. Ly *et al*., "Role of synaptic variability in spike-based neuromorphic circuits with unsupervised learning," IEEE International Symposium on Circuits and Systems (ISCAS), 2018.

[22] T. Diokh *et al*., "Investigation of the Impact of the Oxide Thickness and RESET conditions on Disturb in HfO2-RRAM integrated in a 65nm CMOS Technology," IEEE International Reliability Physics Symposium (IRPS), 2013.

[23] G. Sassine *et al.*, "Sub-pJ Consumption and Short Latency Time in RRAM Arrays for High Endurance Applications," IEEE International Reliability Physics Symposium (IRPS), 2018.

[24] A. Grossi *et al*., "Experimental Investigation of 4-kb RRAM Arrays Programming Conditions Suitable for TCAM," accepted for inclusion in a future issue of IEEE transactions on very large scale integration (VLSI) systems journal.

[25] A. Levisse *et al*., "Architecture, Design and Technology Guidelines for Crosspoint Memories," IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH), pp. 55-60, 2017.