Write Termination circuits for RRAM : A Holistic Approach From Technology to Application Considerations

Alexandre Levisse, Marc Bocquet, Marco Rios, Mouhamad Alayan, Mathieu Moreau, Etienne Nowak, Gabriel Molas, Elisa Vianello, David Atienza, Jean-Michel Portal

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 27, 2020, accepted May 24, 2020, date of publication June 8, 2020, date of current version June 23, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3000867

Write Termination Circuits for RRAM: A Holistic Approach From Technology to Application Considerations

ALEXANDRE LEVISSE[1], (Member, IEEE), MARC BOCQUET[2], MARCO RIOS[1], MOUHAMAD ALAYAN[2], MATHIEU MOREAU[2], ETIENNE NOWAK[3], (Member, IEEE), GABRIEL MOLAS[3], (Senior Member, IEEE), ELISA VIANELLO[3], (Member, IEEE), DAVID ATIENZA[1], (Fellow, IEEE), AND JEAN-MICHEL PORTAL[2]

[1]Embedded Systems Laboratory (ESL), EPFL, 1015 Lausanne, Switzerland
[2]CNRS, IM2NP, Aix Marseille University, Université de Toulon, 13453 Marseille, France
[3]CEA-LETI, Minatec Campus, 38054 Grenoble, France

Corresponding authors: Alexandre Levisse (alexandre.levisse@epfl.ch) and Jean-Michel Portal (jean-michel.portal@univ-amu.fr)

This work was supported in part by the ERC Consolidator Grant Computing Server Architecture with Joint Power and Cooling Integration at the Nanoscale (COMPUSAPIEN) under Grant 725657, in part by the EC H2020 Architecting More Than Moore-Wireless Plasticity for Heterogeneous Massive Computer Architectures (WiPLASH) under Grant 863337, and in part by the Agence Nationale de la Recherche grant Réseau Neuronal Binaire à base d'architecture hybride de mémoires intégrant des fonctions de calcul (CMOS/RRAM) pour la fusion de capteurs (NEURONIC) under Grant ANR-18-CE24-0009.

ABSTRACT While Resistive Random Access Memories (RRAM) are perceived nowadays as a promising solution for the future of computing, these technologies suffer from intrinsic variability regarding programming voltage, switching speed and achieved resistance values. Write Termination (WT) circuits are a potential solution to solve these issues. However, previously reported WT circuits do not demonstrate sufficient reliability. In this work, we propose an industrially-ready WT circuit that was simulated with a RRAM model calibrated on real measurements. We perform extensive CMOS and RRAM variability simulations to extract the actual performances of the proposed WT circuit. Finally, we simulate the effects of the proposed WT circuit with memory traces extracted from real Edge-level data-intensive applications. Overall, we demonstrate 2× to 40× of energy gains at bit level. Moreover, we show from 1.9× to 16.2× energy gains with real applications running depending on the application memory access pattern thanks to the proposed WT circuit.

INDEX TERMS RRAM, embedded memories, write termination, memory modeling, low-power design.

I. INTRODUCTION

With the increased need for computing of data analytic and signal processing applications on the edge [1], regular computing architectures appear to not be adapted. The main reason is that they do not provide enough energy efficiency while running such data-intensive workloads, thereby, new architectures and technologies are being explored in the scientific community [2]. An extremely active research field focuses on the integration of emerging non-volatile memory technologies (usually referred as Resistive Random Access Memories or RRAM), as they enable a relatively low cost and fine grain integration with CMOS, a fast switching capability, a higher density and endurance than regular eFlash technologies [3]–[6]. All these characteristics make RRAM technologies an enabler for sub-28nm CMOS technologies microcontrollers [7]–[9] and open the path for new breakthrough memory integration, as data memory or as part of the cache hierarchy for Edge-level computing architectures [10], [11].

However, today the physics of these emerging devices is still not completely known. Thus, circuit designers may propose solutions based on partial knowledge of the device's behavior, which in term might be misleading when benchmarked at the architectural level. Indeed, architectural and application designers have to rely on data and tools released by the technology and circuit community [12]–[14], and usually do not have access to the physical device behavior. It is thus mandatory to close the gap between technology and architecture, and to propose technology aware design and architecture based on accurate device modeling and extensive circuit simulations, taking into account reliability constraints, statistical data about resistance states and timing distributions.

The associate editor coordinating the review of this manuscript and approving it for publication was Cihun-Siyong Gong.

In this work, we propose a holistic approach to RRAM design in which we start from technology characterization and modeling. Based on these characterizations, we then propose an innovative and reliable WT circuit. Since our new WT circuit detects dynamically the switching current with a minimal impact on programming operation, it is mandatory to rely on an extremely well-fitted RRAM compact model. Moreover, this is exacerbated by the need to provide precise simulated energy to the application level. Finally, we assess the gains achieved with this WT circuit while running realistic Edge-level workloads and explore how technology and circuit parameters propagate up to the application level. The main contributions of this work are the following:

- We introduce light modification on the model from [15] and calibrate it with a new extensive set of characterization extracted on a 2k bits 1T1R cell array, with an emphasis on Reset variability.
- We propose, implement and simulate a new WT circuit enabling a dynamic detection and termination of the programming operation. We validate the functionality of the proposed circuit through extensive variability simulations.
- We characterize Edge level applications and simulate the effect of the proposed WT circuit while running these applications in an edge device architecture. We demonstrate from 1.9× to 16.2× energy gains with the considered RRAM technology and proposed WT circuit.
- We explore the design space of RRAM-enabled Edge systems and show that such holistic exploration methodologies can open new perspectives. In particular, we show that depending on the RRAM technology, WT specifications, and application memory access patterns, it is preferable to consider a LRS as a 0 or as a 1 when running a program.

From our previous work, [15] and [16], we upgrade our device model to accurately model variability. Based on this new experimental set of data, we then provide an extensive simulation of our WT circuit and give an evaluation of the proposed WT solution at the architectural level with real application workload.

The remainder of the paper is organized as follows. Section II presents the technology, circuit and architecture background of RRAM-enabled Edge devices. Section III presents the considered RRAM model and its fitting with recent RRAM characterization. Section IV introduces the proposed Write Termination circuit and exhibits simulation results considering CMOS and RRAM variability. Section V presents the Edge level application characterization flow and demonstrates application-level energy gains while using the proposed WT circuit. Finally, Section VI summarizes the main conclusions of the paper.

II. BACKGROUND

In this section, we present the necessary background regarding edge computing paradigm and systems, RRAM technologies and Write Termination circuits.

A. EDGE COMPUTING

Now that it is clear that the current cloud-based infrastructure is not scalable and cannot sustain the rise of deported data-intensive and machine learning workloads [1], new solutions have to be found to be able to run locally these applications. While new computing architectures and accelerators are being explored to cope with the new computing paradigm opened by machine learning workloads [2], [10], [11], [17]–[19], there is no clear solution on how to store the data needed by the application. As a reference, recent Convolutional Neural Networks (CNN) require from few MegaBytes (MB) to hundred of MB [20] of memory to run. While such memory size cannot be efficiently stored in SRAM (due to static leakage) the introduction of a non-volatile memory technology as Storage Class Memory (SCM) inside the memory hierarchy is considered as a potential solution. In this work, we propose to explore an Edge device relying on RRAM-based technology as the main data memory, and we explore the effect of the proposed WT circuit while running real applications.

B. RRAM TECHNOLOGIES

In the last ten years, RRAM technologies have gain interest in the scientific community and are already perceived as a future eFlash technology replacement candidate by several industrials for their advanced technology node microcontrollers [7]–[9], [21], [22]. In this context, three major technology families are under investigation: (i) filamentary RRAM technologies (called Resistive RRAM - ReRAM) including Oxide-Based RRAM (OxRAM), Conductive Bridge RRAM (CBRAM) and hybrid CBRAM-OxRAM technology [3], [5]. (ii) Magnetic Tunel Junction (MTJ)-based RRAM technologies (MRAM) including 2-terminals Spin Transfer Torque (STT) [4], [23] and (iii) Chalcogenide-based Phase Change Memories (PCM) [6]. In this work, we focus on filamentary ReRAM technologies as they feature an extremely cheap co-integration with CMOS thanks to their simple metal-oxide-metal structure. On the other hand, it is important to consider that while there is a huge focus on MRAM technologies lately, recent works actually integrating STT-MRAM technologies [22], [23] do not demonstrate breakthrough performance improvement compared to ReRAM technologies. It should also be noted that RRAM technologies exhibiting low LRS/HRS ratios tend to have a strongly reduced resistance window [24]–[26] and show high sensitivity to CMOS aging effects [27].

Our RRAM technology relies on hafnium oxide (HfO_2) oxide-based resistive Random Access Memory (OxRAM) from [3], [28]. The device stack is composed of a HfO_2 layer and a titanium layer. Both layers have a thickness of ten nanometers, sandwiched between two titanium nitride (TiN) electrodes. Our memory devices are processed within the back-end-of-line of a commercial 130 nanometer CMOS logic process, on top of the fourth metallic layer, as illustrated in Figure 1. We fabricated a 2k bits 1T1R cell array allowing

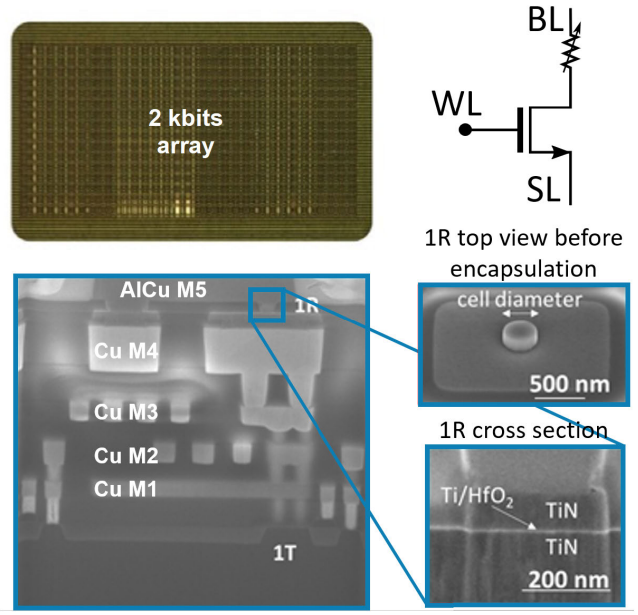


FIGURE 1. Micro-photography of the manufactured 2k bits array used for characterization purpose [15] and scanning electron microscopy image of a cross section of TiN/HfO2/Ti/TiN OxRAM device processed between metal 4 and metal 5 in the back-end-of-line of a commercial 130 nm commercial process [28] .

on one side direct access to each cell in 1T1R configuration and on the other side a differential operating behavior. In this latter case, a bitcell is built with two 1T1R cells programmed in complementary states. The array in its 1T1R configuration is composed of 64 Bit-Line (BL), 32 Word-Line (WL) and 32 Source-Line (SL), providing direct access to each cell (1T1R) for characterization purpose. Indeed decoder logic routes global WL, BL and SL signal, controlled by measurement analyzer, to the selected cell. In this work, extensive characterization measurements are performed on this test chip in 1T1R configuration to calibrate our RRAM compact model.

C. WRITE TERMINATION CIRCUITS

Programming operations in RRAM technologies can be described in the following way: (i) a set operation which corresponds to a switch from a High Resistance State (HRS) down to a Low Resistance State (LRS). (ii) a reset operation that corresponds to a switch from a LRS to a HRS. The electro-forming operation (usually called forming), needed at the beginning of the device lifetime, could be described as a stronger set operation. The main limitations regarding these operations are related to the switching time and to the HRS variability, which tends to exceed several decades. For example, temporal variability is discussed and described in [29], while HRS variability has been extensively studied [3], [26], [28]. Thereby, it is now clearly admitted that RRAM adoption roadblock is mainly related to switching time variability, whatever the programming operation (Forming/Set/Reset) and High Resistance State (HRS) variability due to an inherent stochastic Reset process [28], [30], [31]. To overcome these limitations, numerous research efforts have been spent on technology optimization and recently toward circuit and system-level solutions, with the so-called Write Termination (WT), Self Termination (ST) or Auto-Programming (AP) solutions [21], [32]–[34]. In [32], the demonstrated current-mode WT circuit consists of a 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 [33], [34], voltage-mode WT circuits are proposed. These designs are based on the detection of voltage variations that take place on the memory array bit-lines when resistive switching occurs, with possible impact on programming operation-biasing conditions. To avoid any perturbation on the programming path, more classical write-verify strategies are proposed as in [21]. This write-verify strategy is based on an increase of the programming pulse duration when the read-verify process detects a non-successful cell-state change. This strategy may lead to large programming time overhead for hard to program cells, typically 150ns when tree program/verify sequence has to be applied [21]. Therefore, a WT scheme with small area overhead, dynamic switching detection capability, and robustness against process and RRAM variability is required. This scheme should also isolate the sensing path from the programming path to reduce WT impact on programming conditions.

III. RRAM DEVICE COMPACT MODELING

In this section, we propose an extension to the model published in [15] to expend its range of validity and cover variability effects. The model description is proposed in the next subsection, with emphasis on the light modifications introduced mainly at the reduction and oxidation rate level and at the current calculation level. The second part of this section focuses on the characterization of our 2k bits 1T1R cells array from [35] and model fit to achieve a new model card extraction.

A. RRAM MODEL DESCRIPTION

Forming converts a highly resistive pristine oxide into a switchable sub-oxide region. After this step, standard Set and Reset operations may then occur. Due to the higher voltage bias required during Forming, with respect to Set operation, a CF is generally formed concomitantly to the sub-oxide region after Forming. The growth of the sub-oxide region is controlled by Eq. 1, where $E_{a_{Form}}$ is the activation energy for Forming and τ_{Form0} the nominal Forming rate. Moreover, the Forming process, as an initial step, determines the limit of the switchable sub-oxide region, defined as r_{CFmax} in equation (Eq. 2).

$$\tau_{Form} = \tau_{Form0} \cdot e^{\frac{Ea_{Form} - q \cdot \alpha_{Form} \cdot V_{RRAM}}{k_b \cdot T}} \quad (1)$$

$$\frac{dr_{CF_{max}}}{dt} = \frac{r_{work} - r_{CF_{max}}}{\tau_{Form}} \quad (2)$$

where V_{RRAM} is the voltage applied between the top and the bottom electrodes, q is the elementary charge of the electron, k_b is the Boltzmann constant, T is the temperature in the device.

Similarly to Forming operation, Set operation relies on an electrochemical reaction whose charge transfer rate can be described by the Butler-Volmer equation [36]. From this equation the electrochemical reduction rate τ_{Red} (Eq. 3) (resp. oxidation rate τ_{Ox} (Eq. 4)) can be derived, here k_b denotes the Boltzmann constant, Ea_{Red} (resp. Ea_{Ox}) an activation energy, α_{Red} (resp. α_{Ox}) the charge transfer coefficient (ranging between 0 and 1) and τ_{RedOx} the nominal redox rate.

$$\tau_{Red} = \tau_{RedOx} \cdot e^{\frac{Ea_{Red} - q \cdot \alpha_{Red} \cdot V_{RRAM}}{k_b \cdot T}} \quad (3)$$

$$\tau_{Ox} = \tau_{RedOx} \cdot e^{\frac{Ea_{Ox} + q \cdot \alpha_{Ox} \cdot V_{RRAM}}{k_b \cdot T}} \quad (4)$$

$$\frac{dr_{CF}}{dt} = \frac{r_{CF_{max}} - r_{CF}}{\tau_{Red}} - \frac{r_{CF}}{\tau_{Ox}} \quad (5)$$

The growth/destruction of the filament then results from the interplay between both redox reaction velocities through the master equation Eq. 5, with the CF radius r_{CF} ranging from 0 to r_{CFmax}.

Current computation in the proposed RRAM compact model is based on the sum of three contributions (Eq. 9), namely the pristine current (Eq. 6), the current through the filament (Eq. 8) and the suboxide region current (Eq. 7).

$$I_{Pris} = A_{Pris} \cdot F \cdot \exp\left(B_{Pris} \cdot |F|\right) \quad (6)$$

Before the Forming operation, only the pristine current (Poole-Frenkel) contributes to the overall current in the RRAM, since CF and suboxide region creation is linked to the Forming process.

$$I_{SubOx} = \pi\left(r_{CF_{max}}^2 - r_{CF}^2\right) \cdot A_{SubOx} \cdot F \cdot \exp\left(\alpha_{SubOx} \cdot |F|\right) \quad (7)$$

$$I_{CF} = F \cdot \pi \cdot r_{CF}^2 \cdot \sigma_{CF} \quad (8)$$

$$I_{RRAM} = I_{CF} + I_{SubOx} + I_{Pris} \quad (9)$$

with $S_{SubOx} = \pi r_{CF_{max}}^2$ area of sub-oxide region and $F = \frac{V_{RRAM}}{L_X}$ average electric field.

After the Forming operation, the suboxide region is determined by r_{CFmax} and a Poole-Frenkel current I_{SubOx} is considered in this region, whereas an ohmic current I_{CF} takes place in the CF. Knowing the I_{CF} current, it is possible to express the evolution of the temperature T in the device. It is important to note that the temperature plays a major role in the redox reaction.

$$\frac{\partial T}{\partial t} + T \cdot \frac{R_{th}}{C_{th}} = \frac{I_{CF} \cdot V_{RRAM} \cdot L_X}{S_{SubOx} \cdot C_{th}} \quad (10)$$

B. RRAM ELECTRICAL CHARACTERIZATION AND MODEL CARD EXTRACTION

We performed an extensive electrical characterization of the fabricated 2k bits cells array. The considered bitcells are based on a standard one transistor & one RRAM (1T1R) structure. First, we perform quasi-static electrical characterizations and extract I-V characteristics of the RRAM cells during Forming, Set and Reset operations and reported Figure 2.a. The model (line on Figure 2.a) accurately fits characterized data (symbol on Figure 2.a) on the I-V curves. Then, we performed dynamic characterization to take into account for the voltage - timing dependencies of the RRAM cell for each operation.Figure 2.b, presents the Forming time (t_{form}) versus the Forming pulse amplitude applied on the 1T1R bitcell. The solid line shows simulation data and the box-plot characterized measurements. Here also, the model shows a perfect fit with experimentally extracted data behavior. Furthermore, it follows the classical negative exponential law of the Forming time with the increase of the applied voltage (Vapp Forming) described in [15]. The dependencies of the Set and respectively Reset voltages versus pulse duration are also perfectly modeled as represented Figure 2.c and respectively Figure 2.d.

TABLE 1. Set of model card parameters, defined to fit accurately all reported measurement and to ensure robust simulation convergence.

$L_X = 5nm$	$\sigma_{CF} = 10^5 m \cdot S$	$\sigma_{SubOx} = 1m \cdot S$
$\tau_{Form} = 10^{-9} s^{-1}$	$Ea_{Form} = 1.22eV$	$\alpha_{Form} = 0.33$
$\tau_{Red} = 10^{-13} s^{-1}$	$Ea_{Red} = 1.2eV$	$\alpha_{Red} = 0.9$
$\tau_{Ox} = 5 \cdot 10^{-11} s^{-1}$	$Ea_{Ox} = 0.82eV$	$\alpha_{Ox} = 0.3$
$R_{th} = 80W/(K \cdot m)$	$r_{work} = 10nm$	$\alpha_{SubOx} = 2$
$A_{Pris} = 8 \times 10^{-10} A \cdot nm/V$		$B_{Pris} = 20.0nm/V$
$A_{SubOx} = 2 \cdot 10^{-6} A/(V \cdot nm)$		$C_{th} = 10^{-12} J/(K \cdot m)$

Our solution deals with energy reduction through Write Termination during Set and Reset operations, thus variability of both operations must be considered. Knowing that the LRS variability is mainly due to selector whereas HRS variability is mainly due to a stochastic behavior of the RRAM, a strong characterization and modeling effort is conducted on HRS variability extraction. First of all, the influence of the compliance current during Forming, is studied to extract the evolution of the high resistive state resistance versus the voltage applied to the bitcell structure during Reset. It is worth to note that whatever the applied Reset voltage, the HRS resistance value range is more spread with higher values when a low compliance current is used during Forming. Here also, the model (line) greatly fit the measurement (box-plot), as reported Figure 3.a. Finally, both extracted and simulated HRS distributions are reported on Figure 3.b, for two couple of compliance currents used during Forming operation (55 μA and 200 μA) and applied Reset voltages. These results clearly show the availability of the model to simulate our RRAM technology including HRS variability.

All the simulation results are obtained with a single model card, given Table 1, which is extracted to perfectly capture the RRAM quasi-static and dynamic behavior including the variability of the HRS resistance value.

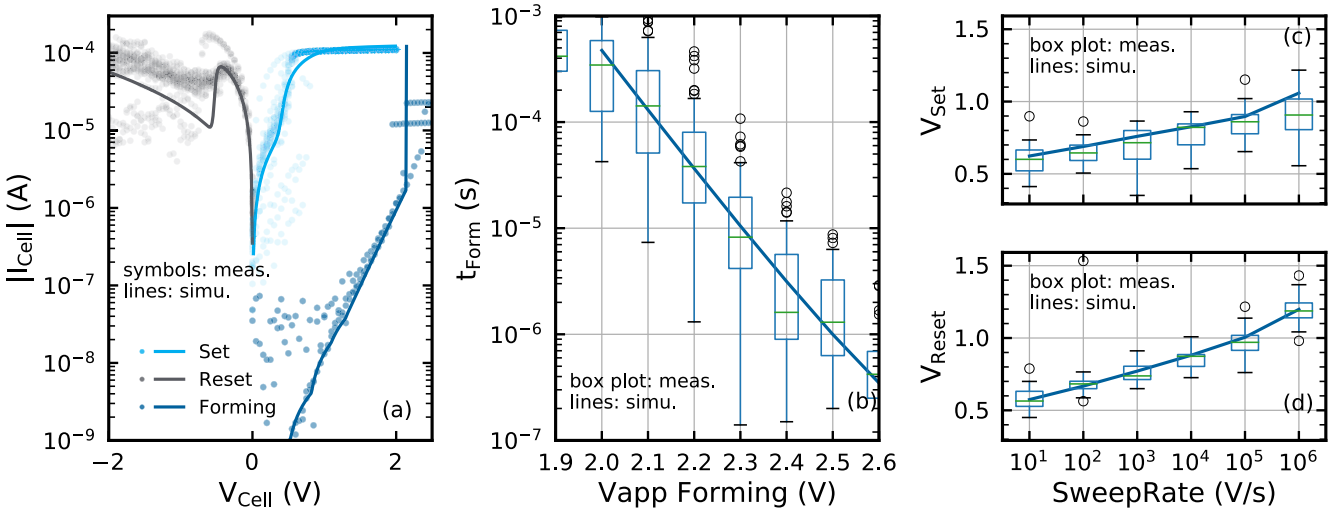


FIGURE 2. (a) Quasi-static measurements (symbol) and simulation (line) of the 1T1R cell current $I_{Cell}(A)$ versus applied voltage $V_{Cell}(V)$ for Forming, Set and Reset operations. (b) Forming time $t_{Form}(s)$ measurements (symbol) and simulation (line) versus applied Forming voltage Vapp Forming (V) (64 devices are considered for each Forming conditions). (c) and (d) are respectively measurements (symbol) and simulation (line) of the threshold voltages $V_{Set}(V)$ and $V_{Reset}(V)$ versus applied SweepRate (V/s) (with 21 and 22 measurements respectively for each ramp speed conditions).

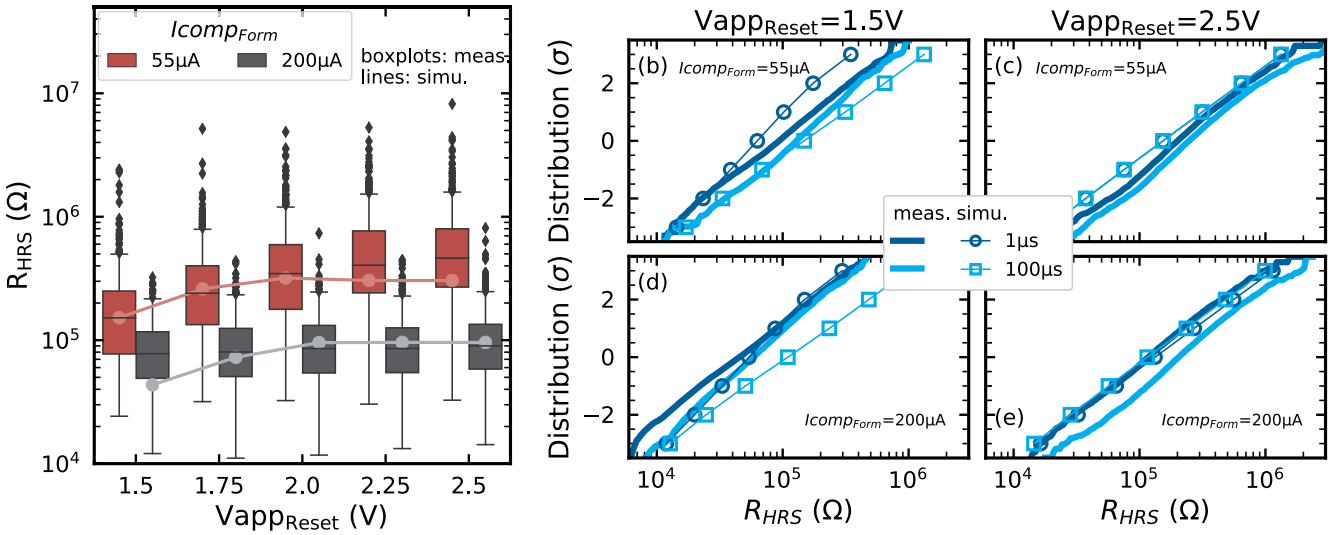


FIGURE 3. (a) $R_{HRS}(\Omega)$ versus applied Reset voltage $Vapp_{Reset}(V)$ (512 devices are measured for each Reset conditions) with measurements (box-plots) and simulations (line) for two different compliance currents (forming and set operation). (b), (c), (d) and (e) are $R_{HRS}(\Omega)$ distributions, with measurements (bold line) and simulations (symbol), for various compliance and Reset voltage conditions.

IV. PROPOSED WRITE TERMINATION CIRCUIT

This section describes the proposed WT circuit architecture, the simulation methodology and simulation results considering the model introduced in Section III.

A. CIRCUIT ARCHITECTURE

The proposed WT circuit and scheme are presented Figure 4. The circuit is connected on one side to the Bit Line (BL) and on the other side to the Source Line (SL), considering a 1T1R array with the bitcells connected between BL and SL. The cell selection is performed in applying a given voltage *Vgate* to the gate of the selection transistor (i.e. the WL).

The proposed WT circuit aims to stop dynamically the Forming, Set or Reset operations when a switching event is detected. Therefore, the current flowing through the programmed RRAM cell is copied and monitored. As the considered RRAM cells are based on a bipolar technology, the WT structure is symmetric. Indeed, one part is activated only during Set (light grey on Figure 4), while the other part is activated only during Reset. Each part of the circuit is composed of:

- one switch transistor M1 for the Set part (respectively M2 for the Reset part),
- one current mirror to monitor the current through the selected cell with a minimal impact on the programming conditions (M6,M7 for the Set/Forming part and M3, M4 respectively for the Reset part)
- one level detection transistor M8, controlled by a trimmed voltage V_{TRIMS}, to take into account change

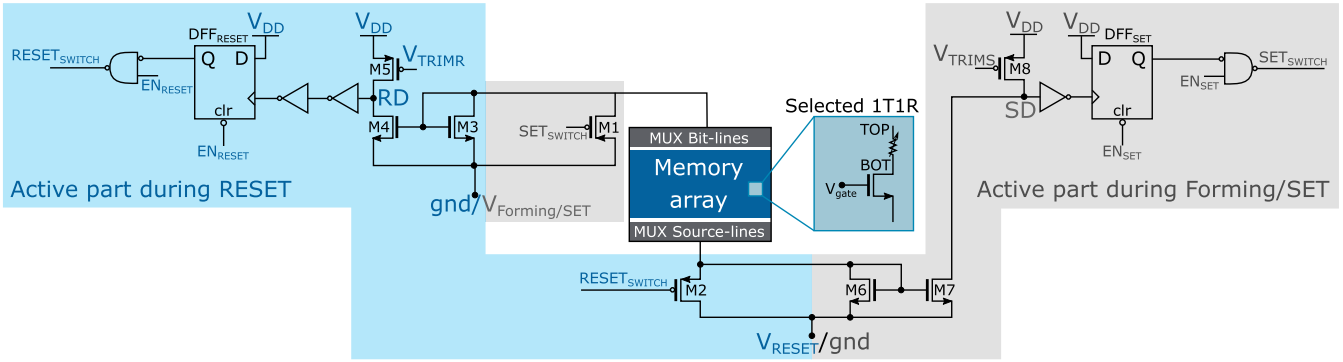


FIGURE 4. Schematic of the write termination circuit associated with a RRAM memory array.

in compliance current condition for the Set part (M5 respectively for the Reset part, with another trimmed voltage V_{TRIMR} to take into account change in Reset conditions and thus HRS level change)
- one D Flip-Flop and a NAND gate to detect a level change on the net SD for the Set part (respectively RD for the Reset part) to switch off the switch transistor M1 for Set part (respectively M2 for the Reset part).

Forming process usually takes place during the Electrical Wafer Sort (EWS) test, after production, thus energy consumption is less of a constraint and the circuit may be not used. However, since Forming is similar to a Set operation but with medium voltage, it must be acknowledged that in case of Forming operation in the field, the circuit can be activated as for the Set operation with a comparable energy gain.

Trimming voltages, namely V_{TRIMS} and V_{TRIMR} are defined at design time for the standard programming conditions given Table 2. Doing so, M8 and respectively M5, are biased in the sub-threshold conduction regime. However, if during the EWS test, programming conditions have to be modified due to light process drift, thus changing programming current; V_{TRIMS} and V_{TRIMR} might be slightly adjusted to keep the same circuit response level. Indeed, if programming currents increase and V_{TRIMS} and V_{TRIMR} are kept unchanged, the WT circuit could stop the programming operation below compliance current. Thus losing the benefit of increasing programming current to adjust LRS level, for example,

Since programming voltages are above the nominal voltage of most of the advanced CMOS technologies, 500nm gate length thick oxide transistors are used for all the elements of the WT circuit. Consequently, the use of level-shifters is avoided. Conjointly, the global circuit V_{DD} is Set to 5V in order to have minimal response time, even with thick oxide transistor. The WT circuit is duplicated depending on the number of bits to be written in parallel, which may have various impact on the area overhead, depending on the multiplexing factor, as reported in our previous work [16]. Finally, during either stand-by or read phase, the WT circuit is power-gated to avoid extra power consumption due to leakage in the structure.

B. FUNCTIONAL VALIDATION

Before describing the WT circuit functionality, it is important to note that the circuit is located between:

- On one side, the BL drivers and the memory array. The BL drivers apply, on the input of the WT circuit connected to the selected BL, either the Forming/Set voltages (through the transistor M1) for a Forming/Set operation or, the ground (through the transistor M3) for a Reset operation.
- On the other side, the SL drivers and the memory array. The SL drivers apply, on the input of the WT circuit connected to the selected SL either, the Reset voltage (through the transistor M2) for a Reset operation, or the ground (through the transistor M6) for a Forming/Set operations.

Moreover, as presented in Figure 4, D Flip-Flops have a clear signal active at a low level, which means that the clear is always activated, except during a Forming/Set operations for the D Flip-Flop belonging to the Set part of the circuit (DFF_{SET}), and respectively during a Reset operation for the D Flip-Flop belonging to the Reset part of the circuit (DFF_{RESET}). This is illustrated in Figure 5.a, with a full view of all the voltage node and the current in the RRAM cell for a Set operation followed by a Reset operation.

The functionality of the structure is described in detail first for a Set operation. The Set operation starts by pulling the node EN_{SET} at V_{DD}, which deactivates the clear input of the DFF_{SET} and pulls down also the node SET_{SWITCH} to the ground). Thus, the Set voltage can be applied on the top electrode (BL) of the selected RRAM cell through M1. Moreover, the transistor M3 is shorted by the transistor M1 and the node EN_{RESET} is grounded, freezing completely the voltages in the part of the circuit activated during the Reset operation. The ground is applied to the node SL, through the transistor M6 and the transistor M2 is cut-off with the $RESET_{SWITCH}$ at V_{DD} (the node EN_{RESET} is grounded). As long as the RRAM cell does not switch, it remains in HRS and the current flowing through the cell remains low. Transistor M7 copies this low current and is biased in the sub-threshold region; the net SD is equal to V_{DD} and the clock input of the D Flip-Flop remains low. When the cell switches to the LRS value the

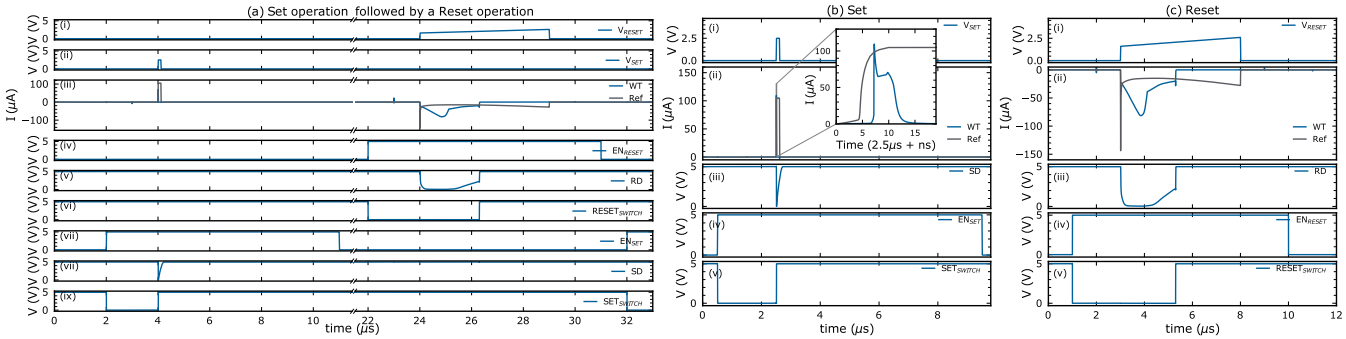


FIGURE 5. Capture of transient simulation waveforms, with all voltage node and current through the selected RRAM during a Set operation followed by a Reset operation (a), focus on a Set operation (b) focus on a Reset operation (c) with (WT) and without write termination (Ref).

current increases to the compliance current limited by the selection transistor (biased at a constant voltage $V_{G_{Set}}$). This current increase is copied by the current mirror (transistors M6, M7), and the net SD voltage decreases; when it reaches the threshold voltage of the inverter a clock edge is generated, thus cutting-off transistor M1 (the node SET_{SWITCH} is set to V_{DD}). The transient simulation waveform of the Set voltage, the current through the cell, the node SD voltage, the node SET_{SWITCH} voltage and the node EN_{SET} voltage are reported Figure 5.b to assess the WT circuit functionality during a Set operation. Comparing the evolution of the current through the cell, with (WT) and without (Ref) WT activation, hence a very high efficiency of the solution is obtained, with a fast dynamic cut-off of transistor M1 in 2.8ns.

In a similar way than for the Set operation, a Reset operation starts with the rise of the signal EN_{RESET} in order to ground the gate of the access transistor M2 ($RESET_{SWITCH}$ is grounded) and to deactivate the input clear of the DFF_{RESET}. In doing so, the Reset voltage can be applied, through the transistor M2, and transmitted to the bottom electrode of the selected RRAM device through the select transistor (NMOS). Moreover, the transistor M6 is shorted by the transistor M2 and the node EN_{SET} is grounded, freezing completely the voltage levels in the part of the circuit activated during Forming/Set operations The ground is applied to the top electrode of the RRAM through transistor M3 and M1 is cut-off since the node SET_{SWITCH} is at V_{DD} with EN_{SET} grounded. As long as the cell does not switch, the resistance value remains low and the current, through the cell, increases with the Reset ramp voltage. Please note, that during a Reset operation the current through the cell is not limited by the selection transistor (biased at a constant voltage $V_{G_{Reset}}$). Since the cell current increase is copied, the node RD is lowered down to the ground. When the cell switch to the HRS value, the current starts to decrease in the cell in a self-limited manner, this implies that RD voltage node rises and when the threshold value of the inverter is reached, a clock edge is generated cutting-off transistor M2. It is important to note, that the Reset process is self-limited by the augmentation of the RRAM resistance value. However depending on the technology, the consumption may remain high. This is typically the case with the RRAM technology considered in our study with R_{HRS} in a range of tens of k-ohms up to M-ohms. Thus, the WT circuit still reduces the energy consumption but with a lower extent than for the Set operation. The transient simulation waveform of the Reset voltage, the current through the cell, the node RD voltage, the node $RESET_{SWITCH}$ and the node EN_{RESET} are reported Figure 5.c to assess the WT circuit functionality during a Reset operation. In comparison to the Set operation, the dynamic cut-off of transistor M2 is slower, due to the slow and self-limited current change during Reset, the cut-off time is 1.4μs. It is important to note that, if we consider a resistive memory technology with R_{HRS} value above 1 Mega ohms, the self-limiting Reset operation cancels the advantages brought by the WT circuit in terms of energy reduction as it will be discussed in the result section.

TABLE 2. Programming conditions applied during the simulation of the circuit with and without WT enabled.

Programming conditions	Value
V_{FORM}	4.5 V
V_{Set}	2.5 V
V_{Reset}	ramp from 1.6 V to 2.6 V
$V_{G_{FORM}}$	2 V
$V_{G_{Set}}$	2.2 V
$V_{G_{Reset}}$	5 V
$Settime$	100 ns
$Resettime$	5 µs

C. RESULTS

To investigate the efficiency of the proposed WT solution, we performed Monte Carlo simulations (1000 runs), while considering CMOS variability (global and local) as well as RRAM variability based on the model depicted section III. Simulations conditions are given Table 2, which includes timing as well as voltage conditions. The efficiency of the solution is established by comparing simulation results of the circuit without WT (Ref) and the circuit with WT (WT) enabled on the following metrics:

- Energy consumption during a Set operation
- Energy consumption during a Reset operation
- LRS distribution

- HRS distribution
- Energy consumption during a programming operation that Set a cell already in LRS
- Energy consumption during a programming operation that Reset a cell already HRS

Moreover, to have a fair comparison between both operation modes (with and without WT) a similar programming window is targeted. Figure 6.a gives the Set energy cumulative distribution of the circuit with WT enabled (WT) and without (Ref). A substantial energy reduction (i.e., energy is divided by more than 39 times considering median value) is obtained when the WT circuit dynamically stops the application of the programming pulse. Knowing that Set variability increases with fast switching time (i.e., the variability of the Set voltage necessary to switch a cell to LRS as described Figure 2.c, section III), some safety margins have to be taken on pulse length. Here Set pulse duration is chosen to be 100ns in order to successfully Set all cells. One can note that this duration is coherent with the state of the art [21]. Figure 6.b shows the Reset energy cumulative distribution for both operation modes. The situation, in this case, is slightly different than for the Set, since the Reset process is self-limited. The activation of the WT circuit brings only an energy reduction of 1.4 times for the median value. To analyze this result, Figure 7 shows the Reset energy versus the ratio of R_{HRS}/R_{LRS} and exhibits two different cell populations. Considering the Ref, the first population is defined by a RRAM cell having a low HRS/LRS ratio (below 50). This population due to low HRS has a poor current self-limitation capability after the switching event and thus WT is efficient to limit the consumption by stopping dynamically the Reset ramp. Considering also the Ref, the second population is identified by cells having a large HRS/LRS ration (above 50) with a strong current self-limitation ability. In this case the WT does not bring significant advantage and is even costly for 20% of the cells with the lowest consumption. The different efficiency of the WT circuit on the two populations also explain the bend on the Reset energy distribution in Figure 6.b. Finally, in the resistance distributions presented in Figure 8, the WT has a negligible impact on LRS distribution, but allows to have a more sharp HRS distribution, opening the way to a more homogeneous stress during Set and thus a potential more homogeneous aging.

D. MULTIPLE PULSE PROGRAMMING

To have the full picture of the WT efficiency, we also have to consider write access to a cell already in the targeted state. In other words, we simulated the proposed WT circuit when a Set operation is performed on a cell already in LRS and when a Reset operation is performed on a cell already in HRS. Interestingly for both cases, similar gains are obtained in Figure 8 compared to the switching case in Figure 6. Indeed, if the cell is already in LRS the current reaches directly the compliance current and the WT is directly activated. On the other hand, if the cell is already in HRS either the current reaches the WT threshold and the WT is activated, either the current level is already below the threshold and Reset operation further increases the HRS value self-limiting the consumption. The energy consumption obtained from circuit-level simulations is the basement for the estimation of real workload applications, as described in the next section.

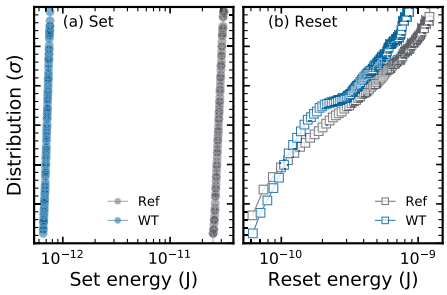


FIGURE 6. (a) and (b) are respectively the Set energy (J) distribution and the Reset energy (J) distribution with (WT) and without write termination (Ref).

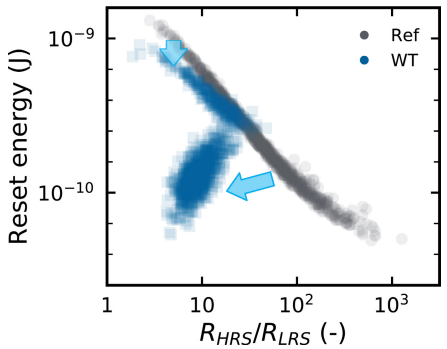


FIGURE 7. Reset energy (J) versus R_{HRS}/R_{LRS} ratio, with (WT) and without write termination (Ref).

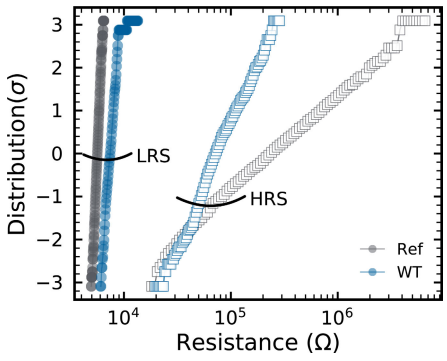


FIGURE 8. $R_{HRS}(\Omega)$ and $R_{LRS}(\Omega)$ distribution with (WT) and without write termination (Ref).

V. REAL WORKLOAD EXPLORATION

In order to assess the potential performance improvement of the proposed WT programming methodology, we perform application-level simulations considering real Edge-level applications workloads. Hence, we propose the following methodology: (i) We extract memory traces from real Edge-level applications. (ii) We profile the memory traces to identify different access patterns. (iii) We extract the energy gains considering the system with and without the proposed

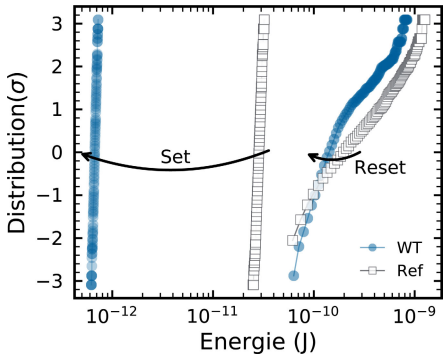


FIGURE 9. Set and Reset energy distribution with (WT) and without write termination (Ref), when the bitcell is already in the targeted state LRS or HRS.

WT, and explore trade-off regarding the programming conditions and the application memory access pattern.

A. APPLICATION CHARACTERIZATION

We consider an edge-level architecture embedding an ARM-based low-power processor and a RRAM-based memory. We thereby extract real memory traces from edge level applications. To do so, we run real C-Code applications and track the variables kept in memory with a methodology analogous to [37]. We take here the assumption that the local variables can be handled by the core registers (e.g., there are 12 general purpose registers in an ARM M0 core) and we neglect the effect of the save and restore of the stack. The different applications considered in this work are described in the following:

- Data compression algorithm: Compress Sensing (CS) algorithm is a highly efficient compression algorithm widely used in biosignal processing applications [38] for data storage. We consider here the compression of a 3seconds 1-lead Electrocardiogram. The interest of this application is that it shows a balanced read-write pattern and a random memory access pattern.
- Machine learning algorithm: Epilepsy Seizures Detection [39] is an industry-ready algorithm running on connected glasses and determining from Electroencephalogram (EEG) waves a soon-to-occur epileptic seizure. It relies on a two-steps processing: (i) a Feature Extraction (FE) on 4 seconds of EEG signal and (ii) a Decision Tree (DT) classification based on a random forest algorithm.
- Application-specific kernels: We consider two family of data-intensive-specific application kernels widely used in both machine learning and signal processing [40], [41]. We consider a matrix multiplication algorithm involving random 30 × 30 arrays and a convolution between random 3 × 3 and 30 × 30 arrays. For both, we consider two cases: full and sparse. That way, we can compare the effect of data sparsity [42] on the system execution.

Figure 10 presents the proportions of read and write operations for the applications considered in this work. While most of these applications exhibit around 20% of write operations,

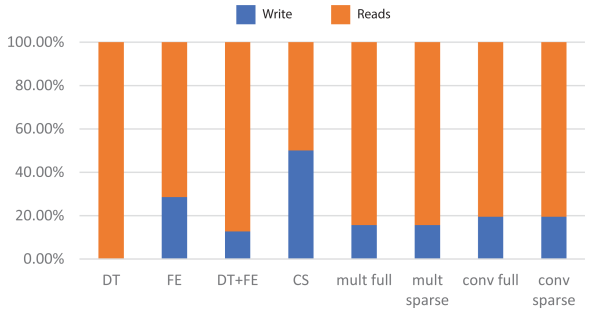


FIGURE 10. Normalized ratio of reads and write in the memory for the considered edge-level applications.

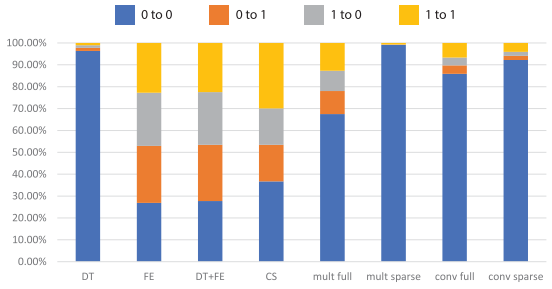


FIGURE 11. Normalized share of written bits during the execution of the considered Edge-level applications.

CS exhibits an equalized read/write pattern of 50%. On the other hand, the decision tree (DT) application shows around 0.3% of write operations. Based on these data, we could expect application-level gains to follow the read/write ratio, and we could expect very limited gains for the DT application, while CS would show stronger energy gains.

Figure 11 shows the details of the write operations for the aforementioned applications. for each memory write operation, we compute the number of bit switches and generate statistics. Thereby, for 32bits words, we represent in blue (respectively in yellow) the amount of non-changed bits from 0 to 0 (respectively 1 to 1). And we represent the switching bits in orange (respectively grey) from 0 to 1 (respectively 1 to 0). With this approach, we exhibit a high variability from one application to another. As an example, FE, DT+FE and CS show a highly balanced memory write pattern between the different types of write operations. On the other hand, the DT application, matrix multiplication kernel and convolution kernel show a high majority of 0 to 0 write operations. For application-specific kernels, a clear difference is visible between full and sparse data. Furthermore, kernels running with sparse data exhibit more 0 to 0 writes than full data as they induce smaller numeric values, i.e., fewer MSBs are changed to ones in full data.

Figure 12 presents the gains achieved by the proposed WT circuit when running the considered applications. Two different gain curves are presented: (i) in grey, the gains when LRS are considered as ones (respectively HRS as zeros). (ii) In green, the gains when LRS are zeros (respectively HRS as ones). It is interesting to note here that the presented trend is highly correlated to the considered memory technology and WT circuit. In this work, as extensively described in

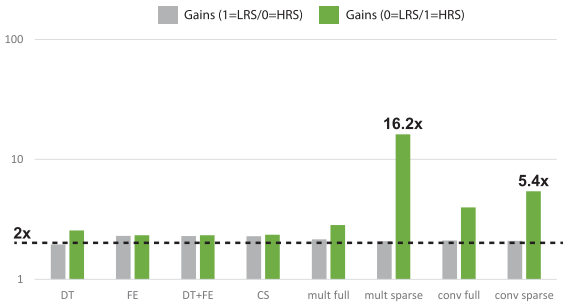


FIGURE 12. Gains provided by the WT circuit versus solution without WT.

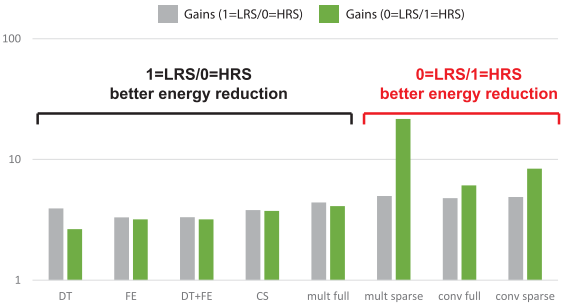


FIGURE 13. Gains provided by the WT circuit when considering balanced set/reset operation energy consumption for various applications patterns.

the first part of this paper (section IV), programming operations are highly unbalanced (i.e., set and reset operations exhibit extremely different energy and time). In that sense, as described Figure 12, in order to accommodate for a bad reset operation, it is preferable to keep the most used memory state (0 in the considered applications) as a LRS. Thereby, it enables up to 16.2× for specific application cases, while the gains would be limited to 2× if 0 is considered as a HRS.

Then, as an exploration case, Figure 13 shows a more balanced case for which reset operations are less pessimistic. In this situation, it appears that the optimal case shifts and it becomes more profitable to consider 1 as a LRS for DT, FE, DT+FE and when performing full matrix multiplication. On the other hand, sparse matrix multiplication and convolutions still show better energy reduction due to the huge proportion of 0 to 0 operations these applications exhibit. Such configurations open the way for adaptive memory configuration enabling a switch from one configuration to another depending along with the application execution.

VI. CONCLUSION

In this work, we have proposed an innovative and reliable Write Termination (WT) circuit for RRAM technologies. We have benchmarked its performance through a holistic methodology going from technology characterization and modeling, toward circuit design and simulation, and finally, application memory traces extraction, characterization and simulation with the proposed WT circuit. Overall, in this work, we have characterized a 2kbit RRAM memory array and fitted a model accounting for RRAM variability. Then, based on this model, we have demonstrated that the proposed WT circuit can improve the energy efficiency of bit-level set (respectively reset) operation by 40x (respectively 2x) and that such gains can be visible while running real applications. Also, we have shown that, depending on the application memory access pattern, the gain can vary from 1.9x up to 16.2x. Finally, we have shown that such a cross-layer study can pave the way for new optimization methodologies mixing technology, circuit and application considerations.

ACKNOWLEDGMENT

(Alexandre Levisse and Marc Bocquet contributed equally to this work.)

REFERENCES

[1] B. Reese. (2019). *Ai At the Edge: A GigaOm Research Byte*. GigaOm. [Online]. Available: https://gigaom.com/report/ai-at-the-edge-a-gigaom-research-byte/
[2] J. Chen and X. Ran, "Deep learning with edge computing: A review," *Proc. IEEE*, vol. 107, no. 8, pp. 1655–1674, Aug. 2019.
[3] E. Vianello, O. Thomas, M. Harrand, S. Onkaraiah, T. Cabout, B. Traore, T. Diokh, H. Oucheikh, L. Perniola, G. Molas, P. Blaise, J. F. Nodin, E. Jalaguier, and B. De Salvo, "Back-end 3D integration of HfO_2-based RRAMs for low-voltage advanced IC digital design," in *Proc. Int. Conf. IC Design Technol. (ICICDT)*, May 2013, pp. 235–238.
[4] D. Apalkov, B. Dieny, and J. M. Slaughter, "Magnetoresistive random access memory," *Proc. IEEE*, vol. 104, no. 10, pp. 1796–1830, Oct. 2016.
[5] H.-S. P. Wong, H.-Y. Lee, S. Yu, Y.-S. Chen, Y. Wu, P.-S. Chen, B. Lee, F. T. Chen, and M.-J. Tsai, "Metal–oxide RRAM," *Proc. IEEE*, vol. 100, no. 6, pp. 1951–1970, Jun. 2012.
[6] H.-S. P. Wong, S. Raoux, S. Kim, J. Liang, J. P. Reifenberg, B. Rajendran, M. Asheghi, and K. E. Goodson, "Phase change memory," *Proc. IEEE*, vol. 98, no. 12, pp. 2201–2227, Dec. 2010.
[7] F. Disegni *et al.*, "Embedded PCM macro for automotive-grade microcontroller in 28nm FD-SOI," in *Proc. Symp. VLSI Circuits*, Jun. 2019. pp. C204–C205.
[8] *Reram Embedded Super Low-Power Consumption MCU MN101L*. Accessed: Jun. 10, 2020. [Online]. Available: https://industrial.panasonic.com/ww/products/semiconductors/microcomputers/mn1011
[9] A. Kawahara, K. Kawai, Y. Ikeda, Y. Katoh, R. Azuma, Y. Yoshimoto, K. Tanabe, Z. Wei, T. Ninomiya, K. Katayama, R. Yasuhara, S. Muraoka, A. Himeno, N. Yoshikawa, H. Murase, K. Shimakawa, T. Takagi, T. Mikawa, and K. Aono, "Filament scaling forming technique and level-verify-write scheme with endurance over 107 cycles in ReRAM," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2013, pp. 220–221.
[10] A. Pullini, D. Rossi, I. Loi, A. Di Mauro, and L. Benini, "Mr. Wolf: A 1 GFLOP/s energy-proportional parallel ultra low power SoC for IOT edge processing," in *Proc. IEEE 44th Eur. Solid State Circuits Conf. (ESSCIRC)*, Sep. 2018, pp. 274–277.
[11] E. Flamand, D. Rossi, F. Conti, I. Loi, A. Pullini, F. Rotenberg, and L. Benini, "GAP-8: A RISC-V SoC for AI at the edge of the IoT," in *Proc. IEEE 29th Int. Conf. Appl.-Specific Syst., Archit. Processors (ASAP)*, Jul. 2018, pp. 1–4.
[12] X. Dong, C. Xu, Y. Xie, and N. P. Jouppi, "NVSim: A circuit-level performance, energy, and area model for emerging nonvolatile memory," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 31, no. 7, pp. 994–1007, Jul. 2012.
[13] H. Labs. *Cacti an Integrated Cache and Memory Access Time, Cycle Time, Area, Leakage, and Dynamic Power Model*. Accessed: Jun. 10, 2020. [Online]. Available: https://www.hpl.hp.com/research/cacti/
[14] D. M. Mathew, A. L. Chinazzo, C. Weis, M. Jung, B. Giraud, P. Vivet, A. Levisse, and N. Wehn, "RRAMSpec: A design space exploration framework for high density resistive ram," in *Embedded Computer Systems: Architectures, Modeling, and Simulation*, D. N. Pnevmatikatos, M. Pelcat, and M. Jung, Eds. Cham, Switzerland: Springer, 2019, pp. 34–47.
[15] M. Bocquet, D. Deleruyelle, H. Aziza, C. Muller, J.-M. Portal, T. Cabout, and E. Jalaguier, "Robust compact model for bipolar oxide-based resistive switching memories," *IEEE Trans. Electron Devices*, vol. 61, no. 3, pp. 674–681, Mar. 2014.
[16] M. Alayan, E. Muhr, A. Levisse, M. Bocquet, M. Moreau, E. Nowak, G. Molas, E. Vianello, and J. M. Portal, "Switching event detection and self-termination programming circuit for energy efficient ReRAM memory arrays," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 66, no. 5, pp. 748–752, May 2019.

[17] A. S. J. Levisse, M. A. Rios, W. A. Simon, P.-E. J. M. Gaillardon, and D. A. Alonso. (2019). *Functionality Enhanced Memories for Edge-Ai Embedded Systems*. p. 4. [Online]. Available: http://infoscience.epfl.ch/record/272717

[18] W. A. Simon, Y. M. Qureshi, M. Rios, A. Levisse, M. Zapater, and D. Atienza, "An in-cache computing architecture for edge devices," *IEEE Trans. Comput.*, early access, Feb. 10, 2020, doi: 10.1109/TC.2020.2972528.

[19] ARM. *ARM Neon Technology*. Accessed: Jun. 10, 2020. [Online]. Available: https://developer.arm.com/architectures/instruction-sets/simd-isas/neon

[20] S. Bianco, R. Cadène, L. Celona, and P. Napoletano, "Benchmark analysis of representative deep neural network architectures," *IEEE Access*, vol. 6, pp. 64270–64277, Oct. 2018.

[21] P. Jain, U. Arslan, M. Sekhar, B. C. Lin, L. Wei, T. Sahu, J. Alzate-vinasco, A. Vangapaty, M. Meterelliyoz, N. Strutt, A. B. Chen, P. Hentges, P. A. Quintero, C. Connor, O. Golonzka, K. Fischer, and F. Hamzaoglu, "13.2 a 3.6Mb 10.1Mb/mm2 embedded non-volatile ReRAM macro in 22nm FinFET technology with adaptive forming/set/reset schemes yielding down to 0.5 V with sensing time of 5ns at 0.7 V," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2019, pp. 212–214.

[22] L. Wei *et al.*, "13.3 A 7Mb STT-MRAM in 22FFL FinFET technology with 4 ns read sensing time at 0.9 V using write-verify-write scheme and offset-cancellation sensing technique," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2019, pp. 214–216, doi: 10.1109/ISSCC.2019.8662444.

[23] Q. Dong, Z. Wang, J. Lim, Y. Zhang, Y.-C. Shih, Y.-D. Chih, J. Chang, D. Blaauw, and D. Sylvester, "A 1Mb 28nm STT-MRAM with 2.8ns read access time at 1.2 V VDD using single-cap offset-cancelled sense amplifier and *in-situ* self-write-termination," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2018, pp. 480–482.

[24] C. Y. Chen, A. Fantini, L. Goux, R. Degraeve, S. Clima, A. Redolfi, G. Groeseneken, and M. Jurczak, "Programming-conditions solutions towards suppression of retention tails of scaled oxide-based RRAM," in *IEDM Tech. Dig.*, Dec. 2015, pp. 10.6.1–10.6.4.

[25] C. Nail *et al.*, "Understanding RRAM endurance, retention and window margin trade-off using experimental results and simulations," in *IEDM Tech. Dig.*, Dec. 2016, pp. 4–5.

[26] A. Grossi, E. Vianello, C. Zambelli, P. Royer, J.-P. Noel, B. Giraud, L. Perniola, P. Olivo, and E. Nowak, "Experimental investigation of 4-kb RRAM arrays programming conditions suitable for TCAM," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 26, no. 12, pp. 2599–2607, Dec. 2018.

[27] A. Levisse, M. Rios, M. Peon-Quiros, and D. Atienza, "Exploration methodology for BTI-induced failures on RRAM-based edge AI systems," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2020.

[28] A. Grossi, E. Nowak, C. Zambelli, C. Pellissier, S. Bernasconi, G. Cibrario, K. El Hajjam, R. Crochemore, J. F. Nodin, P. Olivo, and L. Perniola, "Fundamental variability limits of filament-based RRAM," in *IEDM Tech. Dig.*, Dec. 2016, pp. 4.7.1–4.7.4.

[29] G. Sassine, C. Nail, L. Tillie, D. A. Robayo, A. Levisse, C. Cagli, K. El Hajjam, J.-F. Nodin, E. Vianello, M. Bernard, G. Molas, and E. Nowak, "Sub-pJ consumption and short latency time in RRAM arrays for high endurance applications," in *Proc. IEEE Int. Rel. Phys. Symp. (IRPS)*, Mar. 2018, pp. P-MY.2-1–P-MY.2-5.

[30] N. Raghavan, R. Degraeve, A. Fantini, L. Goux, D. J. Wouters, G. Groeseneken, and M. Jurczak, "Stochastic variability of vacancy filament configuration in ultra-thin dielectric RRAM and its impact on OFF-state reliability," in *IEDM Tech. Dig.*, Dec. 2013, pp. 21.1.1–21.1.4.

[31] S. Ambrogio, S. Balatti, A. Cubeta, A. Calderoni, N. Ramaswamy, and D. Ielmini, "Statistical fluctuations in HfOx resistive-switching memory: Part I—Set/reset variability," *IEEE Trans. Electron Devices*, vol. 61, no. 8, pp. 2912–2919, Aug. 2014.

[32] X. Xue, W. Jian, J. Yang, F. Xiao, G. Chen, S. Xu, Y. Xie, Y. Lin, R. Huang, Q. Zou, and J. Wu, "A 0.13 μm 8 Mb logic-based Cu_x Si_yO ReRAM with self-adaptive operation for yield enhancement and power reduction," *IEEE J. Solid-State Circuits*, vol. 48, no. 5, pp. 1315–1322, May 2013.

[33] M.-F. Chang, J.-J. Wu, T.-F. Chien, Y.-C. Liu, T.-C. Yang, W.-C. Shen, Y.-C. King, C. J. Lin, K.-F. Lin, Y.-D. Chih, and J. Chang, "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 J. Solid-State Circuits*, vol. 50, no. 11, pp. 2786–2795, Nov. 2015.

[34] W.-H. Chen, W.-J. Lin, L.-Y. Lai, S. Li, C.-H. Hsu, H.-T. Lin, H.-Y. Lee, J.-W. Su, Y. Xie, S.-S. Sheu, and M.-F. Chang, "A 16Mb dual-mode ReRAM macro with sub-14ns computing-in-memory and memory functions enabled by self-write termination scheme," in *IEDM Tech. Dig.*, Dec. 2017, pp. 28.2.1–28.2.4.

[35] M. Bocquet, T. Hirztlin, J.-O. Klein, E. Nowak, E. Vianello, J.-M. Portal, and D. Querlioz, "In-memory and error-immune differential RRAM implementation of binarized deep neural networks," in *IEDM Tech. Dig.*, Dec. 2018, pp. 20.6.1–20.6.4.

[36] A. J. Bard and L. R. Faulkner, *Electrochemical Methods: Fundamentals and Applications*, 2nd ed. Hoboken, NJ, USA: Wiley, 2000, p. 864.

[37] A. Bartzas, M. Peon-Quiros, C. Poucet, C. Baloukas, S. Mamagkakis, F. Catthoor, D. Soudris, and J. M. Mendias, "Software metadata: Systematic characterization of the memory behaviour of dynamic applications," *J. Syst. Softw.*, vol. 83, no. 6, pp. 1051–1075, Jun. 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0164121210000026

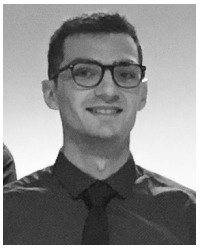
[38] J. Constantin, A. Dogan, O. Andersson, P. Meinerzhagen, J. N. Rodrigues, D. Atienza, and A. Burg, "TamaRISC-CS: An ultra-low-power application-specific processor for compressed sensing," in *Proc. IEEE/IFIP 20th Int. Conf. VLSI Syst.-on-Chip (VLSI-SoC)*, Oct. 2012, pp. 159–164.

[39] D. Sopic, A. Aminifar, and D. Atienza, "E-glass: A wearable system for real-time detection of epileptic seizures," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2018, pp. 1–5.

[40] A. Vasudevan, A. Anderson, and D. Gregg, "Parallel multi channel convolution using general matrix multiplication," in *Proc. IEEE 28th Int. Conf. Appl.-Specific Syst., Archit. Processors (ASAP)*, Jul. 2017, pp. 19–24.

[41] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[42] R. Spring and A. Shrivastava, "Scalable and sustainable deep learning via randomized hashing," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*. New York, NY, USA: Association for Computing Machinery, Aug. 2017, pp. 445–454, doi: 10.1145/3097983.3098035.



ALEXANDRE LEVISSE (Member, IEEE) received the Ph.D. degree in electrical engineering from CEA-LETI, France, and Aix-Marseille University, France, in 2017. He is currently a Postdoctoral Researcher with the Embedded Systems Laboratory, Swiss Federal Institute of Technology Lausanne (EPFL). His research interests include circuits and architectures for emerging memory and transistor technologies, 3D stacked architectures, and in-memory computing and accelerators.

MARC BOCQUET received the M.S. and Ph.D. degrees in electrical engineering from the University of Grenoble, France, in 2006 and 2009, respectively. He is currently an Associate Professor with the Institute of Materials, Microelectronics, and Nano-Sciences of Provence (IM2NP), Univeriste Aix-Marseille. His research interests include memory model, memory design, characterization, and reliability.

MARCO RIOS received the master's degree in computer science and electronics for embedded systems from Université Grenoble Alpes, in 2018. He is currently pursuing the Ph.D. degree in electrical engineering with the Embedded Systems Laboratory, Swiss Federal Institute of Technology Lausanne (EPFL). His research interests include design of integrated systems and circuits, in-SRAM computing, 3D stacked technologies, and the system impact of emerging memories.

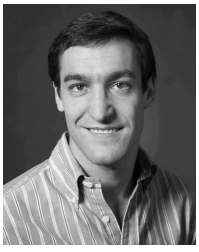
MOUHAMAD ALAYAN received the Ph.D. degree in nanoelectronics and nanotechnologies from the University of Grenoble Alpes, France. He is currently a Memory Design Engineer with ARM, France. He also works as a Researcher with the IM2NP Laboratory, Marseille, France. His research interests include modeling and characterization of emerging nonvolatile memory devices, and memory circuits design.

MATHIEU MOREAU received the Master of Science and Ph.D. degrees in micro and nanoelectronics from Aix-Marseille University, France, in 2007 and 2010, respectively. His Ph.D. research at the Institute of Materials Microelectronics and Nanosciences of Provence (IM2NP) covered numerical simulation and compact modeling of advanced nano-devices, like FinFET, based on new materials (high-k and III–V semiconductors). From 2010 to 2011, he was a Teaching Assistant with polytech'Marseille. Since 2012, he has been an Associate Professor with Aix-Marseille University and conducts his research at IM2NP in the field of hybrid circuit design based on emerging non-volatile memories (ReRAM, MRAM, ...). He was a recipient of the Newcas 2013 Best Paper Award and the 2017 IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS Guillemin-Cauer Best Paper Award.

ETIENNE NOWAK (Member, IEEE) received the M.Sc. degree in microelectronics from Grenoble University, Grenoble, France; Polito di Torino, Turin, Italy; and the École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, in 2007, and the Ph.D. degree from the Institute National Polytechnique de Grenoble, Grenoble, in 2010. From 2010 to 2014, he was a Senior Engineer with the Semiconductor Research and Development Center, Samsung Electronics, Hwaseong, South Korea, where he was involved in the first generations of vertical NAND flash memory. He joined CEA-LETI, Grenoble, in 2014, as a Project Manager on emerging nonvolatile memory. He has published over 80 publications and holds two patents on these topics. Since 2017, he has been appointed as the Head of the Advanced Memory Device Laboratory, CEA-LETI, dedicated to nonvolatile memory backend technologies.

GABRIEL MOLAS (Senior Member, IEEE) received the B.S. and M.S. degrees in physics engineering, with microelectronics specialization, and the Ph.D. degree in micro- and nano-electronics from the Polytechnics Institute of Grenoble, France, in 2001 and 2004, with a thesis on few electron memories. He joined the Laboratory of Electronics, Technology and Instrumentation (LETI), CEA, Grenoble, as a Research Engineer, in 2004. He is/was responsible of various industrial and institutional projects. Since 2016, he has been an LETI Senior Expert on non-volatile memories. Since 2018, he has been in charge of LETI RRAM activity. In 2020, he was also nominated scientific delegate of microelectronic service of LETI. He is the author or a coauthor of more than 140 publications in international conferences, including 23 IEDM and more than 17 invited talks, one shortcourse (IEDM), one tutorial, 47 articles in refereed journals, four book chapters, and 20 patents. He is a Reviewer of several international journals (IEEE journals, Elsevier...). He is/was a member of several technical committees (IEDM, ESSDERC, and IRPS) and organizing committees (Co-Organizer of Fall MRS Memory Symposium, from 2015 to 2016, 2016–2019 IEEE IMW Organizing Committee, the 2018 General Chair, and the ESSDERC 2020 Financial Chair).

ELISA VIANELLO (Member, IEEE) received the Ph.D. degree in electrical engineering from the University of Udine, Italy, and the University of Grenoble, France, in 2010. She is currently the Scientist of CEA-LETI Grenoble. She is also a coordinator of the MeM-Scales h2020 European project focused on the joint co-development of a novel class of algorithms, devices, and circuits that reproduce multi-timescale processing of biological neural systems. She is the author or a coauthor of more than 100 technical articles and four book chapters. Her current research interests include development of new technologies for bio-inspired neuromorphic computing, with a special focus on new resistive memory devices. She was in the Technical Committee of the IRPS Symposium, from 2013 to 2014, and in the TCP of the IEDM Circuit Device Interaction Subcommittee, from 2016 to 2017. She is in the TPC of the ESSDERC Conference and the Cass Nano-Giga TC. She is also an Associate Editor of the APL Special Issue on Emerging Materials in Neuromorphic Computing (February 2020) and of the incoming IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—II: EXPRESS BRIEFS, from 2020 to 2021.



DAVID ATIENZA (Fellow, IEEE) received the Ph.D. degree in computer science and engineering from UCM, Spain, and IMEC, Belgium, in 2005. He is currently an Associate Professor of electrical and computer engineering and the Director of the Embedded Systems Laboratory, Swiss Federal Institute of Technology Lausanne (EPFL), Switzerland. His research interests include system-level design methodologies for high-performance multi-processor system-on-chip (MPSoC) and low-power Internet-of-Things systems, including new 2-D/3-D thermal-aware design for MPSoCs and many-core servers, ultralow power edge AI architectures for wireless body sensor nodes and smart consumer devices. He has coauthored over 300 articles in peer-reviewed international journals and conferences, several book chapters, and seven patents. He received the 2018 DAC Under-40 Innovators Award, the 2018 IEEE TCCPS Mid-Career Award, the 2016 ERC Consolidator Grant, the 2013 IEEE CEDA Early Career Award, the 2012 ACM SIGDA Outstanding New Faculty Award, and the Faculty Award from Sun Labs at Oracle, in 2011. He has served as the DATE 2015 Program Chair and the DATE 2017 General Chair, an ACM Distinguished Member, and the IEEE CEDA President (for the period 2019–2020).

JEAN-MICHEL PORTAL received the degree in electronic engineering, in 1996, and the Ph.D. degree in computer sciences, in 1999. He is currently a Full professor in electronics with Aix-Marseille University, where he heads the Electronic Department, Institute of Materials Microelectronics and Nanosciences of Provence (IM2NP). His research interests include emerging non-volatile memory design and neuromorphic applications. He is the author or a coauthor of more than 200 articles in international refereed journals and conferences. He is a co-inventor of six patents. He has supervised 20 Ph.D. students. He was a recipient of the NanoArch 2012, the Newcas 2013, and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS Guillemin-Cauer 2017 Best Paper Awards.