Experimental investigation of RRAM programming conditions suitable for TCAM applications in 4kbit arrays

Alessandro Grossi, Cristian Zambelli *Member, IEEE*, Elisa Vianello, Pablo Royer, Jean-Philippe Noel, Bastien Giraud, Etienne Nowak, Jean François Nodin, Piero Olivo, and Luca Perniola

***Abstract*—While Resistive RAM (RRAM) are seen as an alternative to NAND Flash, their variability and cycling understanding remain a major roadblock. Extensive characterizations of multi-kbits RRAM arrays during Forming, Set, Reset and cycling operations are presented allowing to investigate the relationships between programming conditions, memory window and endurance features. The experimental results are then used to perform variability-aware simulations of a RRAM-based ternary content-addressable-memory (TCAM) 128 bit macro with different operating conditions.**

***Index Terms*—RRAM, HfO_2, deposition, variability, process, reliability, performance**

I. Introduction

Even if RRAM is seen as a possible replacement for Flash memories due to manufacturing process simplicity, easy integration with logic, lower voltage operation, and good cycling with sufficient retention capability [1]–[3], the intrinsic cell-to-cell variability is still preventing its adoption as traditional data storage media [4]–[6]. Beyond that, RRAM gathered interest for several other applications such as neuromorphic systems [7], [8], content-addressable memories (CAM) and nonvolatile SRAM [9], [10]. CAM memory systems allow searching by content as opposed to searching by address [11], [12]. This approach avoids frequent and expensive memory accesses in applications where searching operations among a large amount of data are required, such as pattern recognition, routing tables and branch prediction in a processor [10]. CAMs are further divided in two categories: binary CAM (BCAM) and ternary CAM (TCAM): while BCAM allows storage of two states (0,1), TCAM allows to handle partial matches using an additional "don't care" state denoted by an "X". Such additional feature is required for example to mask certain bits of an IP address in network routers and makes TCAM more flexible than BCAM at the cost of an additional bit per cell. Conventionally CAM systems are SRAM-based, however the replacement of SRAM with RRAM could reduce the search latency/power and the leakage current while providing higher storage density and non-volatility of the stored information.

The expected impact of RRAM in TCAM applications has been already evaluated at first glance by using an analytical model of RRAM cells [13], [14]. In this work, to gather a deeper understanding of RRAM variability impact in TCAM applications the experimental results obtained from the characterization of 4kbits OxRAM array are used. A full characterization of Forming, Set and Reset is performed and the variability of each operation is extracted. After that, the experimental results are used to perform variability-aware simulations of a 128 bit RRAM-based TCAM macro in order to identify operating conditions suitable for TCAM applications with different search latency, power consumption and endurance properties.

A. Grossi, E. Vianello, P. Royer, J.-P. Noel, B. Giraud, E. Nowak, J. F. Nodin, and L. Perniola are with CEA-Leti, Minatec Campus, Grenoble, France (e-mail: alessandro.grossi@cea.fr).

C. Zambelli and P. Olivo are with Dipartimento di Ingegneria, Università degli Studi di Ferrara, Via Saragat 1, Ferrara, Italy.

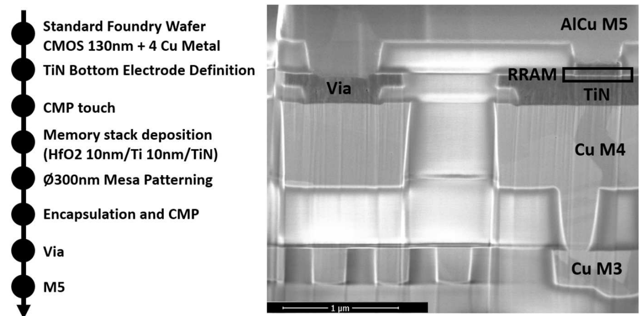


Fig. 1. Description of the integration flow and TEM cross section of the integrated $TiN/HfO_2/Ti/TiN$ OxRAM [2]

A. Experimental Setup

RRAM technology has been integrated on 130nm CMOS logic. On top of Cu Metal 4, a TiN bottom electrode is defined. Then a CMP touch is done and an HfO_2 10nm/Ti 10nm/TiN stack is deposited. Main integration steps and cross section of a 300nm diameter integrated device are described in Fig. 1.

The measurements in this work have been performed on 4kbits 1T-1R array developed within this platform. The word-line (WL) is connected to the NMOS gate, setting the current compliance I_{CC}. Forming and Set operations are performed by applying a positive voltage pulse on the bit line (BL) that is also the RRAM top electrode, whereas Reset is performed by applying a voltage pulse on the source line (SL) whose contact is on the transistor side (bottom electrode). The select transistor features large width in order to assess cell operation on a large current range (1μA - 5mA) and to ensure minimal dispersion impact (σ_T at read = 4 Ω). As a result, all measured resistance dispersion on the array can be always considered as intrinsic to the memory element. The standard deviation σ used

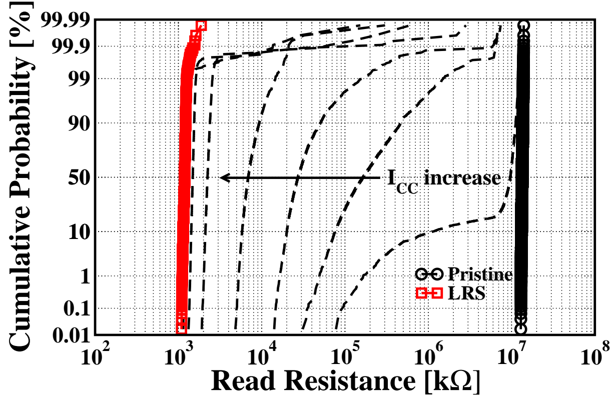


Fig. 2. Forming with increasing I_{CC} and V_{BL}=4V: read resistance distributions evolution with T_{PULSE}=100ns Dotted lines show the distribution evolution.

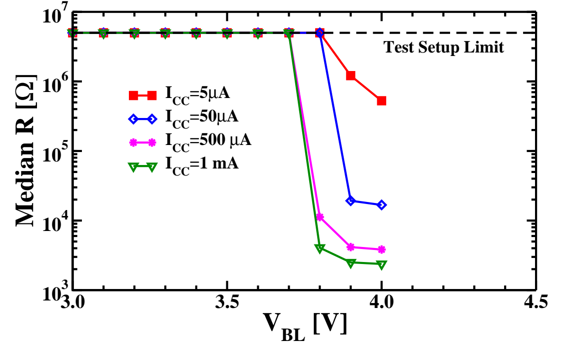
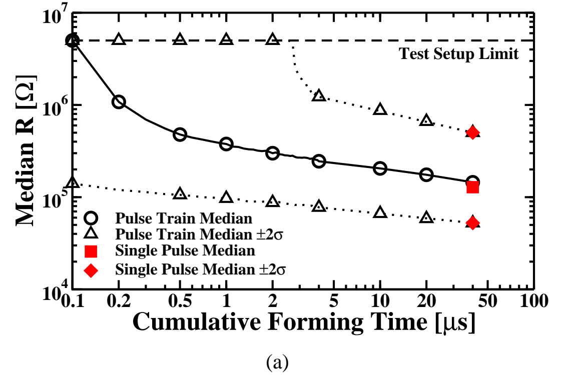


Fig. 3. Forming with V_{BL} increasing, T_{PULSE}=100ns and different I_{CC} [2].



(a)

Fig. 4. Median resistance $\pm 2\sigma$ during pulse train Forming procedure with I_{CC}=5μA, V_{BL}=4V, T_{PULSE}=100ns and after Forming with a single pulse of the same cumulative time T_{PULSE}= 40μs [2].

for this analysis is calculated on the array after excluding the extreme outliers point due to extrinsic causes. The 2σ bounds are plotted for clarity and consistency.

B. Forming characterization

Incremental current compliance pulse operations were performed to investigate the impact of I_{CC}, V_{BL} and pulse length (T_{PULSE}) on Forming. When the compliance current increases, also the maximum current flowing through the transistor increases lowering minimum resistance achievable after Forming (Fig. 2). The filament creation process is found to be purely field driven and happens around V_{BL}=3.8V using 100ns pulses for this stack. Further, we confirmed that the median resistance final value is directly related to I_{CC} (Fig. 3). Fig. 4 shows the time impact on the formed cells creation, pointing out that a single pulse has the same impact of a train pulse with equivalent duration. While 50% of the cells are formed within 200 ns, it required 40μs to form about a 3σ range of the cells population. This very large time dispersion between fast and slow formed cells can be detrimental if a too large I_{CC} is used. In this latter case, the fast formed cells are stuck to low resistance state (LRS) before the slow formed cells are formed. To avoid degradation during Forming, a low compliance current must be used: this also removes the needing of complex and time consuming verification algorithms during RRAM operations.

C. Set/Reset characterization

The maximum R_{HRS}/R_{LRS} ratio reachable is higher than 10^3 (limited by our array measurement setup). Fig. 5 shows the impact of I_{CC}, V_{BL} and T_{PULSE} on the median and dispersion. After Forming, V_{BL} for Set has a threshold-like behavior requiring 1V to set at least 50% of the cells and 1.6V for 2σ. For Reset, as stated earlier, the I_{CC} is the most important, and V_{SL} has no impact beyond 2.5V. The time impact is small from 100ns to 10μs demonstrating a fully operational RRAM array using 100ns pulses.

D. Endurance characterization

2 million cycles test has been performed on all cells of a 4kbits array with no-verify condition for Set and Reset operations (Fig. 6). A separation of HRS and LRS states can be obtained at 2σ up to 500k cycles. Distributions show that the failure is due to stuck cells at low resistance. However, it is observed that starting at the very first cycle, the median resistance and the dispersion increase continuously before reaching a maximum: afterwards the operating windows decrease rapidly.

The stronger the Reset condition, the higher R_{HRS} and the shorter the cycling performance can be obtained. The tradeoff R_{HRS}/R_{LRS} vs. cycle is comparable to state of the art published filamentary RRAM data (Fig. 7).

II. RRAM-BASED TCAM CELL AND ARRAY STRUCTURES

The RRAM-based TCAM cell structure simulated in this work is composed of two transistors and two RRAM resistors (2T2R), denoted as R1 and R2 (see Fig. 8). Forming, Set and Reset operations can be performed as in single 1T-1R RRAM cells by applying the required top electrode voltage on the match line (ML), the bottom electrode voltage on the bit line (BL) and the gate voltage on SLT/SLF while keeping SLF/SLT at 0V in order to activate each 1T-1R structure independently. For writing a logic ”1” on the TCAM cell it is necessary to perform a Set operation on R1 and a Reset on R2, whereas a logic state ”0” can be obtained performing a Set on R2 and

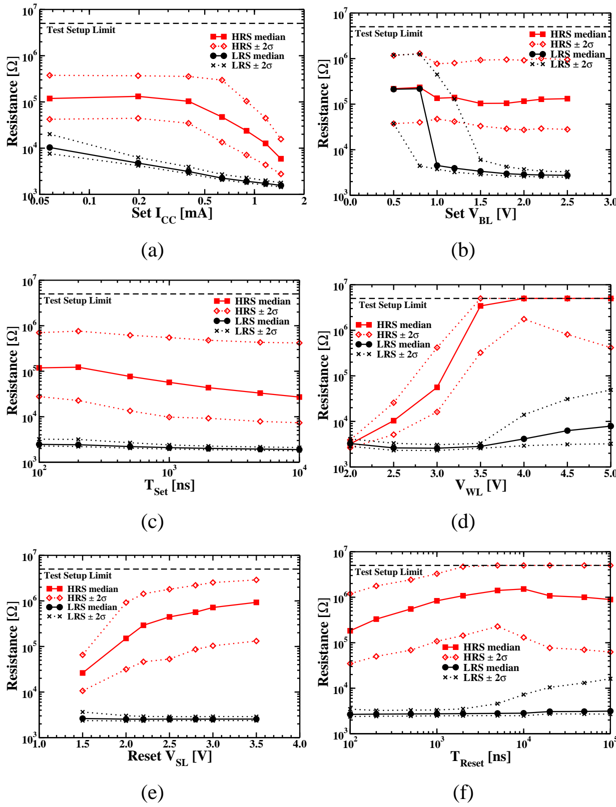


Fig. 5. Impact of Set I_{CC} (a), V_{BL} (b) and T_{PULSE} (c) on Set and Reset (median $\pm 2\sigma$), with $V_{SL,Reset}$=2.5V, $V_{WL,Reset}$=3V and $T_{PULSE,Reset}$=100 ns. Set I_{CC}=0.4 mA, V_{BL}=2V and T_{PULSE} = 100ns were used when varying other parameters. Impact of Reset V_{WL} (d), V_{SL} (e) and T_{PULSE} (f) on Set and Reset (median $\pm 2\sigma$), with $V_{BL,Set}$=2V, $I_{CC,Set}$=0.4mA and $T_{PULSE,Set}$=100 ns. Reset V_{WL}=3V, V_{SL}=2.5V and T_{PULSE} = 100ns were used when varying other parameters [2].

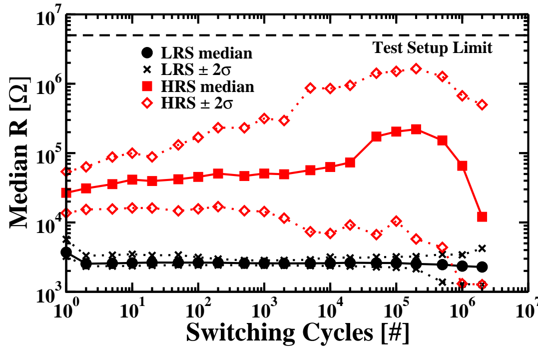


Fig. 6. LRS and HRS median $\pm 2\sigma$ resistances measured during 2M cycles performed with Set I_{CC}=0.4mA, V_{BL}=2V, T_{Set}=100ns and Reset V_{WL}=3V, V_{SL}=2.5V, T_{Reset}=100ns [2].

a Reset on R1. In order to obtain the "don't care" state "X", Reset must be performed on both R1 and R2. The TCAM cell's states and the correspondent resistive state of the two RRAM cells are summarized in Tab. I.

During a read/search operation ML is initially precharged, then it is left floating while the two transistors of the TCAM structure are activated as a function of the search pattern by applying a voltage on SLT and SLF. In case of a search-1 operation (SLT=1 and SLF=0 logic states) a pulse is applied on the gate of the transistor below R1: in case R1 is in LRS

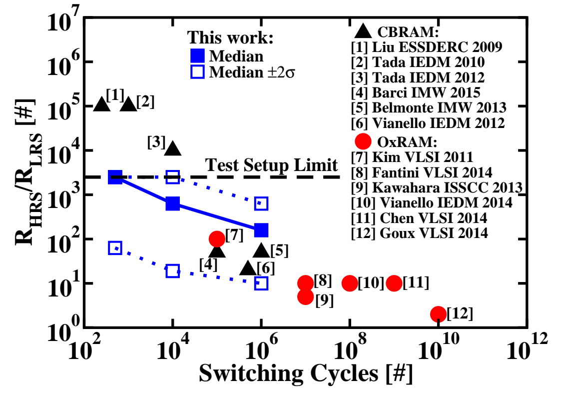


Fig. 7. R_{HRS}/R_{LRS} vs. switching cycles. Median resistance $\pm 2\sigma$ results are reported and compared with literature. Limits in the test setup bound R_{HRS}/R_{LRS} ratios measurement to 2500 [2].

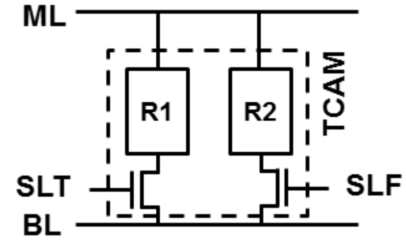


Fig. 8. Standard TCAM cell schematic

TABLE I
DEFINITION OF TCAM STATES

State	R1	R2
X	HRS	HRS
0	HRS	LRS
1	LRS	HRS

(mismatch), the ML voltage will be pulled down quickly, whereas if R1 is in HRS (match) it will be pulled down slower. In case of a search-0 operation (SLT=0, SLF=1 logic states), the same considerations can be derived depending on R2 resistive state. After a specific period the final voltage of ML is detected to evaluate if the search operation output is a match or a mismatch.

The schematic of an $n \times m$ RRAM-based TCAM macro is reported in Fig. 9 composed of n words of m TCAM cells. During a search operation all the match lines are precharged and then left floating while the search address is driven through the search lines. Then, all MLs are connected to an encoder which identify the matched line and produces as output the corresponding address of the matched entry. In this work, only single 2T2R TCAM cells and a word of 128 TCAM cells will be simulated to understand the shortest discharge time achievable (corrensponding to the case of the single TCAM cell) and the impact of the parasitic capacitances on the discharge time when a line of 128 TCAM cells is considered.

III. RRAM-BASED TCAM SIMULATIONS

The performance of the RRAM-based TCAM cell and 128 bit word are evaluated through SPICE simulations considering LRS and HRS extracted from the experimental results

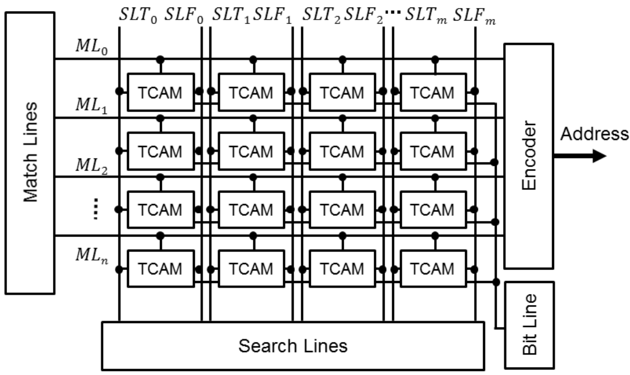


Fig. 9. Schematic of RRAM-based TCAM macro.

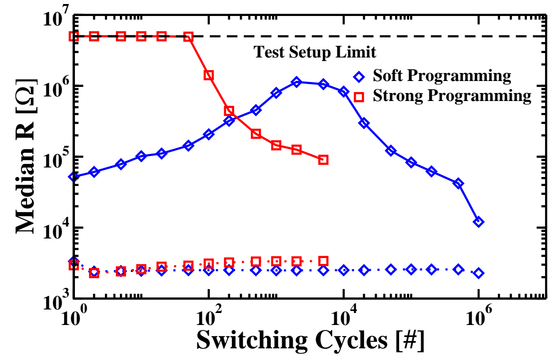


Fig. 10. Comparison of LRS (dotted lines) and HRS (full lines) median resistances in cycling with Soft and Strong Programming conditions.

TABLE II
PROGRAMMING CONDITION PARAMETERS

Parameter	Soft	Strong	OTP
VML,Set	2V	2V	4V
VSLT/SLF,Set	1.6V	1.6V	2V
VBL,Set	0V	0V	0V
VML,Reset	0V	0V	-
VSLT/SLF,Reset	3V	3.5V	-
VBL,Reset	2.5V	2.5V	-

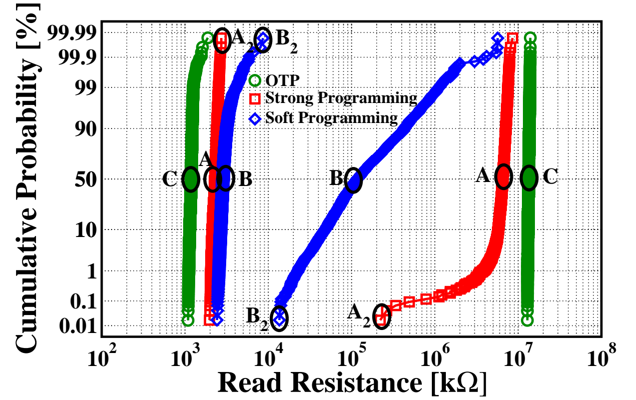


Fig. 11. LRS and HRS cumulative distributions measured on 4 kbits array after Set and Reset with soft, strong and OTP programming conditions.

with different programming conditions. The simulations were performed on a 130nm CMOS technology, where the transistor parameters (W=6.7μm, L=0.5μm) in the 2T2R cells are the same of the 1T1R cells previously characterized in the 4kbit RRAM arrays. As previously observed, RRAM programming condition strongly impact the memory window and the endurance. In case of strong programming conditions it is possible to obtain a large memory window at a cost of shorter endurance, whereas with soft programming conditions the memory window is reduced but the endurance is improved as shown in Fig. 10. In addition a One-Time-Programming (OTP) case can be considered, providing the largest memory window possible on the technology (from pristine state to irreversible breakdown of the oxide). The programming parameters for soft, strong and OTP conditions are summarized in Tab. II, with all pulses featuring a 100 ns duration. Reset conditions are not reported for OTP since a single Forming/Set operation is performed to bring the cells from pristine state to breakdown.

The cumulative distributions of LRS and HRS obtained after a single operation with soft, strong and OTP programming conditions are reported in Fig. 11. In the following simulations the median resistances obtained after programming will be used to evaluate the discharge time in case of match and mismatch for single cells and a 128 bit word. In addition to the median LRS and HRS resistances, also the minimum HRS and maximum LRS (extracted at +/-2 sigma of the population) will be considered as worst-case scenario for soft and strong programming conditions, in which a high cell-to-cell variability is observed. The resistance values extracted from the experimental results and considered in the simulations are summarized in Tab. III.

The relationship between LRS and HRS resistances for all

TABLE III
SET OF LRS AND HRS COMBINATIONS SELECTED FROM CHARACTERIZATION RESULTS.

Condition	R_{HRS}	R_{LRS}	Description
A	5MΩ	2.3kΩ	Large window
A_2	223kΩ	2.8kΩ	Large window -2σ
B	116kΩ	3.1kΩ	Small window
B_2	12kΩ	9kΩ	Small window -2σ
C	1.2kΩ	12MΩ	One-Time Programming

considered resistive states is shown in Fig. 12: in order to have fast search operation a low LRS is required in case of mismatch, while a high HRS is necessary to guarantee a slow discharge during match and therefore successfully discriminate between match and mismatch. Condition C shows the optimal combination of LRS and HRS resistance. The effects of Set and Reset failures or degradation on LRS and HRS are indicated, which cause an increase of LRS and decrease of HRS.

The match line discharge time through HRS has been simulated for a single 1T1R in the TCAM structure by putting SLT in logic state 1 and SLF in logic state 0 (in all possible search configurations, SLT and SLF are never at 1 in simultaneously). ML was initially set at 0.5V and then left floating while the gate of the transistor was put at 1.6V and the time required to reach 0.25V on ML, hereafter indicated as search time latency, was extracted. The search time latency was evaluated while

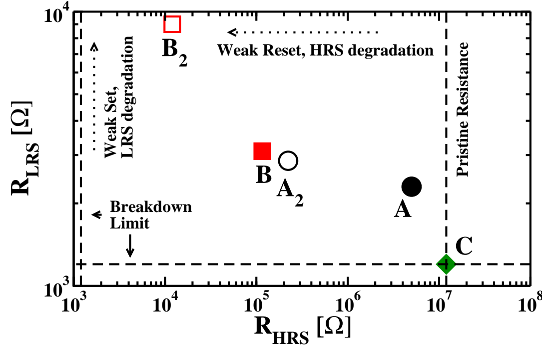


Fig. 12. Galaxy plot of LRS,HRS $\pm 2\sigma$ for the selected characterization results.

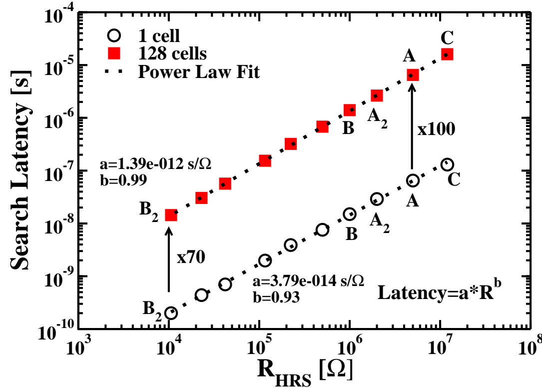


Fig. 13. Search latency evaluation in case of discharge through HRS in case of 1 single TCAM cell and a word of 128 TCAM cells.

varying the HRS resistance from 10kΩ to 20MΩ, for a single TCAM cell and in case of a 128 bit word of TCAM cells with SLT in logic state 1 and SLF in logic state 0 on all cells simultaneously. When the 128 bit TCAM word is considered, the worst case scenario is when all resistors are set at the lowest HRS since the presence of cells with higher HRS would make the discharge time increase, hence all resistances were set to the same HRS value in the TCAM word simulation. The simulation results are reported in Fig. 13: the search latency increases when HRS increase on both single cell and 128 bit word and the parasitic capacitance of the 128 bit word causes an increase of the discharge time varying from 70 to 100 when compared to the single cell. The search latency timings for conditions A, B and C are indicated in the figure. The relationship between search latency and HRS follows a power law, whose fitting parameters are shown in the figure.

The same simulation approach has been used to evaluate the search latency when R1 is at LRS, assuming SLT in logic state 1 and SLF in logic state 0 during the search operation. Again in the case of 128 bit word TCAM all cells were considered at same LRS since the presence of cells with lower LRS would make the discharge faster, hence this represents the worst case scenario. The results are reported in Fig. 14, where the resistance values corresponding to conditions A, B and C are indicated. When considering conditions A (for both median and worst-case) and C the search latency is in the order on few ns for a 128 bit TCAM word, which is comparable

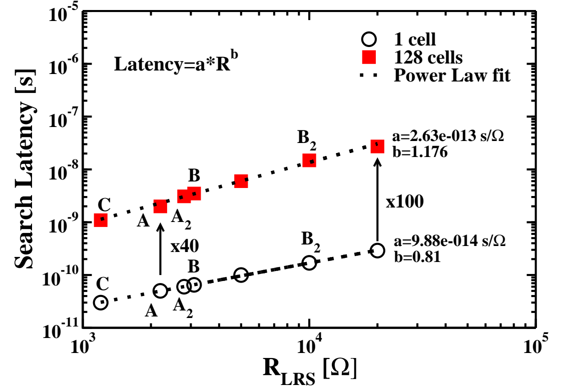


Fig. 14. Search latency evaluation in case of discharge through LRS in case of 1 single TCAM cell and a word of 128 TCAM cells.

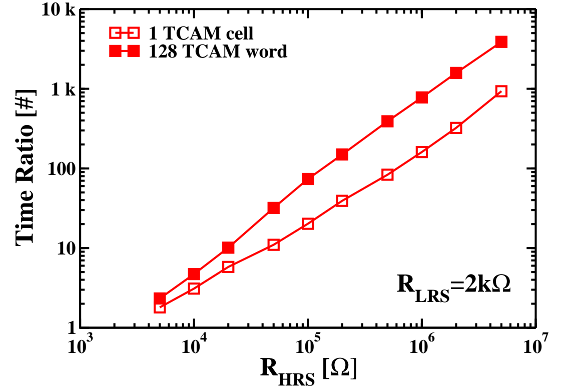


Fig. 15. Evaluation the time ratio between match and mismatch as function of HRS resistance when LRS is 2kΩ.

with SRAM. The power law fitting parameters are reported in figure.

While a low LRS is necessary in order to have a search latency comparable with SRAM, a high HRS is required to have a sufficient ratio between match and mismatch discharge time and correctly identify the matched entry. In Fig. 15 the ratio between the discharge time obtained with LRS=2kΩ (which is sufficient to have a search latency comparable to SRAM cells) and different HRS is reported for both single TCAM cell and a 128 bit TCAM word, showing that in the case of a 128 bit TCAM word in order to have at least a ratio of 10 it is required to have HRS higher than 20kΩ.

A final comparison of programming conditions strong, soft and OTP in terms of average and minimum memory window (MW), endurance, programming energy (calculated as the sum of Set and Reset energies), maximum search latency and time ratio is summarized in Tab. IV, also comparing with state of the art 32nm SRAM [15] in terms of endurance, programming energy and search latency. It must be pointed out that even if there's not an advantage when compared to SRAM programming, the non volatile nature of RRAM-based TCAM will strongly reduce the static power dissipation as well as the search energy.

TABLE IV
PROGRAMMING CONDITION COMPARISON

Parameter	A	B	C	SRAM
MW (avg.,min)	2.2e4,80	40,1.3	1e4,6e3	-
Endurance	100	1e6	1	1e16
Programming Energy	1.24nJ	1.1nJ	20nJ	0.58fJ
Search Latency (max)	2ns	3.5ns	1.1ns	1ns
Time ratio (avg.,min)	3e3,150	80,4	1.6e4,1e4	-

IV. CONCLUSIONS

An extensive study of variability of multi-kbits RRAM arrays over the full operation range has been presented. In contrast to all papers based on single cell analysis with extensive cycle-to-cycle analysis, OxRAM array distributions are extremely stable and predictable. The experimental results were used to perform variability-aware simulations of a single RRAM-based TCAM cell and of a 128 bit RRAM-based TCAM macro, showing that in order to obtain search latency times comparable with SRAM a LRS of around 2kΩ is required while HRS higher than 20kΩ is required to have at least a decade between match and mismatch discharge time on a 128 bit TCAM word. Such LRS and HRS conditions can be satisfied with strong and OTP programming conditions at cost of short endurance, making RRAM-base TCAM interesting candidates for high-speed networking applications such as routing tables and database acceleration where the memory is mainly read and the stored informations are not updated frequently. Soft programming condition can be used up to 1e6 cycles if the a minimum time ratio of 4 between match and mismatch can be accepted. Possible applications achievable with such condition could be neuromorphic associative memories and reconfigurable computing. In these applications the pattern to be matched is varied frequently and the match does not need to be perfect, since a hierarchical search approach can be used where only part of the search pattern needs to be matched. Such approach can be implemented varying the number of bits in "don't care" state in the search pattern to modify the hamming distance requirements between searched and matched words and is particularly suitable for neural pattern recognition.

REFERENCES

[1] L. Perniola, G. Molas, G. Navarro, E. Nowak, V. Sousa, E. Vianello, and B. D. Salvo, "Universal Signatures from Non-Universal Memories: Clues for the Future..." in *IEEE 8th International Memory Workshop (IMW)*, May 2016, pp. 1–3.

[2] A.Grossi, E. Nowak, C. Zambelli, C. Pellissier, S. Bernasconi, G. Cibrario, K. E. Hajjam, R. Crochemore, J. Nodin, P. Olivo, and L. Perniola, "Fundamental Variability Limits of Filament-based RRAM," in *IEEE Int. Electron Devices Meeting (IEDM)*, Dec 2016, pp. 4.7.1–4.7.4.

[3] L. Zuolo, C. Zambelli, A. Grossi, R. Micheloni, S. Bates, and P. Olivo, "Memory System Architecture Optimization for Enterprise All-RRAM Solid State Drives," in *IEEE 8th International Memory Workshop (IMW)*, May 2016, pp. 1–4.

[4] A. Grossi, D. Walczyk, C. Zambelli, E. Miranda, P. Olivo, V. Stikanov, A. Feriani, J. Sune, G. Schoof, R. Kraemer, B. Tillack, A. Fox, T. Schroeder, C. Wenger, and C. Walczyk, "Impact of Intercell and Intracell Variability on Forming and Switching Parameters in RRAM Arrays," *IEEE Transactions on Electron Devices*, vol. 62, no. 8, pp. 2502–2509, Aug 2015.

[5] 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. on Electron Devices*, vol. 61, no. 8, pp. 2912–2919, 2014.

[6] C. Cagli, G. Piccolboni, D. Garbin, A. Grossi, G. Molas, E. Vianello, C. Zambelli, and G. Reimbold, "About the intrinsic resistance variability in hfo2-based rram devices," in *2017 Joint International EUROSOI Workshop and International Conference on Ultimate Integration on Silicon (EUROSOI-ULIS)*, April 2017, pp. 31–34.

[7] T. Werner, E. Vianello, O. Bichler, A. Grossi, E. Nowak, J. F. Nodin, B. Yvert, B. DeSalvo, and L. Perniola, "Experimental demonstration of short and long term synaptic plasticity using oxram multi k-bit arrays for reliable detection in highly noisy input data," in *2016 IEEE International Electron Devices Meeting (IEDM)*, Dec 2016, pp. 16.6.1–16.6.4.

[8] E. Vianello, T. Werner, O. Bichler, A. Valentian, G. Molas, B. Yvert, B. D. Salvo, and L. Perniola, "Resistive memories for spike-based neuromorphic circuits," in *2017 IEEE International Memory Workshop (IMW)*, May 2017, pp. 1–6.

[9] R. Dhiman, M. Kaur, and G. Singh, "Memristor based ternary content addressable memory (mtcam) cell," in *2015 2nd International Conference on Recent Advances in Engineering Computational Sciences (RAECS)*, Dec 2015, pp. 1–6.

[10] R. Karam, R. Puri, S. Ghosh, and S. Bhunia, "Emerging trends in design and applications of memory-based computing and content-addressable memories," *Proceedings of the IEEE*, vol. 103, no. 8, pp. 1311–1330, Aug 2015.

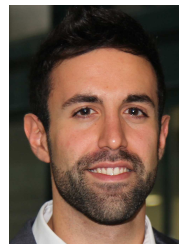
[11] S. Paul and S. Bhunia, "Reconfigurable computing using content addressable memory for improved performance and resource usage," in *2008 45th ACM/IEEE Design Automation Conference*, June 2008, pp. 786–791.

[12] K. Pagiamtzis and A. Sheikholeslami, "Content-addressable memory (cam) circuits and architectures: a tutorial and survey," *IEEE Journal of Solid-State Circuits*, vol. 41, no. 3, pp. 712–727, March 2006.

[13] Z. Jiang, Y. Wu, S. Yu, L. Yang, K. Song, Z. Karim, and H. S. P. Wong, "A compact model for metal-oxide resistive random access memory with experiment verification," *IEEE Trans. on Electron Devices*, vol. 63, no. 5, pp. 1884–1892, May 2016.

[14] L. Zheng, S. Shin, S. Lloyd, M. Gokhale, K. Kim, and S. M. Kang, "Rram-based tcams for pattern search," in *2016 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2016, pp. 1382–1385.

[15] I. Arsovski, T. Hebig, D. Dobson, and R. Wistort, "A 32 nm 0.58-fj/bit/search 1-ghz ternary content addressable memory compiler using silicon-aware early-predict late-correct sensing with embedded deep-trench capacitor noise mitigation," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 4, pp. 932–939, April 2013.



Alessandro Grossi received the M.Sc. and Ph.D. degrees in electronic engineering from Universitá degli Studi di Ferrara, Ferrara, Italy, in 2013 and and 2017, respectively. Since 2017, he has been a Post-Doc researcher at Cea-Leti, Grenoble, France. His main research interests include the characterization, physics, modeling and reliability of emerging non-volatile memory devices.

Cristian Zambelli (S'08-M'12) received the M.Sc. and Ph.D. degrees (with honors) in electronic engineering from Universitá degli Studi di Ferrara, Ferrara, Italy, in 2008 and 2012, respectively. Since 2015, he has been an Assistant Professor with Dipartimento di Ingegneria, Universitá degli Studi di Ferrara. His main research interests include the characterization, physics, and modeling of the reliability of nonvolatile memory devices, and algorithmic solutions for the reliability-performance tradeoff exploitation in solid-state drives.

Piero Olivo received the M.Sc. degree in electronic engineering and the Ph.D. degree from the University of Bologna, Bologna, Italy, in 1980 and 1987, respectively. Since 1994, he has been a Full Professor of Electronics with Dipartimento di Ingegneria, Universitá degli Studi di Ferrara, Ferrara, Italy. His scientific activity concerns the theoretical and experimental aspects of microelectronics, with an emphasis on the physics, reliability, and characterization of electron devices and nonvolatile memory devices.