Experimental demonstration of Single-Level and Multi-Level-Cell RRAM-based In-Memory computing with up to 16 parallel operations

Eduardo Esmanhotto, T. Hirtzlin, N. Castellani, S. Martin, B. Giraud, F. Andrieu, J F Nodin, D. Querlioz, J-M. Portal, E. Vianello

Experimental demonstration of Single-Level and Multi-Level Cell RRAM-based In-Memory-Computing with up to 16 parallel operations

E. Esmanhotto[1], T. Hirtzlin[1], N. Castellani[1], S. Martin[1], B. Giraud[2], F. Andrieu[1], J.F. Nodin[1], D. Querlioz[4], J-M. Portal[3] and E. Vianello[1]

[1]CEA-Leti, Université Grenoble Alpes, Grenoble, France, email: Eduardo.Esmanhotto@cea.fr, Elisa.Vianello@cea.fr
[2]CEA-List, Université Grenoble Alpes, Grenoble, France, [3]Aix-Marseille Université, IM2NP, Marseille, France,
[4]Université Paris-Saclay, CNRS, C2N, Palaiseau, France.

Introduction – In today's data-intensive applications, data transfer is the main contributor to the power consumption of computing systems [1]. To tackle this issue, IMC [2] is a major lead. Different RRAM-based IMC logic solutions have been studied so far [3][4][5]. However, RRAM variability is a considerable challenge, and only a few have demonstrated experimental results, with limited parallelism [3]. These works use RRAM as a single-bit memory to perform IMC, giving up to a major ability of RAMS: Multi-Level Cell (MLC).

In this work, we demonstrate experimentally an RRAM-based IMC logic concept with strong resilience to RRAM variability, even after one million endurance cycles. Our work relies on a generalization of in-memory Scouting Logic, and we demonstrate it experimentally with up to 16 parallel devices, a new milestone for RRAM in-memory logic. Moreover, we combine IMC with a new smart Multi-Level Cell (MLC) programming algorithm and demonstrate experimentally, for the first time, an IMC RRAM-based MLC 2-bit adder.

Our technique belongs to the Non-Stateful Logic IMC category, meaning that the results of the operations are obtained after a read operation on multiple cells in parallel that preserves memory endurance [7]. To overcome the RRAM variability issue, we proposed a new smart programming method suitable for binary and Multi-Level Cell solutions.

The experiments are performed on a custom HfO_2 crossbar 1T1R array built with 130 nm CMOS (Fig. 1a). This array, incorporating the CMOS periphery, operates on both Memory Mode (Fig. 1b) and IMC (Fig. 1c), thanks to the parallel read operation.

New smart programming strategies – MLC Programming suffers from short-term (relaxation) and long-term (retention) conductance drift, a severe limitation for MLC storage and MLC/binary IMC [8]. We propose a new programming method to avoid relaxation called Full-Correction Smart Programming (FC-SP). It is an extension of the more standard algorithm of [8], here called Partial-Correction Smart Programming (PC-SP). Both algorithms use RRAM cycle-to-cycle variability to program the devices in a specific conductance range. The main addition of FC-SP is a wait time Δt (Fig. 2) after the SET operation. This waiting time enables the algorithm to take into account conductance relaxation. Fig. 3 shows a conductance level programmed with PC-SP: the conductance relaxation causes the initial distribution to spread over time. By contrast, the FC-SP strategy (Fig. 3) results in conductance levels that remain stable and the conductance relaxation effect is negligible after one hour (less than 1% of cells are out of the target range for the corresponding level 0 in Fig. 4).

Fig. 6 shows the number of devices out of the target range (Bit Error Count - BEC) through successive iterations. The FC-SP programming strategy needs more iterations to achieve the same programming performance of PC-SP. The BEC of the most critical conductance level of an MLC (level 0 in Fig. 4) rapidly increases during the first 10 s if the cells are programmed with PC-SP due to conductance relaxation, while it remains low for FC-SP (Fig. 7). The retention data of three programmed levels with FC-SP is evaluated over a month (Fig. 8). FC-SP MLC levels are resilient to both short-term relaxation and long-term data retention.

Multi-Level Cell and binary IMC – We combine MLC with parallel read to achieve complex logic functions in-memory. We demonstrate an RRAM-based Multi-Level In-Memory 2-bit adder. Fig. 9a shows the conventional 2-bits adder with 2 inputs (A_0A_1 and B_0B_1), with classic logic gates (equivalent to 48 transistors using CMOS technology). In [9], a CMOS-based IMC is proposed; however, it requires additional latch circuits to store intermediate data. Our solution proposes to combine parallel read of two devices storing 4 conductance levels (2 bits) to perform the 2-bits adder operation (2 transistors and 2 RRAM, Fig. 9b). The RRAM Based Multi-Level In-Memory 2-bit adder is implemented naturally in-memory using the crossbar architecture of Fig. 1b following the table in Fig. 9b.

Fig. 10 shows the distributions of the 7 logic outputs of the proposed Multi-Level In-Memory 2-bit adder. The relaxation causes overlaps for PC-SP, while for FC-SP the overlap effect is minimal (Fig. 11). We also show IMC based on binary RRAMs by extending the scouting logic concept [6] to up to 16 parallel devices (operands) with an increased experimental success rate when FC-SP is adopted (Fig. 12).

Conclusions – This work shows experimentally a programming strategy that controls conductance relaxation in MLC RRAM and stabilizes programmed levels up to more than one month. This solution is fundamental for multi-level IMC. By combining MLC and IMC, we show experimentally, for the first time, an RRAM-based Multi-Level In-Memory 2-bit adder. These results highlight the potential of RRAM IMC logic, and bring this field beyond purely circuit level and architecture studies.

ACKNOWLEDGMENT: This work is supported by the ECSEL TEMPO project (826655) and the ANR grant NEURONIC (ANR-18-CE24-0009).

REFERENCES: [1] M. Howoritz, ISSCC, 2014. [2] H. Wong et al, Nature, 2018. [3] W. -H. Chen *et al*., IEDM, 2017. [4] Kvatisnky et al, VLSI, 2014. [5] S. Gupta et al, DATE, 2020. [6] Lei Xie, H.A. et al, VLSI, 2017.[7] J. Reuben et al, PATMOS, 2017 [8] E. Esmanhotto et al, IEDM, 2020. [9] D. Fan. Et al. ASPDAC, 2019. [10] J. Han et al, ETS, 2013

I – Introduction

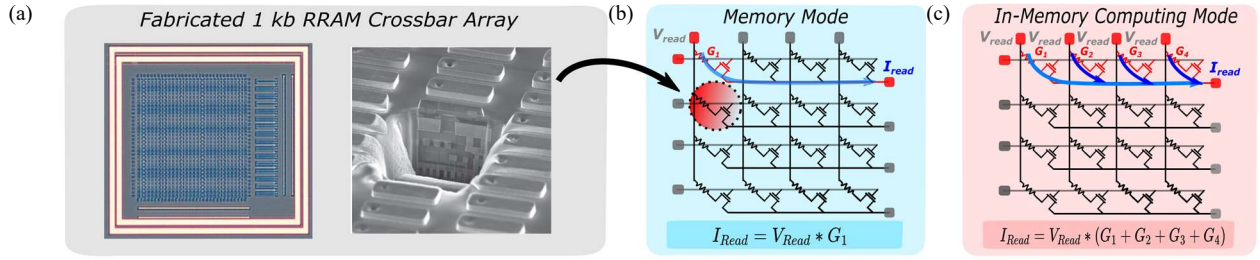

Fig. 1: Fabricated 1 kb RRAM crossbar array based on 1T1R HfO_2 stack (a). The proposed architecture can operate on both Memory Mode (b) and In-Memory Computing Mode (c). In Memory Mode, a single device or the full column is addressed and the current is sensed along the row using the current I_{read}. On the In-Memory Computing Mode, several devices of the same line are selected simultaneously.

II – New Smart Programming Strategies

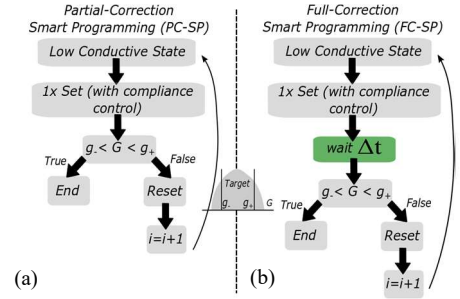

Fig. 2: MLC Programming strategies flow: Partial Correction Smart Programming (PC-SP) [8] (a) and Full Correction Smart Programming (FC-SP) (b).

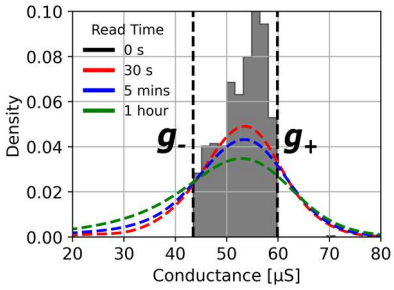

Fig. 3: Conductance distribution programmed with PC- SP strategy. The dashed lines represent the relaxation over different read times.

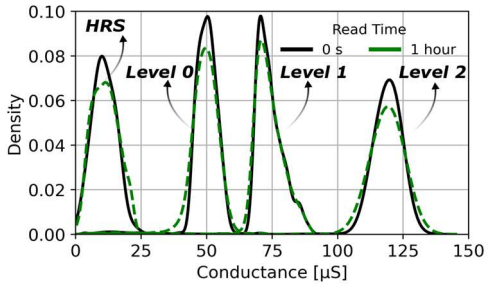

Fig. 4: Three conductance levels programmed with FC-SP strategy and the High Resistive State (HRS) just after programming (black) and after 1 hour (green).

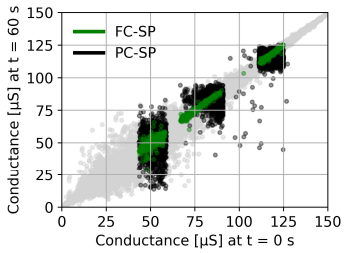

Fig. 5: Relaxation effect for FC-SP and PC-SP after 60 seconds. Grey color represents distributions without smart programming.

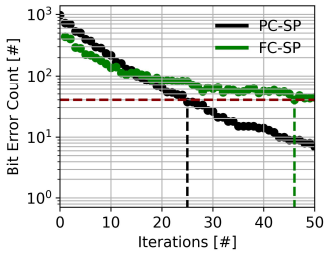

Fig. 6: Bit Error Count (BEC) as a function of the number of iterations during PC-SP and FC-SP.

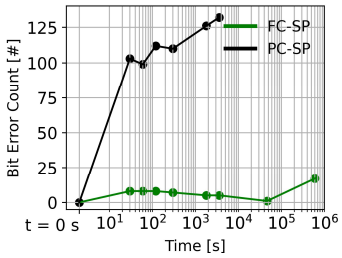

Fig. 7: Bit Error Count on time for the most error prone lvel (level 0 in Fig. 4) programmed with FC-SP and PC-SP.

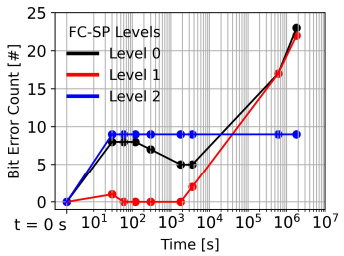

Fig. 8: Bit Error Count on time (up to one month) for the 3 FC-SP programmed levels.

III –Multi-Level Cell and binary IMC

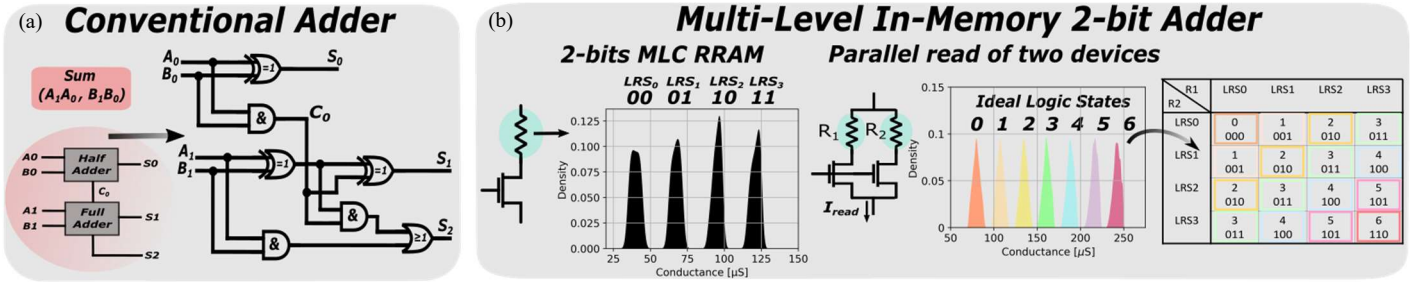

Fig. 9: Conventional architecture needed for a 2-bit adder (a) and proposed Multi-Level In-Memory 2-bit adder based on two parallel RRAMs (b).

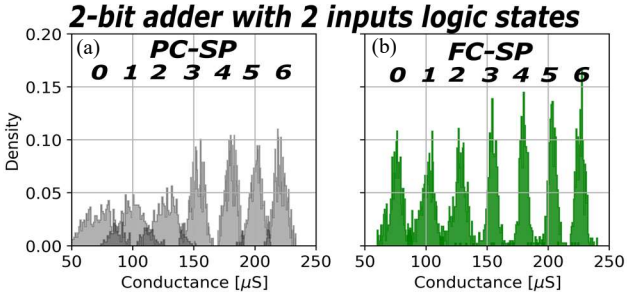

Fig. 10: Experimental distributions of the In-Memory 2-bit Adder logic states based on conductance distributions obtained with PC-SP (a) and FC–SP (b).

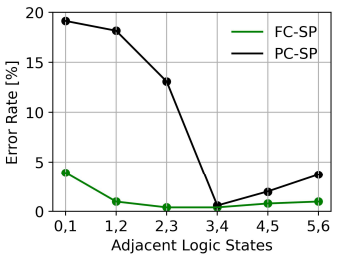

Fig. 11: Error rate between adjacent logic states for the proposed 2-bit adder for PC-SP and FC-SP.

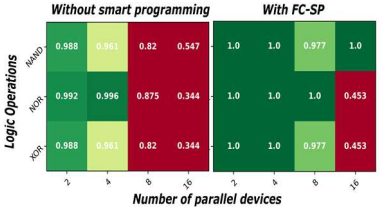

Fig. 12: Experimental success rate for NAND, NOR and XOR scouting logic operations without smart programming and with FC-SP. FC-SP renders IMC possible with up to 16 parallel operands.