Experimental Demonstration of Memristor Delay-Based Logic In-Memory Ternary Neural Network

A. Renaudineau, K.-E. Harabi, C. Turck, A. Laborieux, E. Vianello, Marc Bocquet, Jean-Michel Portal, Damien Querlioz

Experimental Demonstration of Memristor Delay-Based Logic In-Memory Ternary Neural Network

A. Renaudineau[1], K.-E. Harabi[1], C. Turck[1], A. Laborieux[1], E. Vianello[2], M. Bocquet[3], J.-M. Portal[3], D. Querlioz[1]
[1]Université Paris-Saclay, CNRS, C2N, Palaiseau France. [2]CEA-LETI, Univ. Grenoble- Alpes, Grenoble, France.
[3] Aix-Marseille Univ., CNRS, IM2NP, Marseille France. Email: adrien.renaudineau@universite-paris-saclay.fr

Abstract — **We present a fabricated hybrid CMOS/memristor integrated circuit for efficient implementation of Ternary Neural Networks. Our approach overcomes memristor resolution limitations and uses a simple sense amplifier that simultaneously reads memristor states and performs the TNN multiplication operations. The test chip validates our scheme, paving the way for energy-efficient, nanosecond-latency TNNs.**

I. Introduction

Memristors are a promising technology for low-power artificial intelligence (AI) applications, although their limited resolution poses a significant challenge [1]. Ternary neural networks (TNNs), which constrain synapses and neurons to only three possible values (-1, 0, +1), have demonstrated their ability to achieve remarkable accuracy AI while utilizing minimal resources [2]. As such, TNNs present an attractive solution to address the resolution limitations of memristors. In this paper, we present the fabrication of a hybrid CMOS/memristor integrated circuit designed for the efficient implementation of TNNs. Our approach combines simple sense amplifiers with a novel arrangement of ideas proposed in previous works [2-4], enabling the simultaneous reading of memristor states and execution of multiplication operations.

II. Memristor Ternary Neural Network

Our test chip, fabricated in a 130-nm process (Fig. 1a,b), combines a 128-memristor array (Fig. 1b) with CMOS periphery for programming memristors and performing logic-in-memory operations (Fig. 2). The memristors (Fig. 1c) consist of a 10-nm atomic-layer-deposition HfO_x layer and a 10-nm Ti layer, between metal levels 4 and 5 of the CMOS backend of line. For flexibility, the test chip relies on an external Keysight B1530 pulse generator to provide the sense read pulses.

Weights are programmed using two memristors per weight (Fig. 2c). Each column features a precharge sense amplifier (PCSA, Fig. 2b), charging all voltages to the supply voltage during read operations. Discharge rates then reveal memristor states: slow discharge occurs when both memristors are in high resistance state (HRS), representing zero weight (Fig. 3a); if one memristor is in low resistance state (LRS) and the other in HRS, either output Q or Qb discharges in under 1 ns depending on the programmed weight being 1 or -1 (Fig. 3b). This behavior enables logic-in-memory operations in two cycles (Fig. 4): the sense amplifier output reflects the product of input IN and the programmed weight.

III. Results

All memristors were initially formed using a 1µs, 2.5V pulse. SET and RESET operations applied 3V/3.5V and 5V/4.5V to the word line (WL) and source line (SL), respectively, while logic-in-memory used a 1.0 or 1.2V power supply. The PCSA discharge time and supply voltage VDD were critical for optimal success rates, while the 30ns PCSA charge time had minimal impact on the reading circuit. Fig. 5a shows the mean logic-in-memory success rate at 1.2V supply voltage. Operations with 1 and -1 weights were successful, barring two difficult-to-program devices. Operation with 0s posed a challenge due to the PCSA convergence speed. Reducing VDD to 1V (Fig. 5b) and using a discharge time of 4ns significantly increased operations with 0 weight success rates, with -1s and 1s rates remaining relatively constant. Shorter discharge times would further enhance 0-weight operation success rates; however, the B1530 pulse generator cannot produce pulses shorter than 4ns. Our future design will feature on-chip one-nanosecond pulse generation to optimize success rates in all cases.

IV. Conclusion

Our demonstration of a memristor delay-based logic-in-memory TNN shows promising results. Our circuit provides an efficient solution for implementing TNNs, where logic and memory are integrated, with the promise of extremely energy-efficient AI avoiding the von-Neumann bottleneck. The concept can also be extended by using delays not only for multiplication, but also accumulation.

Acknowledgment

This work was supported by ANR grants (ANR-18-CE24-0009 and ANR-22-PEEL-0010) and ERC grant NANOINFER (715872).

References

[1] D. Ielmini, HSP Wong, Nat. Electr. 1, 333, 2019.
[2] A. Laborieux et al., IEEE TCAS I 68, p. 138, 2021.
[3] W. S. Zhao et al., IEEE TCAS I 61, p. 443, 2014.
[4] M. Bocquet et al. In IEDM Tech. Dig., 20.6.1, 2018.

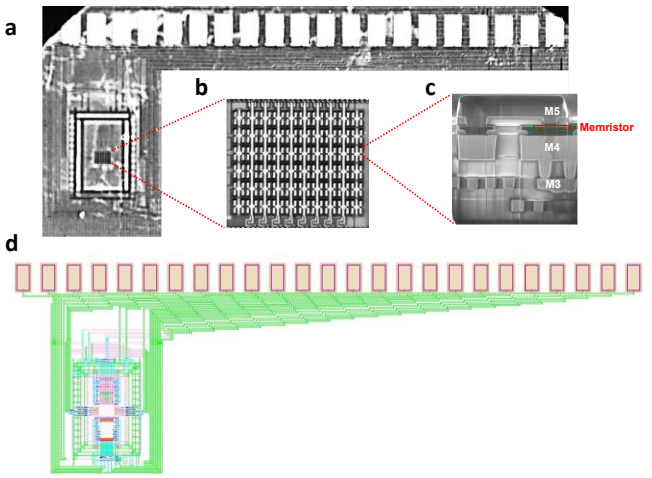

Fig. 1 **a** Optical photography image of the fabricated hybrid CMOS/HfO_x memristor test chip. **b** Zoom on memristor array. **c** electron microscopy image. **d** View of the test chip mask design (GDS file).

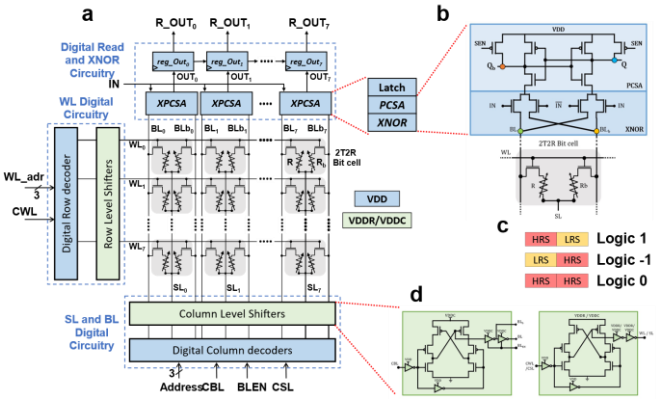

Fig. 2 **a** Schematic of the test chip of Fig. 1. **b** Detailed schematic of the logic-in-memory precharge sense amplifier used to read weight and perform multiplication. **c** Programming scheme for the ternary weight. State LRS/LRS is not used. **d** Detailed schematic of the level shifters used to program the memristors.

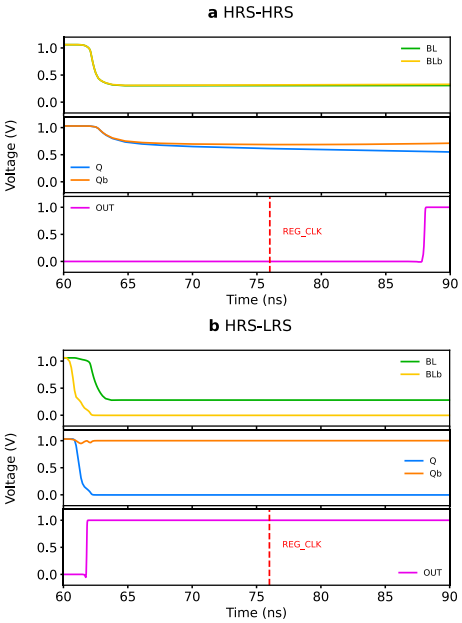

Fig. 3 Spice simulation illustrating the principle of delay-based sensing. The output registers are loaded with OUT when REG_CLK is pulsed. **a** In HRS-HRS (100kΩ-100kΩ) state, outputs Q and Qb remain high. **b** In HRS-LRS (100kΩ-10kΩ here) state, Q quickly discharges to 0 (for an LRS-HRS state, Qb discharges to 0).

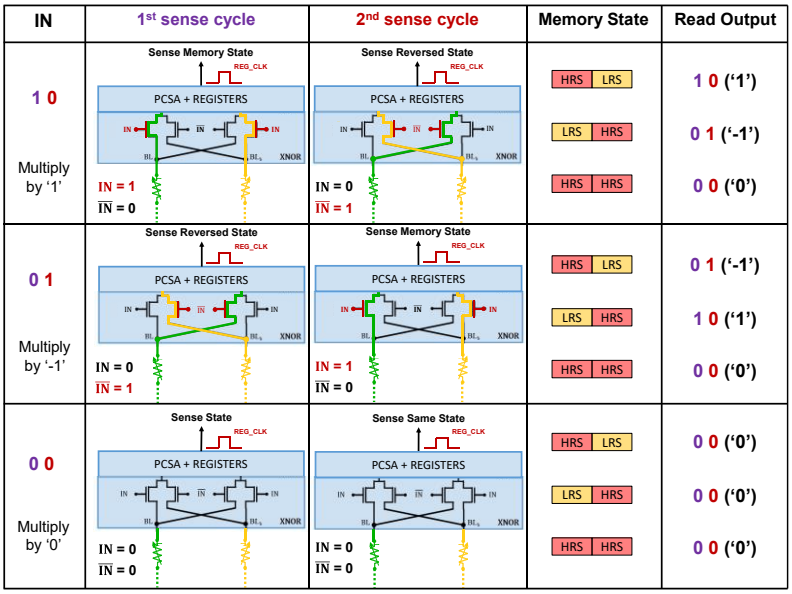

Fig. 4 Principle of logic-in-memory ternary input-weight multiplication using two cycles. The input is voltage IN, the weight is programmed into the memristors using the scheme of Fig. 2c. The ternary value is determined by the successive outputs of the register.

a

Value \ $T_{discharge}$	4ns	14ns	24ns	54ns	504ns
1	96.4%	96.4%	96.4%	96.4%	96.4%
-1	96.4%	96.4%	96.4%	96.4%	96.4%
0	20 .0%	6.6%	9.2%	10.9%	9.8%

b

Value \ $T_{discharge}$	4ns	14ns	24ns	54ns	504ns
1	96.4%	96.6%	96.4%	96.4%	96.6%
-1	100%	99.3%	98.9%	98.9%	99.3%
0	51%	3.9%	2.3%	3.7%	3.7%

Fig. 5 Experimentally-measured results, on the test chip of Fig. 1 for **a** VDD=1.2V. **b** VDD=1V. The tables list the success rate of logic-in-memory ternary multiplications, depending on the programmed weight values, and the chosen discharge time. Lower discharge time are needed for superior accuracy on zero-weight operations.