A 28nm 129.5fJ/Bit Embedded RRAM Macro with Cross-Coupled Sensing Scheme for Robust Near-Threshold Reading

Chenyang Li[1#], Keji Zhou[1,2#], Jin Xin[1], Wenkui Zhang[1], Chixiao Chen[1], Jianguo Yang[2], Xumeng Zhang[1*], Qi Liu[1*], Ming Liu[1]

[1]State Key Laboratory of Integrated Chips and Systems, College of Integrated Circuits and Micro-Nano Electronics, Frontier Institute of Chip and System, Fudan University, Shanghai, China; [2]Zhangjiang Laboratory, Shanghai, China;
[#]These authors contributed equally to this work. Email: xumengzhang@fudan.edu.cn, qi_liu@fudan.edu.cn

***Abstract*—Embedded RRAM is attractive for always-on and wake-up edge systems, yet low-VDD operation is fundamentally limited by the selector-dominated readout of conventional 1T1R arrays, which severely collapses the read margin in the near-threshold regime. This paper presents a low power RRAM macro with cross-coupled sensing and boosting pre-charge technique. Unlike conventional margin-limited sensing, the proposed cross-coupled structure exploits positive feedback to amplify resistance-dependent voltage development, thereby enhancing discrimination between different RRAM resistance states under scaled voltage. To further compensate for the reduced sensing margin at low VDD, a boosting pre-charge assist technique is introduced to strengthen the initial sensing trajectory. In addition, the peripheral driving circuits are optimized so that all IO devices can share the same voltage domain as core devices in read cycles, alleviating the burden on the external power management module. Fabricated in 28nm technology, the proposed macro achieves robust read at 0.5V supply and a measured read energy of 129.5fJ/bit at 0.6V, representing more than 50% improvement over prior work.**

Keywords*—RRAM, embedded non-volatile memory, cross-coupled sensing, boosting pre-charge.

I. Introduction

Low power wake-up and always-on edge systems require embedded non-volatile memory (NVM) that remains energy-efficient under aggressive supply-voltage scaling [1]. Among emerging memory technologies, resistive random access memory (RRAM) has emerged as a promising candidate due to its CMOS compatibility, high endurance, and low-programming voltage [2], [3]. However, existing RRAM macro designs have primarily focused on density [4] and high speed [2], while robust read operation at low supply voltage remains a critical challenge. As illustrated in Fig. 1, although VDD scaling is highly attractive for edge systems due to its quadratic power benefit [4], it introduces two key difficulties for embedded RRAM. First, conventional RRAM macros rely on dual voltage domains: core voltage and I/O voltage (required for high voltage write operations) [3]. Second, the on-resistance of the selectors in RRAM arrays becomes increasingly dominant as VDD scales down. Consequently, in the near-threshold region, the read current of a typical 1T1R cell is determined by the transconductance of the selector, rather than the RRAM resistance. The increasing on-resistance of the selector severely degrades the sensing margin and can cause read bit errors.

This work was supported by the National Key Research and Development Program (2024YFA1208800); the National Natural Science Foundation of China (Nos. 62488101, T2293732, and 62374040), the Key Research and Development Program of Shanghai (25LN3201200), the Shanghai Rising-Star Program (No.24QA2700400), the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (JYB2025XDXM120); the AI for Science Foundation of Fudan University (FudanX24AI036), and the National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, Xi'an Jiaotong University (HMHAL-202415).

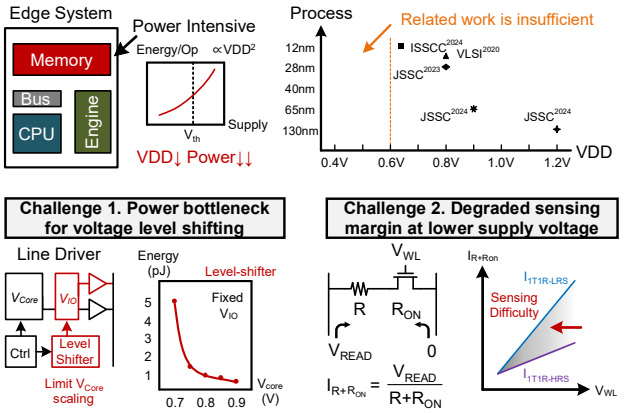


Fig. 1 Challenges for low VDD non-volatile memory design.

To address these challenges, this work presents a low VDD_{min} RRAM macro featuring:

- A 2T1R RRAM array architecture with a cross-coupled sensing scheme to enhance the read window of a single-level cell.
- A boosting pre-charge assist technique to further compensate for the read margin loss induced by VDD scaling.
- Customized level-shifters and peripheral drivers to minimize energy overhead during near-threshold reading while supporting high-voltage programming.

II. Proposed RRAM Macro Architecture

The proposed RRAM macro in Fig. 2 consists of a 32Kb 2T1R RRAM array partitioned into 32 subarrays, each with 128 column-wise selectors and 8 row-wise sensing units. The selectors' gate nodes are connected to line drivers to enable vertical access to the RRAM array. Every eight sensing units share a dedicated read-assist module, which generates a boosting pre-charged voltage applied to their gate nodes through multiplexing. A voltage sense amplifier (VSA) samples the voltage changes at this boosted node, determined by the selected RRAM's resistance, to perform 1-bit readout. In write cycles, address information is passed to level-shifters, which enable high-voltage access to target

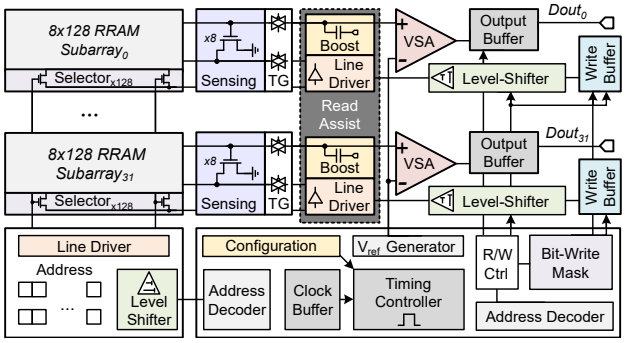


Fig. 2 Proposed RRAM macro architecture.

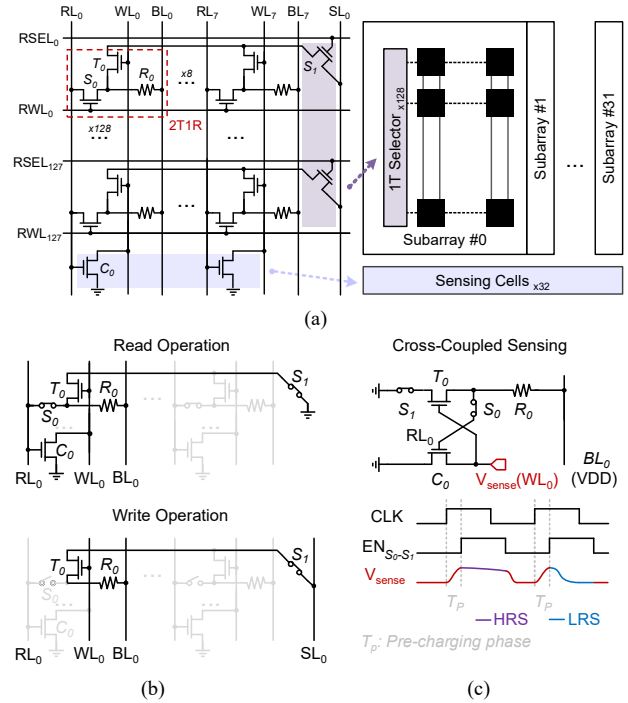


Fig. 3. (a) 2T1R array architecture; (b) Schematic of selected memory cell in read/write operations; (c) Cross-coupled sensing principle in read operations.

memory cells. A bit-write mask supports selective write operations on specific cells within a column.

A. *Cross-Coupled Sensing RRAM Array*

Previous studies [6]-[8] predominantly adopt one-transistor-one-resistor (1T1R) cells combined with current-mode sense amplifiers (CSAs) in macro design, since this implementation provides high memory density and fast access speed at nominal supply voltage. However, accurate current sensing incurs additional area overhead to generate an adaptive clamping voltage and a suitable reference current [3]. Furthermore, when VDD scales down, the limited voltage headroom [9] in the CSA sensing path and increased on-state resistance of MOS devices will severely corrupt 1T1R's sensing margin.

In order to ensure reliable read operations under VDD scaling, a cross-coupled sensing RRAM array is proposed in Fig. 3. Each subarray features 8x128 quasi-2T1R units: every eight 2T1R bitcells share a selecting transistor. A standard 2T1R structure is highlighted in Fig. 3(a): Transistor T_0 and RRAM R_0 form a 1T1R pattern while S_0 and S_1 serve as pass

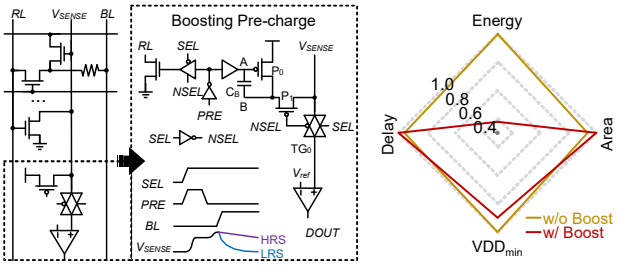


Fig. 4 Schematic of the proposed boosting pre-charge technology (left) and comparison of normalized key metrics including access delay, energy consumption and area overhead (right).

transistors to provide dedicated driving strength for each bitcell. In read cycles, the common node between T_0 and R_0 is connected, via S_0, to the gate node of the sensing unit C_0. The drain node of C_0 is in turn connected to the gate node of T_0, establishing a cross-coupled feedback structure. A simplified schematic for cell read operation is shown in Fig. 3(c). The voltage division between T_0 and R_0 determines the conduction state of C_0, thereby yielding different discharging levels on the sensing line (WL_0).

If R_0 is programmed to the low-resistance state (LRS), the reduced voltage drop across the RRAM raises the gate potential of C_0, which gradually pulls down V_{SENSE}. Meanwhile, the discharged V_{SENSE} weakens T_0's conduction state, creating a positive feedback loop to further elevate C_0's gate voltage.

If R_0 is programmed to the high-resistance state (HRS), the gate voltage of C_0 drops to a level insufficient to discharge V_{SENSE}.

Consequently, within a fixed access time, the two resistance states of the RRAM are converted into nearly rail-to-rail voltage (0-VDD) swings at the VSA input. The write operation is illustrated in Fig. 3(b) (bottom), where S_1 remains on while the sensing transistor C_0 and the selector S_0 are turned off. Each bitcell is configured as a 1T1R structure: set and reset pulses are applied to BL_0 and SL_0, accompanied by the appropriate WL_0 bias.

B. *Boosting pre-charge read-assist technology*

The effectiveness of the cross-coupled sensing scheme fundamentally depends on the voltage division between the RRAM and its series-connected transistor. When VDD scales to lower levels, the on-resistance of MOS devices can become comparable to, or even exceed, the RRAM's high resistance state, which may lead to sensing failure. Since the establishment of cross-coupled positive feedback is determined by the initial voltage division upon cell access, transiently elevating the transistor's gate voltage can significantly enhance the read margin. As shown in Fig. 4, a boosting capacitor C_B is coupled to device P_0. The combination of P_1 and TG_0 serves as a column multiplexer to select which cell's sensing node is pre-charged and amplified. During the initial phase of a read cycle, the signal PRE is high, pulling the V_{SENSE} node to VDD. Concurrently, the top plate (Node A) of C_B is grounded while its bottom plate (Node B) is clamped to VDD. The cross-coupled path remains inactive since the signal BL is held low. Upon the falling edge of PRE, Node A is driven to VDD, which bootstraps Node B to a higher potential via charge conservation. By applying this boosted voltage to V_{SENSE}, the memory cell is subjected to an equivalent increase in its supply voltage. Finally, BL is rapidly asserted high, triggering the voltage sensing phase.

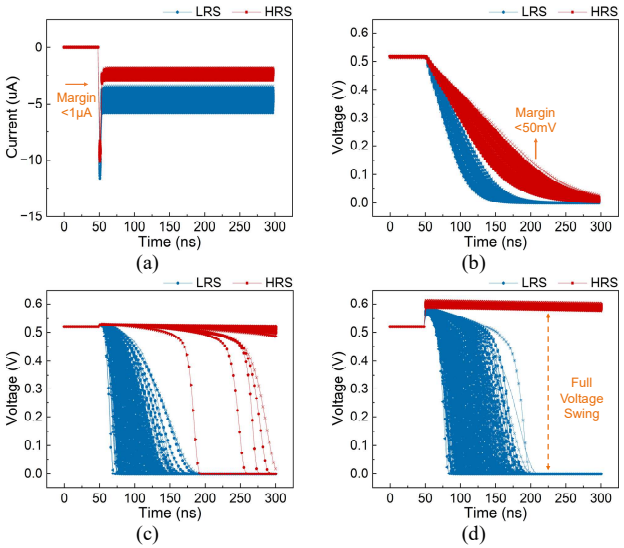


Fig. 5 1000 Monte Carlo simulation results of read margin at VDD=0.52V: (a) Current sensing for 1T1R; (b) Voltage sensing for 1T1R; (c) Proposed cells without boosting pre-charge; (d) Proposed cells with boosting pre-charge.

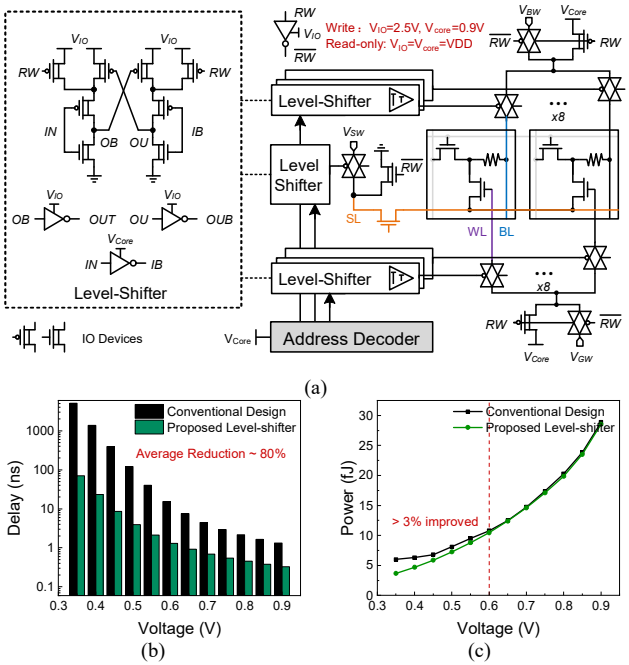
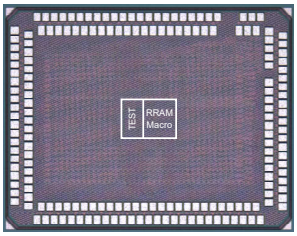


Fig. 6 (a) Proposed level-shifters and schematic for RRAM read/write paths; Comparison of (b) delay and (c) power with conventional level-shifters, under different VDD conditions (Core devices and IO devices are both biased at VDD).

The introduction of the boosting pre-charge technology enables reliable read operations at a lower VDD, with less than 5% area overhead and 1.6% delay penalty compared to normal driving circuits. It achieves a nearly 10% reduction in minimum operating voltage and reduces power consumption by over 60%. Fig. 5 compares the sensing margin of the conventional 1T1R and the proposed design, based on 1000 Monte Carlo simulation points. To better emulate the non-ideal behavior of RRAM devices, the test configuration uses LRS and HRS values of 20kΩ and 60kΩ, respectively. At 0.52V supply voltage, the read window for 1T1R cell is severely compressed regardless of whether current sensing or voltage sensing is employed. In contrast, the proposed bitcell



Technology	28nm
Die Area	2x1.55mm²
Capacity	32Kb
Cell Area	0.0962μm²/1b
VDD_{min}(read)	V_{IO}=V_{core}=0.5V
Frequency	5MHz@0.5V
	20MHz@0.6V
Read Energy	207.2fJ@0.5V
	129.5fJ@0.6V
Leakage	11.2μW@0.5V
	20.3μW@0.6V

Fig. 7 Die micrograph (left) and summary of chip parameters (right).

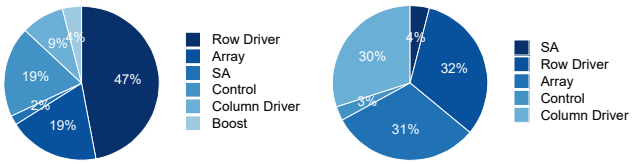


Fig. 8 RRAM macro (a) area breakdown; (b) power breakdown.

with boosting pre-charge technique achieves a nearly rail-to-rail read window, demonstrating an improved margin.

C. *Peripheral line drivers and level-shifters*

The physical programming mechanism of RRAM demands high voltages for write operations, requiring the array peripherals to incorporate IO devices as pass transistors and related control logic for high voltage transmission. However, since read operations are executed at nominal core voltage, the continued operation of IO devices during read cycles can introduce unnecessary power overhead. Level-shifters, in particular, impose stringent constraints on voltage scaling due to their feedback structure. To overcome these issues, this work proposes a modified level-shifter featuring a pair of gated switches, as depicted in Fig. 6(a). The control signal RW, operating in the IO voltage domain, is toggled according to the access mode. In write cycles, RW is driven to 2.5V, with the IO and core voltages maintained at 2.5V and 0.9V, respectively. This ensures normal control of the column multiplexers and provides sufficient write driving capability. In read cycles, RW is pulled to 0V. By disabling the internal feedback path, the level-shifter is reconfigured into two simple inverters. Under this condition, the supply voltage across the entire RRAM macro can be unified, allowing the IO and core domains to share a common VDD. Fig. 6(b) and (c) provide the simulated signal delay and power consumption. The results reveal that the introduction of gated transistors reduces the propagation delay by an average of 80% across the entire voltage range, thereby yielding power savings of 3% at 0.6V and 30% at 0.35V.

III. Measurement Results

Fig. 7 presents the chip micrograph and a summary of key parameters. Fabricated in a 28 nm CMOS process, the RRAM array achieves a cell area of 0.0962μm².

The macro area breakdown is shown in Fig. 8(a). Owing to the improved read margin from cross-coupled sensing scheme, the sense amplifiers require neither special design nor offset calibration methods. As a result, the read circuits occupy only a small fraction of the chip area. It is noted that the line drivers account for more than 50% of the macro area, as they contain a large number of IO devices and require upsizing to maintain sufficient driving speed at low supply voltages. This area overhead presents an opportunity for further optimization in future designs. Fig. 8(b) presents the

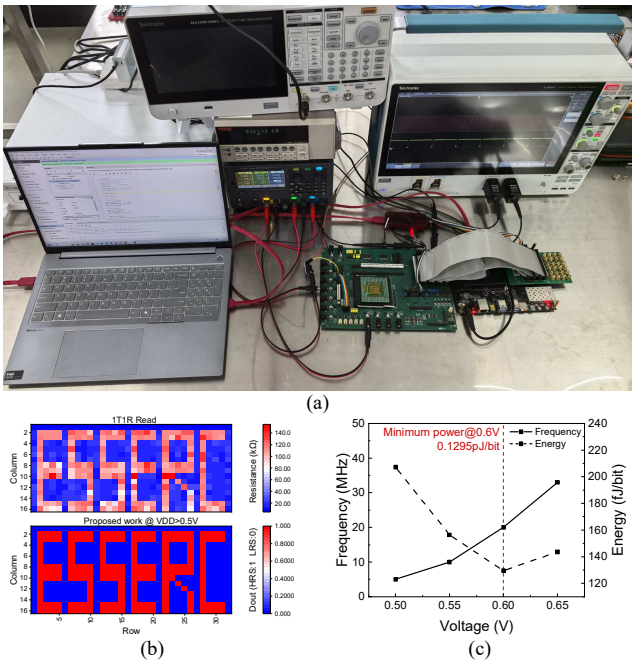


Fig. 9 (a) Test Setup; (b) Evaluation of read results using 1T1R and 2T1R Structures; (c) Measured performance and power consumption at different supply voltage.

macro read energy breakdown. The majority of the power is consumed during the establishment of cross-coupled positive feedback in the RRAM array, while the VSAs contribute to less than 5% of the total energy.

A testing board was built to verify the functionality of the proposed RRAM array. Since each cell supports standard 1T1R operation, the board integrates off-chip analog-to-digital converters (ADCs) to accurately extract the resistance of selected RRAM cells. The test setup is depicted in Fig. 9(a). The procedure begins with programming a subset of cells and acquiring their resistance values via the ADCs. Then, the ADCs are powered down and the chip switches to cross-coupled read mode independent of any off-chip circuitry. Fig. 9(b) illustrates the output matrix generated from the two RRAM read operations. The results indicate that the 2T1R structure maintains excellent LRS/HRS discrimination under low-voltage read conditions, even when some HRS resistances degrade to levels approaching the LRS range. The measured frequency and power consumption are shown in Fig. 9(c). The proposed structure achieves its optimal power efficiency at 0.6V, operating at 20 MHz with a read energy of 129.5 fJ/bit.

Table I compares this work with other RRAM macro designs. Although prior works achieve better performance in terms of memory density and access speed, this work exhibits a significant advantage in low-voltage read robustness. The read energy per bit is superior to that of other designs, achieving an improvement of at least 50%.

IV. Conclusion

This work demonstrates an RRAM macro architecture optimized for near-threshold sensing. By overcoming the intrinsic voltage scaling limits of conventional 1T1R arrays, the proposed design enables robust and energy-efficient operation at low supply voltage. The implementation of a cross-coupled feedback mechanism significantly enhances the read margin of RRAM cells under low voltage conditions. Meanwhile, with the assistance of the boosting pre-charge technique, the RRAM array achieves a rail-to-rail read window. The extra burden imposed on peripheral circuits due to such aggressive voltage scaling is effectively alleviated by the customized level-shifters, ensuring that the peripheral circuits can match the operational requirement of RRAM array. The final test results demonstrate robust RRAM read operations down to 0.5V supply. The optimal power consumption is achieved at 0.6V with a read energy of 129.5 fJ/bit, representing a nearly 50% improvement in power efficiency over prior arts.

Table I. Comparison of prior RRAM macros

	This Work	ISSCC[7] 2024	ISSCC[3] 2024	JSSC[6] 2024	JSSC[8] 2022
Process	28nm	40nm	12nm	40nm	40nm
Bitcell	2T1R	1T1R	1T1R	1T1R	1T1R
Cell area (μm^2)	0.0962	-	0.0249	0.286	0.364
Voltage (V) (V_{Core}/V_{IO})	V_{Core}=V_{IO} 0.5-0.9	0.8/3.3	1.62-1.98/ 0.63-0.77	0.9/2.5	1.1/-
Frequency (MHz)	20@0.6V	80	200	100	50
Read Energy (pJ/b)	0.2072@0.5V 0.1295@0.6V	0.256	-	1 for 1b 1.1 for 2b	1

References

[1] C. Y. Tsai, S. F. Liu, J. H. Hsuen, et al., "A CMOS-Compatible 12nm 8Mb MLC RRAM Enabling Producible 2-Bit Per Cell for High Energy Efficiency Compute-In-Memory in Edge AI Applications," in *Proc. Symp. VLSI Circuits*, Jun. 2025, pp. 1-3.

[2] J. Mu, L. Lu, J. E. Kim, et al., "A 1-Mb RRAM Macro With 9.8 ns Read Access Time Utilizing Dynamic Reference Voltage for Reliable Sensing Operation," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 71, no. 5, pp. 2564-2568, Mar. 2024.

[3] Y. C. Huang, S. H. Liu, H. S. Chen, et al., "15.7 A 32Mb RRAM in a 12nm FinFet Technology with a 0.0249 μm^2 Bit-Cell, a 3.2 GB/S Read Throughput, a 10K Cycle Write Endurance and a 10-Year Retention at 105°C," in *IEEE Int. Solid-State Circuits Conf*, Feb. 2024, pp. 288-290.

[4] W. Yue, Z. Jing, L. Ye, et al., "A 40nm 48.7 F^2/bit 1T2R Resistive Memory Bitcell Array for Adaptive Vector-Symbolic in-Memory Computing," in *IEEE European Solid-State Electronics Research Conference*, Sep. 2025, pp. 241-244.

[5] C. Xue, X. Qiao, X. Ren, et al., "A 768.7-2124.2 TOPS/W time-domain computing-in-memory macro with low static leakage and precision-configurable TDC," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 71, no. 5, pp. 2789-2793, May. 2024.

[6] A. Levy, L. R. Upton, M. D. Scott, et al., "EMBER: Efficient multiple-bits-per-cell embedded RRAM macro for high-density digital storage," *IEEE J. Solid-State Circuits*, vol. 59, no. 7, pp. 2081-2092, Jul. 2024.

[7] S. D. Spetalnick, A. S. Lele, B. Crafton, et al., "30.1 A 40nm VLIW edge accelerator with 5MB of 0.256 pJ/b RRAM and a localization solver for bristle robot surveillance," in *IEEE Int. Solid-State Circuits Conf*, Feb. 2024, pp. 482-484.

[8] M. Giordano, K. Prabhu, K. Koul, et al., "CHIMERA: A 0.92 TOPS, 2.2 TOPS/W edge AI accelerator with 2 MByte on-chip foundry resistive RAM for efficient training and inference," *IEEE J. Solid-State Circuits*, vol. 57, no. 4, pp. 1013-1026, Apr. 2024.

[9] M. F. Chang, J. J. Wu, T. F. Chien, et al., "Low VDD_{min} 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. 2024.