Resistive Random-Access Memory (ReRAM) Circuit Design Optimization

Rui Liu *

Department of Computer Science and Technology, Xi'an Jiaotong-Liverpool University, Suzhou, China

* Corresponding Author Email: Rui.Liu23@student.xjtlu.edu.cn

Abstract. With the fast development and broad application of artificial intelligence, mobile terminals, and autonomous driving, the advanced System-on-Chip (SoC) design faces unprecedented memory demand, especially under the two core requirements for High-Density Integration and Low-Power Operation. Under such circumstances, Resistive Random-Access Memory (RRAM) emerges as a promising candidate to replace traditional memories, owing to its high storage density and low energy consumption, excellent non-volatile characteristics, and good compatibility with the standard CMOS manufacturing process. This paper systematically sorts out and reviews a series of significant progress and research threads in recent years in RRAM technology, covering multiple levels, from process optimization at the device level and design for pure storage applications to near-memory computing architectures and more innovative in-memory computing architectures. From this overall technical perspective, this paper carries out an in-depth analysis and discussion on the huge application potential of RRAM in building the next-generation, high-energy-efficiency SoC system, the key technological paths to be broken through, and the broad application prospects and development directions that may be realized in the future, providing a systematic reference for subsequent research in this field.

Keywords: RRAM, Low power, In-memory and near-memory computing.

1. Introduction

With smartphones, wearable devices, and AI edge devices becoming more prevalent, the need for high-density, low-power System on Chips (SoCs) has been rising. Memory is a critical module for data storage and interaction in the SoC design process, and its area efficiency and energy consumption directly affect the chip's integration level and the device's battery life. Therefore, research on high-density, low-power memory development has been highly focused in many fields.

From the perspective of today's mainstream memory technologies, there is a significant difference in performance, power, and integration among various types of memory; thus, each has its strengths and weaknesses. For instance, due to the fast read/write speed and low power consumption, Static Random-Access Memory (SRAM) is the first choice for fast caches. However, its six-transistor structure leads to extremely low storage density, and leakage is one of the problems. Compared with SRAM, Dynamic Random-Access Memory (DRAM) enhances the storage density significantly due to the single-transistor-capacitor structure, but leakage is still a problem. Besides, it also needs to be refreshed periodically to keep the data, which increases the power consumption. Flash can retain the data without continuous power; however, the write power consumption is much higher than the read power consumption. Most importantly, it is hard for CMOS processes to achieve high-density integration in SoCs with 22nm and below, so they cannot meet the storage requirement. Under such circumstances, Resistive Random-Access Memory (RRAM) is one of the most promising new memory technologies due to its outstanding overall performance. It has extremely high storage density, orders of magnitude lower read-write power than Flash, and almost no leakage problems. RRAM thus has irreplaceable value in high-density and low-power SoC applications, representing a core direction in research on new memory technology. This paper discusses RRAM circuit design in three aspects: advanced memory devices, pure storage, in-memory and near-memory computing.

2. Preliminary

In the general operation principle of RRAM, the fundamental memory cell (bit cell) stores data by changing the resistance state. The process can be broadly divided into three main stages: Forming, Writing, and Reading.

Forming: Initially, the material in an RRAM cell is in the High Resistance State (HRS). Applying relatively high voltages (generally >3V) creates a conductive filament in the resistive material, switching the cell into a low-resistance state (LRS). Generally, this process is irreversible.

Write: The write operation controls the resistance state using voltages of opposite polarities. Thus, a positive voltage is used in the SET operation, which writes "1" in the cell and decreases its resistance value. In contrast, a negative voltage is used in the RESET operation, which writes "0" in the cell and returns it to a high resistance state.

Read: A small voltage (0.3-0.5 V) is applied to measure the current and determine the resistance state. A low voltage is used to ensure that the stored state is not disturbed. A low resistance state implies that "1" is stored, and a high resistance state indicates that "0" is stored.

3. The Design of RRAM

3.1. Advanced RRAM Devices

Peking University has proposed an advanced logic-process-compatible RRAM architecture, providing a record high memory density of 15.43 Mb/mm² while keeping a data retention of 10 years at 150 °C. A dynamic gate and self-passivation sidewall bit cell design without Shallow Trench Isolation (STI) significantly improve storage density and reliability while providing good retention performance at elevated temperatures, thus making this device a strong candidate for future high-performance computing (HPC) and memory systems [1].

Forschungszentrum Jülich GmbH reports in IEEE Electron Device Letters 2025 an optimized TaO_x ReRAM bitcell with a Ti (2 nm)/Ta (3 nm) bilayer ohmic electrode (OE). Compared with single-layer OE devices, the forming voltage is lower, the SET switching speed is ~11× faster, the RESET switching speed is ~6× faster, and the switching energy is ~6× lower. In addition, the device demonstrates good endurance up to 10^7 cycles with ~10× lower resistance drift, indicating high stability and efficiency. This highlights bilayer OE engineering's enormous potential to push ReRAM to higher densities [2].

The University of Chinese Academy of Sciences proposes a high-area-efficiency (6 TOPS/mm2) multimodal neuromorphic computing system based on 3D multifunctional RRAM arrays. It can realize low-power, high-speed, highly integrated neuromorphic computing through the combination of non-volatile and dynamic RRAM devices. RRAM has been proven to have excellent computing performance and area efficiency and is at the forefront of visual and auditory perception. Advanced 3D stacking technology breaks through the limitations of the traditional structure to realize high throughput and low energy consumption, which further promotes RRAM for neuromorphic computing applications [3].

Table 1 depicts recent research work on three RRAM categories in comparison from different perspectives, including Technology, Memory Type, Size, Throughput, Supply Voltage (VDD), and Write Endurance. As presented, the 1 Mb 1T1R array in Ref. [1] is integrated with a 40-nm-CMOS process while it works with 0.9–1.1 V. An RRAM array with an endurance of more than 10 million cycles, higher than the other works is reported in Ref. [2]. In addition, a 3D multifunctional RRAM array with neuromorphic computing abilities and an area efficiency of 6 TOPS/mm 2 in throughput is reported in Ref. [3].

Table 1. Comparison of Three Types of RRAM

Source	IEDM'23[1]	EDL'25[2]	IEDM'23[3]
Technology	40nm CMOS	N/A	3D multifunctional RRAM array
Memory Type	1T1R	RRAM array	3D multifunctional RRAM
Size	1 Mb	N/A	N/A
Throughput	N/A	N/A	6 $TOPS/mm^2$
VDD	0.9-1.1 V	N/A	N/A
Write Endurance	>12k cycles	$>10^7$cycles	N/A

3.2. With Storage as the Core

IMECAS presents 14-nm FinFET 1-Mb embedded 1T1R RRAM with a 0.022-μm2 cell for high-density eNVM integration. The self-adaptive delayed-termination (SADT) and multiple-dummy-cell reference are adopted to enhance the post-write filament stability and improve the read reference statistical reliability. These methods enlarge the effective read window and suppress the reference collapse caused by tail bits. Reliable reading is supported by a minimum 0.5-V supply from −40 °C to 125 °C, and the measured read access time is 9.5 ns at 0.8 V and 21 ns at 0.4 V (room temperature). Compared with the prior self-write termination, SADT improves the write endurance by an order of magnitude, decreases the Ron/Roff retention-failure rate by 87.7%/68.6%, and increases the Roff/Ron window by 3×, showing the characteristics of high-density, low-voltage, and reliable nonvolatile storage for embedded RRAM at advanced nodes [4].

TSMC presented a 32 Mb 1T1R RRAM macro in 12-nm FinFET at ISSCC 2024. The design, with a 0.0249 μm2 bit cell and a pipelined read architecture, reaches 200 MHz read frequency at 0.6 V and 3.2 GB/s read throughput. The dual-path WL driver decouples the read/write biasing; a high VWWL is used for writing and a low VRWL for reading. Measured results showed 104-cycle write endurance and 10-year retention at 105 °C, demonstrating RRAM's "high-density × high-bandwidth × low-voltage" advantages for deep logic-memory integration [5].

Tsinghua University presented a 4-Mb embedded 1T1R RRAM IP in commercial 28-nm CMOS with a reliability-enhanced DTCO flow. From the storage perspective, the IP achieves 4-bit/cell MLC, ≥10^8 read-disturb immunity, and >10-year retention at 125 °C; write endurance is 10^7 cycles at a 6-Kb sub-macro and 10^6 cycles at the chip level, showing the robust characteristics of nonvolatile storage and potential for scaling into advanced technology nodes [6].

TSMC demonstrates a high-density embedded 3-D stackable Via RRAM in TSMC's 16-nm FinFET CMOS logic process. Using a 1T10R vertical cell, it achieves 0.1 Gb/mm2 in the array —> 2.2× that of a 2-D 1T1R in the same process. The device demonstrates storage robustness, >10^4 switching-cycle endurance, and stable retention in a 150 °C baking test. Without any additional mask, the Via RRAM is logic-compatible and can be directly integrated into SoCs/MCUs as an embedded NVM [7].

Table 2 summarizes recent research on pure storage RRAM from different perspectives, including Technology, Memory Type, Size, Supply Voltage (VDD), Density, Frequency, and Retention. As can be seen, the device in Ref. [5] is fabricated using a 12-nm FinFET process. The RRAM in Ref. [7] prints the largest storage density, with a memory size of 128 Mb and an areal density of 0.1 Gb/mm 2. It also has good high-temperature retention of more than 10 years at 150 °C, which is the best among the other works.

Table 2. Comparison of Four Types of RRAM (Storage)

Source	ISSCC'21[4]	ISSCC'24[5]	IEDM'24[6]	TED'24[7]
Technology	14nmFinFET	12nmFinFET	28nmCMOS	16nmFinFET
MemoryType	1T1R	1T1R	1T1R	1T1R
Size	1Mb	32Mb	4Mb	128Mb
VDD	0.5V	0.6 V	N/A	0.8 V
Density	0.0220μm^2/cell	0.0249μm^2/cell	N/A	0.1Gb/mm^2

Frequency	N/A	200MHz	N/A	N/A
Retention	N/A	10 years at 105°C	>10 years at 125°C	>10 years at 150°C

3.3. In-Memory Computing and Near-Memory Computing

National Tsing Hua University presents a 22-nm 8-Mb ReRAM-CIM macro based on foundry logic-compatible 1T1R SLC ReRAM. The overall energy consumption is reduced by 5.15× compared with that of the conventional current-mode CIM under 8-b input × 8-b weight/19-b output by employing a direct-current-free time-space in-memory computing (DCFTS-IMC) scheme to remove the DC current from the cell array and an integration-based voltage-to-time converter (IVTC) to expand the time-step margin by 1.58×. The computing latency is 14.4 ns, and the energy efficiency reaches 21.6 TOPS/W energy efficiency [8].

At ISSCC 2021, Georgia Tech and TSMC, together with collaborators, demonstrated a 40-nm 64-Kb digital ReRAM CIM macro. It can also perform in-situ post-write verification with read-disturb monitoring and write-back capability for improved stability with a voltage-based active-feedback readout. With 1–9-row parallelism, it offers 100-MHz evaluation clock peak energy efficiency of 56.67 TOPS/W (average 4.15 TOPS/W) to demonstrate the possibility of RRAM for high-efficiency CIM [9].

Georgia Tech and TSMC at ISSCC 2022 Report a 40nm 64Kb RRAM Binary/CIM Macro. The design uses HV-domain word-line encoding + voltage-regulated current sensing to provide >75% ADC dynamic range usage and 45-75mV voltage domain separation. Integrating 8×1-4b read channels, HV write drivers, and self-biased/reference circuits into 0.027mm2 yields 2.37Mb/mm2. For 9-row parallel 1×1b MAC at 0.83V, 64MHz, measured peak/average energy efficiency is 26.56/5.63 TOPS/W. This result exemplifies the strength of RRAM-based high-density CIM in read chain linearization and area/energy efficiency for system integrability [10].

Applied Materials, Inc. presents a forming-free, operando bulk-switching b-ReRAM. At 0.1 V, 25k devices on 15 dies are programmed stably to 128 levels with analog-like, low-noise behavior and wafer-scale uniformity. This BEOL memory is integrated on a 65-nm CMOS SoC in a 4 × (64 × 64) 1T1R array. In keyword spotting, measured accuracy matches that of software quantization, and compute efficiency reaches 98.5 TOPS/W. The read-disturb shows no discernible change beyond $>10^6$ cycles, and the cells are stable following 10^3 program cycles, highlighting the use of b-ReRAM for low-current multi-level weights for near-memory computing [11].

TSMC and collaborators report in ISSCC 2024 a 40-nm VLIW edge accelerator with 5 MB on-chip RRAM tightly coupled for near-memory computing. The RRAM bank is floor planned into 18 modular sub-banks for APR and read/write routing. Measurements show an access cost of 0.256 pJ/bit (VMIN) and an NVM bandwidth of 12.8 GB/s (VMIN), with representative operating points of 0.8 V/80 MHz on the measured shmoo. These results demonstrate the energy- and bandwidth-efficiency advantages of RRAM-based near-memory computing in a VLIW edge SoC [12].

Stanford University demonstrates monolithic 3D integration of HfOx RRAM and CNFETs directly on top of Si CMOS in a ≤415 °C BEOL process developed in the SkyWater 90/130 nm foundry. The BEOL CNFET+RRAM stack achieves iso-footprint (1.06× of FEOL Si FET+RRAM) and iso-performance (near-identical read power, write energy/latency) compared with a conventional FEOL Si FET+RRAM baseline, and $>10^6$ set/reset endurance, >10 year projected retention at 125 °C, and the first BEOL demonstration of 2-bit-per-cell operation. The results on the same wafers and node enable >5× system-level EDP benefits in iso-footprint, iso-capacity M3D M3D architectures [13].

As summarized in Table 3, all five in-/near-memory computing RRAM designs are fabricated with a CMOS technology node in the range of 22–65 nm, and the memory topology is 1T1R. The memory sizes of the implemented designs are 64 Kb to 8 Mb, and one of them reaches 5 MB, which illustrates the trend toward large-scale integration with computing ability on-chip. The operating voltages usually remain near 1 V for low-power operation. The range of energy efficiency is wide. The highest

point can be asserted to be 98.5 TOPS/W, while one design also achieves an operational frequency of up to 210 MHz.

Table 3. Comparison of Five Types of RRAM

Source	JSSC'23 [8]	ISSCC'21 [9]	ISSCC'22 [10]	IEDM'22 [11]	ISSCC'24 [12]
Technology	22nmCMOS	40nmCMOS	40nmCMOS	65nmCMOS	40nmCMOS
MemoryType	1T1R	1T1R	1T1R	1T1R	1T1R
Size	8Mb	64 Kb	64 Kb	N/A	5MB
VDD	0.8/0.9V	1.0V	0.83V	N/A	0.8-1.1V
Frequency	N/A	N/A	N/A	N/A	80-210MHz
Energy Efficiency	21.6TOP/W	56.67TOPS/W (peak)	26.56TOPS/W (peak)	98.5TOPS/W	0.84TOPS/W

4. Conclusion

In summary, from device optimization through pure memory arrays to in-memory computing, this RRAM circuit design research maps out a clear trend: RRAM evolves from a simple storage element into a core component of intelligent energy-efficient computation. Through co-optimization of materials, devices, and circuits, RRAM technology is expected to cross the memory wall and redefine future computing systems, especially for data-intensive edge and artificial intelligence applications.

References

[1] Wang Qishen, Yang Yuhang, Wang Zongwei, et al. "A Logic-Process Compatible RRAM with 15.43 Mb/mm2 Density and 10years@150°C retention using STI-less Dynamic-Gate and Self-Passivation Sidewall", in Proc. International Electron Devices Meeting, 2023, 1-4.

[2] Paul G, Glint T, Menzel S, et al. Bilayer Ohmic Electrode Engineering in TaOX ReRAM Devices. Electron Device Letters, 2025, 46(9): 1537-1540.

[3] Sun Wenxuan, Li Yi, Zhang Woyu, et al. "High Area Efficiency (6 TOPS/mm2) Multimodal Neuromorphic Computing System Implemented by 3D Multifunctional RRAM Array", in Proc. International Electron Devices Meeting, San Francisco, 2023, 1-4.

[4] Yang Jianguo, Xue Xiaoyong, Xu Xiaoxin, et al. "24.2 A 14nm-FinFET 1Mb Embedded 1T1R RRAM with a 0.022μm2Cell Size Using Self-Adaptive Delayed Termination and Multi-Cell Reference", in Proc. International Solid-State Circuits Conference, 2021, 336-338.

[5] Huang Y.C, Liu S.H, Chen H.S, et al. "15.7 A 32Mb RRAM in a 12nm FinFet Technology with a 0.0249μm2 Bit-Cell, a 3.2GB/S Read Throughput, a 10KCycle Write Endurance and a 10-Year Retention at 105°C", in Proc. International Solid-State Circuits Conference, San Francisco, 2024, 288-290.

[6] Zhang Junyang, Ma Xiangchao, Xi Yue, et al. "A 28nm 4Mb Embedded RRAM IP with Record-High Endurance of 107 Cycles and 10 Years@125°C Retention through Reliability-Enhanced Design-Technology Co-Optimization", in Proc. International Electron Devices Meeting, 2024, 1-4.

[7] Huang Y.H, Hsieh Y.C, Lin Y.C, et al. High-Density Embedded 3-D Stackable Via RRAM in 16-nm FinFET CMOS Logic Process. Transactions on Electron Devices, 2024, 71(6): 3614-3619.

[8] Hung J.M, Wen T.H, Huang Y.H, et al. 8-b Precision 8-Mb ReRAM Compute-in-Memory Macro Using Direct-Current-Free Time-Domain Readout Scheme for AI Edge Devices. Journal of Solid-State Circuits, 2023, 58(1): 303-315.

[9] Yoon J.H, Chang Muya, Khwa W.S, et al. "29.1 A 40nm 64Kb 56.67TOPS/W Read-Disturb-Tolerant Compute-in-Memory/Digital RRAM Macro with Active-Feedback-Based Read and In-Situ Write Verification", in Proc. International Solid-State Circuits Conference, San Francisco, 2021, 404-406.

[10] Spetalnick S, Chang Muya, Crafton B, et al. “A 40nm 64kb 26.56TOPS/W 2.37Mb/mm2RRAM Binary/Compute-in-Memory Macro with 4.23x Improvement in Density and >75% Use of Sensing Dynamic Range”, in Proc. International Solid-State Circuits Conference, San Francisco, 2022, 1-3.

[11] Wu Y, Cai F, Thomas L, et al. “Demonstration of a Multi-Level μA-Range Bulk Switching ReRAM and its Application for Keyword Spotting”, in Proc. International Electron Devices Meeting, San Francisco, 2022, 18.4.1-18.4.4.

[12] Spetalnick S, Lele A, Crafton B, et al. “30.1 A 40nm VLIW Edge Accelerator with 5MB of 0.256pJ/b RRAM and a Localization Solver for Bristle Robot Surveillance”, in Proc. International Solid-State Circuits Conference, San Francisco, 2024, 482-484.

[13] Srimani T, Yu A.C, Radway R.M, et al. “Foundry Monolithic 3D BEOL Transistor + Memory Stack: Iso-performance and Iso-footprint BEOL Carbon Nanotube FET+RRAM vs. FEOL Silicon FET+RRAM”, in Proc. Symposium on VLSI Technology and Circuits, Kyoto, Japan, 2023, 1-2.