A 25-Layer 3D Vertical RRAM Array Featuring Separate-Memory-Cell Architecture: High Density (25Mb/mm²) and Enhanced Electro-Thermal Reliability via Design-Technology Co-Optimization

Wenxuan Sun[1=], Yan Yang[2=], Lihua Xu[3=], Jiayi Bao[1], Kaiping Zhang[1], Yu Liu[1], Jinru Lai[1], Danian Dong[1], Xu Zheng[1], Qingyun Zuo[4], Jing Liu[4], Xiaoxin Xu[1*], Qi Liu[5], Ming Liu[1]

[1]State Key Lab of Fabrication Technologies for Integrated Circuits, Institute of Microelectronics of the Chinese Academy of Sciences, Beijing; [2]College of Communication Engineering (College of Microelectronic), Chengdu University of Information Technology, China; [3]School of Integrated Circuits, Anhui University, Hefei, China; [4]Beijing Memsilicon Technology Co., LTD, China; [5]Fudan University, Shanghai, China. =These authors contributed equally to the work. *Email: xuxiaoxin@ime.ac.cn

***Abstract*—For the first time, we demonstrated a 25-Layer 500nm 3D Vertical RRAM (3D VRRAM) array with excellent electro-thermal reliability. We perform Design-Technology Co-Optimization (DTCO) across three aspects: device, circuit, and system. At the device level, we designed Separate-Memory-Cell (SMC) to suppress inter-layer electrical and thermal crosstalk through reduced program voltage and enhanced dielectric isolation. At the circuit level, we designed an adaptive temperature-sensing circuit that dynamically adjusts V_{ref} to compensate for High- and Low-Resistance States (HRS/LRS) thermal drift, which can significantly reduce the bit error rate. At the system level, we proposed a high-temperature protection algorithm that can dynamically adjust the chip throughput to prevent thermal-induced failures. Other outstanding performances of SMC-3DVRRAM array such as: high density (25Mb/mm²), rectification ratio>2000, retention at 100℃>10^4s，endurance>10^6.**

Keywords—3D VRRAM，Separate-Memory-Cell (SMC)，DTCO, Electro-Thermal Reliability.

I. INTRODUCTION

Memory serves as the technological cornerstone and core pillar of the integrated circuit industry. Driven by the AI wave, its status is evolving from traditional "cost component" into "strategic asset" that dictates computing performance. Three-Dimensional Vertical Resistive Random Access Memory (3D VRRAM) has emerged as an ideal candidate for Storage Class Memory (SCM) [1-4]. It uniquely combines the nanosecond-level speed of DRAM, the non-volatility of NAND flash, and density potential that surpasses both. Moreover, 3D VRRAM holds transformative promise for emerging applications such as compute-in-memory (CIM), neuromorphic computing, and edge computing [8-11]. As such, it is widely regarded as a key technology for breaking through the traditional "memory wall" bottleneck and sustaining AI computing in the post-Moore's Law era.

3D VRRAM-based intelligent SOC is an effective solution for high-density on-chip memory (Fig.1a). However, high density and low power 3D VRRAM array faces three key challenges as shown in Fig.1b. **Challenge 1: Write Crosstalk.** Because devices in the vertical direction share the selection layer and resistive switching layer, the electric field of the selected cell can easily radiate to unselected cells, causing write crosstalk. **Challenge 2: Thermal Crosstalk.** While the 3D vertical architecture increases storage density, it significantly worsens thermal dissipation conditions. Moreover, thermal coupling effects can cause heat from a single cell to be transferred to neighboring cells, resulting in thermal crosstalk. **Challenge 3: Resistance Drift.** Thermal accumulation effects and the resulting resistance drift pose a fundamental threat to the accuracy guarantee of compute-in-memory (CIM) applications.

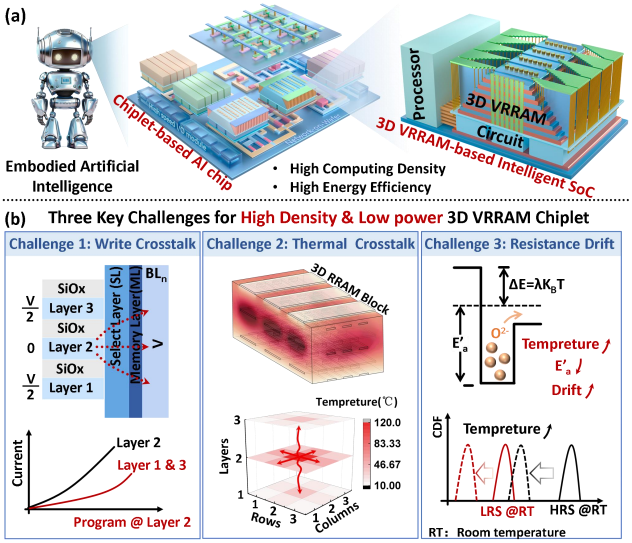

Fig. 1. Motivation: (a) 3D VRRAM chiplets are suitable for artificial intelligence chips with high computing density. (b) Three key challenges faced by high density and low power 3D VRRAM chiplets.

In this work, we provide a solution to the three key challenges faced by 3D VRRAM chiplets (Fig.2). **Highlight 1:**

Design Separate-Memory-Cell (SMC) structure to prevent inter-layer electrical and thermal crosstalk in 3D VRRAM array. **Highlight 2:** The V_{ref} adjusts adaptively when the 3D VRRAM chiplets temperature changes, which effectively mitigates system errors caused by resistance drift. **Highlight 3:** The operating frequency is adjusted based on real-time temperature monitoring of the 3D VRRAM chip to prevent complete failure due to overheating.

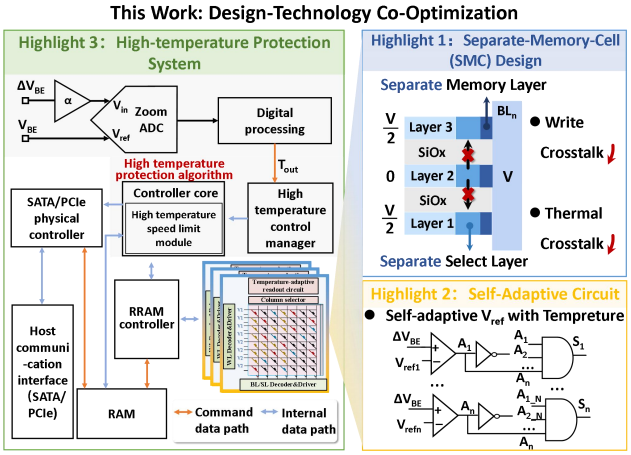

Fig. 2. This work provides a solution to the three key challenges faced by 3D VRRAM chiplets.

II. 25-Layer SMC-3D VRRAM Array

A. Performance of 25-Layer SMC-3D VRRAM Array

The TEM image of 25-layer SMC-3D VRRAM array is shown in Fig.3a, and a brief fabrication flow of our SMC-3D VRRAM is shown in Fig.3b. Multiple TiN (25 nm)/SiO_2 (15 nm) layers are deposited alternately by PVD and PECVD, followed by patterning and one-step etching to form stacked wordlines (WLs). Then, the WLs were laterally etched to create recesses (Fig.3c) and subsequently annealed to form $Ti_xO_yN_z$. HZO as resistive layer was deposited in the recess, followed by one-step etch to be broken (Fig.3d). The top electrode was deposited by ion beam sputtering, followed by lift-off process to form Bit Lines (BL). Fig.3e shows the EDS line of our device.

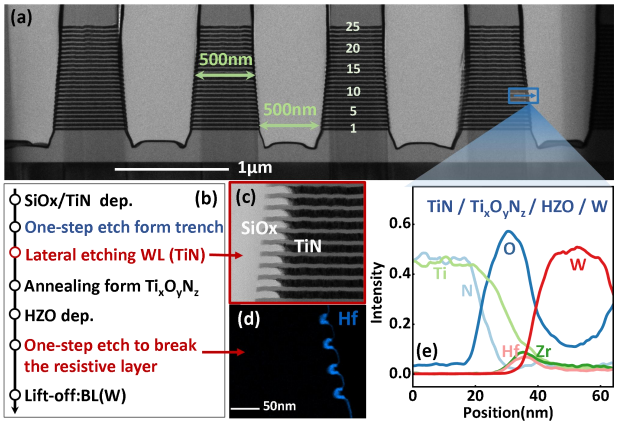

Fig. 3. 25-layer 3D VRRAM array. (a) TEM images. (b)Fabrication flow. (c-d) Morphology of SMC-3D VRRAM. (e) EDS analysis.

The I-V characteristics of vertical RRAM devices from all 25 layers as show in Fig.4a, demonstrating high inter-layer uniformity. The retention could be retained over 10^4 s at 100℃ (Fig.4b). Moreover, the endurance test shows that the device can reliably switch between HRS and LRS for more than 10^6 cycles (Fig.4c). In addition, resistance distributions of 25-layer devices indicate good inter-layer uniformity (Fig.4d).

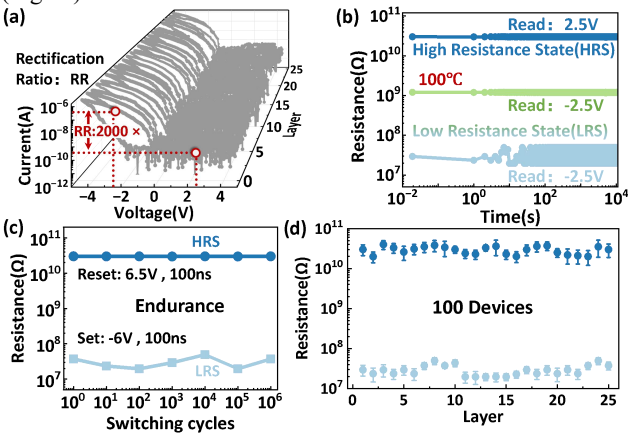

Fig. 4. Characteristics of 25-layer 3D VRRAM: (a) I-V curves. (b) Retention at high temperature. (c) Endurance and (d) LRS / HRS distributions.

B. Electro-Thermal Modeling of SMC-3D VRRAM Array

Compared with the conventional structure, the electric field distribution of the SMC structure is more concentrated, as shown in Fig. 5(a-b). Therefore, for the SMC structure, identical electric field strength can be achieved with a smaller program voltage. When the device in the 2nd layer (L2) is selected, the electric field strength at the unselected devices in the 1st and 3rd layers (L1&3) is significantly reduced (Fig. 5c). As a result, write disturbance in the array is reduced compared to the conventional structure.

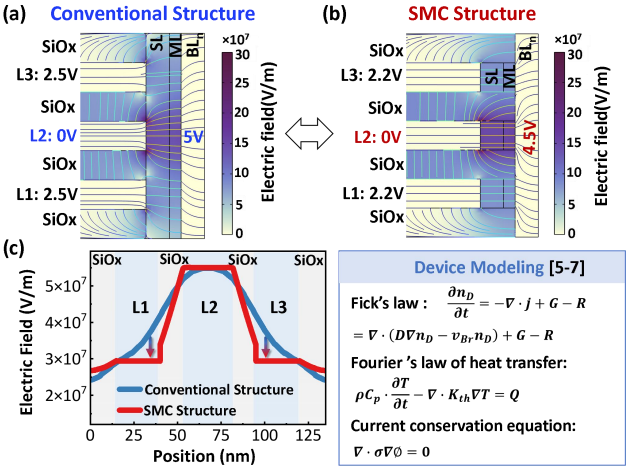

Fig. 5. **Electric field distribution** of (a) conventional structure and (b) SMC structure 3D VRRAM. (c) The significant reduction in write crosstalk can be attributed to the lower program voltage and isolating properties of dielectric layer.

Thermal field distribution of conventional and SMC

structure is shown in Fig.6(a-b). The significant reduction (Fig.6c) in thermal crosstalk can be attributed to the reduction of write crosstalk and isolating properties of dielectric layer.

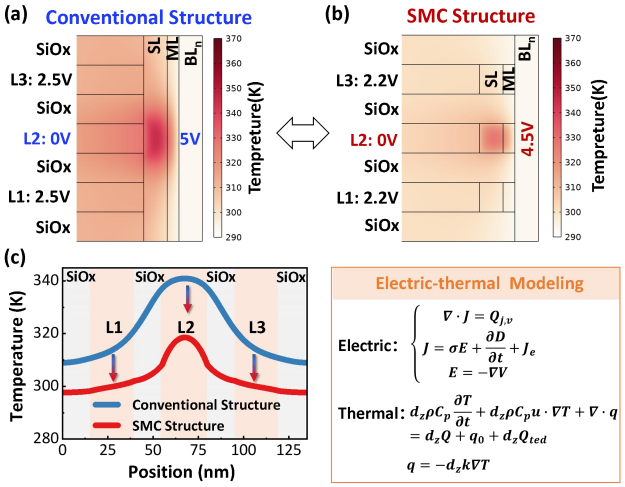

Fig. 6. **Thermal field distribution** of (a) conventional structure and (b) SMC structure 3D VRRAM. (c) The significant reduction in thermal crosstalk can be attributed to the reduction of write crosstalk and isolating properties of dielectric layer.

C. *Vertical Stacking Potential of SMC-3D VRRAM*

Both conventional and SMC structures experience the same electric field strength and feature identical TiN thickness. When the electric field in the unselected layer attains two-thirds of the programming electric field, it is defined as write crosstalk. A 22nm SiOx layer is required in the conventional structure to suppress write crosstalk (Fig.7a-b). However, the SMC structure only needs 13 nm (Fig.7c-d). This can be attributed to the substantially higher resistivity of SiOx compared to memory and selector layers. With the single-step etch depth of 6 μm and TiN thickness of 25 nm, the SMC structure can achieve 158 layers, but the conventional structure is limited to 127 layers (Fig.7e).Compared to 1T1R and conventional 3D VRRAM, the SMC-3D VRRAM array demonstrated a significant density advantage(Fig.8).

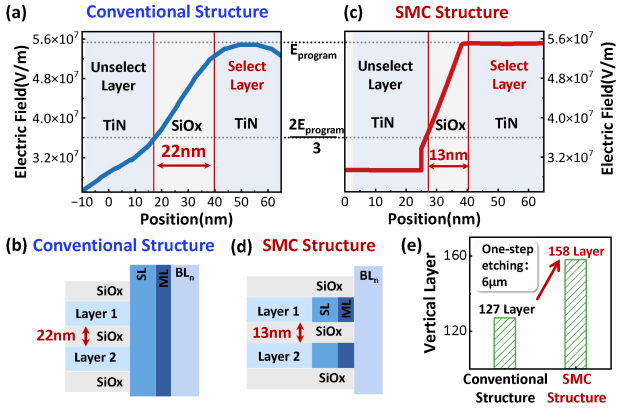

Fig.7. (a-b) 22nm-SiOx layer is required in the conventional structure to suppress write crosstalk. (c-d) The SMC structure only needs 13 nm to suppress write crosstalk. (e) With the single-step etch depth of 6 μm and TiN thickness of 25 nm, the SMC structure can achieve 158 layers.

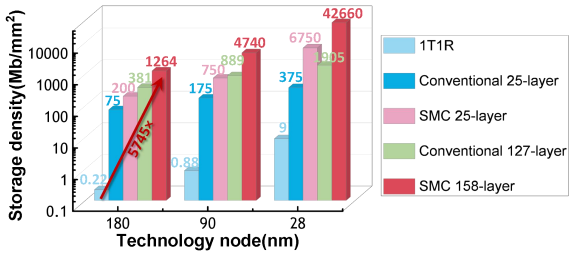

Fig.8. Compared to 1T1R and conventional 3D VRRAM, the SMC-3D VRRAM array demonstrated a significant density advantage.

III. Temperature-Sensing Self-Adaptive Circuit

We proposed a temperature-sensing adaptive circuit based on memristor characteristic tracking as shown in Fig.9. Unlike conventional fixed bandgap references (BGRs), this circuit leverages real-time dual-feedback sensing of both 3D VRRAM resistance and ambient temperature to dynamically adjust the reference voltage, thereby compensating for temperature-induced resistance drift and ensuring stable, reliable output of the 3D VRRAM chip.

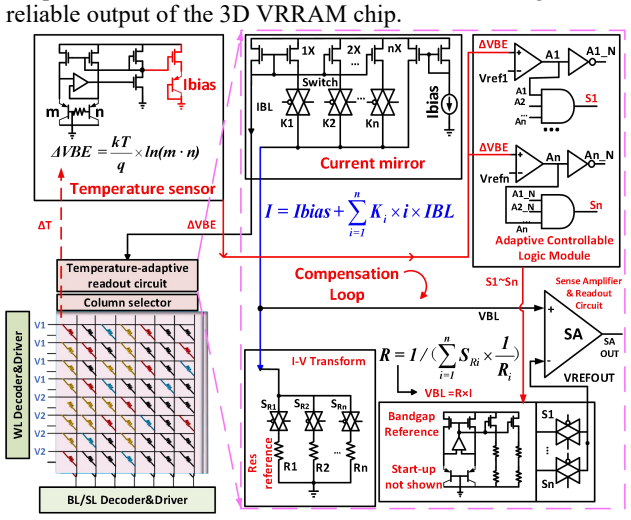

Fig. 9. The temperature-sensing self-adaptive circuit uses dual feedback from the 3D VRRAM's resistance and real-time temperature measurements to dynamically adjust V_{ref}.

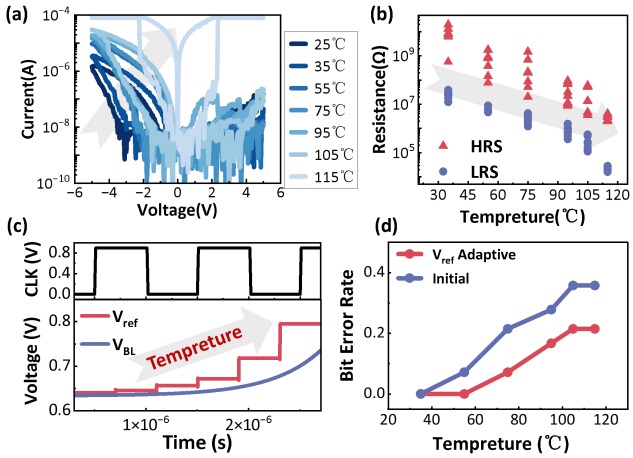

Fig. 10. (a-b) As the temperature rises, both the HRS and LRS decrease,

resulting in indistinguishable resistance states, further leading to bit failures. (c) V_{ref} is adaptively adjusted based on variations in temperature and resistance state, (d) resulting in a significant reduction in bit error rate.

As the temperature rises, both the HRS and LRS decrease, resulting in indistinguishable resistance states(Fig. 10(a-b)), further leading to bit failures. The temperature-sensing adaptive circuit can adaptively adjust V_{ref} based on variations in temperature and resistance states, significantly reducing the bit error rate (Fig. 10(c-d)).

IV. System-Level High-Temperature Protection Algorithm

The system dynamically adjusts data throughput by sensing the temperature deviation from preset threshold and injecting "compensation latency" into the 3D VRRAM operation commands (Fig.11a). Crucially, the high-temperature protection algorithm is executed proactively, prior to the assertion of the operation completion signal (Done). This approach limits the disk bandwidth without incurring any additional command completion time at the system level, thus optimizing the trade-off between performance and efficiency (Fig.11b). Ultimately, by throttling the bandwidth under high-temperature conditions, the system effectively reduces thermal buildup and prevents potential chip or device failure (Fig.11c).

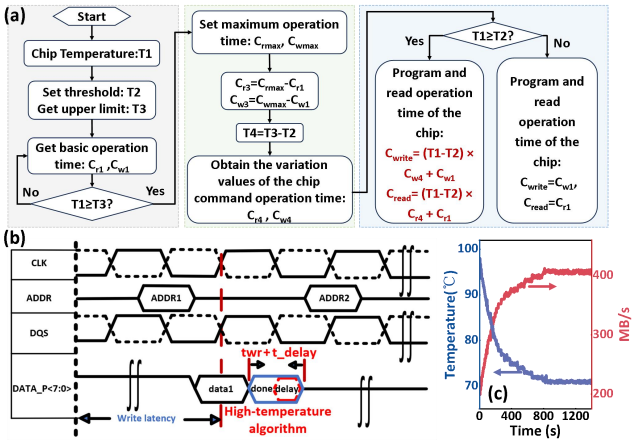

Fig.11. (a-b) System dynamically adjusts data throughput by sensing the temperature deviation from preset threshold and injecting "compensation latency" into the 3D VRRAM operation commands. (c) The write throughput of 3D VRRAM chip is dynamically adjusted via temperature variations.

V. Conclusion

In this work, we demonstrate a high-density 25-layer 500 nm SMC-3D VRRAM array. At the 28 nm technology node, the storage density reaches 42660 Mb/mm^2 (Fig. 8). The system's bit error rate is significantly reduced by dynamically adjusting V_{ref} via temperature-sensing adaptive circuit. Furthermore, system-level high-temperature protection algorithm is proposed to dynamically regulate chip throughput, thereby averting thermal-induced failures. Table I summarizes the performance comparison between this work and previously reported 3D VRRAM arrays.

TABLE I.

Benchmark of 25-layer SMC-3D VRRAM array with other works

	VLSI 2023 [1]	NC 2025 [2]	VLSI 2025 [3]	IEDM 2025 [4]	This work
Stacked Layers	16	1	8	8	25
Size (nm)	10000	3000	4000	4000	500
Storage Density(Mb/mm^2)	0.04	0.03	0.13	0.13	25
Thermal Stability	No	No	No	No	Yes
Stacking Potential (Layers)	127	---	127	30	158
Retention (s)	>10^4 @125℃	---	10^3	>10^8	>10^4 @100℃

Acknowledgment

This work was supported in part by the Innovation 2030 for Science and Technology under Grant No. 2021ZD0201203; in part by the China Postdoctoral Science Foundation under Grant No. 2025M770550; in part by the National Natural Science Foundation of China (NSFC) under Grants 61834009, 62488101, and 62322412; and in part by the Youth Innovation Promotion Association of the Chinese Academy of Sciences.

References

[1] Y. Ding et al., "16-layer 3D Vertical RRAM with Low Read Latency (18ns), High Nonlinearity (>5000) and Ultra-low Leakage Current (~pA) Self-Selective Cells," 2023 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits), Kyoto, Japan, 2023, pp. 1-2.

[2] Zhang, G et al. "Self-rectifying memristors with high rectification ratio for attack-resilient autonomous driving systems," Nat Commun 16, 5759 (2025).

[3] C. Ma et al., "First Implementation of Monolithic Integrated Cim with 1Mb Ultra-High-Density 8-Layer 3D VRRAM, Achieving High Computing Density (204.8GOPs/mm^2) and FoM (2.13×10^{6}GOPS^2 /W/mm^2) for Efficient Scientific Computing," 2025 Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits), Kyoto, Japan, 2025, pp. 1-3.

[4] Y. Zhou et al., "8-Layer Vertical Filament-Free Bulk RRAM with High Dynamic Range and Energy Efficiency for 3D Multilevel Compute-in-Memory," 2025 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2025, pp. 1-4.

[5] Liu, Z et al., "High-performance and environmentally robust multilevel lead-free organotin halide perovskite memristors," Adv. Elect. Materials, 2201005 (2022).

[6] Li, Y et al., "Heat transfer control using a thermal analogue of coherent perfect absorption," Nat. Commun. 13, 2683 (2022).

[7] Dang, B et al., Physically transient optic-neural synapse for secure in-sensor computing, IEEE Electron. Device Lett. 41,1641–1644(2020).

[8] D. Bridarolli et al., "High-Density Multilevel 3D Vertical Resistive Switching Memory (VRRAM) for Massively Parallel in-Memory Computing," 2024 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2024, pp. 1-4.

[9] W. Sun et al., "High Area Efficiency (6 TOPS/mm^2) Multimodal Neuromorphic Computing System Implemented by 3D Multifunctional RRAM Array," 2023 International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2023.

[10] Y.Li et al. "Brain-Inspired In-Memory Data Pruning and Computing with TaOx Mem-Selectors." Adv. Mater.37, no. 43 (2025): e02168.

[11] W. Sun et al., "Investigation on the 3D Memristor Array Architecture for 3D Reservoir Computing System Implementation," in IEEE Electron Device Letters, vol. 45, no. 8, pp. 1445-1448, Aug. 2024.