15.7 A 32Mb RRAM in a 12nm FinFet Technology with a 0.0249μm^2 Bit-Cell, a 3.2GB/S Read Throughput, a 10KCycle Write Endurance and a 10-Year Retention at 105°C

Yi-Cheng Huang, Shang-Hsuan Liu, Hsu-Shun Chen, Hsin-Chang Feng, Chih-Feng Li, Chou-Ying Yang, Wei-Keng Chang, Chang-Feng Yang, Chun-Yu Wu, Yen-Cheng Lin, Tsung-Tse Yang, Chih-Yang Chang, Wen-Ting Chu, Harry Chuang, Yih Wang, Yu-Der Chih, Tsung-Yung Jonathan Chang

TSMC, Hsinchu, Taiwan

2024 IEEE International Solid-State Circuits Conference (ISSCC) | 979-8-3503-0620-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISSCC49657.2024.10454367

Low-power wireless MCU devices for intelligent IoT applications are one of the key drivers for embedded non-volatile memory (eNVM) for technology nodes of 2xnm and beyond; in addition to high-performance advanced CMOS processes with excellent RF/analog devices, there is a need for high read throughput high-density embedded non-volatile memory to store CPU code as well as neural-network models for energy-efficient data-centric machine-learning edge computing. For this purpose, fully logic-compatible TMO-based resistive RAM (RRAM) is a promising candidate [1-3]. In this work, a 32Mb RRAM macro with 0.0249μm^2 bit cells is implemented using a 12nm ultra-low power FinFET technology. Several design solutions are also proposed to address key challenges, including a write-assist scheme to achieve high write endurance and data retention and a pipeline-read scheme for high read throughput. Silicon measurements show 10,000 write-cycle endurance, 10-year retention at 105°C, and a 3.2GB/s read throughput.

Figure 15.7.1 shows the structure of a 12-nm 1T1R bit cell and a block diagram of a 32Mb RRAM macro whose bit-cell size is 0.0249μm^2. Each cell consists of one transistor and one RRAM element (RE). RE is inserted between metal 4 and 5. Metals 2, 3 and 6 are used as routing for the source line (SL), word line (WL) and bit line (BL). The 32Mb RRAM macro is composed of one global analog block and 16 banks of RRAM memory. The analog block includes a bandgap reference, a voltage/current generator, a charge pump, etc., which provide the voltage and current bias needed for bank read/write operations. Each bank contains four 0.5Mb arrays in a butterfly configuration. A sense amplifier and WL driver (WLDRV) are placed in the middle of each bank to achieve a fast read speed. Since the WL driver electrical requirements for read and write are different, and conflict with each other, a dual pull-up-path WL driver is proposed to deliver a high WL voltage (V_{WWL} > 1.5V) for write and a lower WL voltage (V_{RWL} < 1.1V) for a shorter latency for high-speed read. Figure 15.7.2 shows the structure of proposed dual pull-up path WLDRV. The pull-up path contains two branches. The leaf branch with thick-oxide IO devices is used to charge WL to V_{WWL} during a write and the right branch with thin-oxide core devices can rapidly charge WL to V_{RWL} during read. WL is discharged from V_{WWL} and V_{RWL} through a single pull-down path: composed of a cascoded core device to save area. Cascoded core devices are used to speed up the WL rise/fall time for read. We leverage V_{RWL} as a cascoded bias to protect the core devices when the WL drivers operate at a higher voltage during write. The transient simulation plot, shown in Fig. 15.7.2, illustrates WL activation time for V_{RWL} is 1ns faster for high-speed read.

In addition to a dual-path WL driver, a pipeline read is proposed to achieve a 3.2GB/s read speed. Figure 15.7.3 shows the read-sense-amplifier (SA) structure for a pipeline read. The latch-based sense amplifier with BL clamp devices [4] is used to support signal development. The subsequent SR latch and flipflop (FF) store the read data to realize pipeline operation. Since the most time-consuming operations in the read function are the WL activation and development of differential signal, the proposed pipeline read divides the whole read operation into three cycles: address-decoding, signal-development, and data-latching. Figure 15.7.4 illustrates the pipeline read waveform. During the first SE cycle, the first address is decoded to activate the corresponding WL and BL. During the second SE cycle, data develops for the first address, and WL and BL activation for the second address occurs at the same time. During the third SE cycle, the data for the first address is latched by the FF for the DOUT signal, data develops for the second address, and WL and BL activation occurs for the third address. In the following cycles, data outputs successively and achieves a 3.2GB/s read speed.

A high-precision current limiter (CL) and a strong current-driving BL/SL column multiplexer (top and bottom multiplexers) are essential to support a reliable bipolar RRAM write operation [5]. To mitigate the effect of a ground-level gradient and noise coupling in the SoC, a current limiter, implemented by the local current mirror, for SET operation is proposed, as shown in Fig. 15.7.5(top). In this scheme, the bias current (I_{SET}), rather than a bias voltage, is fed into the master device (NMOS) of the local current mirror in each bank; it is used to generate a local bias voltage for better current control of current limiters in each bank. Since the master devices and the slave devices of the current mirror are located in the same bank, the effect of a ground-level gradient on the current mirror can be minimized. Moreover, the parasitic resistance on the write current path, including BL/SL metal-write resistance, can cause a significant voltage drop. As a result, the voltage across RE (V_{RE}) is reduced; impacting the write efficiency. To overcome this challenge, a dual-side BL discharge for RESET operation is proposed: it is shown in Fig. 15.7.5(top). In this scheme, pull-down devices on both the top and bottom of the array are turned on during a RESET operation to reduce the BL resistance to ground; thus, V_{RE} is less sensitive to the BL resistance. The simulation of V_{RE} during a RESET operation for conventional single-side discharge scheme and a dual-side discharge scheme are shown in Fig. 15.7.5(bottom). The V_{RE} in a conventional single-side discharge scheme has a stronger dependency on the WL location along BL; a WL-location-aware compensation is needed to equalize the write voltage across REs [5]. For our proposed scheme V_{RE} is less dependent on the WL location. The improvement to V_{RE} uniformity by dual-side discharge scheme is 10% for 512b/BL and 18% for 1024b/BL. The dual-side discharge provides the potential to extend the BL length to reduce macro area.

A 32Mb embedded RRAM chip has been measured to demonstrate RRAM feasibility in a 12nm FinFET logic process. Double-error correction (DEC) is implemented in the test-chip controller for error correction. Figure 15.7.6(top) is the pipeline read Shmoo plot for V_{DD} at –40, 25 and 125°C. The read frequency measured at 0.6V achieves 200MHz from –40 to 125°C. Figure 15.7.6(bottom) shows that the current difference between the maximum low-resistance state (RL) and the minimum high-resistance state (RH) can be 32.6% of RH after 10k SET+RESET operations at 25°C and 10h of retention bake at 200°C. A summary table of key parameters and the die photograph are shown in Fig. 15.7.7.

Acknowledgement:
We would like to thank all TSMC colleagues that contributed to the N12 RRAM project, including the logistics design, layout, and P&R teams, as well as the test program development, measurement support, reliability section and other teams.

References:

[1] Y. D. Chih, "Design Challenges and Solutions of Emerging Nonvolatile Memory for Embedded Applications," *IEDM*, pp. 2.4.1-2.4.4, 2021.

[2] O. Golonzka, et al., "Non-Volatile RRAM Embedded into 22FFL FinFET Technology," *IEEE Symp. On VLSI Tech.*, pp. *230-231*, 2019.

[3] R. Strenz, "Review and Outlook on Embedded NVM Technologies – From Evolution to Revolution", *IEEE Inter. Memory Workshop*, pp 1-4, 2020.

[4] C-C Chou, et al, "A 22nm 96KX144 RRAM Macro with a Self-Tracking Reference and a Low Ripple Charge Pump to Achieve a Configurable Read Window and a Wide Operating Voltage Range," *IEEE Symp. VLSI Circuits*, pp. 1-2, 2020.

[5] C-A Lai, et al, "Logic Process Compatible 40nm 256K×144 Embedded RRAM with Low Voltage Current Limiter and Ambient Compensation Scheme to Improve the Read Window," *ASSCC*, pp. 13-16, 2018.

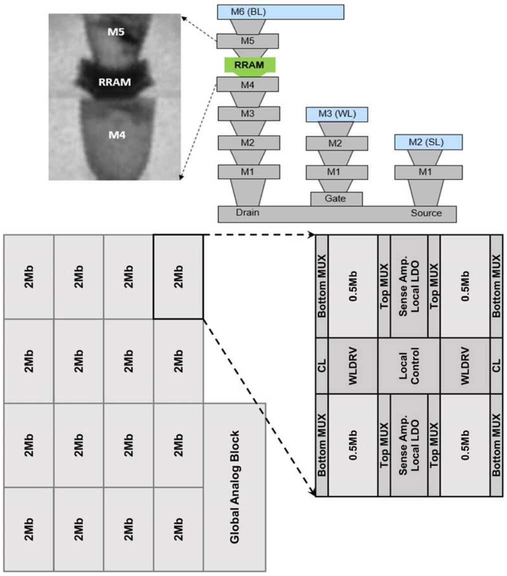

Figure 15.7.1: (Top) Bit cell cross-section and metal scheme. (Bottom) Layout floor plan and block diagram for 12nm 32Mb RRAM macro.

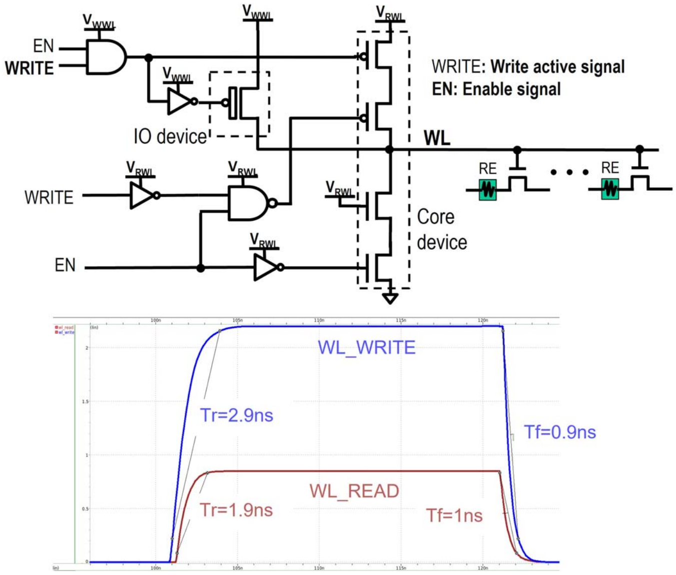

Figure 15.7.2: (Top) Dual-path WL driver for read/write operations using different WL bias. (Bottom) Transient comparison between read and write WL pulses.

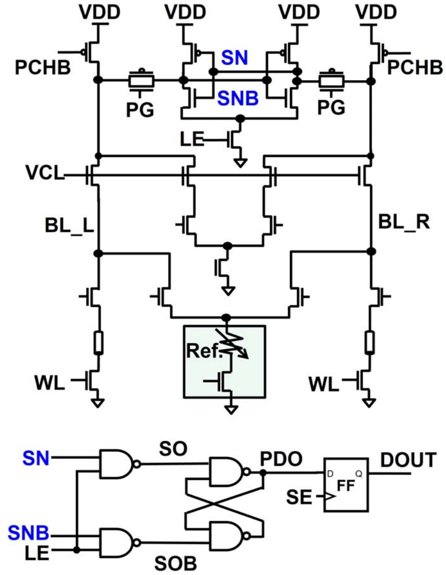

Figure 15.7.3: Read sense amplifier with SR latch and flip flop for pipeline read operations.

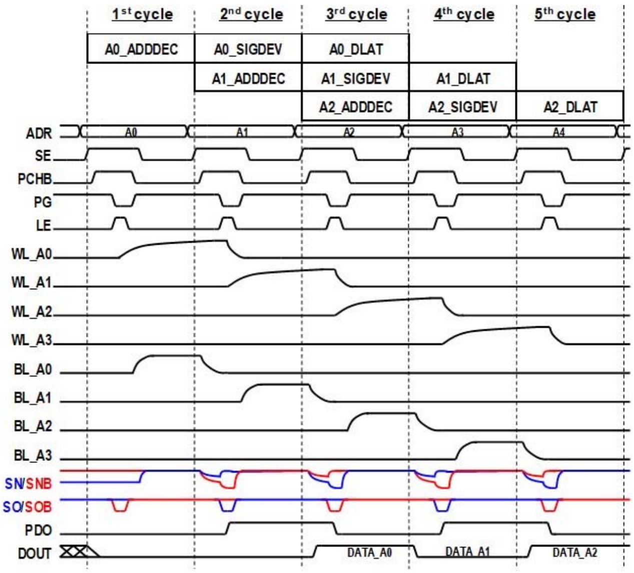

Figure 15.7.4: Pipeline sensing operation timing diagram: ADDDEC - address decode and WL/BL activation; SIGDEV - signal development; DLAT - latch read data.

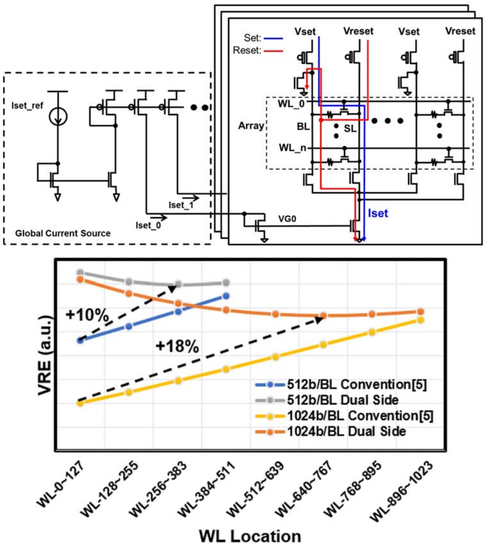

Figure 15.7.5: (Top) Proposed column multiplexer schematic. (Bottom) Simulation results for write voltage across RE for different array structures and pull-down schemes.

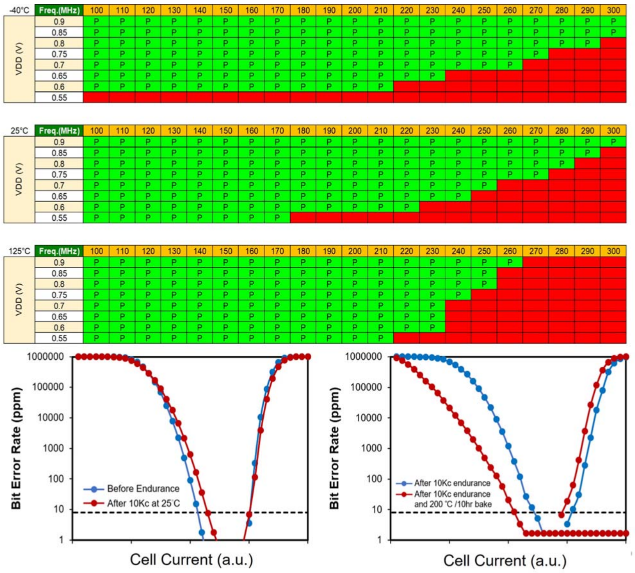

Figure 15.7.6: (Top) Pipeline read Shmoo for –40, 25 and 125°C. (Bottom) Read window before/after 10k cycles and a retention bake.

Technology	12nm
VDD	0.63~0.77V
VDIO	1.62~1.98V
Density	256KX144
Cell size	0.0249um^2
Die size	5000um x 5000um
Read cycle freq. (MHz) @0.6V	200
Read window after retention-after-cycling (RAC)	32.6%

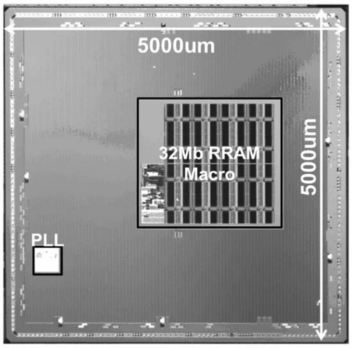

Figure 15.7.7: Summary table and die photo of 12nm RRAM test chip.