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

Jong-Hyeok Yoon[1], Muya Chang[1], Win-San Khwa[2], Yu-Der Chih[3], Meng-Fan Chang[2], Arijit Raychowdhury[1]

[1]Georgia Institute of Technology, Atlanta, GA
[2]TSMC Corporate Research, Hsinchu, Taiwan
[3]TSMC Design Technology, Hsinchu, Taiwan

2021 IEEE International Solid- State Circuits Conference (ISSCC) | 978-1-7281-9549-0/20/$31.00 ©2021 IEEE | DOI: 10.1109/ISSCC42613.2021.9365926

As memory-centric workloads (AI, graph-analytics) continue to gain momentum, technology solutions that provide higher on-die memory capacity/bandwidth can provide scalability beyond SRAM. Resistive RAM (RRAM) owing to (1) higher bit-density (2-4× of SRAM), (2) CMOS process/voltage compatibility, (3) nano-second read (RD) and (4) non-volatility has emerged as a promising candidate [1]. In spite of early prototypes, several technology challenges remain, and need to be addressed through circuit-technology co-design [1]. This paper presents a 64Kb RRAM macro supporting: (1) a programmable (1 to 9) number of row-accesses (N) to enable vector-matrix multiplication (referred to as compute-in-memory, or CIM) for a target algorithm-level inference-accuracy [2-8], (2) voltage-based RD with active feedback, advancing the state-of-the-art current-based RD, targeted for the low ratio between the high-resistance-state (HRS) and low-resistance-state (LRS) in typical RRAM, (3) RD-disturb tolerance under RRAM drift, through embedded RD-disturb monitor and write (WR)-back and (4) in-situ WR verification to enable a tight resistance distribution.

The RRAM cell (Fig. 29.1.1) undergoes: (1) an initial forming process at higher V_{DD}, (2) memory-WR including SET (HRS→ LRS) and RESET (LRS→ HRS) and (3) memory-RD. When configured as an array, it enables parallel access to multiple word-lines (WLs) to sum the resultant voltage/current on the bit-line and perform CIM operation. As the filter size in typical convolutional neural networks (CNNs) is 3×3, we provide programmability to access from 1 WL (full digital access) up to 9 WLs in a single cycle. The macro provides circuit solutions to the following technology challenges. (CH1) When RRAM cells are written at high-V_{DD} to create a large HRS/LRS-ratio (lower RD-failure), the endurance of the array decreases [1], necessitating the use of circuit techniques that provide high RD-margin under low HRS/LRS-ratio with variation. (CH2) A single SET/RESET cycle creates a wide LRS/HRS resistance distribution, and data resolution in CIM is affected. (CH3) Back-to-back RDs at higher temperatures lower HRS resistance (resistance drift) and can eventually cause data corruption/RD-disturb. The architecture of the 64Kb sub-array is shown in Fig. 29.1.1.

Figure 29.1.2 illustrates an array which supports full software-programmability with: (1) per-column ADCs for 1-to-9 single-cycle mixed-signal MAC, (2) pulsed digital inputs on the WL, (3) digital post-MAC shift-and-add to support computation over 1-to-8b inputs/weights and 1-to-20b output over 1-to-8 clock-cycles. In the proposed macro only positive weights and inputs are allowed, but this is not a limiter in machine learning applications where weight normalization and ReLU activation functions can be used to maintain positive operands only [8]. A key CIM requirement is high-resolution, quantization-free RD-out for all input-weight combinations. Prior designs use current-mode sensing and suffer from logic ambiguity in high-endurance RRAM-based CIM [2-5]. Furthermore, traditional voltage-mode sensing with a fixed-current has a narrow sampling margin at the RD-BL voltage (V.RBL) as more LRS cells are activated in parallel. We address this challenge (CH1) via voltage-mode sensing with an RRAM-cell paired with a current source to keep a constant sampling margin, albeit with nonlinear ADC levels. The non-linear levels are linearized using active feedback (FB) control, where a high-gain FB amplifier controls the current of the paired current sources. This enables an input-aware V.RBL which is a linear combination of V_{LRS} and V_{HRS} ($V_{LRS/HRS}$ are V.RBL with 1 LRS and 1 HRS cell) mathematically shown in Fig. 29.1.2. Previous approaches [4] using diode-connected PMOS cannot mitigate the high degree of non-linearity and provide no PVT tolerance, whereas the current design with programmable-gain FB enables linear V.RBL (with 2× increase in sampling margin), PVT tolerance with 50% area reduction (Fig. 29.1.3). The RD-out circuit features a 4b flash ADC with open-loop, strong-arm comparators and current-clamping. Current clamping limits the range of the input-referred offset. This relaxes the range of reference voltages and reduces system power. Monotonicity is maintained in the 8b reference generator (RG) using 3b thermometer and 5b binary control. The programmable RG produces on-die references for 1-to-8b CIM with a semi-uniform distribution that caters to the V.RBL range/resolution for error-free computation. The 4b ADC provides 0.6 lsb redundancy.

An open-loop single-cycle SET/RESET creates a wide distribution of LRS/HRS cells, where the final resistance value depends on the WR PVT conditions, as well as the history effect (i.e., how many previous RDs have happened before the WR). This leads to errors in CIM. In the absence of a hard RESET in high-endurance RRAM, we address the challenge (CH2) using in-situ WR-verification with negligible overhead (Fig. 29.1.4). During WR, after every WR-pulse, the high-resolution RD-circuit enables digital RD-out of the V.RBL representing the resistance state of the individual cell and compares with a target (stored as a digital word). The current RD-out engages FSM logic that either increases/decreases the next programming pulse width (PW) determined from the ADC read-out of the V.RBL, and once the target is reached, it completes the WR process. In the current process, HRS experiences a wider distribution and it is targeted in this design (Fig. 29.1.4). Further, HRS cells are also susceptible to a slow decrease of resistance under back-to-back RDs (CH3). To enable long-term functionality, we address CH3 using an in-situ RD-disturb monitor that monitors the health of each RRAM cell. Each cell is periodically sensed and all HRS cells that have drifted beyond a predefined digital threshold are reset to their original HRS state. A low rate of such RESET operations is sufficient to provide RD-disturb-free operation with no performance/power penalty (RD-disturb occurs only after $>10^6$ RDs). We expect this circuit scheme to address other time/temperature-dependent drifts common in RRAMs.

Correct operation in the RRAM macro is demonstrated in Fig. 29.1.5, where the logic-waveform capture shows correct 1b CIM operation (N=4). The clock frequency is limited by the RD access latency of the bitcell and the peripheral circuits. The WR can take multiple cycles as described above. We indirectly measure the V.RBL for varying N and show measured results for N=1, 9. The FB-amplifier gain is controlled using a bias voltage (V.b). We note that the V.RBL for N=9, with an active FB linearizes the BL voltage providing 5.7×-increased sense margin compared to the case of low V.b (similar to a diode-based BL pull-up). During initialization to known states, the macro undergoes forming, SET and RESET. For 100 cells, we measure the number of operations required to reach target resistance values. We note that while forming is achieved in 2.25 iterations (average), SET and RESET are achieved on an average of 1 and 1.02 iterations, respectively. The closed-loop, iterative WR-with-verification enables a tight HRS distribution and we note a decrease of σ, as measured through V.RBL, from 37.74mV (baseline, non-iterative WR) to 12.78mV (proposed). Fig. 29.1.6 shows a typical run of an HRS cell through back-to-back RDs under accelerated testing conditions (V_{WL}=1.5V, V_{DD}=1V at T=85°C) where the in-situ monitor detects a resistance change below the HRS threshold (-6%) and activates a RESET process. Subsequently, the HRS is restored to its high resistance state and the cycle repeats. For CIM, we note an average (peak) energy-efficiency of 4.15 (56.67) TOPS/W at 100MHz (limited by the low LRS/HRS resistances in the current process) with the largest contributions from the RRAM array and sense/RD circuits. The proposed techniques enable high algorithm-level accuracy across a suite of AI benchmarks. A comparison with the state-of-the-art CIM architectures [2-6] illustrates competitive metrics while addressing key technological challenges. The die-shot and the chip-characteristics are shown in Fig. 29.1.7.

Acknowledgement:
JY, MC, and AR were supported by the Semiconductor Research Corporation under the Center for Brain-Inspired Computing (C-BRIC) under Grant 2777.005 and 2777.006, and the Applications and Systems-driven Center for Energy-Efficient Integrated Nano Technologies (ASCENT) under Grant 2776.037. The authors would also like to thank TSMC for technical discussions and chip fabrication support.

References:

[1] C. Nail et al., "Understanding RRAM Endurance, Retention and Window Margin Trade-Off Using Experimental Results and Simulations," *IEDM*, pp. 4.5.1-4.5.4, 2016.
[2] W.-H. Chen et al., "A 65nm 1Mb Nonvolatile Computing-In-Memory ReRAM Macro with Sub-16ns Multiply-and-Accumulate for Binary DNN AI Edge Processors," *ISSCC*, pp. 494-495, 2018.
[3] C.-X. Xue et al., "A 1Mb Multibit ReRAM Computing-In-Memory Macro with 14.6ns Parallel MAC Computing Time for CNN Based AI Edge Processors," *ISSCC*, pp. 388-389, 2019.
[4] C.-X. Xue et al., "A 22nm 2Mb ReRAM Compute-in-Memory Macro with 121-28TOPS/W for Multibit MAC Computing for Tiny AI Edge Devices," *ISSCC*, pp. 244-245, 2020.
[5] W. Wan et al., "A 74 TMACS/W CMOS-RRAM Neurosynaptic Core with Dynamically Reconfigurable Dataflow and In-situ Transposable Weights for Probabilistic Graphical Models," *ISSCC*, pp. 498-499, 2020.
[6] J. Wang et al., "A Compute SRAM with Bit-Serial Integer/Floating-Point Operations for Programmable In-Memory Vector Acceleration," *ISSCC*, pp. 224-225, 2019.
[7] N. Cao et al., "A 65nm Image Processing SoC Supporting Multiple DNN Models and Real-Time Computation-Communication Trade-Off Via Actor-Critical Neuro-Controller," *IEEE Symp. VLSI Circuits*, 2020.
[8] B. Crafton et al., "Merged Logic and Memory Fabrics for Accelerating Machine Learning Workloads," *IEEE Design & Test*, 2020.

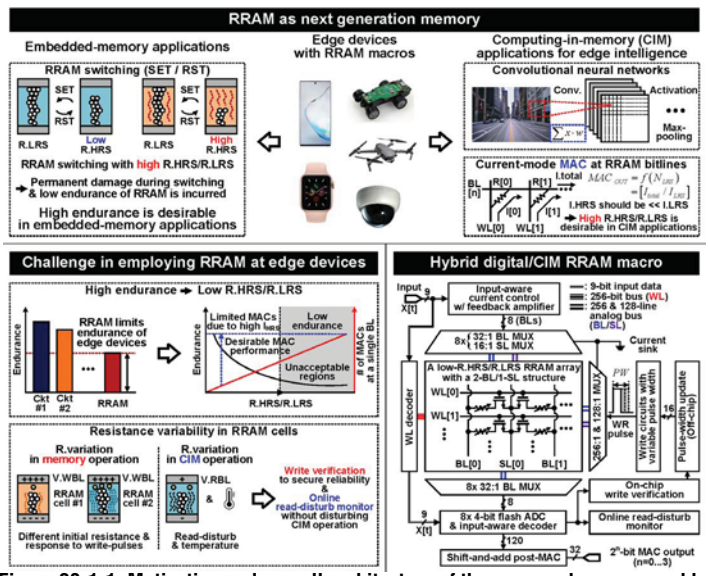


Figure 29.1.1: Motivation and overall architecture of the proposed programmable hybrid digital/CIM RRAM macro.

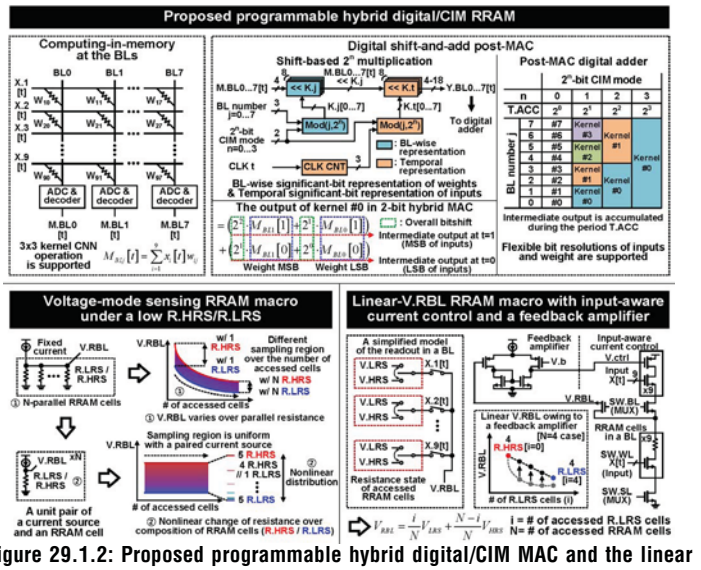


Figure 29.1.2: Proposed programmable hybrid digital/CIM MAC and the linear voltage-mode sensing RRAM macro with input-aware current control and a feedback amplifier.

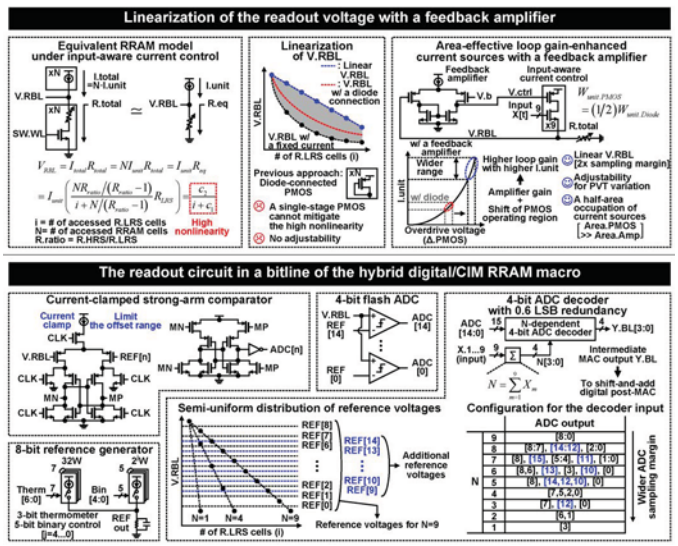


Figure 29.1.3: Linearization of the readout voltage with a feedback amplifier and the readout circuit of intermediate MAC output in the proposed hybrid digital/CIM RRAM macro.

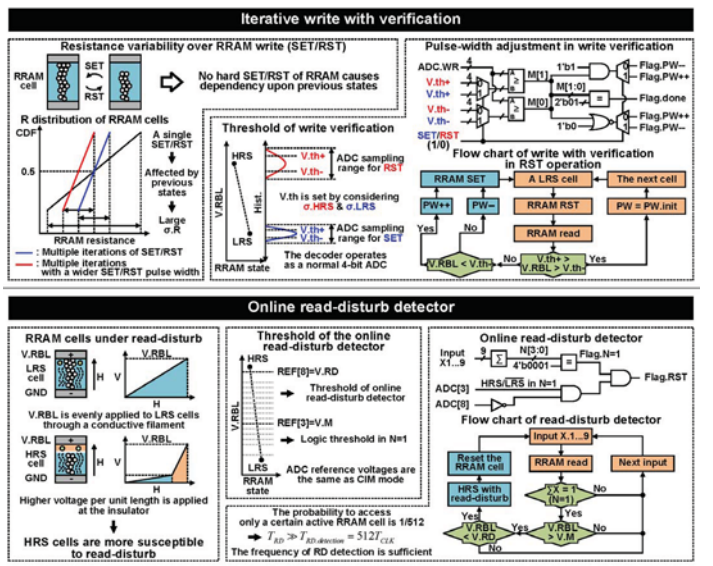


Figure 29.1.4: Iterative write with verification and online read-disturb detector in the proposed hybrid digital/CIM RRAM macro.

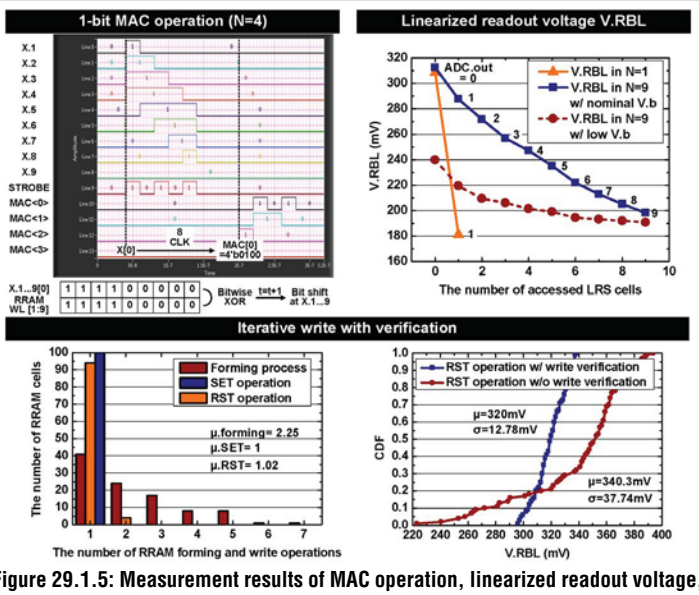
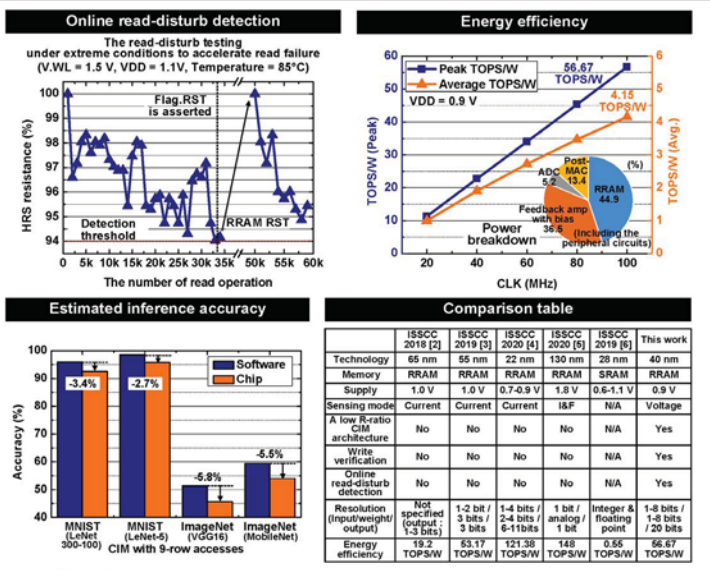
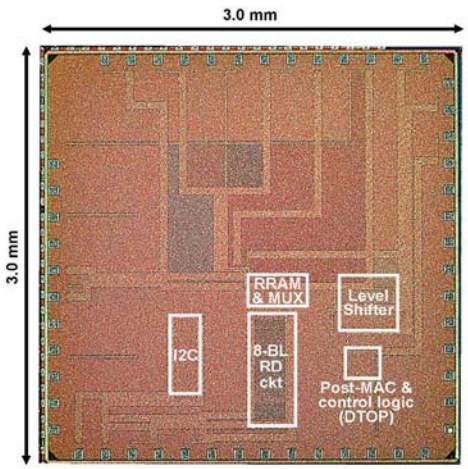


Figure 29.1.5: Measurement results of MAC operation, linearized readout voltage, and iterative write with verification.



	ISSCC 2018 [2]	ISSCC 2019 [3]	ISSCC 2020 [4]	ISSCC 2020 [5]	ISSCC 2019 [6]	This work
Technology	65 nm	55 nm	22 nm	130 nm	28 nm	40 nm
Memory	RRAM	RRAM	RRAM	RRAM	SRAM	RRAM
Supply	1.0 V	1.0 V	0.7-0.9 V	1.8 V	0.6-1.1 V	0.9 V
Sensing mode	Current	Current	Current	I&F	N/A	Voltage
A low R-ratio CIM architecture	No	No	No	No	N/A	Yes
Write verification	No	No	No	No	N/A	Yes
Online read-disturb detection	No	No	No	No	N/A	Yes
Resolution (input/weight/output)	Not specified (output : 1-3 bits)	1-2 bit / 3 bits / 3 bits	1-4 bits / 2-4 bits / 6-11bits	1 bit / analog / 1 bit	Integer & floating point	1-8 bits / 1-8 bits / 20 bits
Energy efficiency	19.2 TOPS/W	53.17 TOPS/W	121.38 TOPS/W	148 TOPS/W	0.55 TOPS/W	56.67 TOPS/W

Figure 29.1.6: Measurement results of online read-disturb detector and energy efficiency, estimated inference accuracy, and comparison table.



Technology	40 nm CMOS & RRAM process
Active area	0.437 mm^2
Frequency	100 MHz
Power	11.05 mW
Digital VDD	0.9 V
Analog VDD	0.8 V
I/O VDD	3.3 V
MAC mode	1 to 8-bit MAC mode
Energy efficiency	56.67 TOPS/W
Package	QFN48

Figure 29.1.7: Microphotograph of the test-chip and summary of performance.

978-1-7281-9549-0/21/$31.00 ©2021 IEEE