# 16.2 A 40nm 64kb 26.56TOPS/W 2.37Mb/mm² RRAM Binary/Compute-in-Memory Macro with 4.23× Improvement in Density and >75% Use of Sensing Dynamic Range

Samuel D. Spetalnick[1], Muya Chang[1], Brian Crafton[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

Compute-in-Memory (CIM) using emerging nonvolatile (eNVM) memory technologies, such as resistive random-access memory (RRAM), has been shown by several implemented macros to be an energy-efficient alternative to traditional von Neumann architectures [1-6]. Since moving data on- and off-chip has a high energy cost, area efficiency is important to the practical utility of CIM with RRAM. Many systems demonstrated so far have not reported area efficiency or addressed the challenges CIM with RRAM presents with respect to practical area-constrained integrated circuits.

Figure 16.2.1 shows the topology of the implemented RRAM macro and presents three challenges. (1) As suggested above, peripheral area overhead for eNVM-based CIM systems can be significant due to the requirement for high-precision analog-to-digital converters (ADCs) to accommodate reduced sensing margin when multiple states are represented on the bitline (BL), and due to the need for level shifters (LSs) and isolation devices to enable safe high-voltage (HV) RRAM writes. To reduce ADC area and power, the readout circuit should maximize the portion of the available voltage headroom which is used for representing output states. (2) The macro should support various applications, which may require different operations (binary vs. MAC) and levels of accuracy and may be optimized at different RRAM operating points. (3) The array macro must be compatible with large integrated systems by including all necessary biasing, having a compact digital interface, and having the ability to power-down with several granularity levels to support power states. The implemented power-switchable macro integrates 8× 1-to-4b read channels, one shared self-biasing and reference-generating circuit per macro, full read and write drivers, and all LSs into a 0.027mm² layout, 16.2× smaller than a similar-featured same-technology array [6]. Using estimated area in the highest-density CIM with RRAM macro recently reported in ISSCC, [4], this macro is 1.56× and 4.23× denser before and after normalizing to the 40nm node.

Figure 16.2.2 details the voltage-regulating current-sense topology designed to address the first two challenges. Sufficient voltage gain $A_0$ provides immunity to process, voltage and temperature (PVT) variation, while sufficient effective regulating transconductance $A_0G_0$ allows RRAM cells to be measured as current sources with ideal current-domain CIM state separation. The explicit current-sensing resistor enables constant, PVT-tolerant current-voltage translation, while the flexible choice of BL target voltage $V_{BLTGT}$ (broken out off-chip for ease of testing) enables compatibility with a wide range of RRAM cell resistances. Analysis of the effects of mismatch in the three critical sensing components motivates geometric matching of the amplifiers and sense resistors to allow read channels to share references. This macro supports a < 6ns binary read mode, 1-to-4b low-power precise cell drift monitoring mode as in [6], 1-to-4b output CIM mode, and 200ns power switching.

Figure 16.2.3 shows transistor-level implementation of the read channels. The BL regulator is implemented as a two-transistor gain-enhanced follower with the voltage-gain and transconductance elements each a single device. All the bias current for each channel flows into the RRAM cell load which counteracts the $I_{LEAK}$ due to off-state measured cells during CIM (increasing effective dynamic range). Since $I_{BIAS}$ flows into the RRAM cell(s), the maximum measurable resistance is roughly limited to $V_{BLTGT}/I_{BIAS}$. To extend the usable maximum RRAM resistance range to ~100kΩ and allow for low-power cell drift monitoring, the bias circuit allows ~10× switching between high- and low-bias modes. When no wordline (WL) is selected, negligible current (in order of 10nA) flows through a diode-connected device to allow the BL to remain in near-regulation. Post-layout simulation shows arbitrarily large and linear state separation, and 5.3ns latency to adequate binary state separation. Low ADC kickback to support shared low-power reference generation is achieved by adding sampling transistors at the comparator inputs. With reduced capacitive loading on the input gates, input common-mode range (ICMR) maximum is reduced due to parasitic common-mode feedback during operation. The addition of PMOSCAPs restores ICMR, resulting in a compact design with > 6.25× reduction in simulated kickback charge. The reference generator uses common-centroid poly resistors for even step size and allows controllable step- and start-voltage using two IDACs for flexibility to application sensing requirements.

Figure 16.2.4 motivates and shows the implementation of the split write/read WL driver architecture for area-efficient WL drive across two power domains. Level-shifting WL drivers are needed to support HV ( >1.5V) WL drive during write, but write speeds are set by the RRAM technology and are at least 100ns. The slower write WL driver is distinct from the faster read WL driver and uses a full decoder-tree in the HV domain. Using pass-transistor logic and a final-drive buffer, the HV driver requires only 10 differential-output LSs (vs. 256 LS without decoder). The separate sub-nanosecond read driver chains each incorporate only a single HV stage, with a thick-oxide NMOS chosen for best speed/area tradeoff. Area is reduced 5× vs. individual LS for each WL at slightly improved overall speed. The read WL drivers are enabled as a block with a single LS, and each WL is individually controlled with a one-hot select signal. During read, the write WL drivers are isolated from the WLs by separately power-gating the decoder tree (ensuring the final drive NMOS is off) and power-gating the final drive PMOS. Post-layout transients confirm sub-1ns read WL driver performance and sub-10ns write WL driver performance.

Figure 16.2.5 shows the measured performance of the read and write circuits. To show readout circuit performance, the reference-ladder is calibrated using an off-chip port then swept through known values to accurately measure on-chip voltages. Linear CIM read with 45-to-75mV state separation (average 3.8mV loss in separation per each additional on-state cell) and >75% usage of available ADC dynamic range is demonstrated using binary cells programmed iteratively with a single pass using a threshold. Binary read of one-shot (no readback/iteration) programmed cells is shown to be robust to cell-resistance variation and voltage shifts due to high open-loop gain in the self-biasing and readout circuits. The readout circuit allows for wide voltage-domain state separation from limited cell-resistance-domain separation, and one-shot write voltages of 1.7V, 2.6V, and 3.1V for SET (LRS), RESET (HRS), and FORM are shown to successfully program cells for binary read.

Figure 16.2.6 shows the area efficiency, measured energy efficiency, and comparison to other works. The 9 WL (parallel) 1×1b MAC operation shows 26.56/5.63 peak/average TOPS/W achieved at 0.83V $AV_{DD}$ and 64MHz. Peak efficiency is sensitive to bias/leakage power, and the shown worst-case figure of 200μW results in nominally worse efficiency compared to prior work [6]. Average efficiency is improved 36% (vs. 4.15TOPS/W), and calibrated simulation shows that biasing the cells for read consumes over 80% of power. The utility of the staged power-on/off is confirmed by measurement. The area-optimized RRAM-based CIM macro achieves 2.37Mb/mm², estimated to be 56% denser than a prior design in 22nm, here with a smaller-capacity sub-array [4]. Compared to similar works in 40nm or larger nodes [2, 6], the effective density is approximately 12.5-to-16.2× higher. This is due to the HV-domain WL signal encoding and the compact readout circuit with >45mV state separation. The readout circuits with biasing represent 2.8% of the total area (9.6% of array area) due to the minimized 2-transistor + bias implementation, with no energy overhead compared to the biased RRAM cells (excepting the single bias arm per macro).

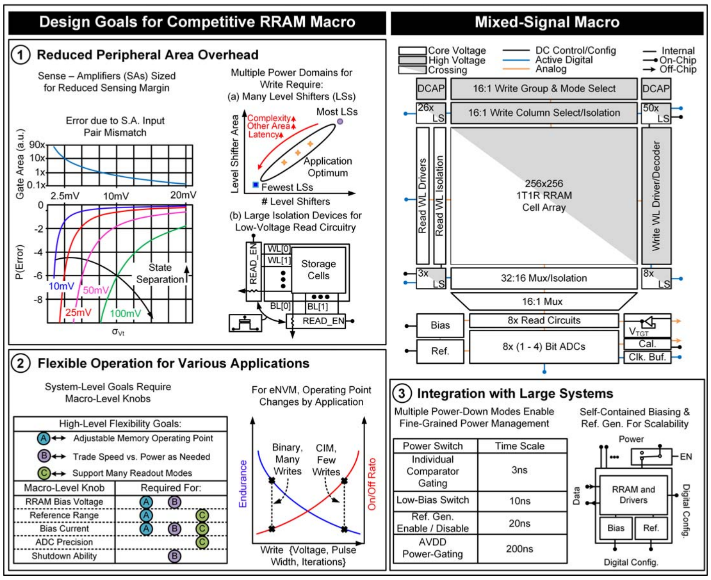
Chip micrograph and chip characteristics are shown in Fig. 16.2.7.

*Acknowledgement*:
SS, MC, BC, and AR were supported by the Semiconductor Research Corporation under the Center for Brain-Inspired Computing (C-BRIC) under Grant 2777.004 and 2777.005, and the Applications and Systems-driven Center for Energy-Efficient Integrated Nano Technologies (ASCENT) under Grant 2776.079. The authors would also like to thank TSMC for technical discussions and chip fabrication support.

*References:*

[1] C. 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. 244-246, 2020.
[2] C. 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. 338-390, 2019.
[3] 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. 494-496, 2018.
[4] C. Xue et al., “A 22nm 4Mb 8b-Precision ReRAM Computing-in-Memory Macro with 11.91 to 195.7TOPS/W for Tiny AI Edge Devices,” *ISSCC*, pp. 245-247, 2021.
[5] Q. Liu et al., “Fully Integrated Analog ReRAM Based 78.4TOPS/W Compute-In-Memory Chip with Fully Parallel MAC Computing,” *ISSCC*, pp. 500-502, 2020.
[6] J. -H. Yoon et al., “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,” *ISSCC*, pp. 404-406, 2021.

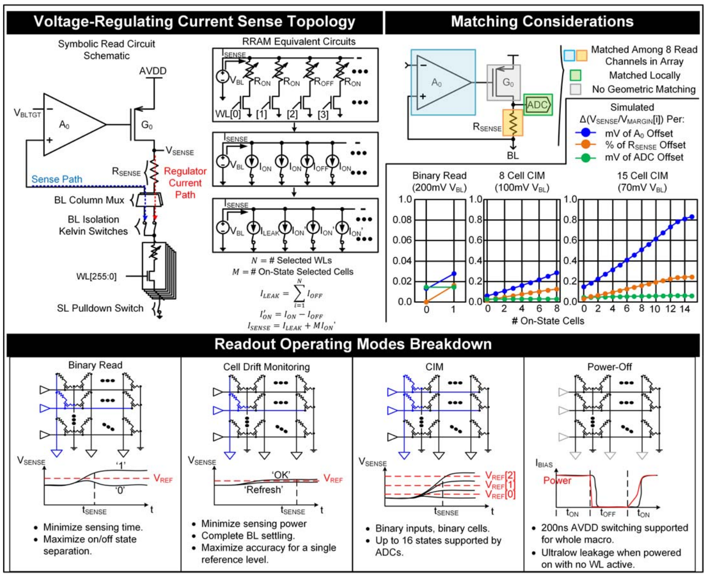
2022 IEEE International Solid- State Circuits Conference (ISSCC) | 978-1-6654-2800-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISSCC42614.2022.9731725


Design Goals for Competitive RRAM Macro
1 Reduced Peripheral Area Overhead
Sense – Amplifiers (SAs) Sized for Reduced Sensing Margin
Error due to S.A. Input Pair Mismatch
Gate Area (a.u.)
90x
10x
1x
0.1x
2.5mV
10mV
20mV
P(Error)
10mV
25mV
50mV
100mV
State Separation
σVt
Multiple Power Domains for Write Require:
(a) Many Level Shifters (LSs)
Level Shifter Area
Most LSs
Complexity
Other Area
Latency
Application Optimum
Fewest LSs
# Level Shifters
(b) Large Isolation Devices for Low-Voltage Read Circuitry
READ_EN
WL[0]
WL[1]
Storage Cells
BL[0]
BL[1]
READ_EN
Mixed-Signal Macro
Core Voltage
High Voltage
Crossing
DC Control/Config
Active Digital
Analog
Internal
On-Chip
Off-Chip
DCAP
16:1 Write Group & Mode Select
DCAP
26x LS
16:1 Write Column Select/Isolation
50x LS
Read WL Drivers
Read WL Isolation
256x256 1T1R RRAM Cell Array
Write WL Driver/Decoder
3x LS
32:16 Mux/Isolation
8x LS
16:1 Mux
Bias
8x Read Circuits
VTGT
Cal.
Ref.
8x (1 - 4) Bit ADCs
Clk. Buf.
2 Flexible Operation for Various Applications
System-Level Goals Require Macro-Level Knobs
For eNVM, Operating Point Changes by Application
Endurance
On/Off Rato
Binary, Many Writes
CIM, Few Writes
Write {Voltage, Pulse Width, Iterations}
3 Integration with Large Systems
Multiple Power-Down Modes Enable Fine-Grained Power Management
Self-Contained Biasing & Ref. Gen. For Scalability
Power
EN
Data
RRAM and Drivers
Bias
Ref.
Digital Config.
Digital Config.

| High-Level Flexibility Goals: | |
|---|---|
| A ↔ Adjustable Memory Operating Point | |
| B ↔ Trade Speed vs. Power as Needed | |
| C ↔ Support Many Readout Modes | |
| **Macro-Level Knob** | **Required For:** |
| RRAM Bias Voltage | A B |
| Reference Range | A C |
| Bias Current | A B C |
| ADC Precision | C |
| Shutdown Ability | B |

| Power Switch | Time Scale |
|---|---|
| Individual Comparator Gating | 3ns |
| Low-Bias Switch | 10ns |
| Ref. Gen. Enable / Disable | 20ns |
| AVDD Power-Gating | 200ns |

**Figure 16.2.1: Design goals for a practical RRAM macro and topology of the implemented macro.**



$$I_{LEAK} = \sum_{i=1}^{N} I_{OFF}$$

$$I'_{ON} = I_{ON} - I_{OFF}$$

$$I_{SENSE} = I_{LEAK} + MI_{ON}'$$

**Figure 16.2.2: Voltage-regulating current-sense readout circuit topology, details about matching requirements, and the supported readout operating modes.**

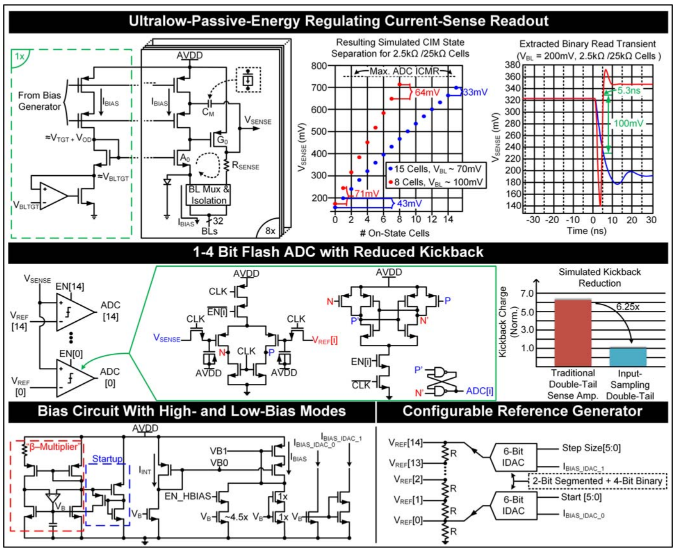


**Figure 16.2.3: Transistor-level schematics and performance simulations of the readout components including voltage-regulating read circuit, flash ADC with reduced kickback, bias circuit, and reference generator.**

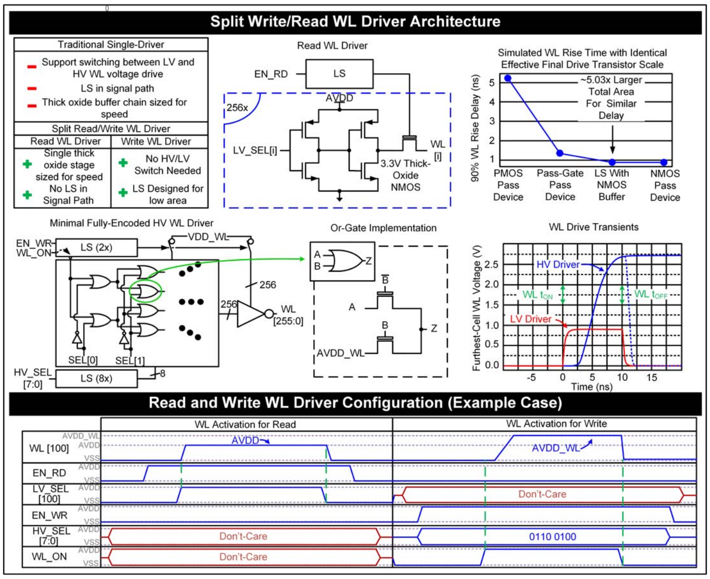
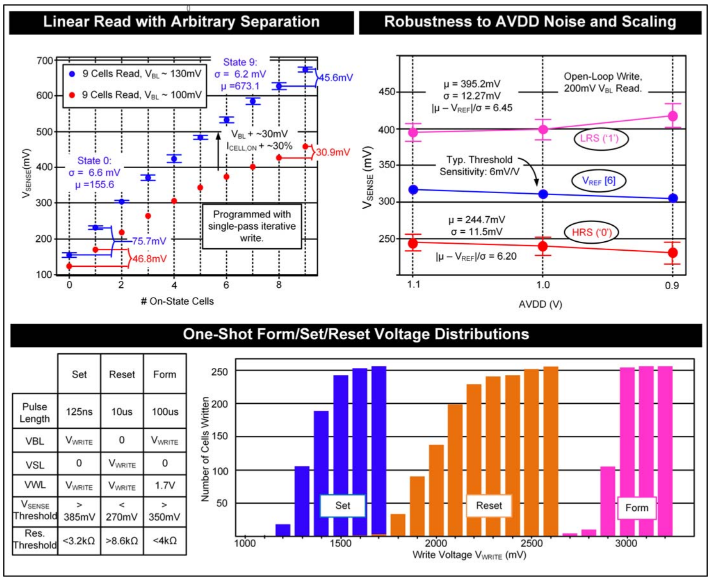
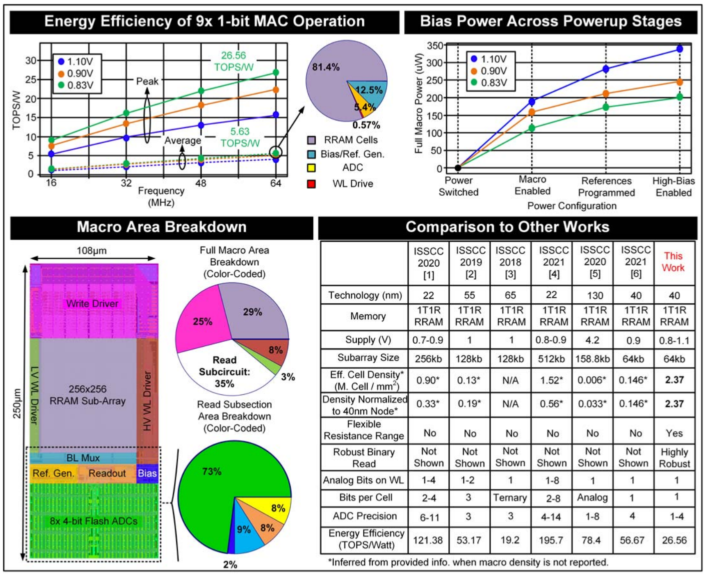


**Figure 16.2.4: Split read/write WL driver motivation, schematics, simulations, and operating waveforms.**



| | Set | Reset | Form |
|---|---|---|---|
| Pulse Length | 125ns | 10us | 100us |
| VBL | $V_{WRITE}$ | 0 | $V_{WRITE}$ |
| VSL | 0 | $V_{WRITE}$ | 0 |
| VWL | $V_{WRITE}$ | $V_{WRITE}$ | 1.7V |
| $V_{SENSE}$ Threshold | > 385mV | < 270mV | > 350mV |
| Res. Threshold | <3.2kΩ | >8.6kΩ | <4kΩ |

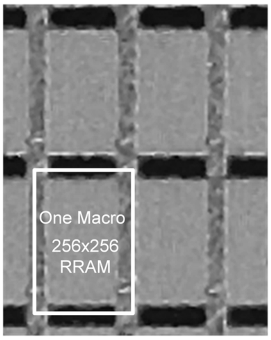
**Figure 16.2.5: Measured behavior of the memory macro detailing performance of CIM read and robust binary read and showing distribution of required write voltages.**



| | ISSCC 2020 [1] | ISSCC 2019 [2] | ISSCC 2018 [3] | ISSCC 2021 [4] | ISSCC 2020 [5] | ISSCC 2021 [6] | This Work |
|---|---|---|---|---|---|---|---|
| Technology (nm) | 22 | 55 | 65 | 22 | 130 | 40 | 40 |
| Memory | 1T1R RRAM | 1T1R RRAM | 1T1R RRAM | 1T1R RRAM | 1T1R RRAM | 1T1R RRAM | 1T1R RRAM |
| Supply (V) | 0.7-0.9 | 1 | 1 | 0.8-0.9 | 4.2 | 0.9 | 0.8-1.1 |
| Subarray Size | 256kb | 128kb | 128kb | 512kb | 158.8kb | 64kb | 64kb |
| Eff. Cell Density* (M. Cell / mm$^2$) | 0.90* | 0.13* | N/A | 1.52* | 0.006* | 0.146* | **2.37** |
| Density Normalized to 40nm Node* | 0.33* | 0.19* | N/A | 0.56* | 0.033* | 0.146* | **2.37** |
| Flexible Resistance Range | No | No | No | No | No | No | Yes |
| Robust Binary Read | Not Shown | Not Shown | Not Shown | Not Shown | Not Shown | Not Shown | Highly Robust |
| Analog Bits on WL | 1-4 | 1-2 | 1 | 1-8 | 1 | 1 | 1 |
| Bits per Cell | 2-4 | 3 | Ternary | 2-8 | Analog | 1 | 1 |
| ADC Precision | 6-11 | 3 | 3 | 4-14 | 1-8 | 4 | 1-4 |
| Energy Efficiency (TOPS/Watt) | 121.38 | 53.17 | 19.2 | 195.7 | 78.4 | 56.67 | 26.56 |

*Inferred from provided info. when macro density is not reported.

**Figure 16.2.6: Measured power and energy efficiency results, area breakdown, and comparison to other CIM RRAM macro implementations.**



| Technology | 40nm CMOS |
|---|---|
| Memory | Foundry 1T1R RRAM |
| Macro Size | 0.027mm$^2$ |
| Capacity | 64Kb |
| AVDD | 0.83 – 1.1V |
| Write Voltages<br>SET<br>RESET<br>FORM | <br>1.6 – 2.2V<br>2.6 – 3.0V<br>3.1 – 3.3V |
| Modes | Binary<br>CIM (4-bit ADC)<br>Monitoring |
| Bias Power:<br>@ 0.83V<br>@ 1.1V | <br>113uW – 201uW<br>189uW – 339uW |
| Energy Eff.:<br>Binary<br>CIM | @ 0.83V:<br>461fJ/Cell<br>26.56TOPS/Watt |

Figure 16.2.7: Die micrograph and chip characteristics.