A 2.38 MCells/mm^2 9.81 - 350 TOPS/W RRAM Compute-in-Memory Macro in 40nm CMOS with Hybrid Offset/I_{OFF} Cancellation and $I_{CELL}R_{BLSL}$ Drop Mitigation

Samuel D. Spetalnick[1], Muya Chang[1], Shota Konno[1], Brian Crafton[1], Ashwin S. Lele[1], Win-San Khwa[2], Yu-Der Chih[3], Meng-Fan Chang[2], Arijit Raychowdhury[1]

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

[3]TSMC Design Technology, Hsinchu, Taiwan. E-mail: spetalnick@gatech.edu

Abstract: A dense compute-in-memory (CIM) macro using resistive random-access memory (RRAM) showing solutions to read channel mismatch, high I_{OFF}, ADC offset, IR drop, and cell resistance variation is presented. By combining a hybrid analog/mixed-signal offset cancellation scheme and $I_{CELL}R_{BLSL}$ drop mitigation with a low cell bias target voltage, the proposed macro demonstrates robust operation (post-ECC bit error rate (BER) < 5×10^{-8} for 8WL CIM) while maintaining an effective cell density 1.03 – 33.1× higher than prior art and achieving 1.74 – 13.35× improved average MAC efficiency relative to the previous highest-density RRAM CIM macro.

Introduction: Analog compute-in-memory (CIM) using resistive random-access memory (RRAM) has emerged as a primitive for computing MACs in resistive arrays at reduced energy and/or higher bandwidth relative to traditional digital systems [1-6]. Despite recent work in this area, practical concerns about accuracy, memory density and efficiency for dense vectors have persisted. Addressing these challenges, the proposed macro features (1) hybrid data-aware analog/mixed-signal offset cancellation to counter channel mismatch and off-state cell current (I_{OFF}) (2) bitline/sourceline (BL/SL) Kelvin sensing to address $I_{CELL}R_{BLSL}$ drop and (3) cell bias $\ll$ 100mV.

Proposed Macro: Shown in Fig. 1, the 64kCell 16-IO macro ingests a 256-wide digital WL data-input vector and outputs a 16×6-wide data-output vector. Read/write control ports and readout circuitry operate at core logic voltage, 0.9~1.1V while write voltages are 1.3V~3.3V. Biasing and analog-to-digital converter (ADC) timing are generated internally. The cell array is a foundry 1T1R array with BL sensing wires added over top. CIM uses current-sensing with RRAM cells biased at a low nominal voltage ($V_{TGT,}$ ~25mV) to improve average-case efficiency and mitigate power transients on wordline (WL) transitions. Shown in Fig. 2, three challenges limit RRAM CIM parallel WL activations: (1) ADC-input-referred gain error during parallel CIM reads due to mismatch between V_{TGT} and the true voltage across the memory cells (V_{CELLS}) caused by (a) variation and noise in the regulator circuit and (b) IR drop due to cell current along the BL and SL; (2) ADC-input-referred offset error due to small sensing margins during CIM, and variable I_{OFF} (due to low R_{OFF}/R_{ON}); and (3) cell-resistance variation. Global variation is solved by adjusting V_{TGT} at each macro, but local variation appears as noise and must be managed *post hoc* with error correction codes (ECC, Fig. 8).

Shown in Fig. 3, the current-sensing front-end addresses (1) and (3). This circuit sets a software-controlled differential V_{TGT} across the selected cells, performs I/V conversion with a polysilicon resistor, and outputs the sensed voltage to the ADC. During a calibration cycle, G_{Cal} cancels G_{In}'s offset and sets a target equivalent offset ($-V_{TGT}$) across G_{In}'s input. An NMOS-follower buffer shifts V_{OFFP} and V_{OFFN} by $V_{GS,N}$ to meet the input common-mode requirement of G_{Cal} and to reduce kickback onto the calibration storage nodes. The storage capacitors are thick-oxide PMOSCAPs for retention, diluting energy and availability costs of calibration cycles. The pre-gain signal path handling V_{CELLS} during operation and V_{TGT} during calibration is differential, and the common-mode input provided by R_{DAC} during calibration is near $0.5\times V_{TGT}$ to match that of the operational voltage across memory cells. 4-terminal sensing of the cell bias V_{CELLS} is achieved by using added BL sensing wires to measure V_{BL} from the north of the array (current flows south to read circuit), while SL Kelvin sensing occurs by tapping the SLs to V_{SS} at the north of the array and sensing V_{SL} at the south. Up to the finite regulator gain this removes distortion from IR drop across the MUXes (fixed-R) and BL/SL metal (variable-R). Monte Carlo shows a working V_{TGT} range of ~20mV to ~80mV. Biasing, 7B IDAC and resistor for V_{TGT} (R_{DAC}) are shared within one macro.

Shown in Fig. 4, the split-DAC array ADC architecture reduces area, yielding a 278um^2 per-lane ADC cost. The ADC uses a shielded 1.7fF unit metal-oxide-metal (MOM) capacitor to fit the 16-column pitch while preserving signal integrity. To address (2), the ADC first senses and stores its intrinsic offset, which is then dynamically trimmed using a lookup table (LUT) based on input data sparsity to counter offset due to I_{OFF}. The 6-bit offset code feeds a ½-LSB weighted built-in offset-trimming CDAC. This allows sub-LSB correction and extends ADC input range. The SAR and calibration logic placement and routing (PnR) is optimized as a unified block, saving 11.67% vs. per-channel logic with the same flow. ECC logic macros (SECDED, TEC [7]) are included on-chip; ECC can efficiently boost accuracy given low (~10^{-2}) raw MAC BER.

Measurements and Results: Measured calibration retention and end-to-end CIM characteristics are shown in Figs. 5 and 6, and measured solutions to challenges are shown in Figs. 7 and 8. The measured analog calibration retention allows ≥10^3 read operations per refresh for a wide range of V_{BL}. For characterization, a checkerboard pattern is programmed using write-readback and pseudorandom test vectors for each mode (8WL, 16WL, …) are generated with 1,000 vectors per MAC output. For ECC testing, synthetic vectors are generated following the measured distribution to query bit error rates (BERs) below 10^{-3}. MAC root-mean-squared error (RMSE) is computed using the decoded test vector results, weighted to a 50% binomial distribution for input and weights bits. RMSE ranges from 0.078 (0.0002 with ECC at 28% access overhead) to 2.245 (units of *decoded LSBs*). The hybrid calibration shows <2% measured ch./ch. read σ/μ using test patterns with 8-32 on-state cells; IR drop is below 0.005% per WL, or 1.28% worst-case error. The ADC offset scheme eliminates the ~0.86 raw LSB per off-state-cell error due to I_{OFF}. 8-WL CIM BER reaches < 5×10^{-8} using TEC ECC. Average efficiency ranges from 9.81TOPS/W in 8-WL mode up to 75.17TOPS/W in 64-WL mode during 900mV/75MHz operation (Fig. 11).

References:[1] J. M. Hung, *ISSCC,* 2022, pp.1-3. [2] S. D. Spetalnick, *ISSCC*, 2022, pp.1-3. [3] J. M. Correll, *VLSI*, 2022, pp.264-265. [4] C. X. Xue, *ISSCC*, 2021, pp.245-247. [5] J. H. Yoon, *ISSCC*, 2021, pp.404-406. [6] C. X. Xue, *ISSCC*, 2020 pp.244-246. [7] B. Crafton, *ASSCC*, 2021, pp.1-3.

Macro Overview and Read Channel Design Highlights

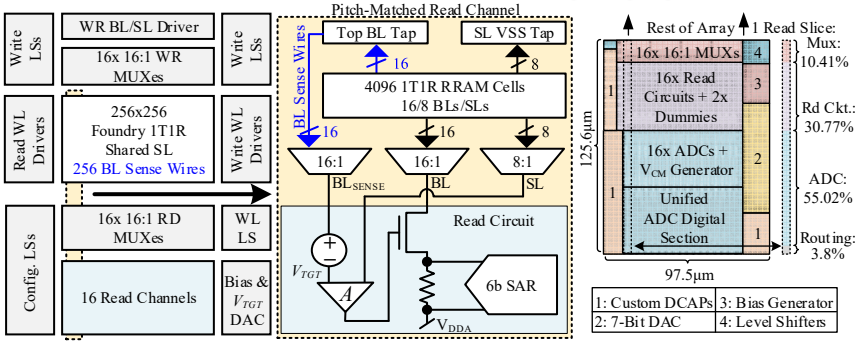

Fig. 1 Topology of the 256x256 RRAM CIM macro with added BL sense wires, summary of a single 16:1 pitch-matched read channel and read-subsection floorplan.

Contributions to Analog CIM with RRAM

Challenge	Importance	Prior Art	Contribution
Macro Channel Mismatch	Address-Dependent MAC Errors	Not Reported	**Hybrid Analog/Digital Offset Cancelling Scheme (Measured)**
Large RRAM I_{OFF}	Data-Dependent MAC Errors	Not Reported / Gain Scaling[5]	
$I_{READ}R_{BL\,SL}$ Drop	Address-Dependent MAC Errors	Not Reported	**Differential Frontend with 4-Terminal V_{CELL} Sensing (Measured)**
Cell/Cell Variation	Noisy MAC Result	Write/Readback	**Write/Readback + CIM ECC (Reported)**
Macro Eff. Cell Density	CIM vs. Data Movement Tradeoff	Low Density / Higher [2,4]	**Highest Density (Post-Shrink Macro Area)**

Fig. 2 Challenges blocking accurate many-WL RRAM CIM, prior art, and contributions.

Offset-Calibrated Differential Current-Sensing Front-End

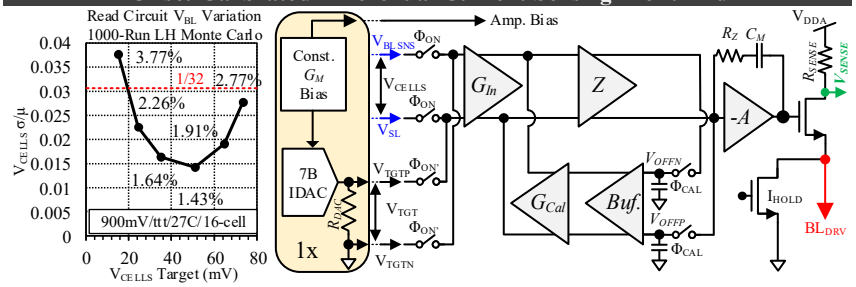

Fig. 3 Topology and simulated input-offset variation of the differential-voltage-regulating current-sensing front-end. One bias arm with V_{BL} DAC is shared in macro.

Offset-Adjusted 6-Bit SAR ADC

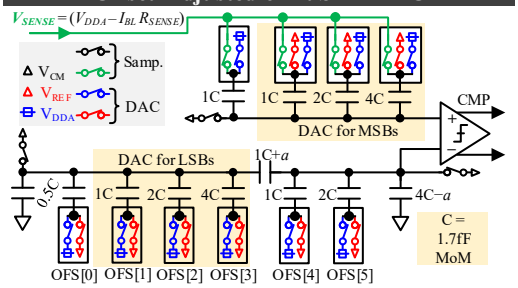

Fig. 4 Implementation of the offset-sensing/offset-adjusted 6-Bit Split-DAC SAR ADC

Measured End-to-End Characteristics of CIM MAC Operation

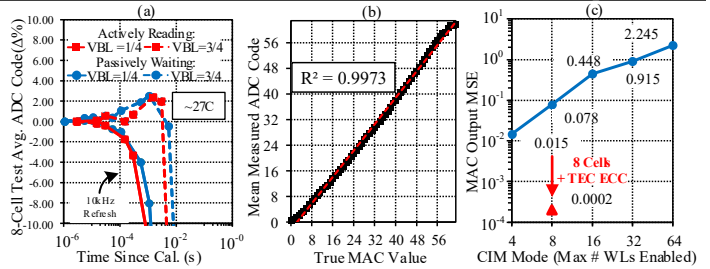

Fig. 5 Measured: (a) analog front-end cal. retention time, (b) CIM MAC linearity in 64-WL mode, (c) MAC result RMS error using processed checkerboard data with 1K pseudorandom test vectors per output state.

Measured Solutions to CIM Challenges: Offset, IR Drop, I_{OFF}, σ_{Cell}

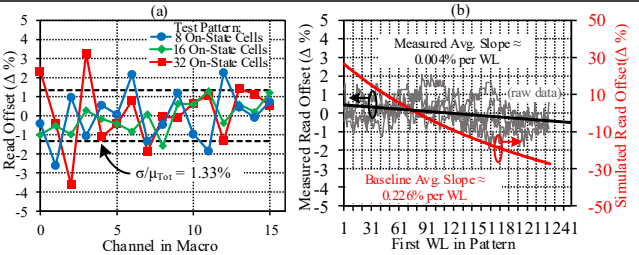

Fig. 7 Measured: (a) calibrated chan./chan. variation (b) change in raw ADC code for test-pattern shifted down BL.

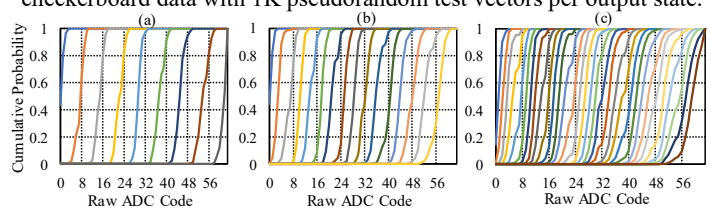

Fig. 6 Meas. cumulative ADC code prob. at each MAC output state for (a) 8, (b) 16, and (c) 32-WL CIM modes. Using raw checkerboard data.

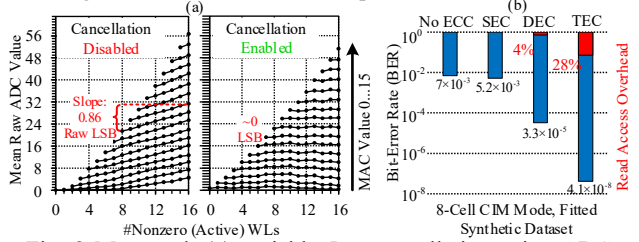

Fig. 8 Measured: (a) variable I_{OFF} cancellation using ADC CDAC, (b) 8-WL CIM BER with and without ECC.

Die Photo and Summary Information

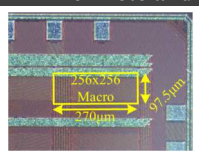

Fig. 9 Die Microphoto.

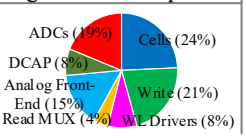

Fig. 10 Area Breakdown.

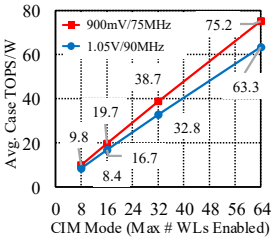

Fig. 11 CIM MAC eff. (50% 1/0 inputs & weights) from meas. and sim.

Tech.	40nm RRAM	CIM Mode:	8	16	32	64
Area	0.0263mm²	Avg. TOPS/W	9.81	19.66	38.73	75.17
Bitcells	65,536	Peak TOPS/W	15.47	30.93	61.87	123.73
RD Volt.	0.9 ~ 1.1	TOPS/mm²	0.87	1.75	3.50	7.01
WR Volt.	1.3 ~ 3.3	MAC RMSE	0.078	0.448	0.915	2.245

Fig. 12 Test chip info. and CIM perf. summary.

Comparison to Prior Compute-in-Memory with RRAM Works

	ISSCC '20 [6]	ISSCC '21 [5]	ISSCC '21 [4]	VLSI '22 [3]	ISSCC '22 [2]	ISSCC '22 [1]	**This Work**
Technology (nm)	22nm	40nm	22nm	65nm	40nm	22nm	40nm
Memory	1T1R RRAM	1T1R RRAM	1T1R RRAM	1T1R RRAM	1T1R RRAM	1T1R RRAM	1T1R RRAM
Supply (V)	0.7-0.9	0.9	0.8-0.9	1.2	0.83-1.1	0.75-0.8	0.9-1.1
I_{OFF} Cancellation	N.R.	Gain Scaling*	N.R.	N.R.	N.R.	N.R.	**To ½ LSB**
$I_{READ}R_{BL}$ Drop	N.R.	N.R.	N.R.	N.R.	N.R.	Cal. Table*	**< 0.005%/WL**
Ch./Ch. Variation	N.R.	N.R.	N.R.	N.R.	N.R.	Cal. Table*	**σ < 2%**
ADC Nom. Precision	6	4	1	8	4	3	**6**
MAC-Level RMSE	App. Only	App. Only	N.R.	App. Only	N.R.	App. Only	**0.078 – 2.245**
Peak (Reported) Eff. (TOPS/Watt)	121.38	56.67	195.7	662.4	26.56	1286.4	**350****
Avg. Eff. (TOPS/W)	N.R.	4.15	N.R.	N.R.	5.63	416.5	**9.81 – 75.17**
Density: M.Cell/mm²	0.9	0.146	1.56	0.072	2.31	0.506	**2.38**
Compute Density (TOPS/mm²)	N.R.	1.437	0.161	8.5	0.303	0.324	**0.87 – 7.008**
ECC Support	N.R.	N.R.	N.R.	N.R.	N.R.	N.R.	**Up to TEC**

N.R.: Not Reported | *Technique Shown, Effect Not Reported | App. Only: Accuracy shown in **Application Only** | Mcell: 2^{20} Cells
**Max. is 256 Active Rows, 75% Sparse, 75MHz @ 900mV with degraded accuracy. Avg. Eff & Compute Density use up to 64WL Mode.

Fig. 13 Taped-out prototype design comparison with prior works.