Efficient & Reliable RRAM-based Compute-in-Memory for Edge Intelligence

iMACAW Next Gen Talk @ DAC

Wantong Li

Jul 09, 2023

ELECTRICAL [+] COMPUTER

ENGINEERING

Georgia Tech School of Electrical and Computer Engineering

Outline

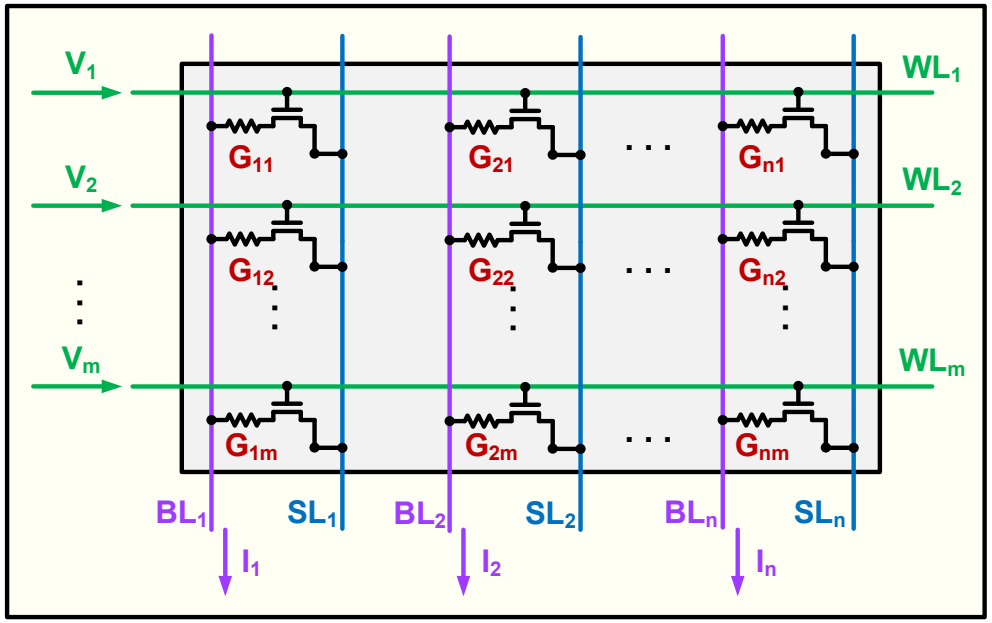
- Overview of Compute-in-Memory
  - Merits of CIM and its design challenges
  - Overview of RRAM-CIM macro tape-out
- Techniques to Address Reliability Challenges
  - On-chip write-verify
  - Temperature-independent ADC references
  - Parallelism-preserving ECC
- Summary and Related Work

Outline

- Overview of Compute-in-Memory
  - Merits of CIM and its design challenges
  - Overview of RRAM-CIM macro tape-out
- Techniques to Address Reliability Challenges
  - On-chip write-verify
  - Temperature-independent ADC references
  - Parallelism-preserving ECC
- Summary and Related Work

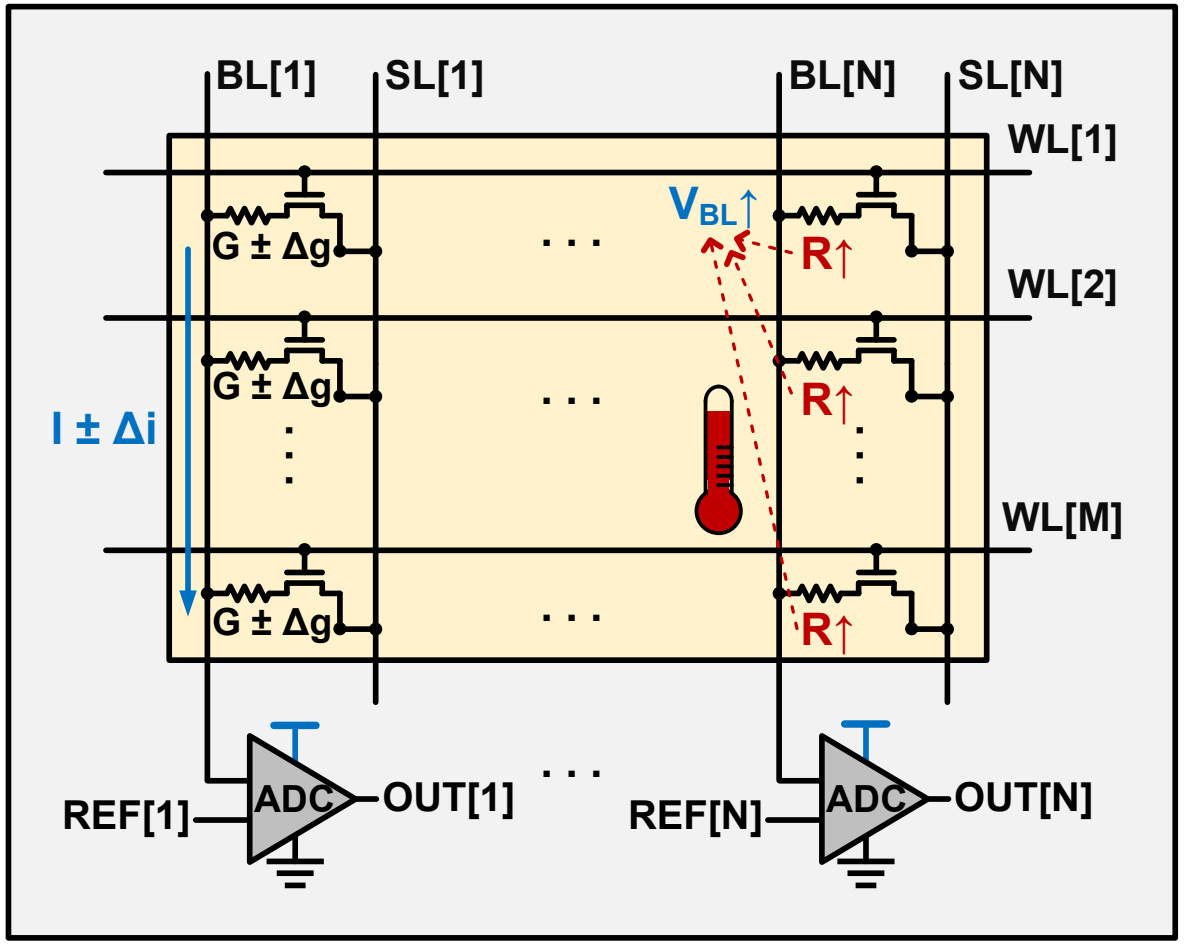
Analog CIM with Non-Volatile Memories

- Motivation of analog compute-in-memory (CIM)
  - Leverage current-summation property of memory arrays for parallel computation
  - Accelerate multiply-and-accumulate operations (MAC) in neural network inference
- Benefits of using non-volatile RRAM for CIM
  - Very low leakage power for standby
  - Smaller cell size compared to SRAM
  - Potentially hold all model weights on-chip
  - Reduce off-chip memory access



$$\begin{pmatrix} I_1 \\ I_2 \\ \vdots \\ I_n \end{pmatrix} = \begin{pmatrix} G_{11} & G_{12} & \cdots & G_{1m} \\ G_{21} & G_{22} & \cdots & G_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ G_{n1} & G_{n2} & \cdots & G_{nm} \end{pmatrix} \begin{pmatrix} V_1 \\ V_2 \\ \vdots \\ V_m \end{pmatrix}$$

Design Challenges for Analog RRAM-CIM



- **Challenge 1: Process**

RRAM conductance variations induce BL current variations

- **Challenge 2: Temperature**

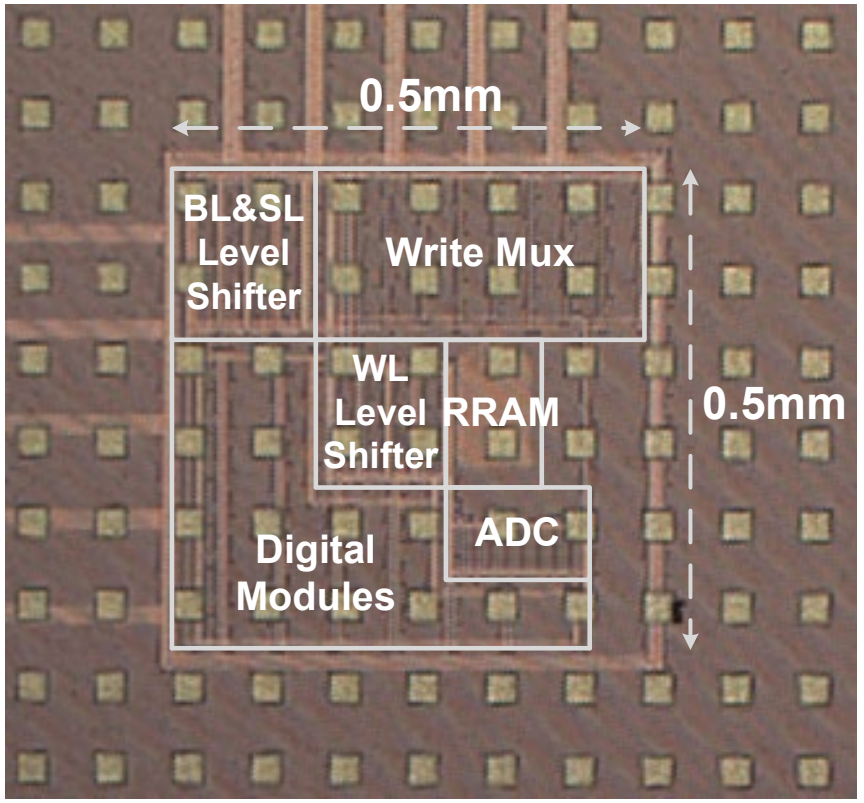
RRAM conductance changes with temperature

- **Challenge 3: Voltage**

Lowering VDD reduces ADC sense margin in voltage-mode CIM

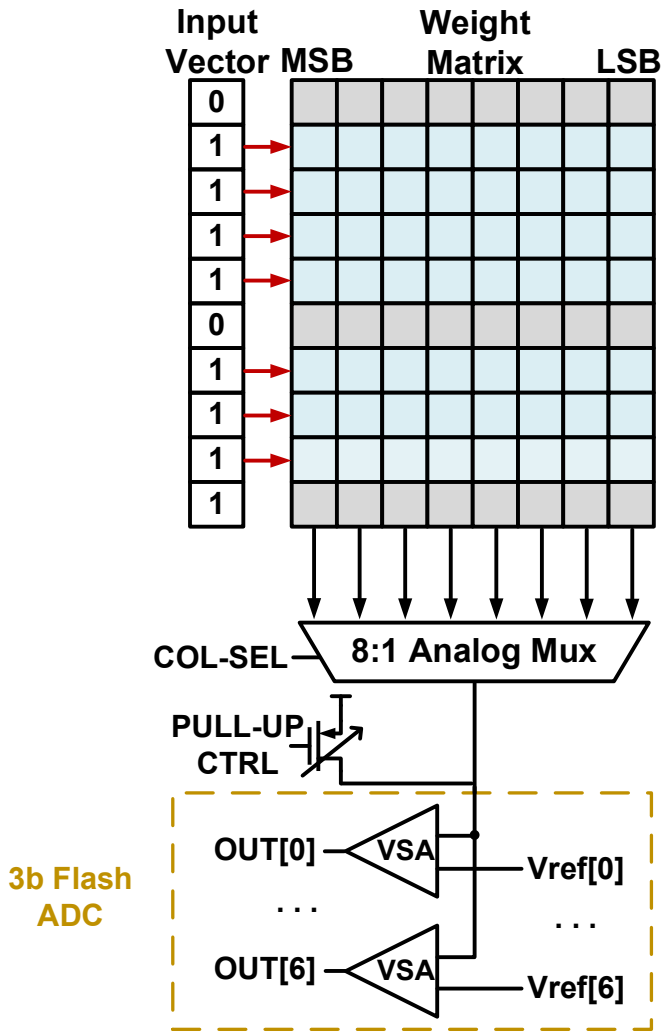
RRAM-CIM Macro Overview

Design Challenges	Proposed Circuit Design Techniques
Process	On-chip write-verify to tighten RRAM distributions
Temperature	RRAM-based temperature-independent ADC references
Voltage	Parallelism preserving in-situ ECC for iso-accuracy voltage scaling



- Test chip taped-out in TSMC 40nm process with 1T1R RRAM

CIM Compute Scheme



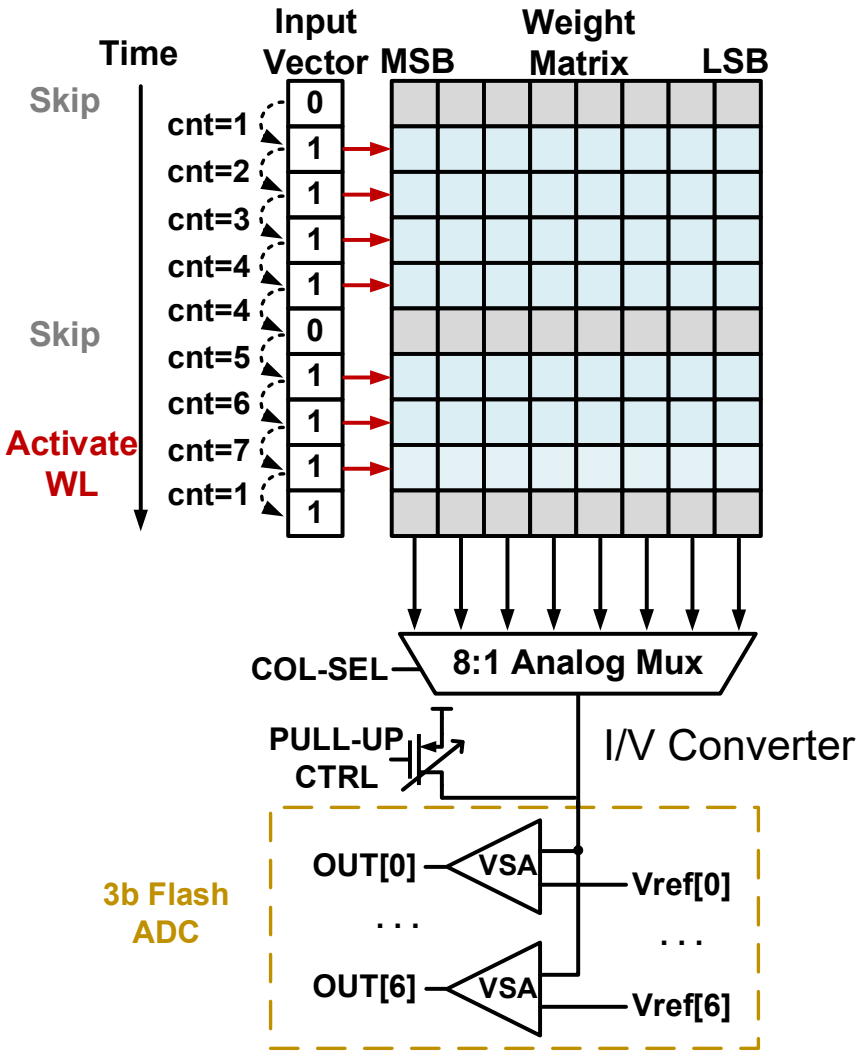
Input Control

- Opening multiple WLs at once enable parallelized MAC operations
- Each digital input bit asserts one WL
- Input controller activates 7 WLs for each set of computation
  - Limited opening of rows due to small on/off ratio

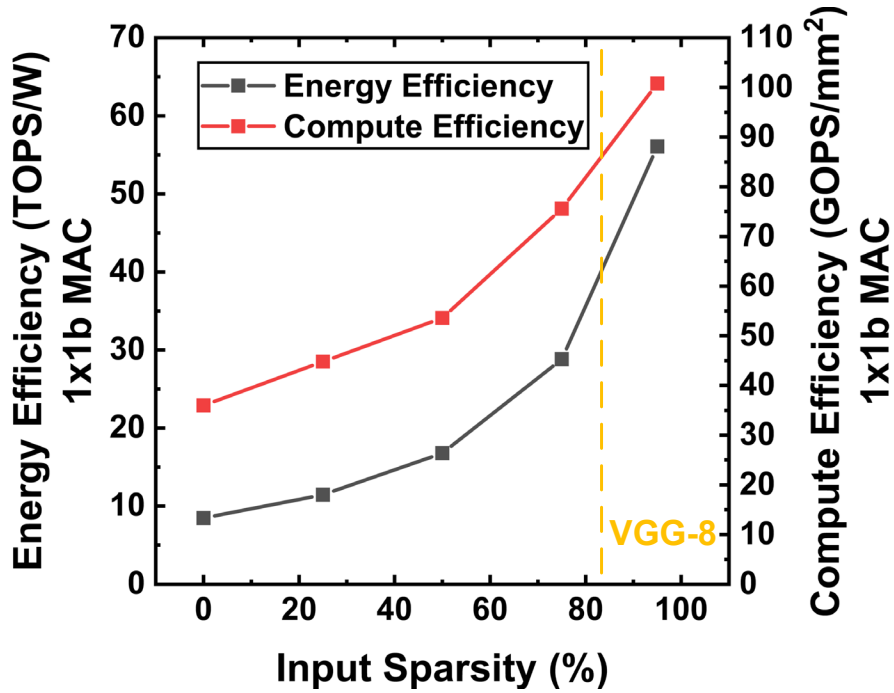
Sensing

- Resistive divider between RRAM cells and pull-up PMOS
  - Converts current to voltage
- BL voltage sensed by 3-bit flash ADC
- 7-bit ADC thermometer output is encoded as 3-bit binary

Sparsity-Aware Input Control



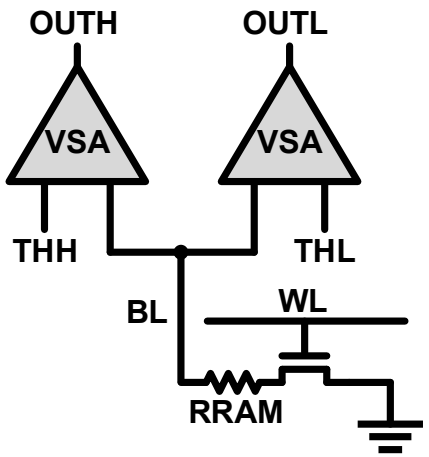
- Sparsity control scans the input vector to skip unnecessary computations with 0's
- Provides 2.3× higher throughput and 4.3× higher energy efficiency for input sparsity at ~80%



Outline

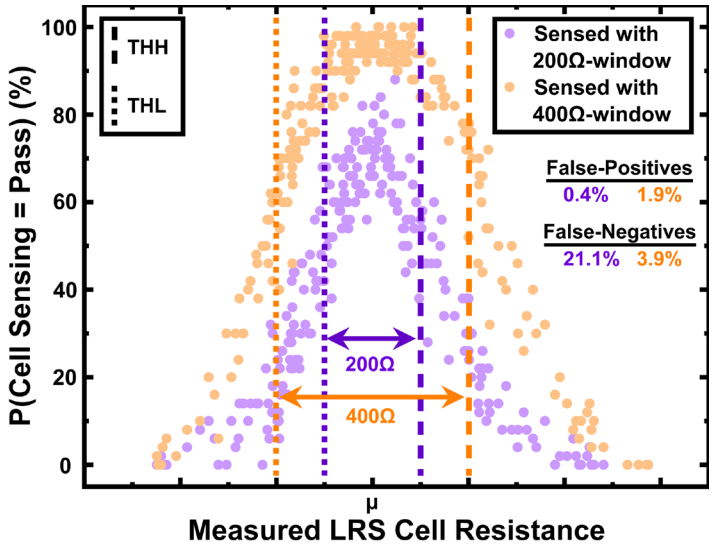
- Overview of Compute-in-Memory
  - Merits of CIM and its design challenges
  - Overview of RRAM-CIM macro tape-out
- Techniques to Address Reliability Challenges
  - On-chip write-verify
  - Temperature-independent ADC references
  - Parallelism-preserving ECC
- Summary and Related Work

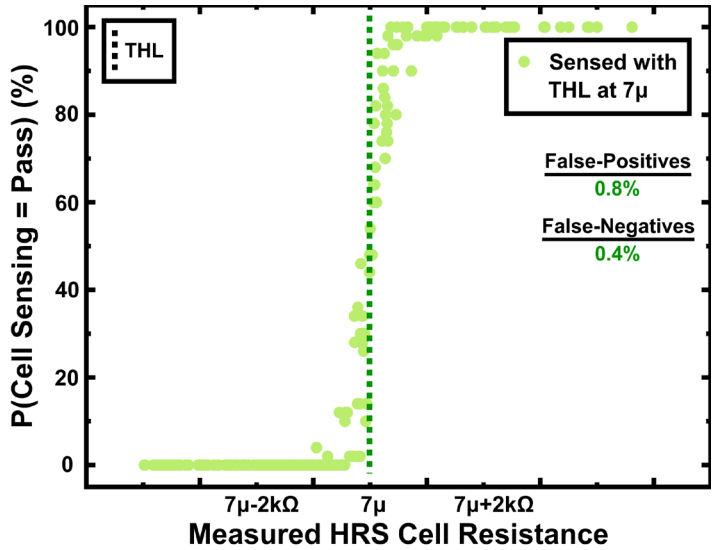
On-Chip Write-Verify



OUTL = BL > THL
OUTH = BL > THH

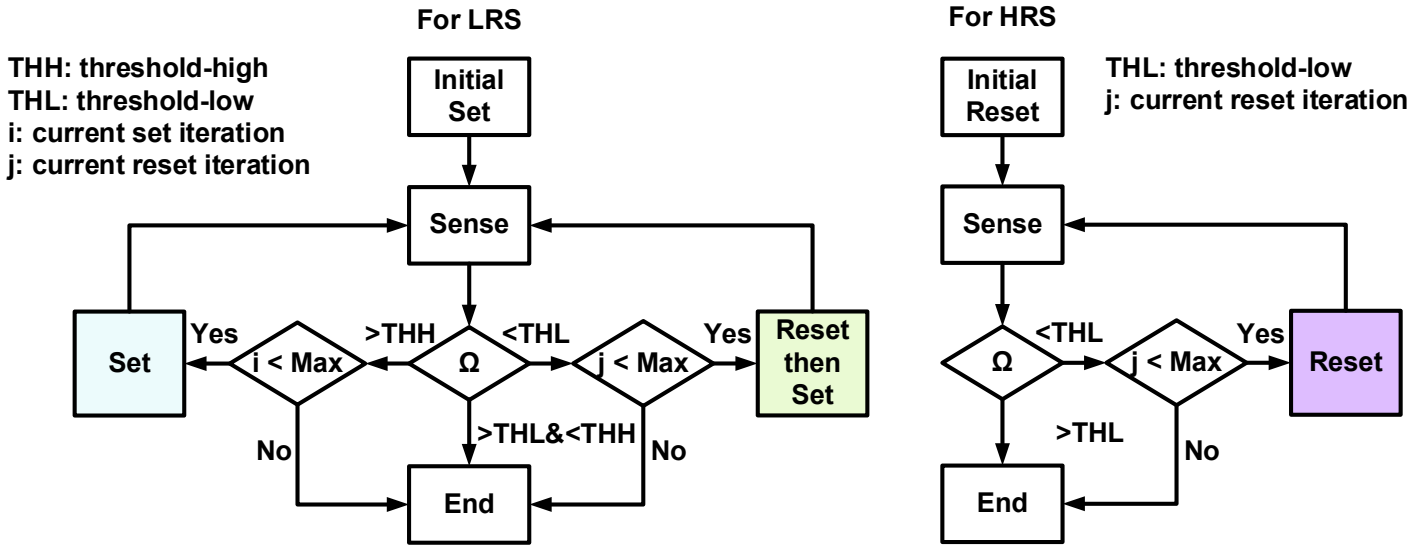
if (OUTL=1 & OUTH=0)
Cell Sensing = Pass
else
Cell Sensing = Fail

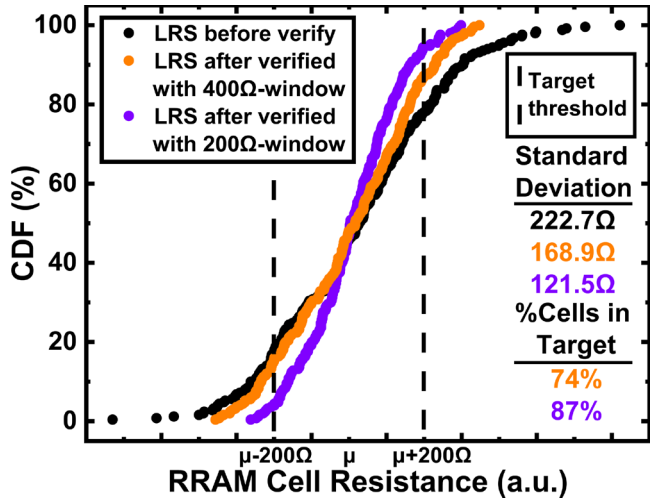


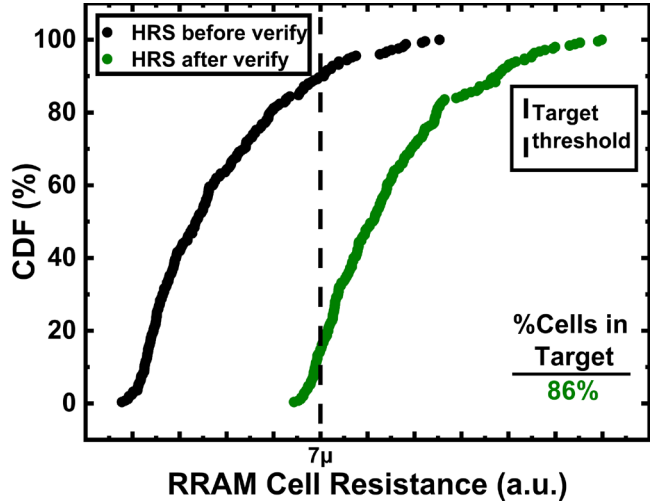


- Dual voltage-sense amplifiers (VSAs) to program RRAM cells for CIM applications
  - Tighten LRS (logic 1) distribution and move HRS (logic 0) ideally to infinity
- Two thresholds (THH & THL) make up a resistance window
- Smaller target window size provides better accuracy, but requires more iterations

Write-Verify Protocol and Measurement Results



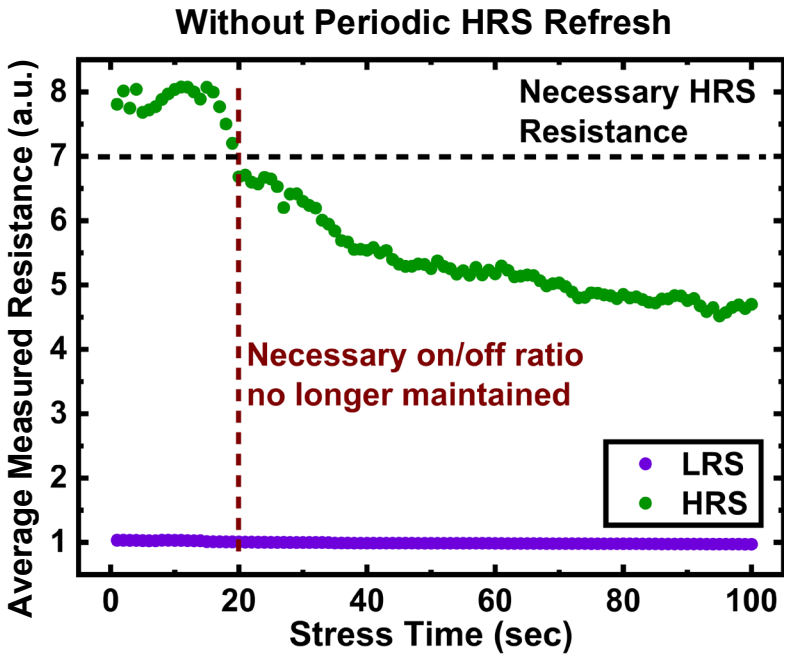


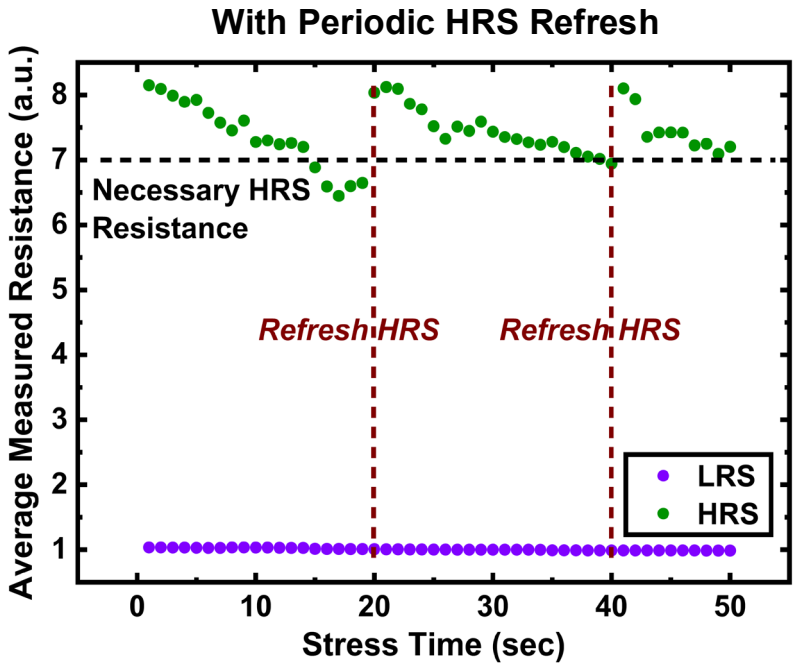


- For HRS, only one VSA is needed
- Goal: program LRS around μ and HRS above 7μ
- >85% cells can be in programmed in target
- 10^5 speedup compared to external equipment

Periodic Cell Refresh

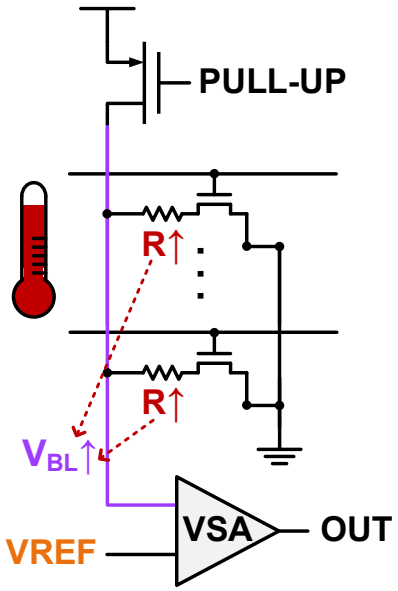
- Compute mode in CIM is similar to high-stress read
  - Can cause downward cell drift, especially for HRS
- Use on-chip write-verify to periodically refresh drifted cells to maintain memory window

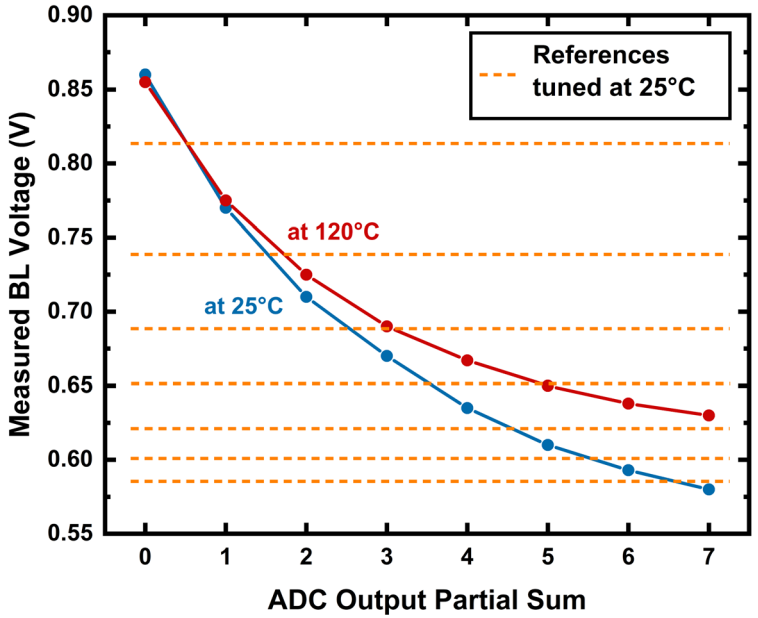




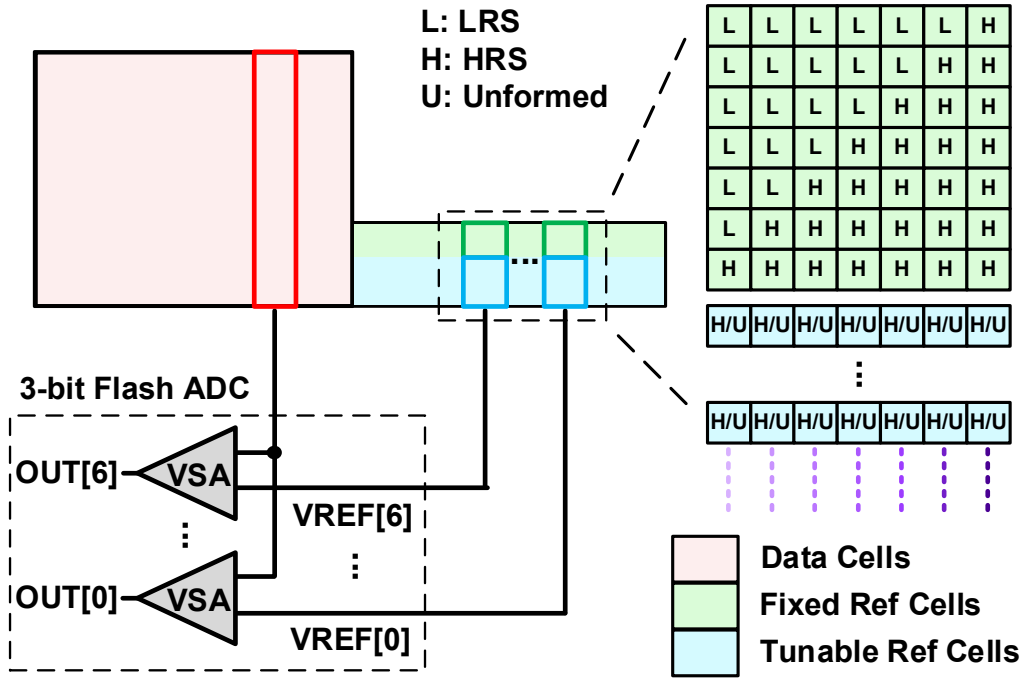
Temperature-Independent ADC References

- V_{BL} is dependent on temperature
  - LRS resistance tends to rise with temperature
  - Rigid references tuned at one temperature works poorly at others



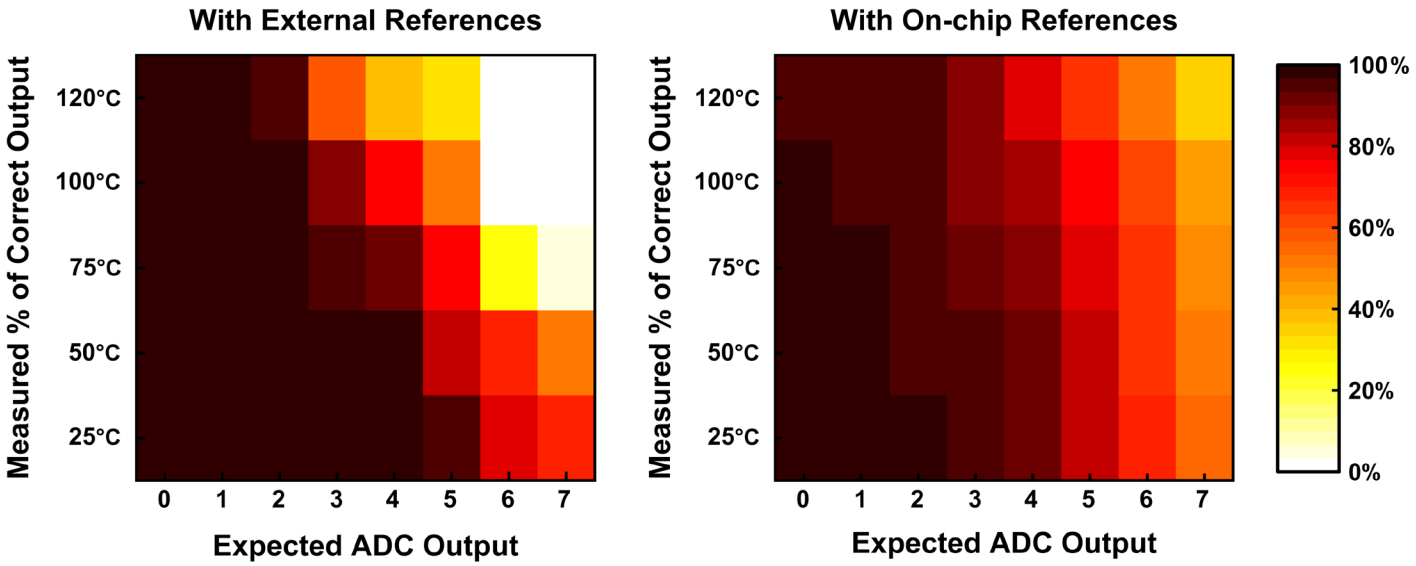


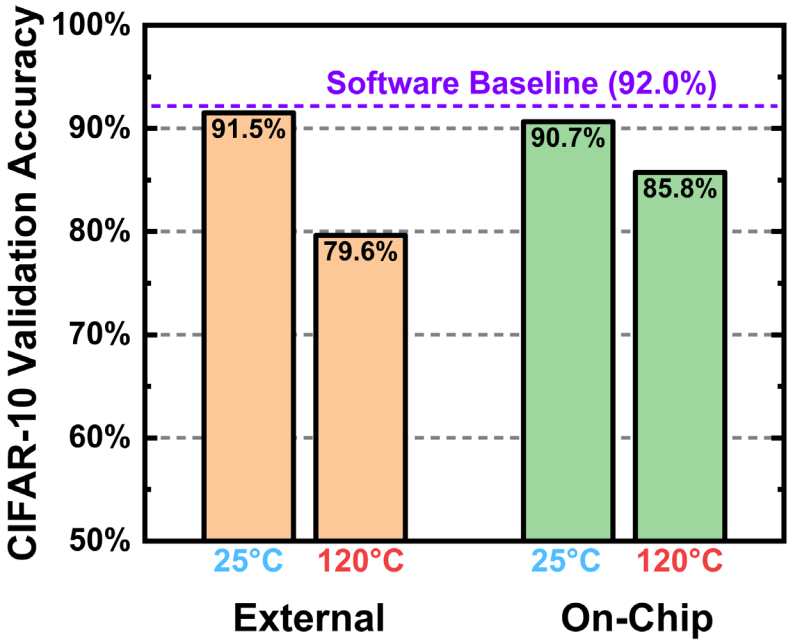
- RRAM-based references provides self-tracking to temperature
  - Data and reference cells move together with temperature



Evaluation of ADC References

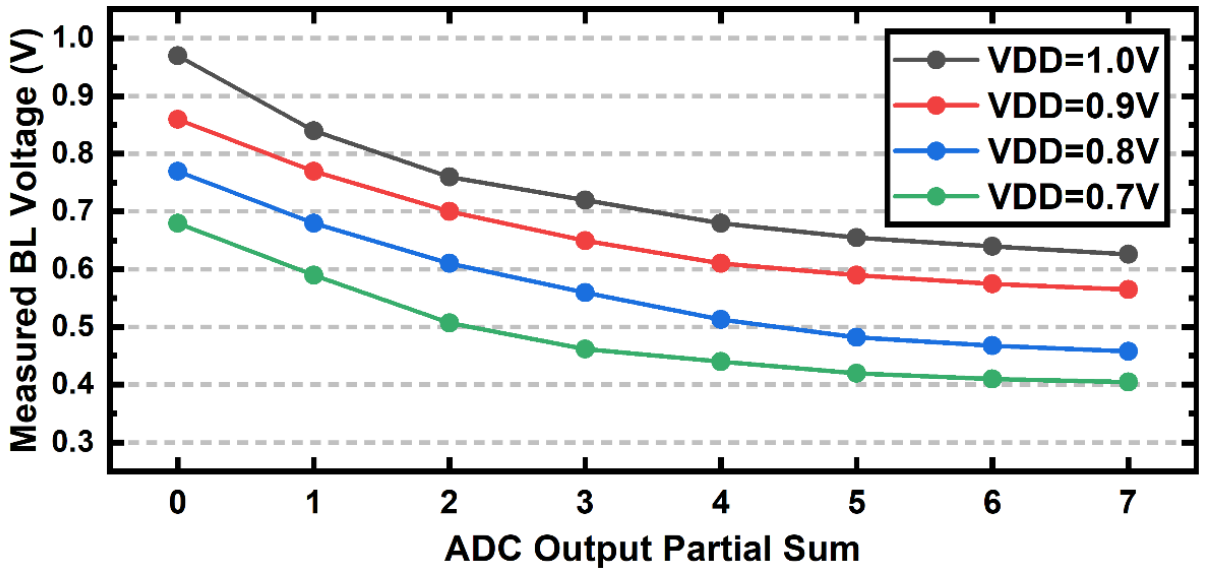
- Measured ADC outputs for each possible partial sum at various temperatures
  - Missing codes in rigid references tuned at one temperature (e.g., at 25°C)
  - All codes are retained with self-tracking on-chip references
- With on-chip references, simulated CIFAR-10 accuracy at 120°C can be recovered by 6% to 85.8%





Effect of Voltage Scaling to CIM

- Voltage scaling is a popular method to toggle between high-performance mode and low-power mode
- Lowering supply voltage is more detrimental to voltage-mode analog CIM
  - 52.4% reduction in ADC sense margins when VDD is lowered from 1V to 0.7V



MAC-ECC: In-Situ SECDED for CIM

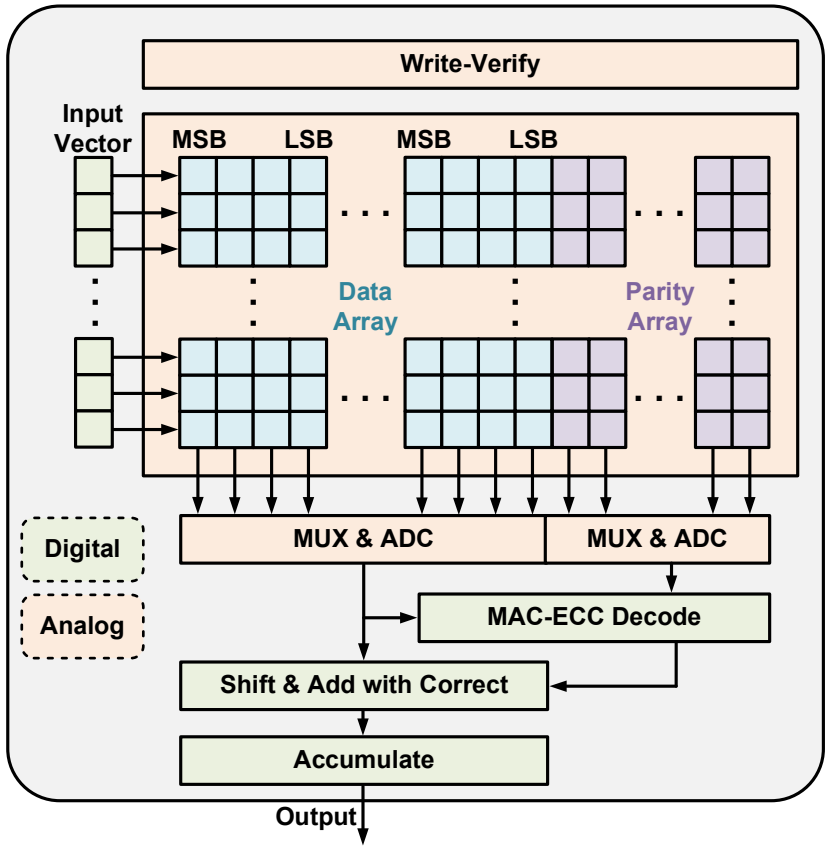
- MAC-ECC corrects digital errors based on arithmetic distance, while preserving CIM row-parallelism
- Example: $(0111)_2$ and $(1000)_2$
  - Hamming distance: 4; Arithmetic distance: 1

Encode off-chip

During compute

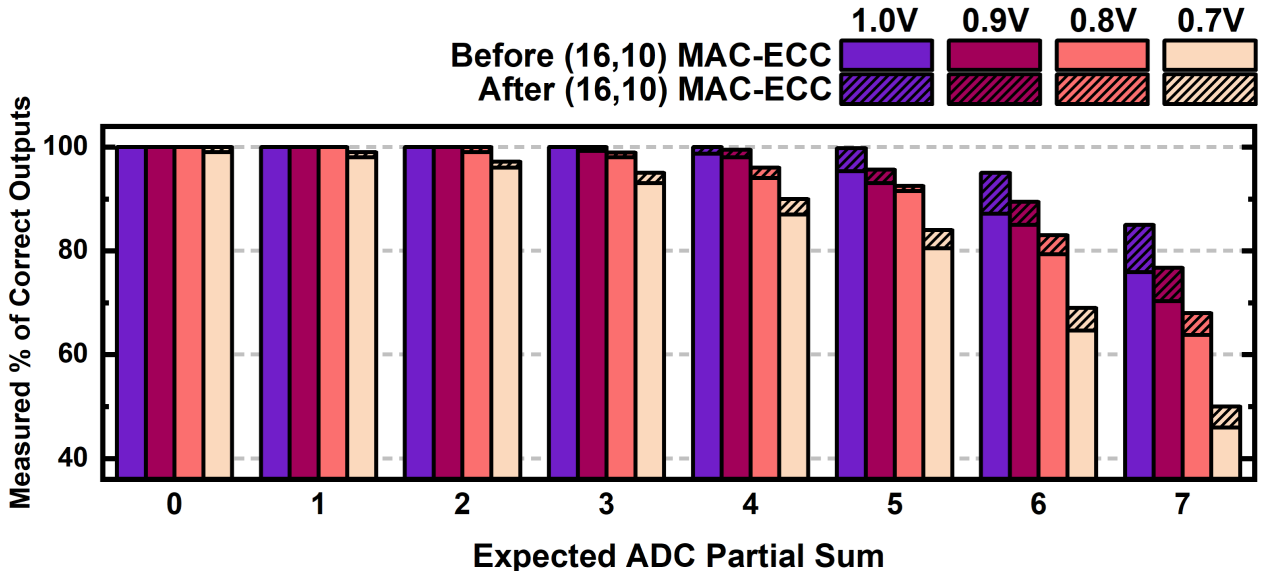
ECC Procedure

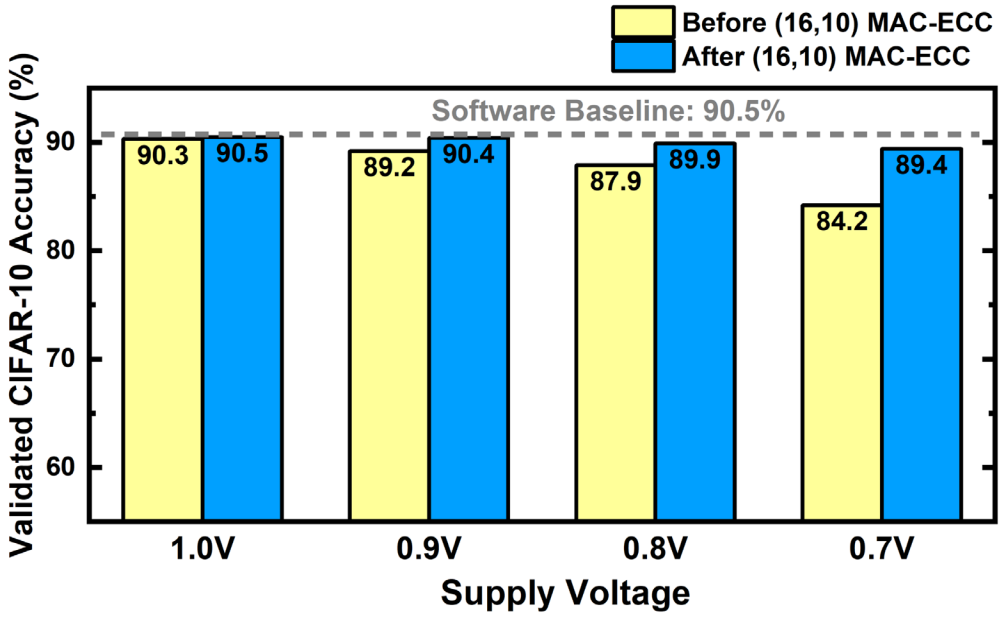
1	Encode Parity	$\sum_{j\in[m\rangle} p_{i,j}\,\alpha_{k+j} = \left(-\sum_{j\in[k\rangle} d_{i,j}\,\alpha_j\right) \bmod \beta$
2	Compute In Memory	Obtain output vector y through asserting WLs with input vector; y may contain one or multiple errors
3	Calculate Syndrome	$s = (y\cdot\alpha) \bmod \beta$ $\bar{s} = \beta - s$
4	Decode	$s = 0$: y is error-free $s = \alpha_j$ for $j\in[n\rangle$: +1 error at column j $\bar{s} = \alpha_j$ for $j\in[n\rangle$: -1 error at column j $s, \bar{s} \neq \alpha_j$ for $j\in[n\rangle$: multiple errors



Measurement and Simulation Results

- ADC errors can be reduced after applying MAC-ECC at each tested VDD
- At 0.7V, original CIFAR-10 accuracy loss of 6.3% may be recovered to 1.1% with MAC-ECC





Iso-Accuracy Voltage Scaling

High Performance

Low Power

VDD	Least costly MAC-ECC for <1% loss	Frequency	Energy Efficiency (TOPS/W)	Compute Efficiency (GOPS/mm^2)	Energy Overhead	Area Overhead
1V	No ECC	115MHz	43.0	112.5	0%	0%
0.9V	(31, 25)	100MHz	46.2	93.1	3.73%	3.1%
0.8V	(25, 19)	90MHz	52.4	82.8	4.95%	4.11%
0.7V	(16, 10)	80MHz	59.1	70.9	7.48%	6.06%

- Target is to keep accuracy loss across different VDDs below 1%
- Iso-accuracy toggling between high performance and low power modes through voltage scaling and reconfigurable MAC-ECC

Outline

- Overview of Compute-in-Memory
  - Merits of CIM and its design challenges
  - Overview of RRAM-CIM macro tape-out
- Techniques to Address Reliability Challenges
  - On-chip write-verify
  - Temperature-independent ADC references
  - Parallelism-preserving ECC
- Summary and Related Work

Summary

- Efficient and PVT-robust RRAM-CIM prototype chip was taped-out in TSMC N40 process

- Address reliability challenges through circuit design innovations
  - Process: On-chip write-verify for tightening distributions and periodic refresh
  - Temperature: Temperature-independent ADC references
  - Voltage: MAC-ECC for iso-accuracy voltage scaling

- Enhance feasibility of analog CIM and provide strong foundation for the building blocks of emerging CIM architecture

References

- W. Li, X. Sun, S. Huang, H. Jiang, S. Yu, “A 40nm MLC-RRAM compute-in-memory macro with sparsity control, on-chip write-verify, and temperature-independent ADC references,” *IEEE Journal of Solid-State Circuits (JSSC)*, 2022.
- W. Li, J. Read, H. Jiang, S. Yu, “MAC-ECC: In-situ error correction and its design methodology for reliable NVM-based compute-in-memory inference engine,” *IEEE Journal on Emerging and Selected Topics in Circuits and Systems (JETCAS)*, 2022.
- W. Li, S. Huang, X. Sun, H. Jiang, S. Yu, "Secure-RRAM: A 40nm 16kb compute-in-memory macro with reconfigurability, sparsity control, and embedded security," *IEEE Custom Integrated Circuits Conference (CICC)*, 2021.
- W. Li, X. Sun, H. Jiang, S. Huang, S. Yu, “A 40nm RRAM compute-in-memory macro featuring on-chip write-verify and offset-cancelling ADC references,” *IEEE European Solid-State Circuits Conference (ESSCIRC)*, 2021.
- W. Li, J. Read, H. Jiang, S. Yu, “A 40nm RRAM compute-in-memory macro with parallelism-preserving ECC for iso-accuracy voltage scaling,” *IEEE European Solid-State Circuits Conference (ESSCIRC)*, 2022.

Acknowledgment

- Ph.D. advisor: Dr. Shimeng Yu
- Collaborators: Xiaoyu Sun, James Read, Hongwu Jiang, Shanshi Huang
- Sponsors: