RRAM-ECC: Improving Reliability of RRAM-Based Compute In-Memory

Zishen Wan*[1], Brian Crafton*[1], Sam Spetalnick[1],

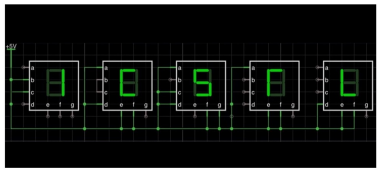
Jong-Hyeok Yoon[2], Arijit Raychowdhury[1]

Georgia Institute of Technology[1]

Daegu Gyeongbuk Institute of Science and Technology[2]

*Equal Contributions

✉ zwan63@gatech.edu, bcrafton3@gatech.edu



22 NVMW

Georgia Institute of Technology

Agenda

1. Motivation & Background
2. RRAM + CIM Measurement
3. CIM-SECDED
4. Successive Correction
5. Summary

Agenda

1. Motivation & Background
2. RRAM + CIM Measurement
3. CIM-SECDED
4. Successive Correction
5. Summary

Compute In-Memory

Advantages

1. Increase Memory Bandwidth
2. Multiply-Accumulate on BL

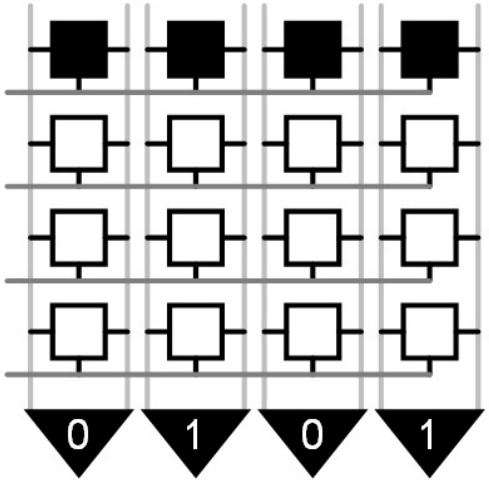
Features

1. Multiple WLs
2. Multiply & Accumulate on BL

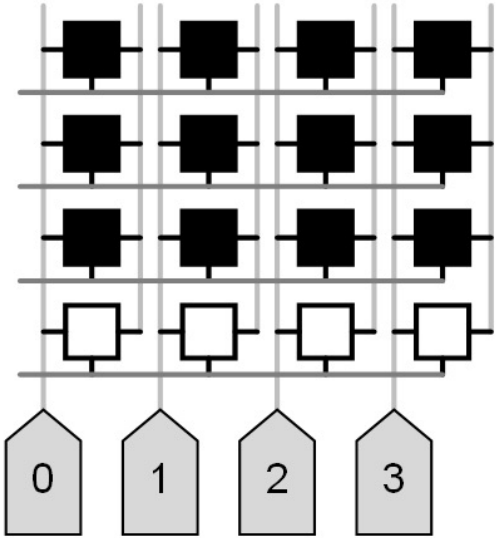
Implications

1. Compute In-Memory →
2. Matrix Multiplication →
3. Deep Learning & AI

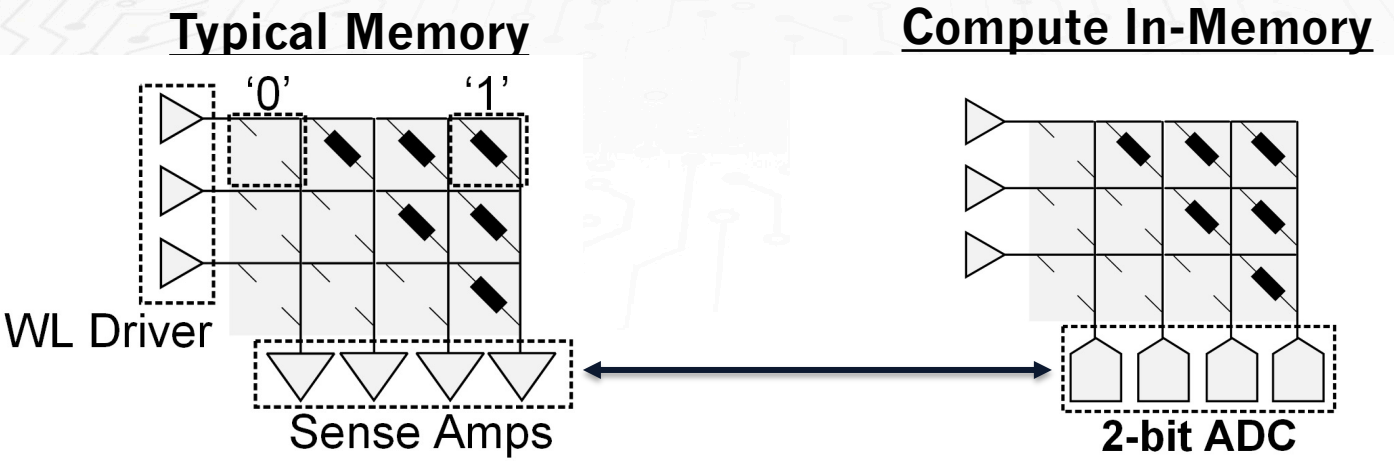
Typical Memory



Compute In-Memory

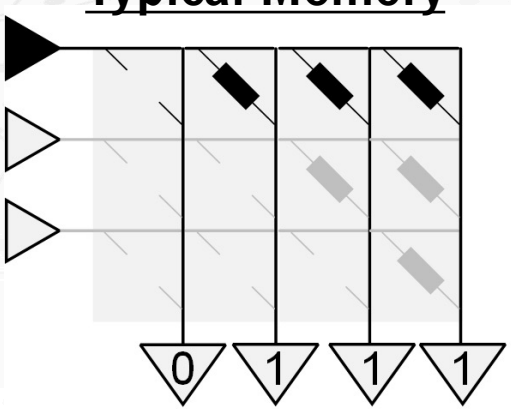


Compute In-Memory

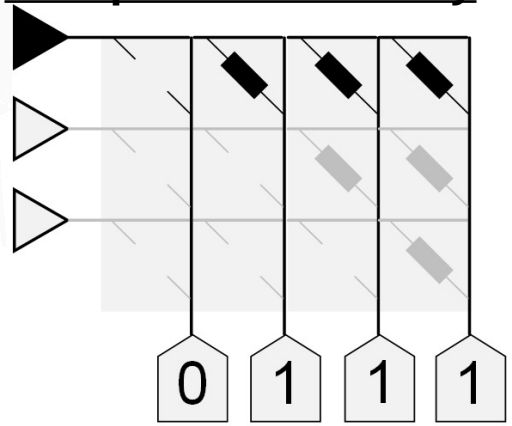


Compute In-Memory

Typical Memory

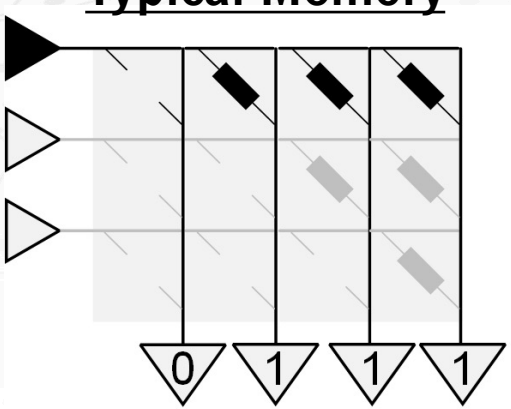


Compute In-Memory

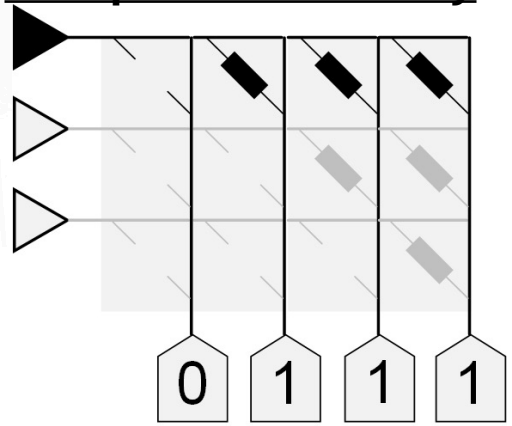


Compute In-Memory

Typical Memory



Compute In-Memory



Advantages:

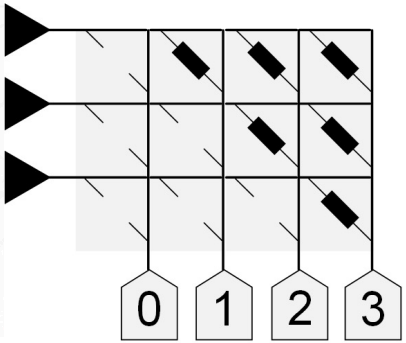
1. ↑Bandwidth (N×)
2. Less communication (N → Log_N)
3. “Free” Compute (N×)

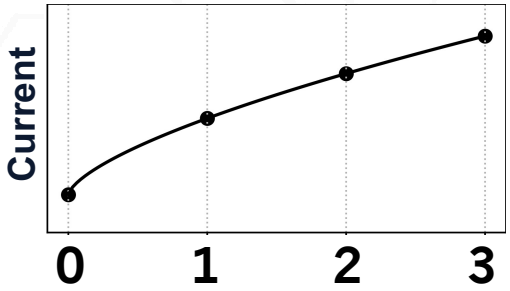
Challenges:

1. ↑Bits → ↑Noise → ↑Error
2. ↑Bits → ↓Headroom → ↑Error

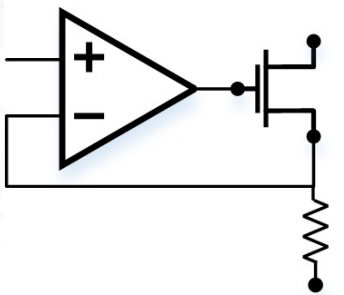
Compute In-Memory

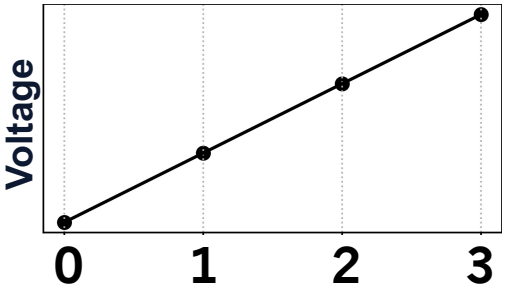
1) Activate WL



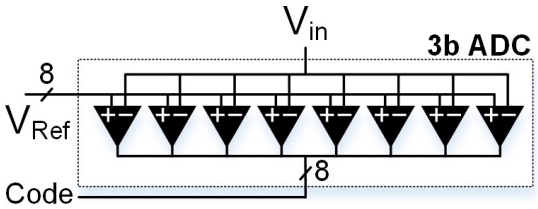


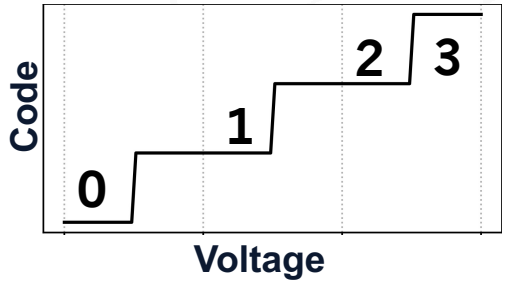
2) Read Circuit





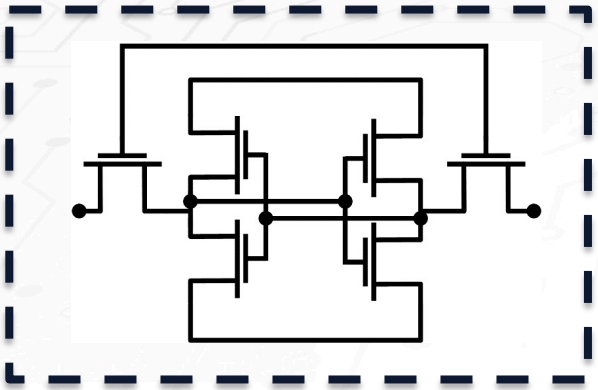
3) Quantize



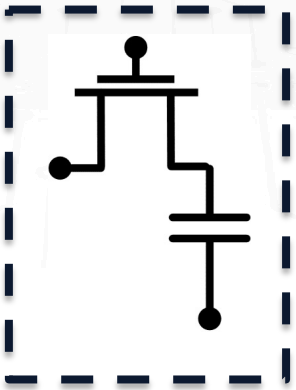


Dense Embedded (On-Chip) Memory

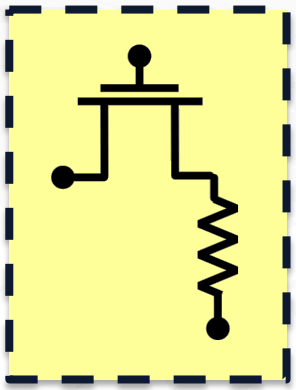
6T SRAM



1T1C DRAM



1T1R RRAM



Memory	SRAM	DRAM	RRAM
Latency	Very Fast	Fast	Fast
Power	Low	Medium	Low
Volatile	Volatile	Volatile	Non-Volatile
Density	Low	Very High	High

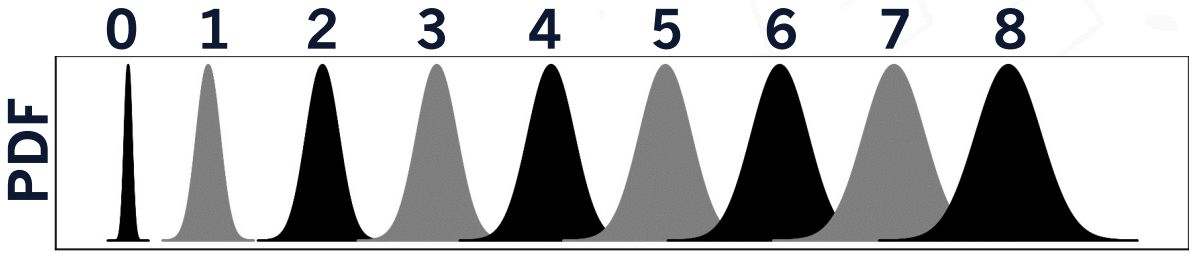
Challenges for RRAM + CIM

RRAM

- HRS: 30KΩ → 0
- LRS: 3KΩ → 1

Challenges for CIM

- Accumulate variation
- Reduced sense margin

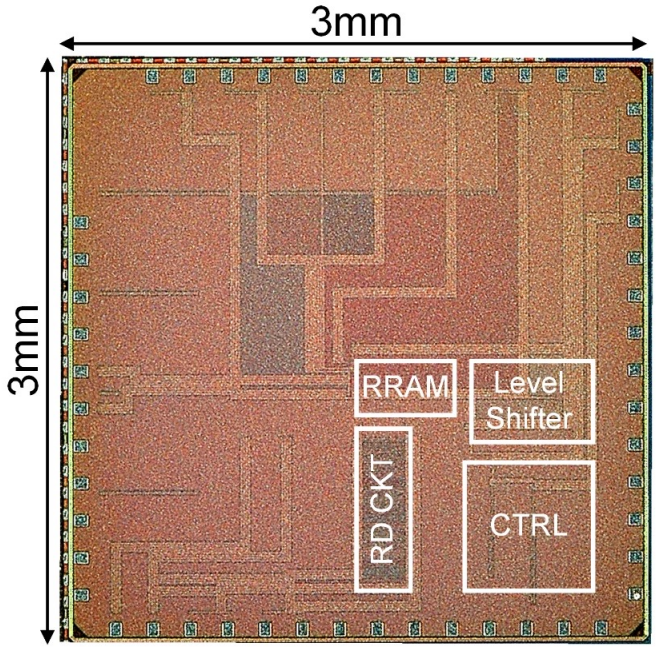


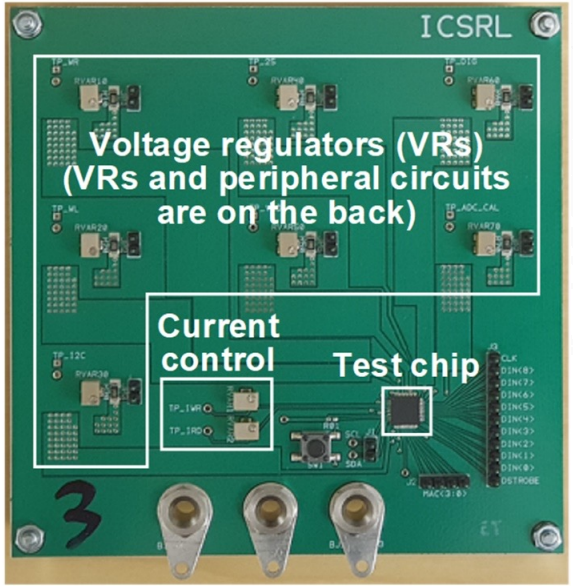
… but no ECC!

Agenda

1. Motivation & Background
2. RRAM + CIM Measurement
3. CIM-SECDED
4. Successive Correction
5. Summary

Die Shot & PCB





Technology	40nm RRAM & CMOS
Frequency	100 MHz
Active Area	0.437 mm^2
Package	QFN48

Digital VDD	0.9V
Analog VDD	0.8V
I/O VDD	3.3V
Power	11.05 mW

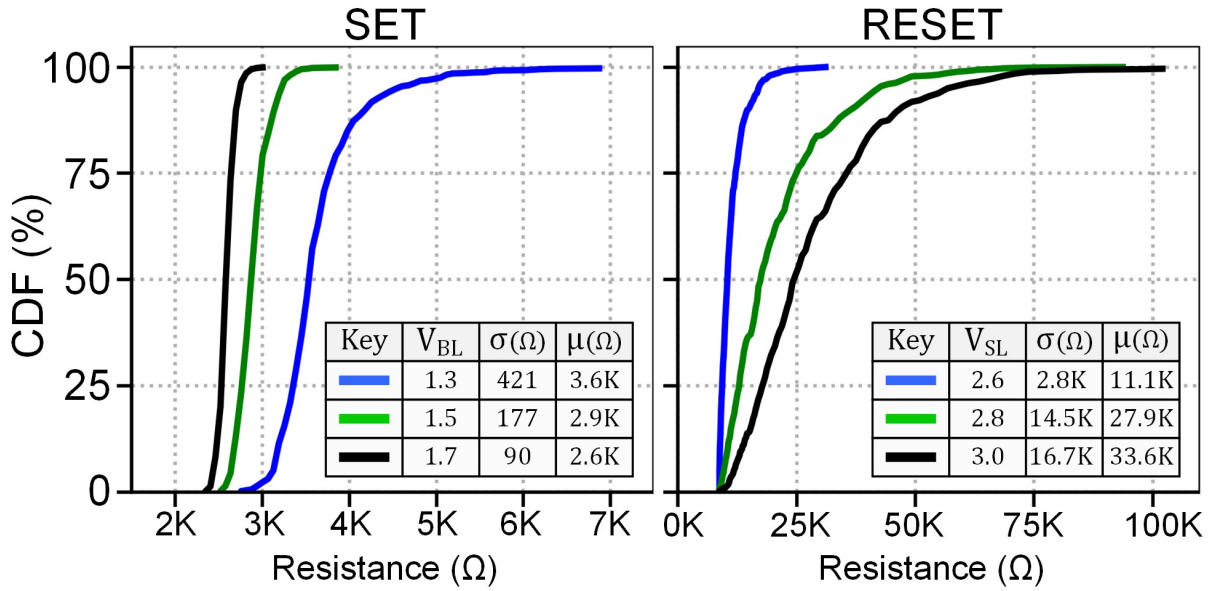
Measurements: Variation

Experiment

- 8192 measurements
- Resistance distributions (CDF)

Observations

- ↑Write Voltage → ↓Variation
- ↑Write Voltage → ↑Ratio
- … and lower endurance



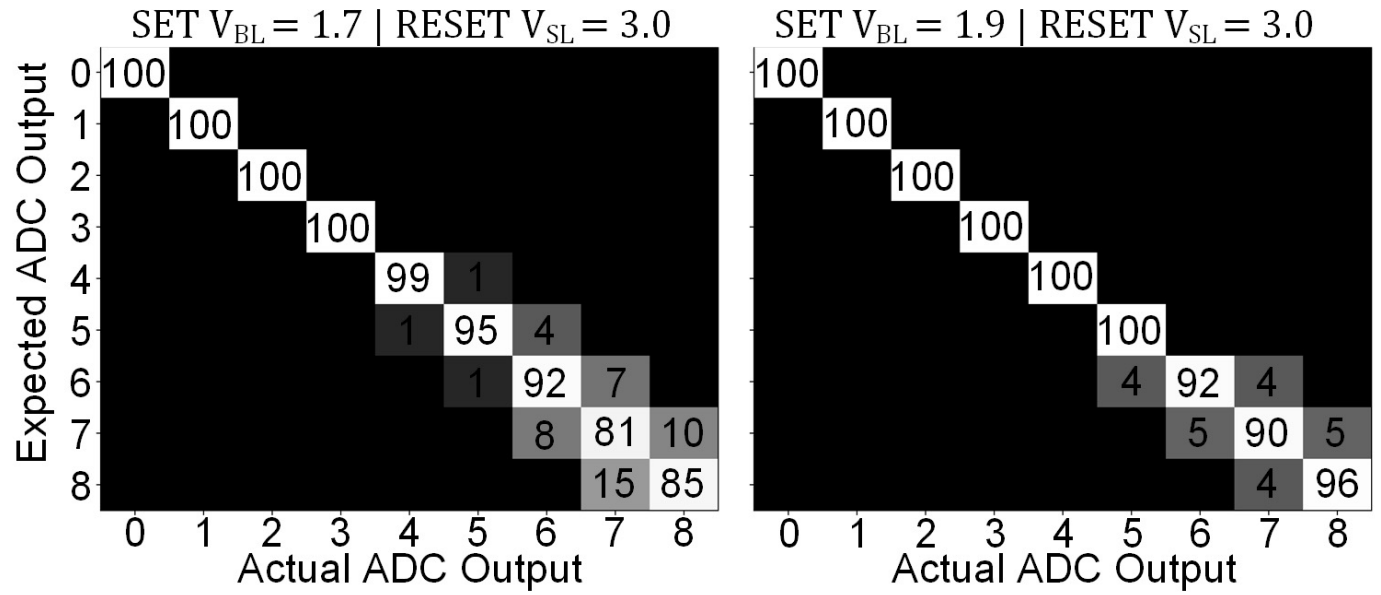
Measurements: BER

Experiment

- 8192 measurements
- Confusion matrix

Observations

- ↑Variation → ↑CIM Error Rate
- ↑LRS → ↑CIM Error Rate



Agenda

1. Motivation & Background
2. RRAM + CIM Measurement
3. CIM-SECDED
4. Successive Correction
5. Summary

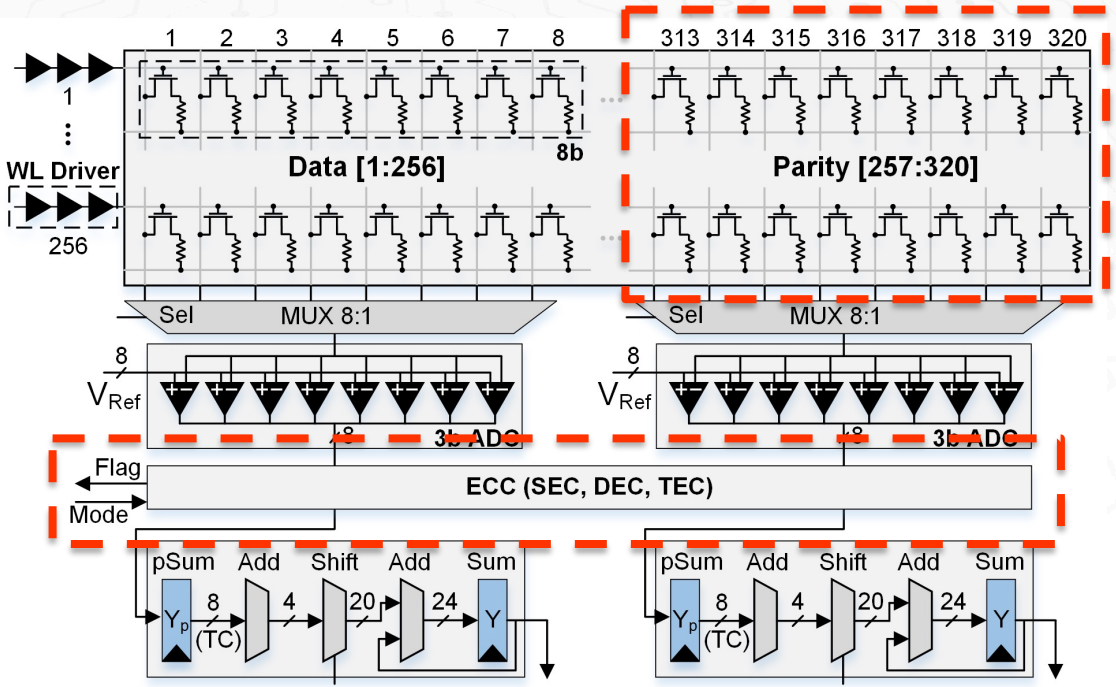
Macro Level Implementation

Specifications

- 256×256 RRAM Array
- 8 WL/cycle, 3b ADC
- Shift + Add logic (VMM)
- **ECC (SECDED)**

ECC

- [32, 8] SECDED code
  - 64 Check bits
- SECDED decoder

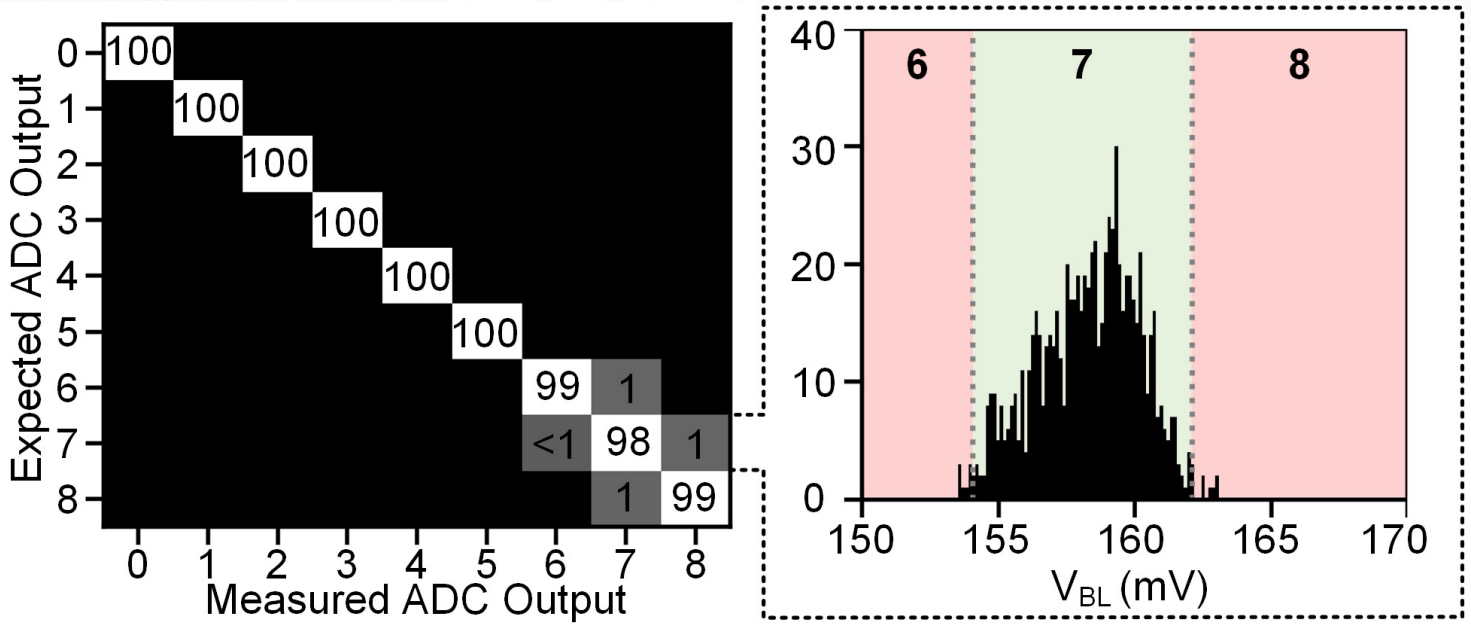


[1] A 40nm 64Kb 56.67TOPS/W Read-Disturb-Tolerant Compute- in-Memory/Digital RRAM Macro, Yoon et al, ISSCC 2021
[2] CIM-SECDED: A 40nm 64Kb Compute In-Memory RRAM Macro with ECC Enabling Reliable Operation, Crafton et al. ASSCC 2021

CIM-SECDED: ECC for CIM

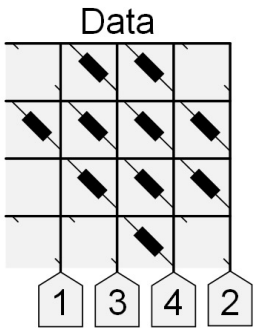
2 Key Observations

- Only ±1 errors from ADC
- Residue arithmetic



CIM-SECDED: ECC for CIM

Encoding

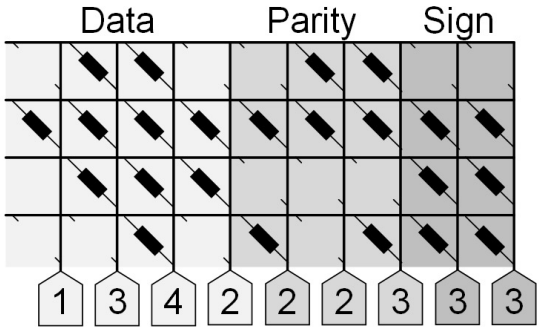


Detection & Localization

Correction

CIM-SECDED: ECC for CIM

Encoding



Sign

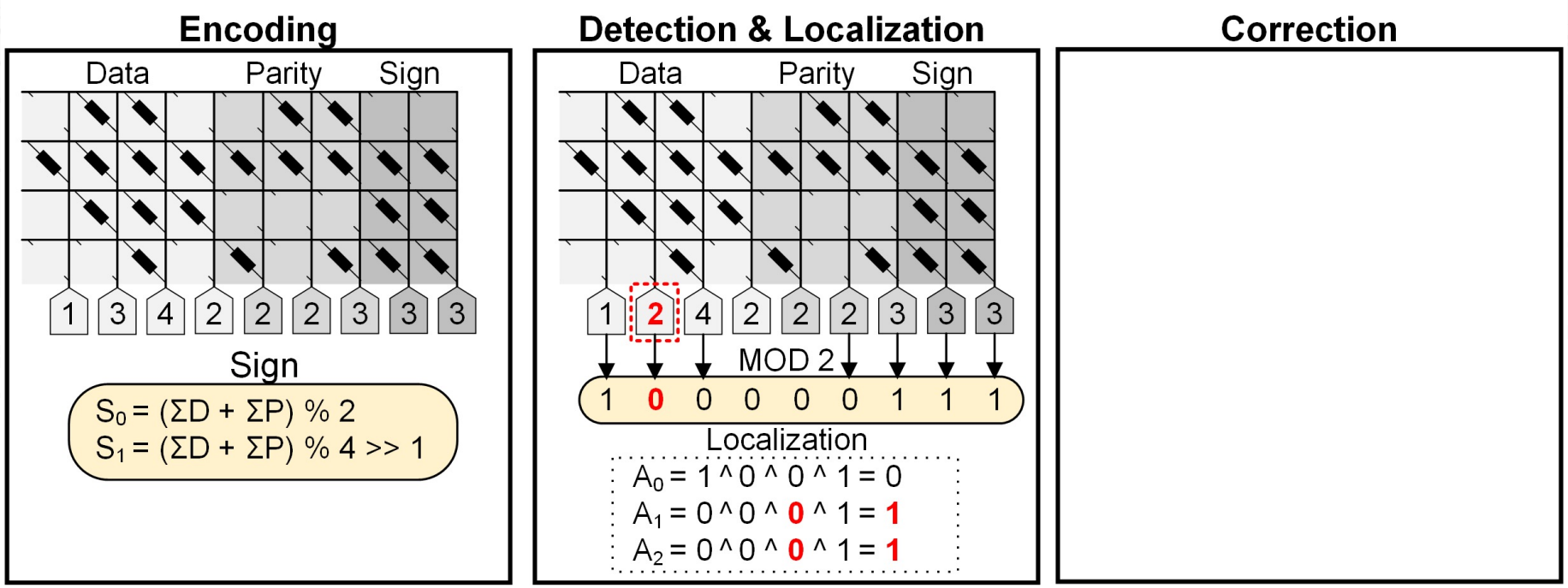
$S_0 = (\Sigma D + \Sigma P) \% 2$

$S_1 = (\Sigma D + \Sigma P) \% 4 >> 1$

Detection & Localization

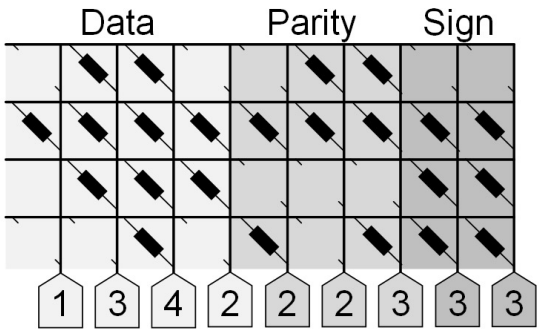
Correction

CIM-SECDED: ECC for CIM



CIM-SECDED: ECC for CIM

Encoding

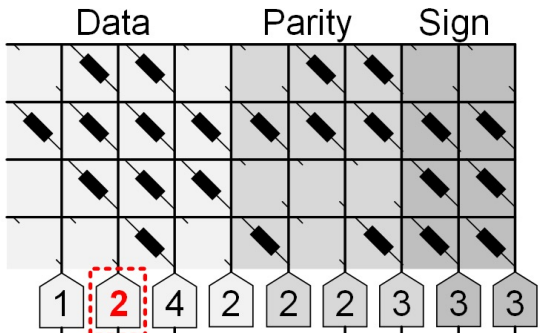


Sign

$S_0 = (\Sigma D + \Sigma P) \% 2$

$S_1 = (\Sigma D + \Sigma P) \% 4 >> 1$

Detection & Localization



Localization

$A_0 = 1 \wedge 0 \wedge 0 \wedge 1 = 0$

$A_1 = 0 \wedge 0 \wedge \mathbf{0} \wedge 1 = \mathbf{1}$

$A_2 = 0 \wedge 0 \wedge \mathbf{0} \wedge 1 = \mathbf{1}$

Correction

$Sum_{calc} = (\Sigma D + \Sigma P) \% 4$

$Sum_{orig} = (2 \cdot S_1 + S_0) \% 4$

$1+2+4+2 + 2+2+3 = 16 \% 4 = \mathbf{0}$

$2 \cdot 3 + 3 = 9 \% 4 = \mathbf{1}$

Sum_{orig}

Sum_{calc}	0	1	2	3
0	0	-1	0	+1
1	+1	0	-1	0
2	0	+1	0	-1
3	-1	0	+1	0

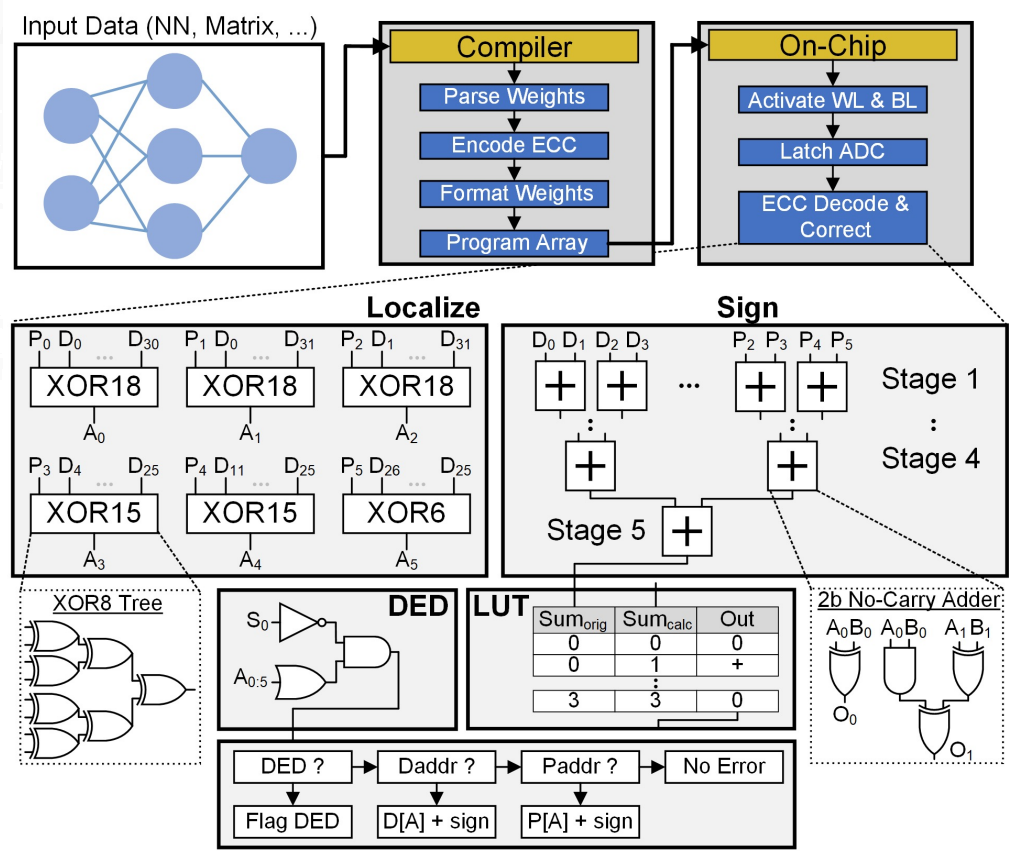
CIM-SECDED Decoder

Implementation:

- Encoder → Compiler
- Decoder → Digital logic

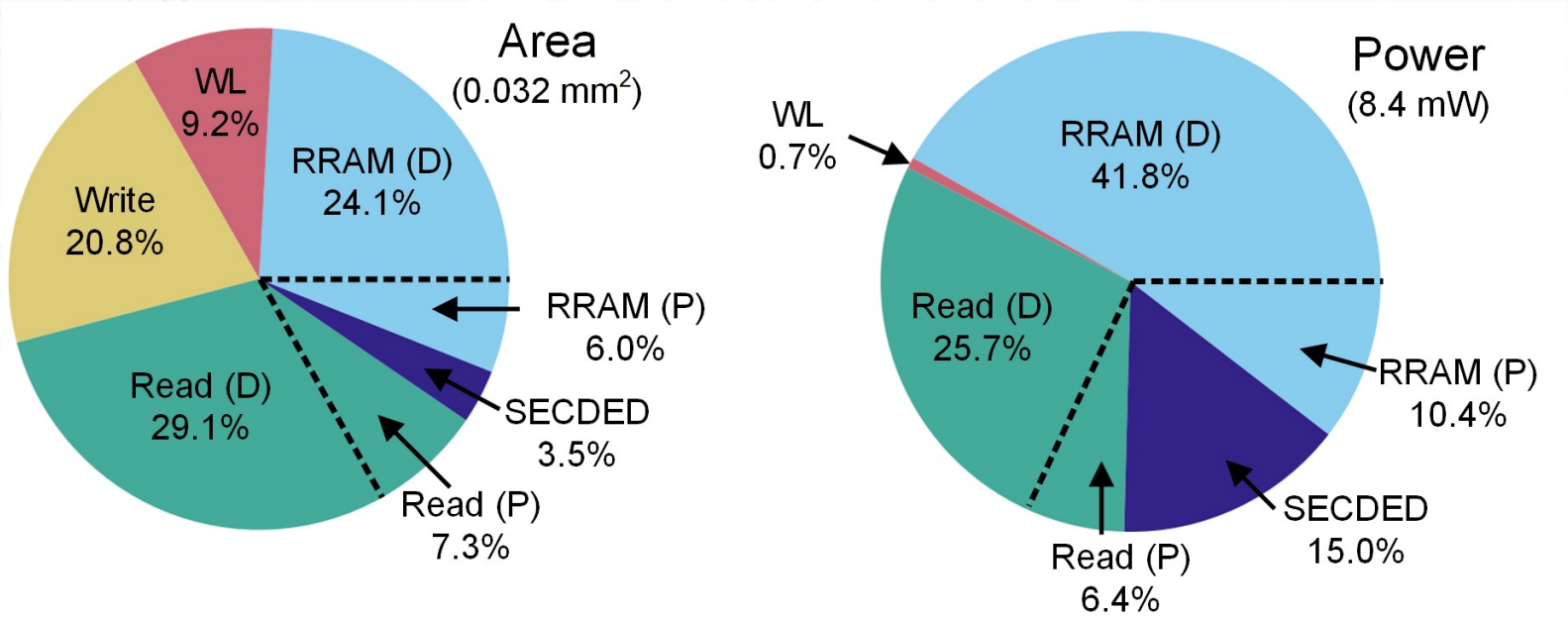
Architecture:

- XOR Tree + DED (SECDED)
- Adder Tree + LUT (NEW)



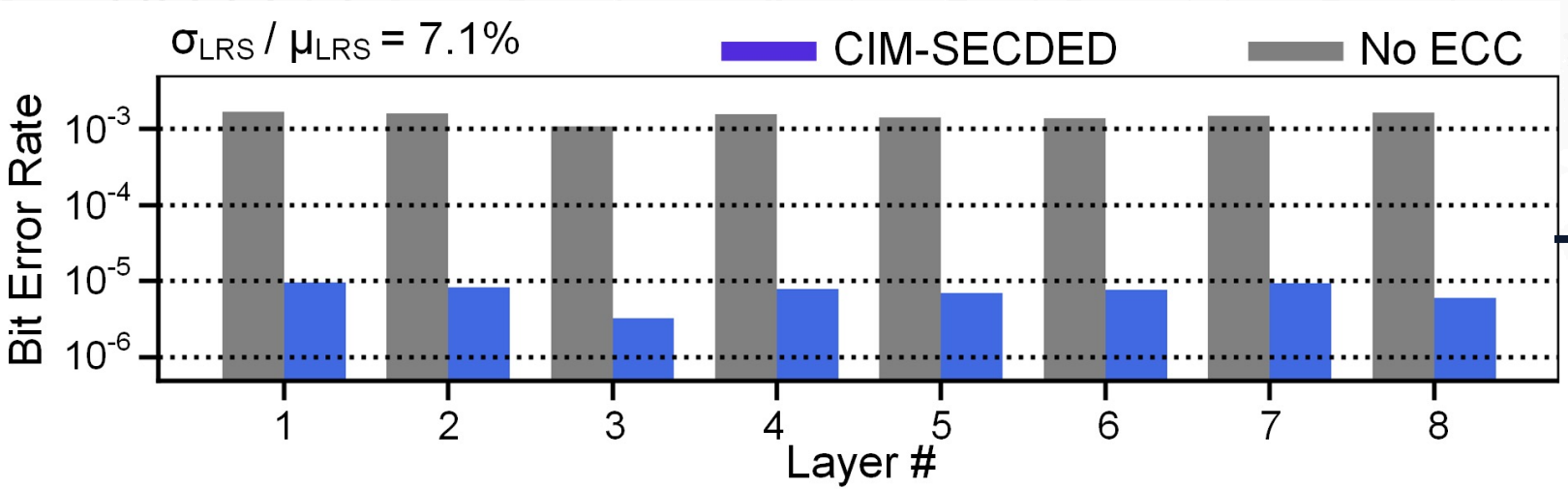
CIM-SECDED Overhead

- 1 Extra parity bit (32/8 vs 32/7)
- 16.8% Area overhead
- 31.8% Power Overhead



CIM-SECDED Results

- 100× reduction in BER
- 3.9% and 16.3% accuracy improvement



Task		Variation (%)		Accuracy Loss (%)	
Dataset	Network	WR Voltage	σ_{LRS} / μ_{LRS}	No ECC	ECC
ImageNet	ResNet18	$V_{BL} = 1.9$	3.7%	3.9%	0%
		$V_{BL} = 1.7$	7.1%	16.7%	0.4%

Agenda

1. Motivation & Background
2. RRAM + CIM Measurement
3. CIM-SECDED
4. Successive Correction
5. Summary

Can We Do Better ?

Observation:

- CIM-SECDED: $10^{-3} \rightarrow 10^{-6}$
- SRAM: 10^{-15}
- DNNs $\rightarrow 10^{-5}$ [1]

Memory	BER
SRAM	1e-15
CIM	1e-3

Experiment:

- DEC $\rightarrow 10^{-9}$
- TEC $\rightarrow 10^{-12}$

ECC	BER
CIM + SEC	$(1e\text{-}3)^2 = 1e\text{-}6$
CIM + DEC	$(1e\text{-}3)^3 = 1e\text{-}9$
CIM + TEC	$(1e\text{-}3)^4 = 1e\text{-}12$

[1] Ares: A framework for quantifying the resilience of deep neural networks, DAC 2018, Reagen et al.

Successive Correction → Triple Error Correction

Observation:

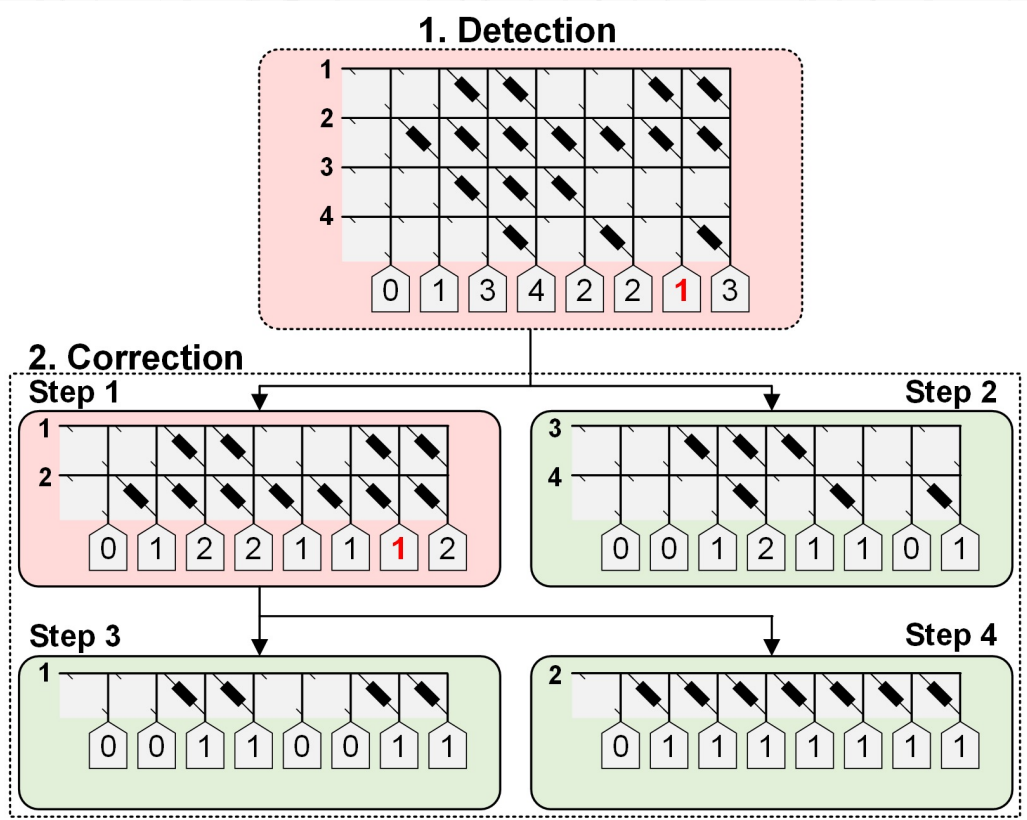
1. ↓WL → ↓BER
2. Can detect 2 errors
   - SEC**DED**

Idea:

- Read **4** WL
- Detect error ?
- Read **2** WL

Result:

- DED → DEC



[1] Improving compute in-memory ecc reliability with successive correction. Crafton et al. DAC 2022.

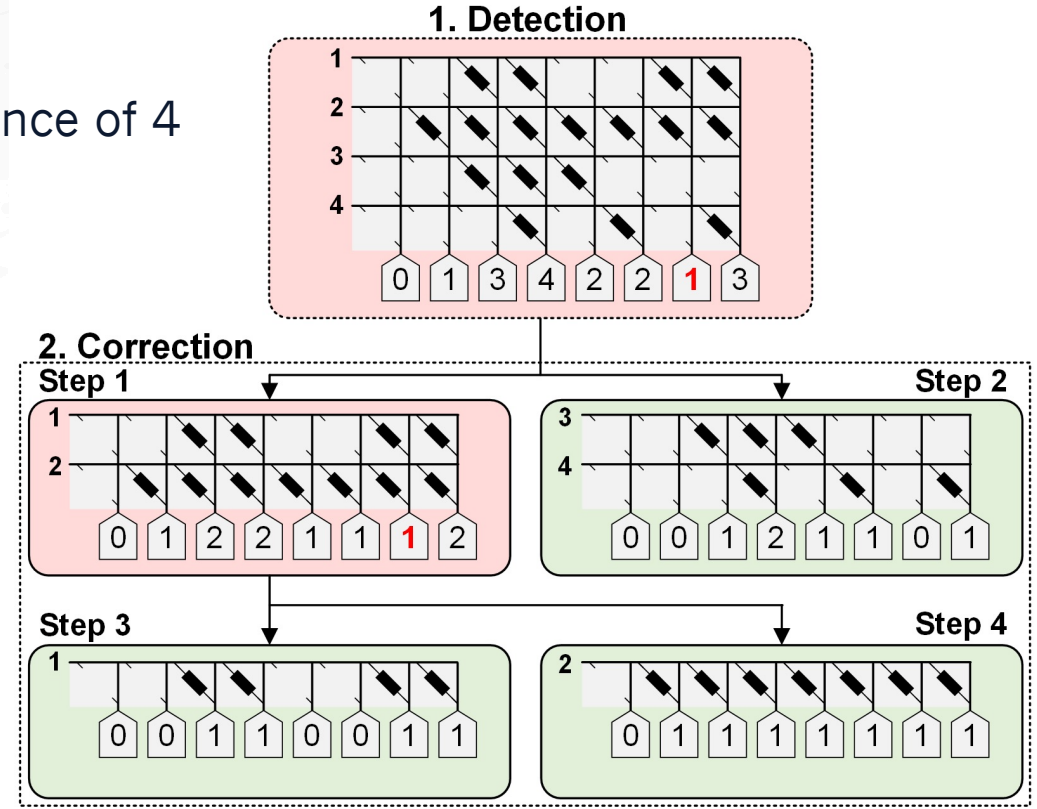
Successive Correction → Triple Error Correction

Observation:

- SECDED → Hamming distance of 4
  - SECDED
  - **TED**

Result:

- TED → TEC

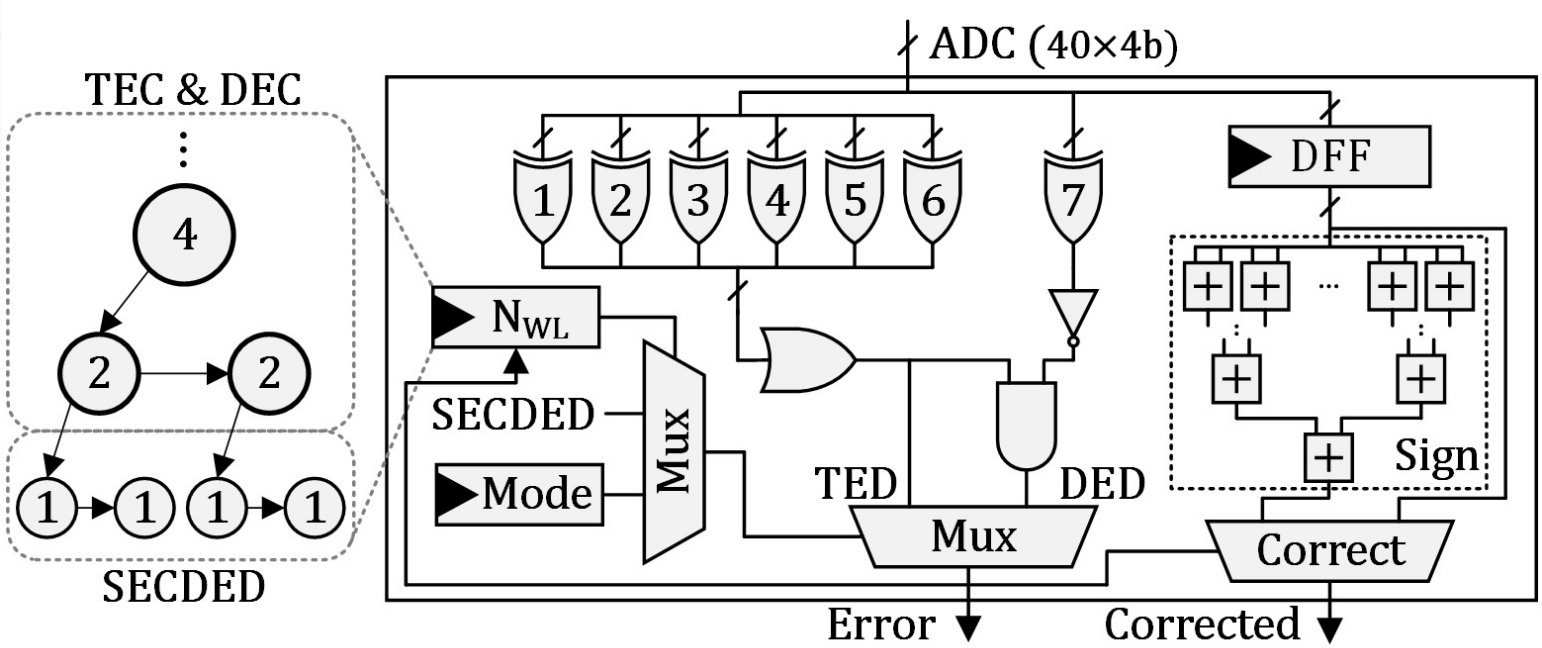


[1] Improving compute in-memory ecc reliability with successive correction. Crafton et al. DAC 2022.

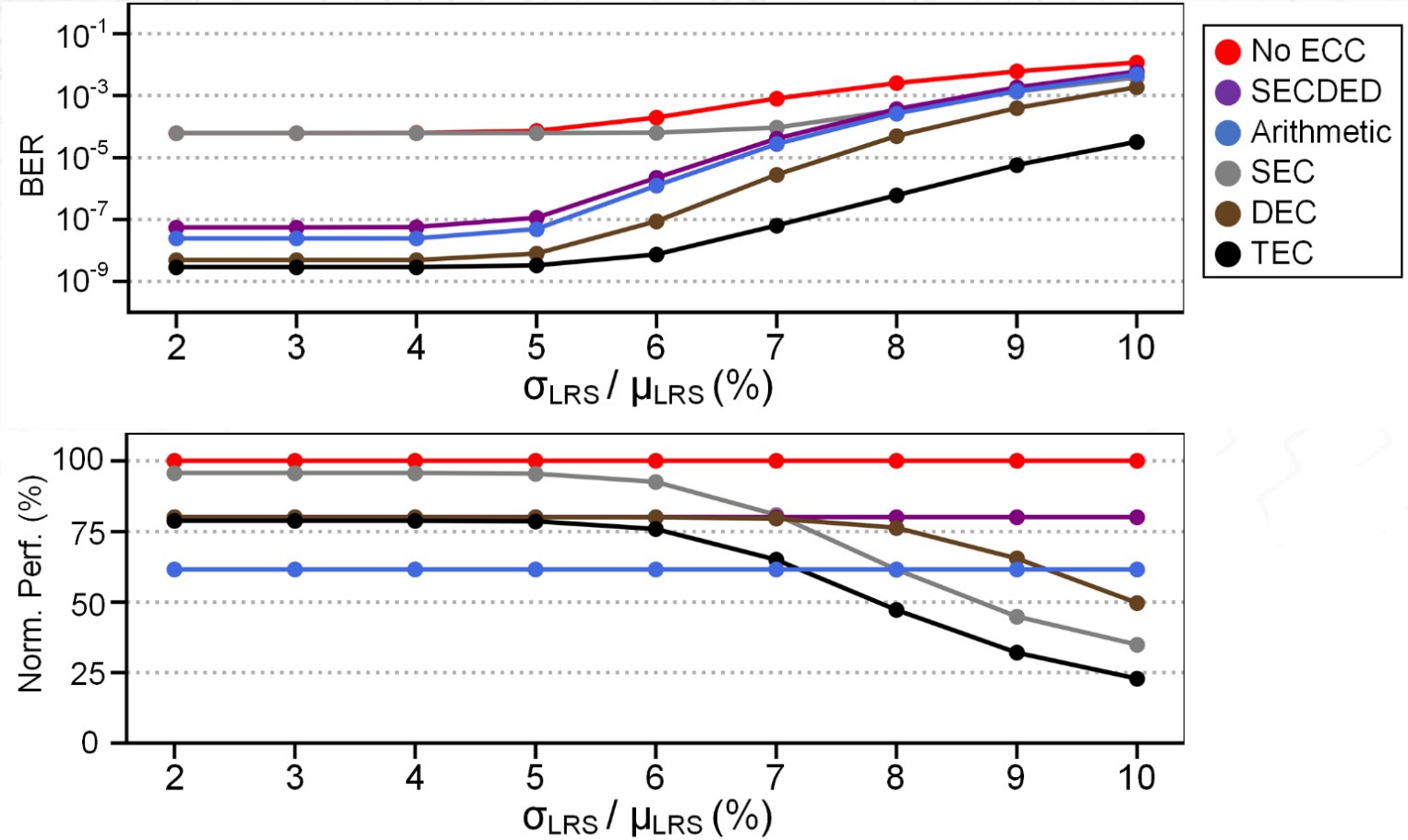
Successive Correction → Triple Error Correction

Implementation:

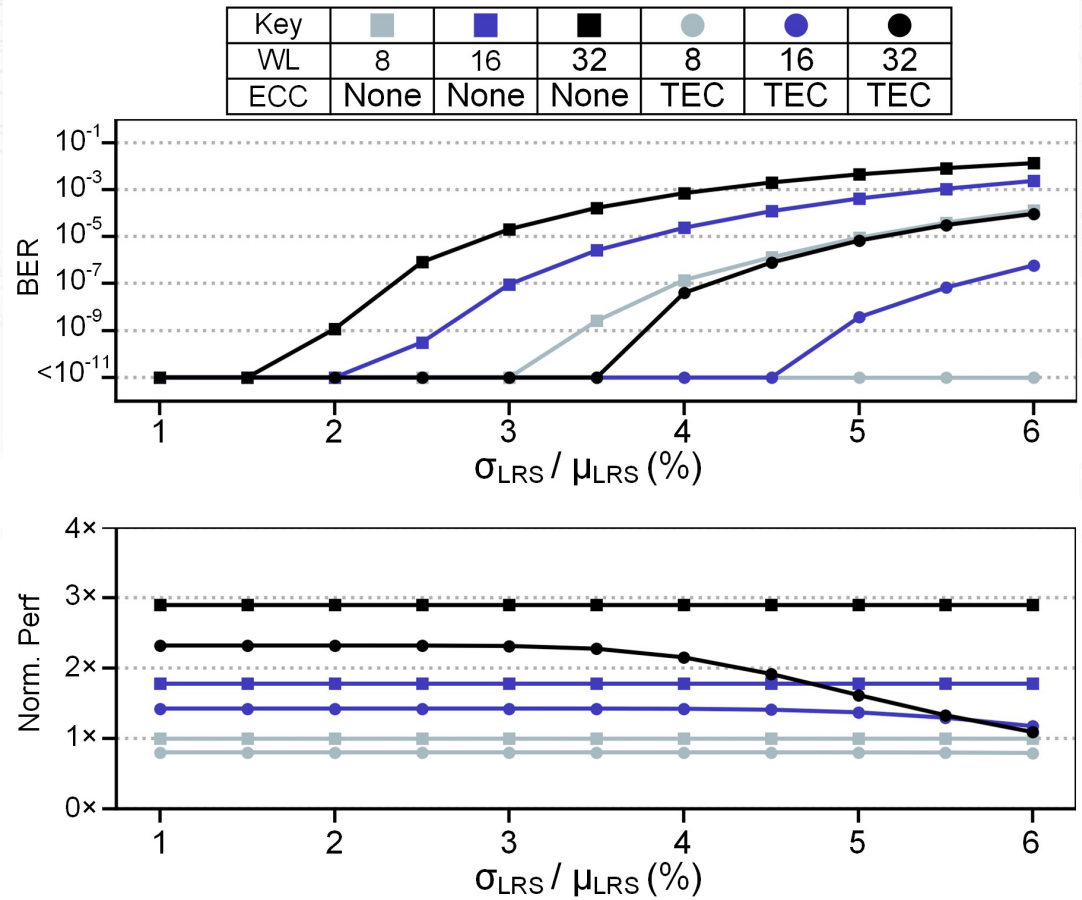
- DEC → Exact same as SECDED
- TEC → 0.1% Area (10um^2)



Successive Correction → Triple Error Correction



Successive Correction → Triple Error Correction



Agenda

1. Motivation & Background
2. RRAM + CIM Measurement
3. CIM-SECDED
4. Successive Correction
5. Summary

Summary

- CIM: ↑WL → ↑Noise + ↓Voltage Range → ↑BER
- Detect error → ↓WL → ↓BER
- >16,000× improvement in BER over No ECC
- 636× ↓ BER @ 5.7% ↓ performance over SOTA

ECC	BER
No ECC	1e-3
SEC	$(1e\text{-}3)^2 = 1e\text{-}6$
DEC	$(1e\text{-}3)^3 = 1e\text{-}9$
TEC	$(1e\text{-}3)^4 = 1e\text{-}12$