IEEE Advancing Technology for Humanity

CAS

Fundamental Limits on the Computational Accuracy of Resistive Crossbar-based In-memory Architectures

Saion K. Roy[1], Ameya Patil[2], and Naresh R. Shanbhag[1]

1: University of Illinois at Urbana-Champaign
2: Amazon Lab126

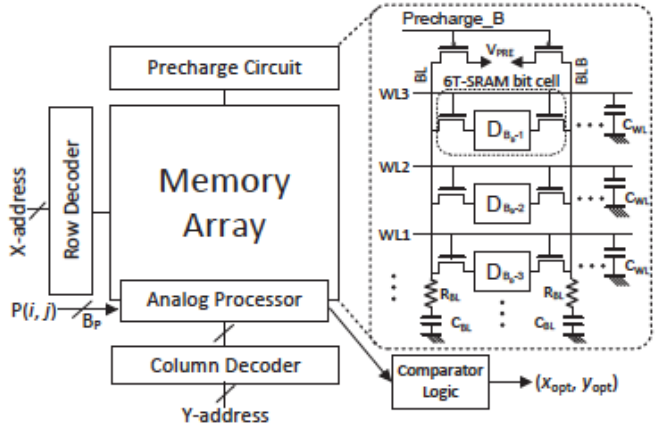
2022 IEEE International Symposium on Circuits and Systems
May 28- June 1, 2022 Hybrid Conference

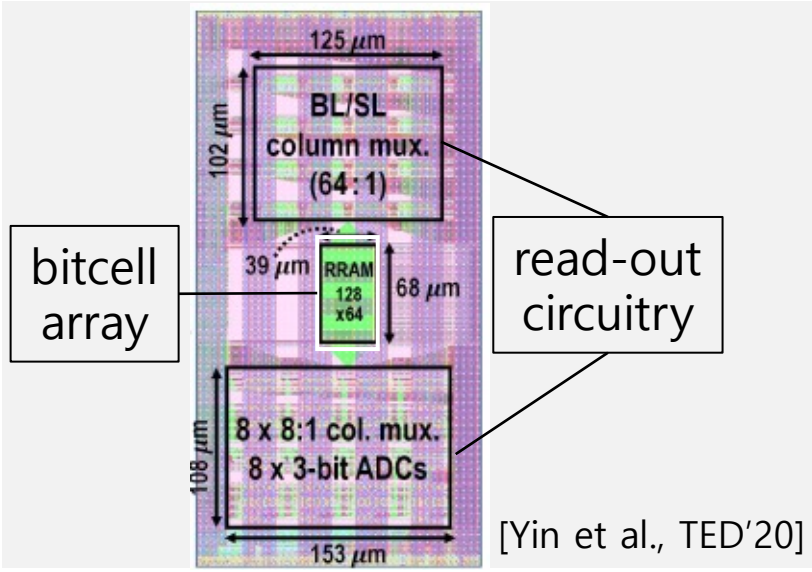
Outline

- Introduction
- Resistive Crossbar Architecture
- Behavioral Modeling
- Simulation Results
  - Model validation
  - Compute SNR analysis for MRAM, ReRAM, and FeFET crossbars
  - System level accuracy of ResNet-20 on CIFAR-10
- Conclusion

In-memory Computing (IMC)

compute memory





[Yin et al., TED'20]

first IMC concept paper (ICASSP 2014)

AN ENERGY-EFFICIENT VLSI ARCHITECTURE FOR PATTERN RECOGNITION VIA DEEP EMBEDDING OF COMPUTATION IN SRAM

Mingu Kang, Min-Sun Keel*, Naresh R. Shanbhag*, Sean Eilert†, and Ken Curewitz†*

*Dept. Electrical and Computer Engineering, University of Illinois at Urbana-Champaign
†Micron Technology, Inc

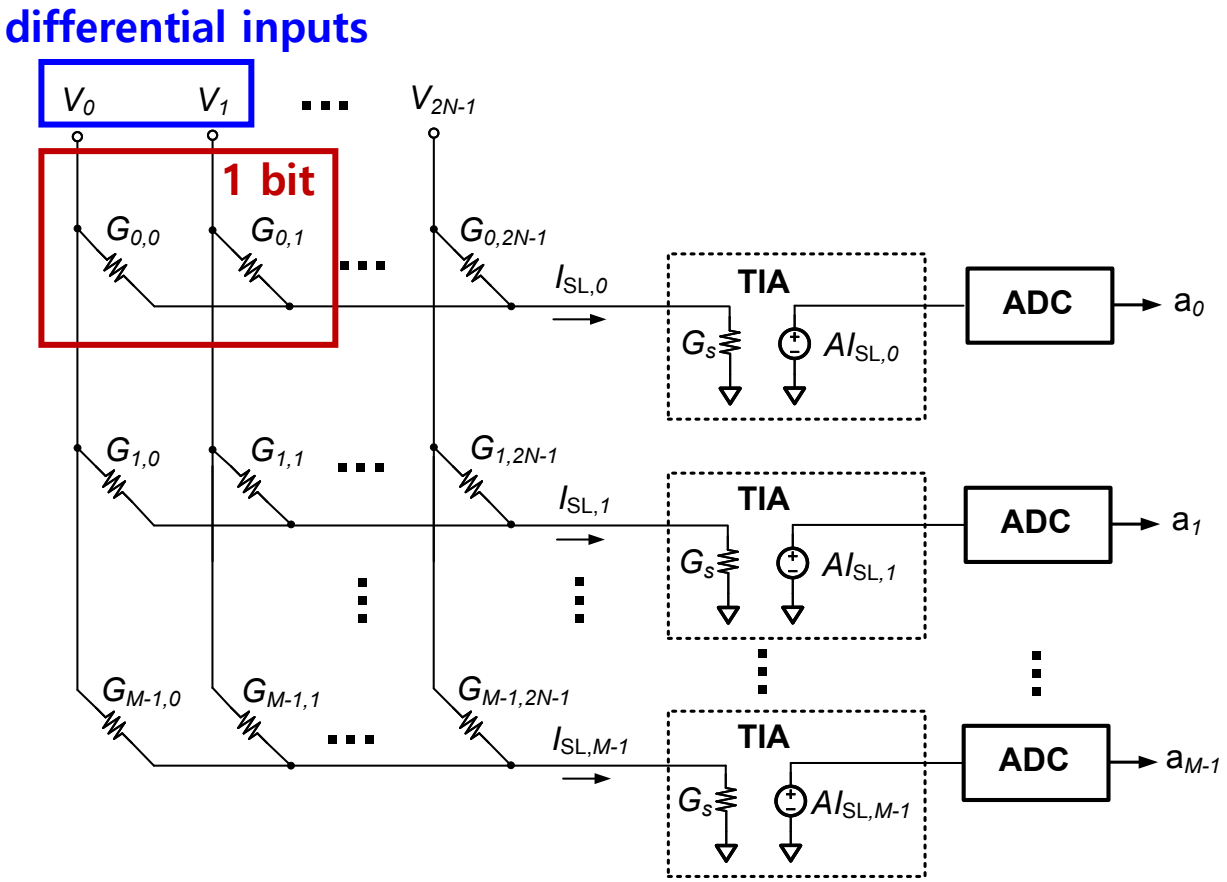
- computes a $M \times N$ matrix-vector multiply (MVM)
- SRAM-based IMC banks are mature → $20\times$ lower energy + $9\times$ higher compute density than digital[1]

- eNVM-based (MRAM/ReRAM) IMCs have potential for high compute density but lags digital due to low compute SNR → this work explains why

[1]N. R. Shanbhag and S. K. Roy, "Comprehending In-memory Computing Trends via Proper Benchmarking," CICC 2022 (invited)

Resistive Crossbar Architecture

voltage-drive current-sensing crossbar



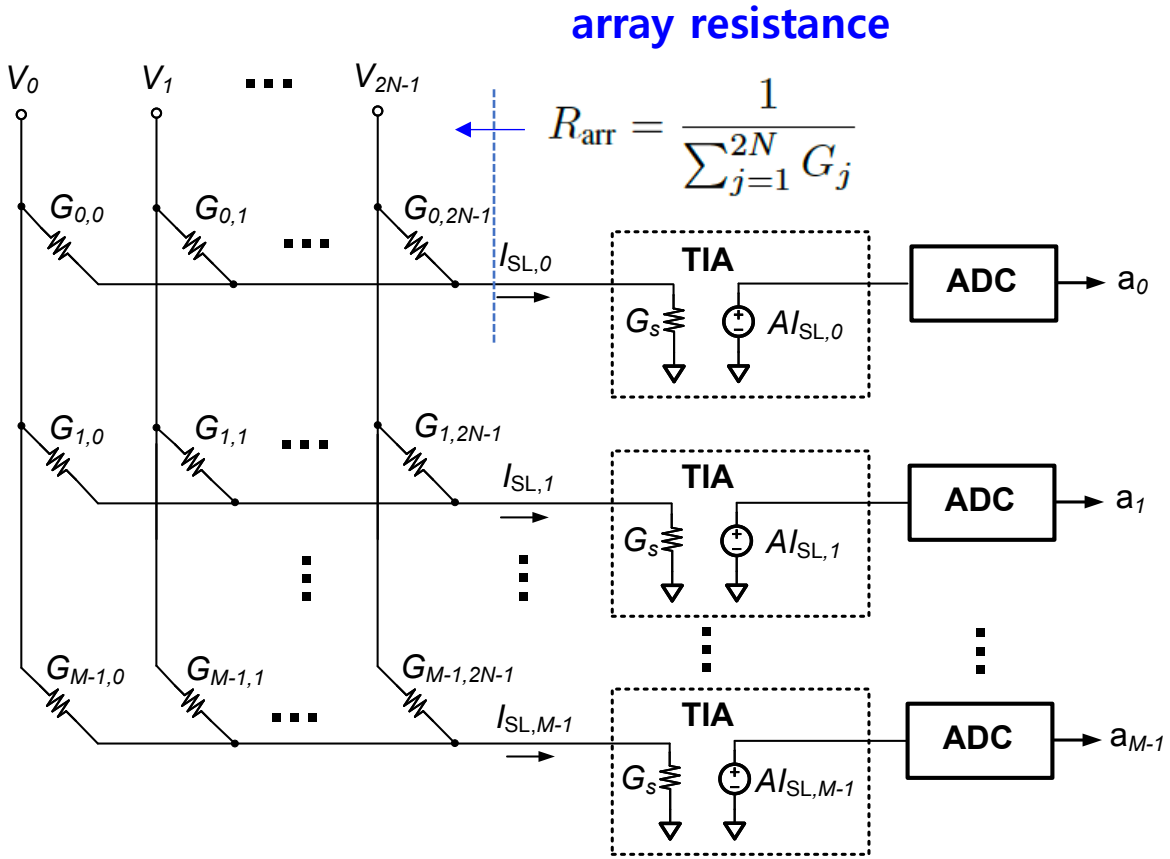
[A. Patil et al., ISCAS 2019]

- computes a $M \times N$ matrix-vector multiply (MVM)
- V-DACs provide differential inputs on BLs ($V_{2k} = -V_{2k-1}$)
- two BCs (G_{2k-1}, G_{2k}) store 1 bit
- current summing & sensing on SLs
- device ***resistive contrast***

$$\rho = \frac{R_{\text{off}}}{R_{\text{on}}}$$

2 (MRAM); 12 (ReRAM); 10^3 (FeFET);

Behavioral Modeling



- signal current in SL

S_I (current scaling factor)

$$I_{\text{sig}} = \left[\frac{R_{\text{arr}}}{R_{\text{arr}} + R_s}\right] \underbrace{\left(\sum_{k=1}^{N} V_{2k}\Delta G_{2k}\right)}_{I_{\text{ideal}}} = S_I I_{\text{ideal}}$$

- total current in SL

quantization noise; DAC mismatch

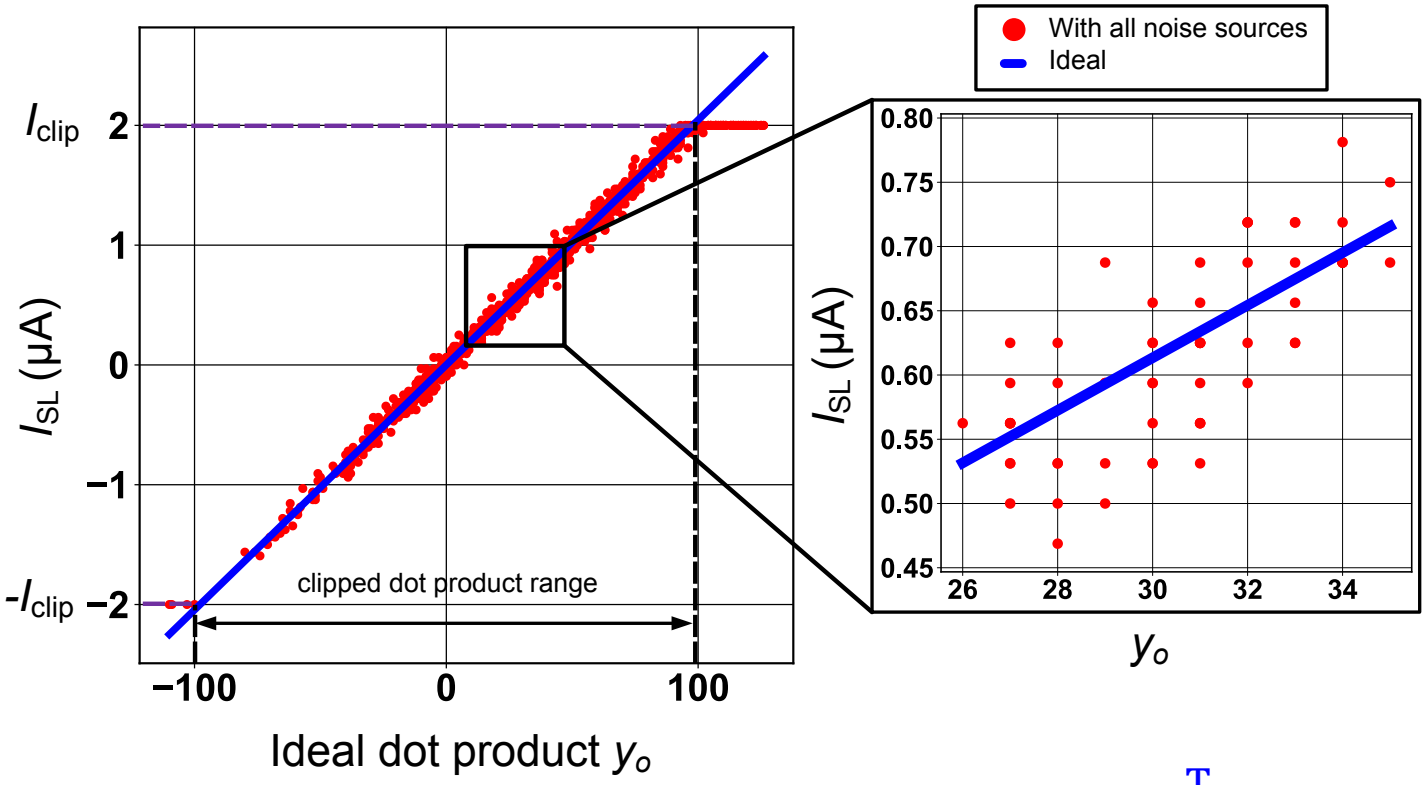
$$I_{\text{SL}} = I_{\text{sig}} + I_{\text{nb}} + I_{\text{nd}} + I_{\text{nc}} + I_{\text{nq}},$$

conductance variation; clipping noise

Results – Model Validation

SPICE simulations in a 22nm process

$N = 512, R_s = 316\Omega$, 6b ADC

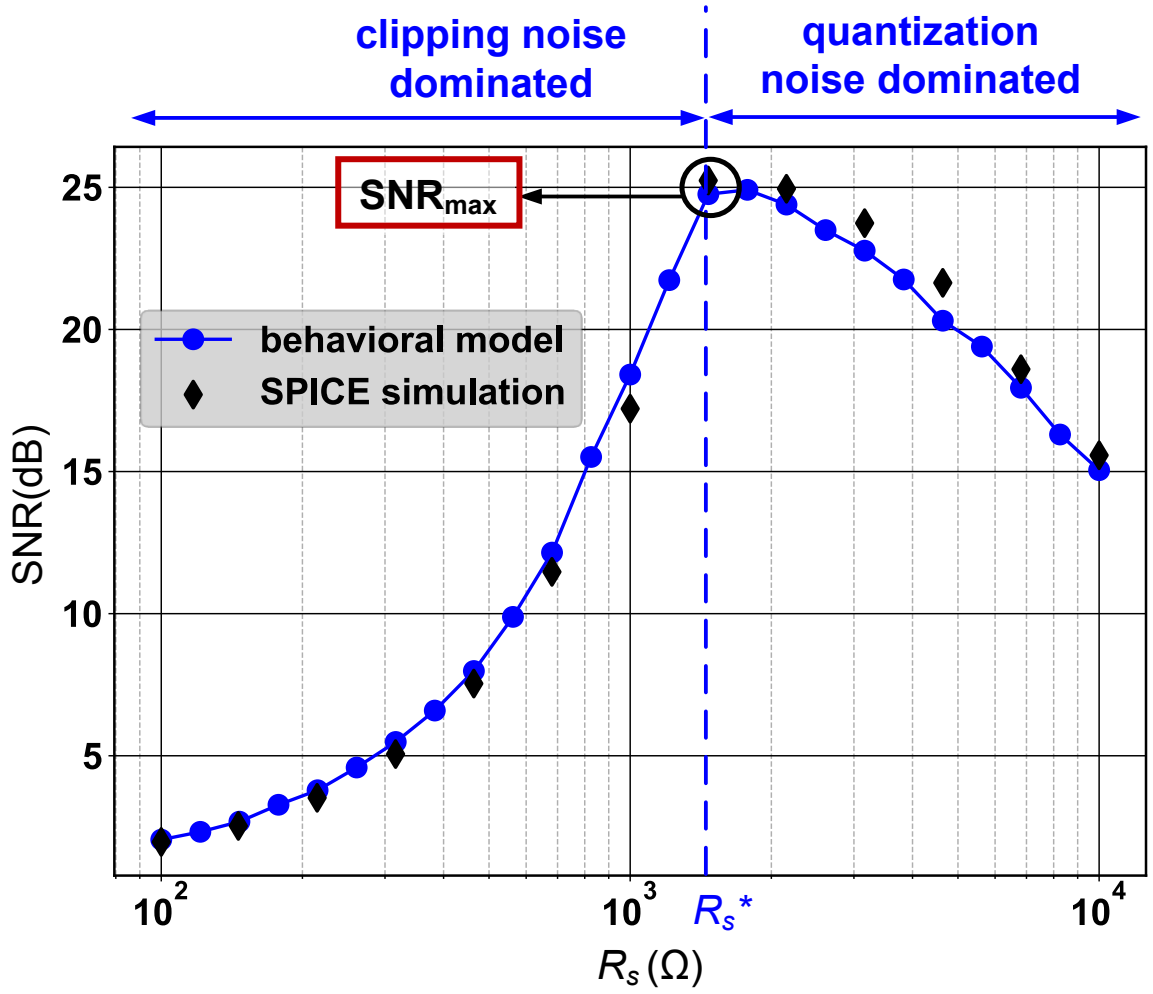


$$y_o = \mathbf{w}^T\mathbf{x}$$

- DAC input: signed 5b with $V_{lsb} = 3\text{mV}$
- DAC mismatch: 4%
- Conductance variation: 4%
- ADC clipping range: $[-2\mu A, +2\mu A]$

- SL current varies due to analog non-idealities

Results – Compute SNR Analysis



TIA input (sensing) resistance

$$\mathrm{SNR} = \frac{\mathbb{E}[I_{\mathrm{sig}}^2]}{\mathbb{E}[I_{\mathrm{nb}}^2] + \mathbb{E}[I_{\mathrm{nd}}^2] + \mathbb{E}[I_{\mathrm{nc}}^2] + \mathbb{E}[I_{\mathrm{nq}}^2]},$$

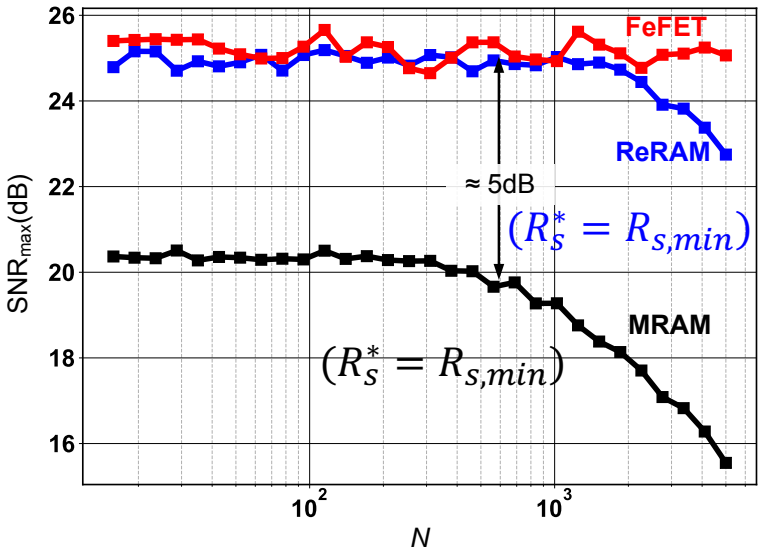
- model and simulations match → further validates model
- ADC clipping vs. quantization noise trade-off:

$$S_I = \left[\frac{R_{\mathrm{arr}}}{R_{\mathrm{arr}} + R_s}\right]$$

- compute SNR maximized if $R_s = R_s^*$ → clipping noise & quantization noise are equal

Results – SNR Dependence

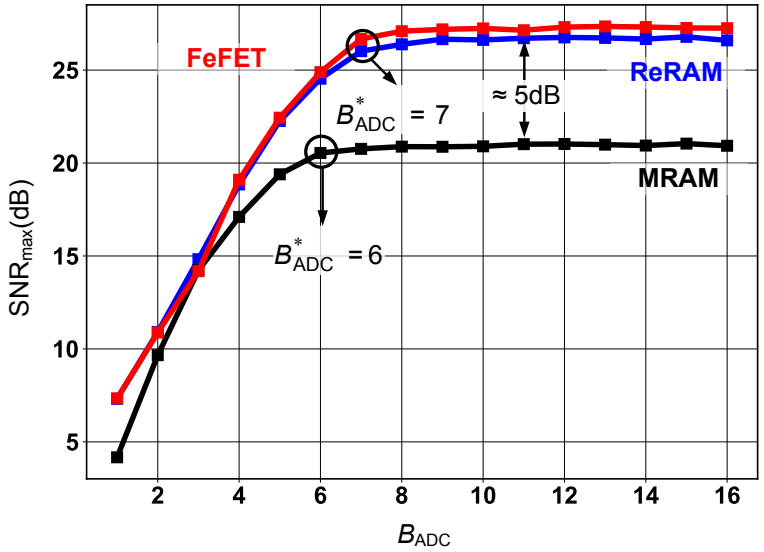
$[B_{ADC} = 6]$ $(R_s^* = 10k\Omega)$



dot product dimension

- SNR$_{max}$ roll-off with higher DP dimension → small R_{arr} → small $R_s = R_{s,\text{min}}(= 1k\Omega) \neq R_s^*$
- higher absolute R_{on}, R_{off} critical for high DP dimension

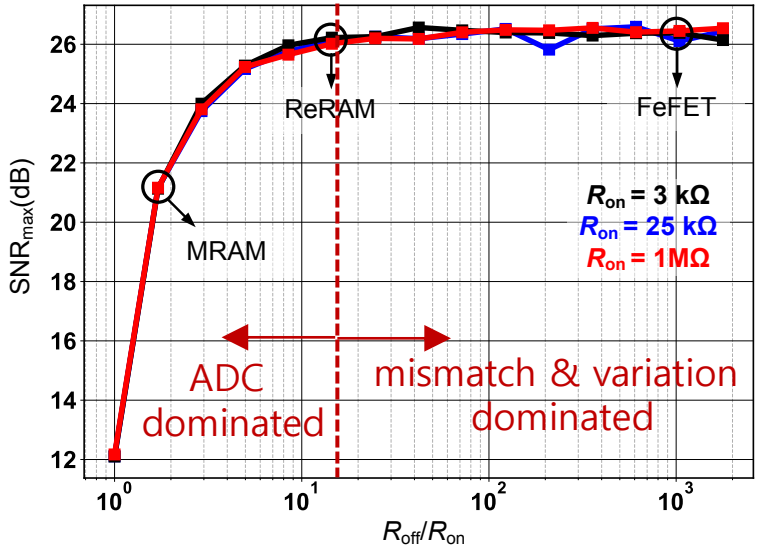
$[N = 512]$



ADC precision

- minimum ADC precision B_{ADC}^* increases with SNR$_{max}$
- SNR$_{max}$ saturates for $B_{ADC} > B_{ADC}^*$ → DAC mismatch and G variations dominate

$[B_{ADC} = 7, N = 512]$

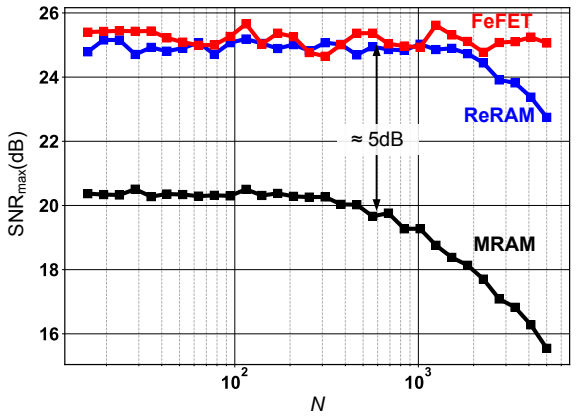


device resistive contrast

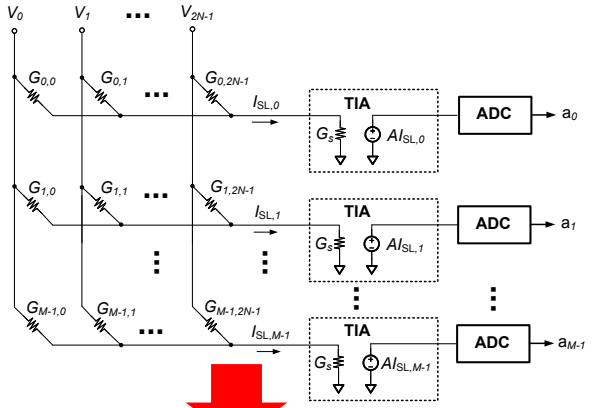
- SNR$_{max}$ improves with higher resistive contrast but..
- increasing device resistive contrast beyond (~12-to-15) is futile

System Level Accuracy Prediction Set-up

Proposed SNR analysis



SNR maximizing Crossbar Designs

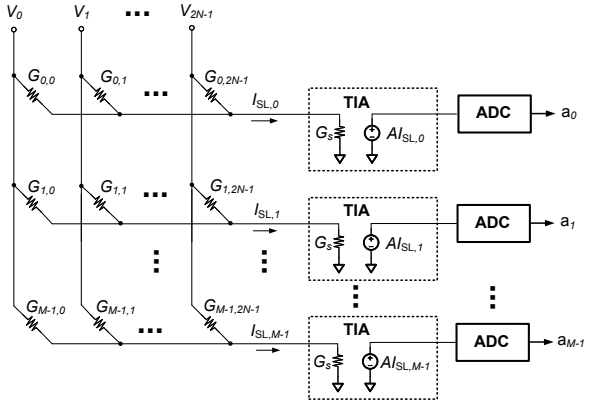


Crossbar Design Parameters

$R_s, B_{\mathrm{ADC}}, N, M, V_{\mathrm{lsb}}$

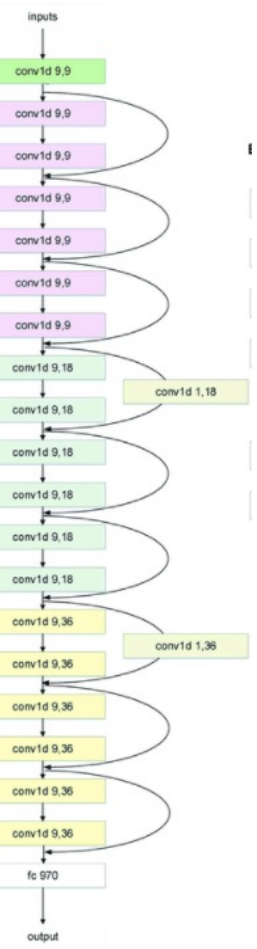
network accuracy comparison

Exhaustive search

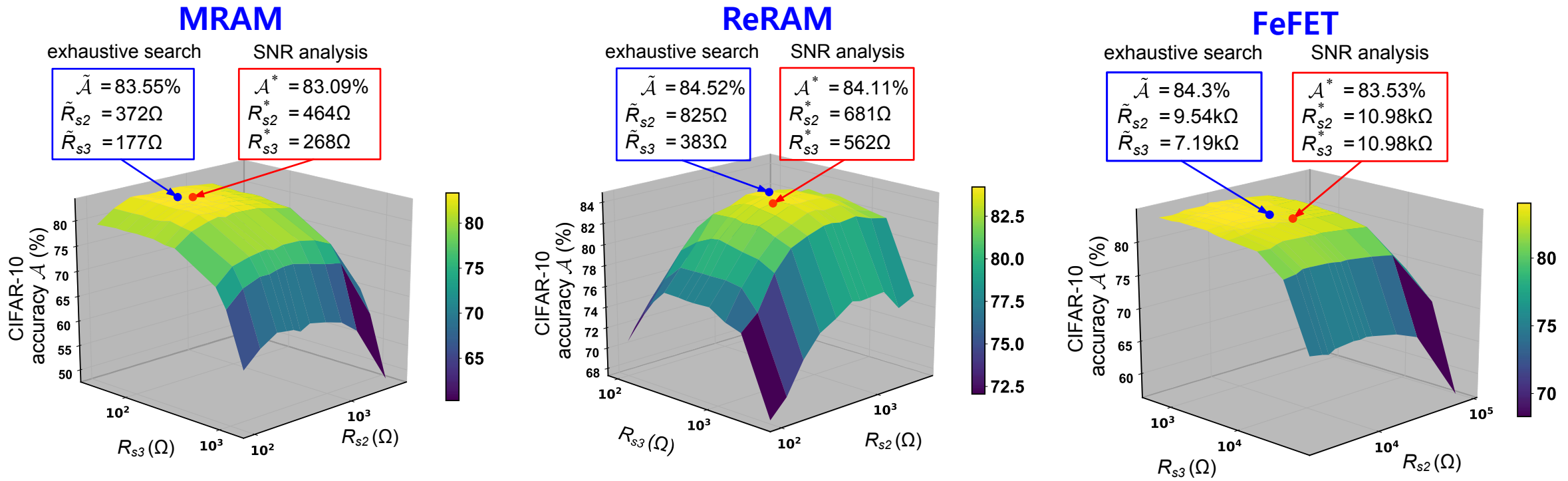


DNN mapping

ResNet-20 on CIFAR-10



Results – System Level Accuracy Prediction



- Baseline ResNET-20 on CIFAR-10: 5b input, ternary weights, accuracy = 84.94% (3-layer network)
- 3 Crossbars: $N = 144, 288, 576$; $R_{s1} = R^*_{s1}$; sweep (R_{s2}, R_{s3})
- SNR analysis predicted crossbar design achieves system-level accuracy to within 1% (exhaustive search) to within 2% of digital baseline value (84.94%)
- bank-level SNR is a good proxy for network level accuracy → SNR analysis bypasses trial & error

Conclusion

- proposed an analytical framework to obtain SNR-optimal resistive crossbar parameters → avoids expensive trial and error
- insights provided by the framework:
  - SNR-optimal sensing resistance R_S^* exists which equalizes the clipping and quantization noise in the column ADCs
  - system level inference accuracy is maximized when bank-level compute SNR is maximized
  - increasing device resistive contrast improves SNR up to a point (~12-15). Diminishing returns due to mismatch (input DACs) and variations (device conductance)
- proposed framework can be extended to other resistive IMC and devices

Thank You

saionkr2@Illinois.edu

Acknowledgement

Work supported by the Defense Advanced Research Agency (DARPA) and the Semiconductor Research Corporation (SRC) via the FRANC Program, and the Center for Brain-inspired Computing (C-BRIC).