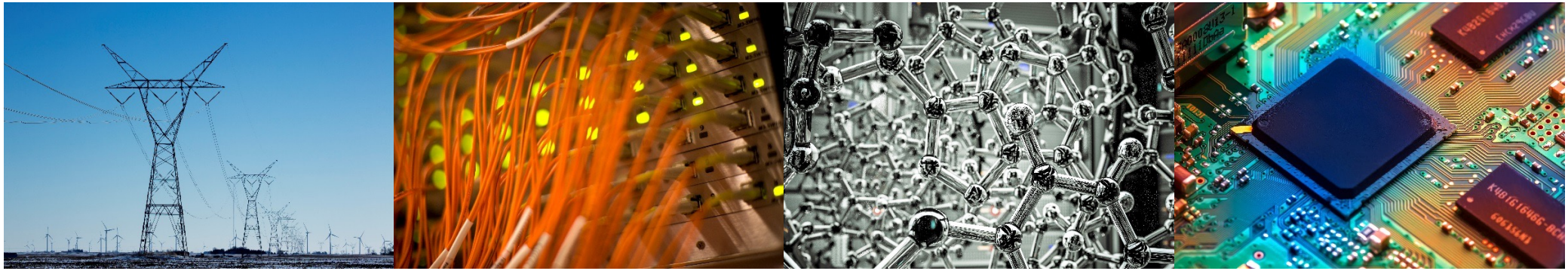
Energy-Accuracy-Security Trade-offs in Resistive In-memory Computing Architectures

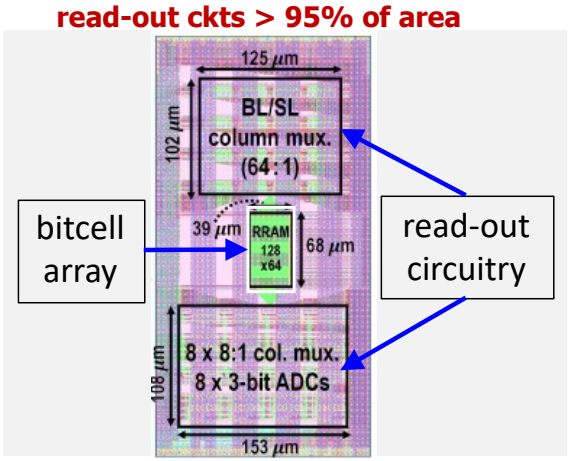


Saion Roy[1,2] and Naresh Shanbhag[1]
[1]University of Illinois at Urbana-Champaign, USA
[2]Northeastern University, USA

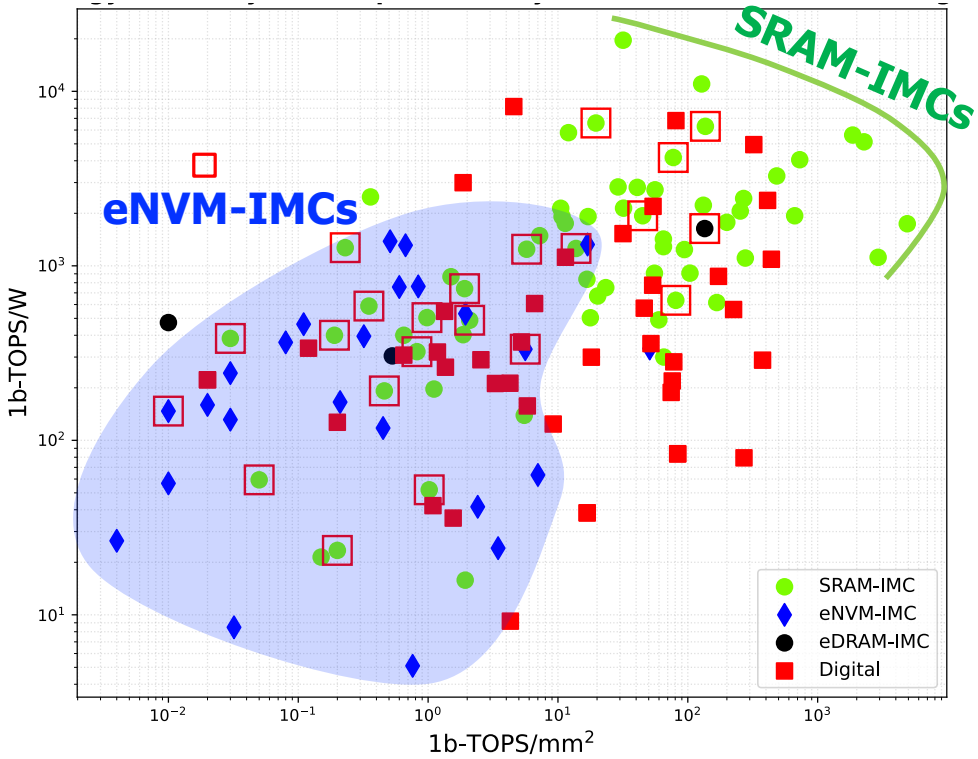
ILLINOIS
Electrical & Computer Engineering
COLLEGE OF ENGINEERING

Resistive IMCs in the Landscape of AI Accelerators

- eNVM-based IMCs lagging in energy efficiency & compute density (why?)
- Reason - low array-level compute accuracy



[Yin et al., TED'20]

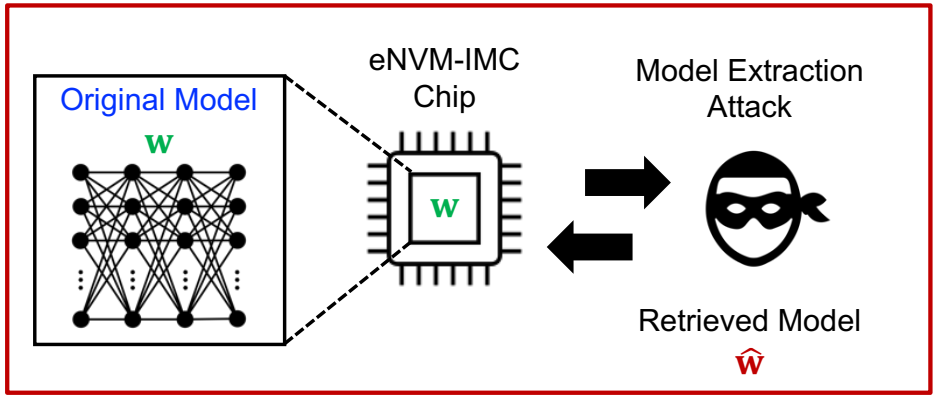


https://github.com/UIUC-IMC/UIUC-IMC-benchmarking

Are eNVM-based IMCs Secure?

Low compute SNDR (Bug) → potential resilience to security attacks (Feature)?

Model Extraction Attacks (MEAs)



leakage of private training data
&
adversarial attacks

Security vulnerability of eNVM IMCs unknown?

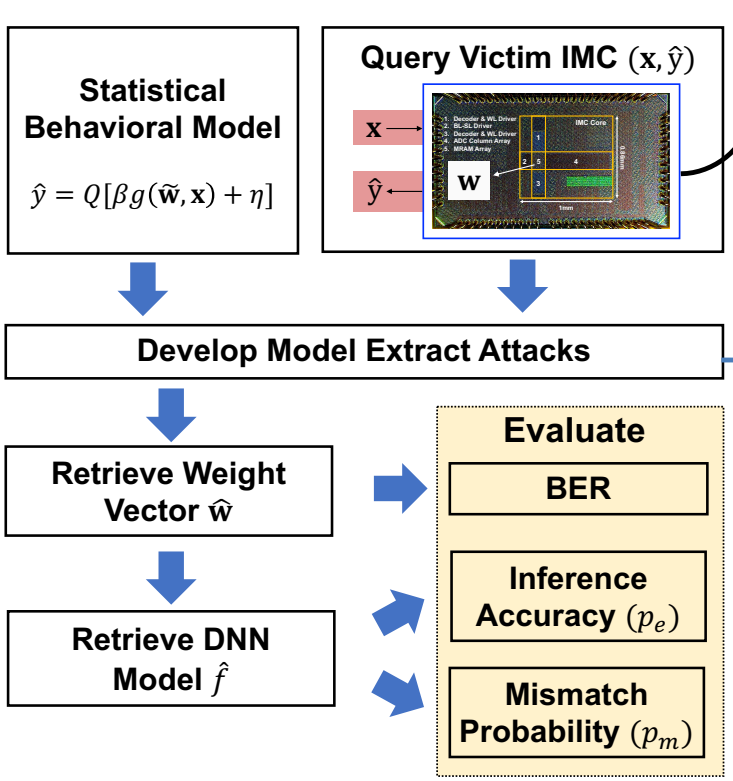
[Roy & Shanbhag, "On the Security Vulnerabilities of MRAM-based In-Memory Computing Architectures against Model Extraction Attacks," *ICCAD,* 2024]
[Roy & Shanbhag, "Energy-Accuracy-Security Trade-off in Resistive In-memory Architectures," *IEEE IEDM,* 2024]

Proposed MEA Construction Framework

statistical model

$$\hat{y} = Q[\beta g(\tilde{\mathbf{w}}, \mathbf{x}) + \eta]$$

- Conductance Variations
- Parasitics Conductance
- Mirroring Mismatch
- ADC Thermal Noise
- ADC Quantization

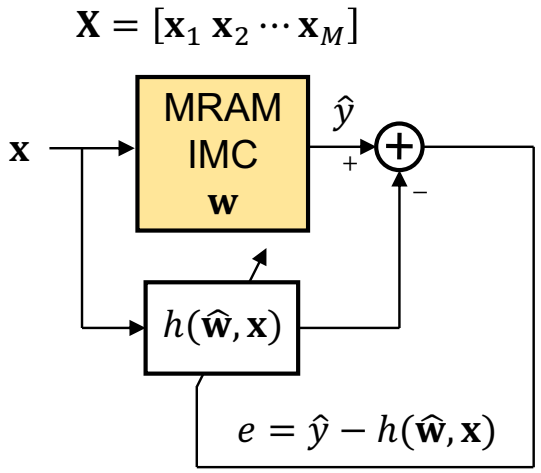


MRAM IMC Prototype
(22nm FD-SOI)
[Roy et al., JSSC'24] with UIUC, Princeton, GF and Raytheon

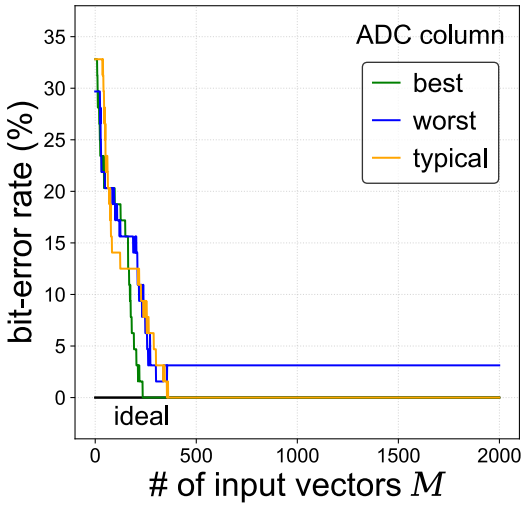
- Linear attacks:
  - basis vectors
  - least squares
- Non-linear SGD attack

Proposed SGD Attack for MEA

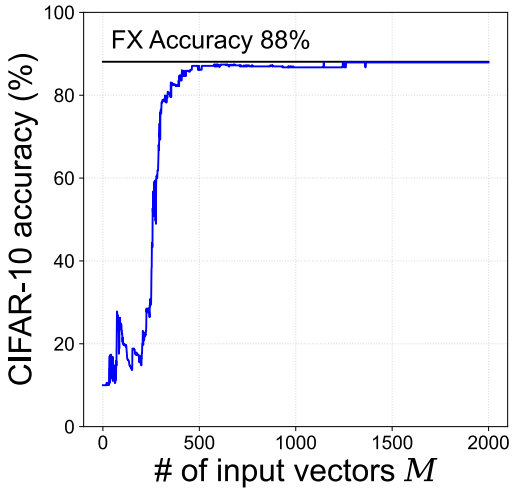
SGD Attack



Bit-error rate (BER)



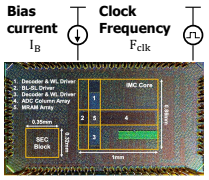
Inference accuracy



Apply MEAs to retrieve weights of ResNet-20 last layer from MRAM-IMC chip

SGD attack requires least number of queries at **high SNDR** to achieve **lowest BER for all ADC columns → inference accuracy within < 0.1% of FX**

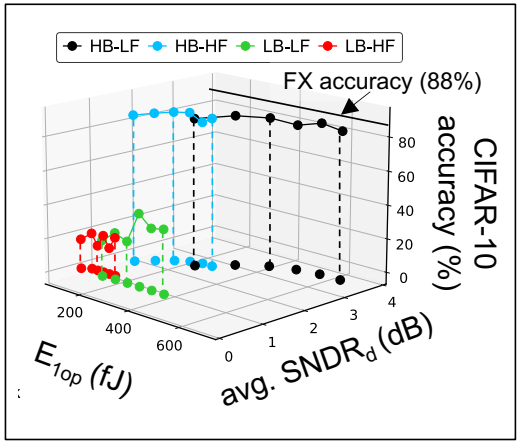
Measured Energy-Accuracy-Security Trade-offs



Previous attacks were performed at high SNDR

What happens at low SNDR?

	High bias (HB) ($I_B = 92\ \mu A$)	Low bias (LB) ($I_B = 67\ \mu A$)
Low frequency (LF) ($F_{clk} = 8.3$ MHz)	**HB-LF (highest SNDR)**	**LB-LF**
High frequency (HF) ($F_{clk} = 16.6$ MHz)	**HB-HF**	**LB-HF (lowest SNDR)**



Strongest attack fails at low-SNDR

Resistive IMCs are vulnerable to model extraction attacks

Low-SNDR settings resilient to MEAs, with need for algorithmic methods to boost inference accuracy under benign scenarios

Thank You!

Support from COCOSYS and CUBIC under JUMP 2.0 (DARPA and SRC)
and DARPA FRANC Program is gratefully acknowledged