Karlsruhe Institute of Technology

29th Asia South Pacific Design Automation Conference
ASP-DAC 2024

ASIA SOUTH PACIFIC DESIGN AUTOMATION CONFERENCE

Power Side-Channel Analysis and Mitigation for Neural Network Accelerators based on Memristive Crossbars

Brojogopal Sapui, Mehdi B. Tahoori

INSTITUTE OF COMPUTER ENGINEERING (ITEC) – CHAIR FOR DEPENDABLE NANO COMPUTING (CDNC)

Agenda

- Introduction
- Problem Definition and Motivation
- Side Channel Analysis for CiM based NN accelerator
- Countermeasures for CiM tailored NN accelerator
- Results and Discussion
- Conclusion

Neural Network (NN) in Modern Technology

- Fastest growing computing paradigm with their ability to **mimic human learning**
- **Efficiency** is a key highlight as they automate and optimize processes

- The structure of NN is:
- **Input Layer:** features from images, text, etc.
- **Hidden Layers:** comprise connected neurons
- **Output Layer:** prediction based on the learned features
- **Connections:** weight, determining its significance in the learning process

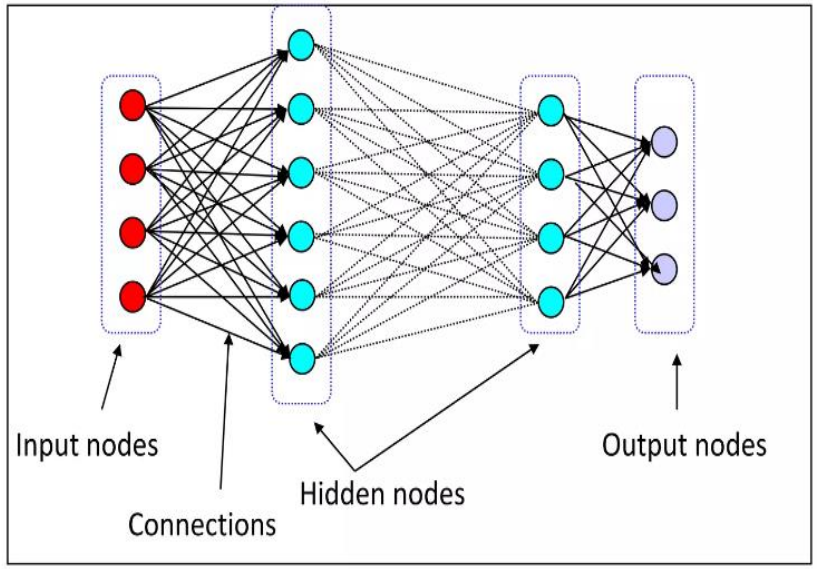
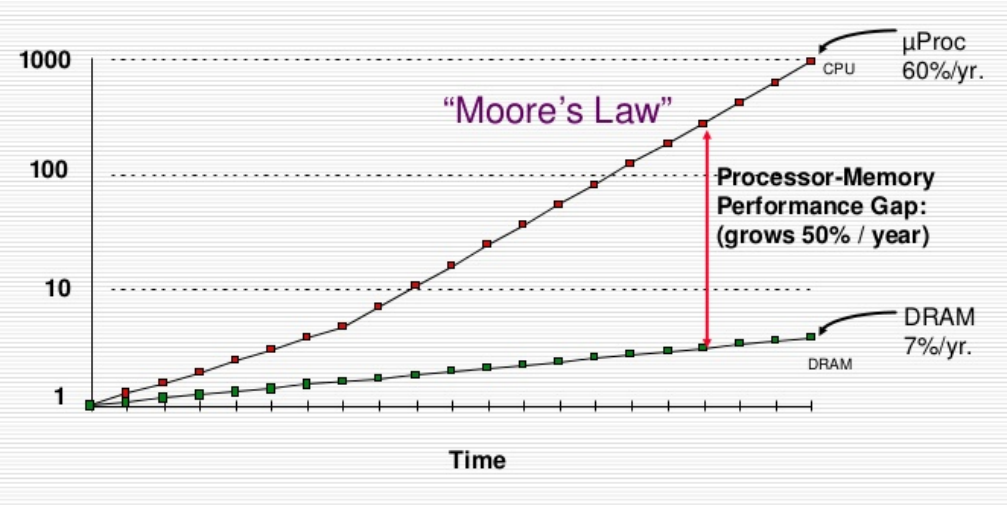
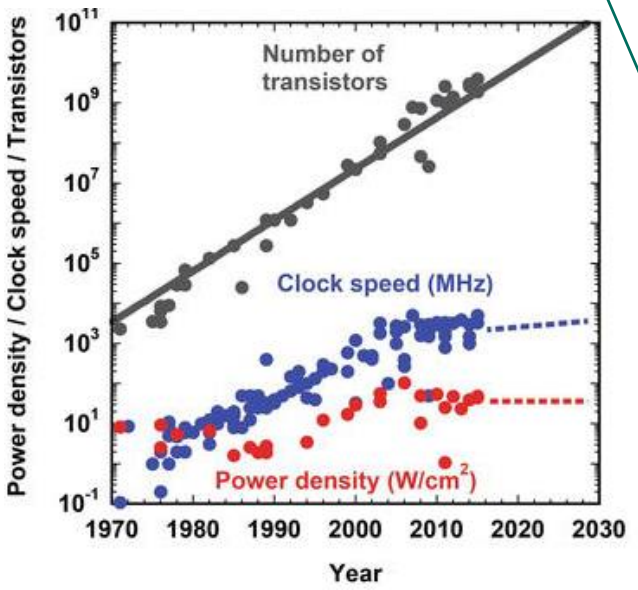
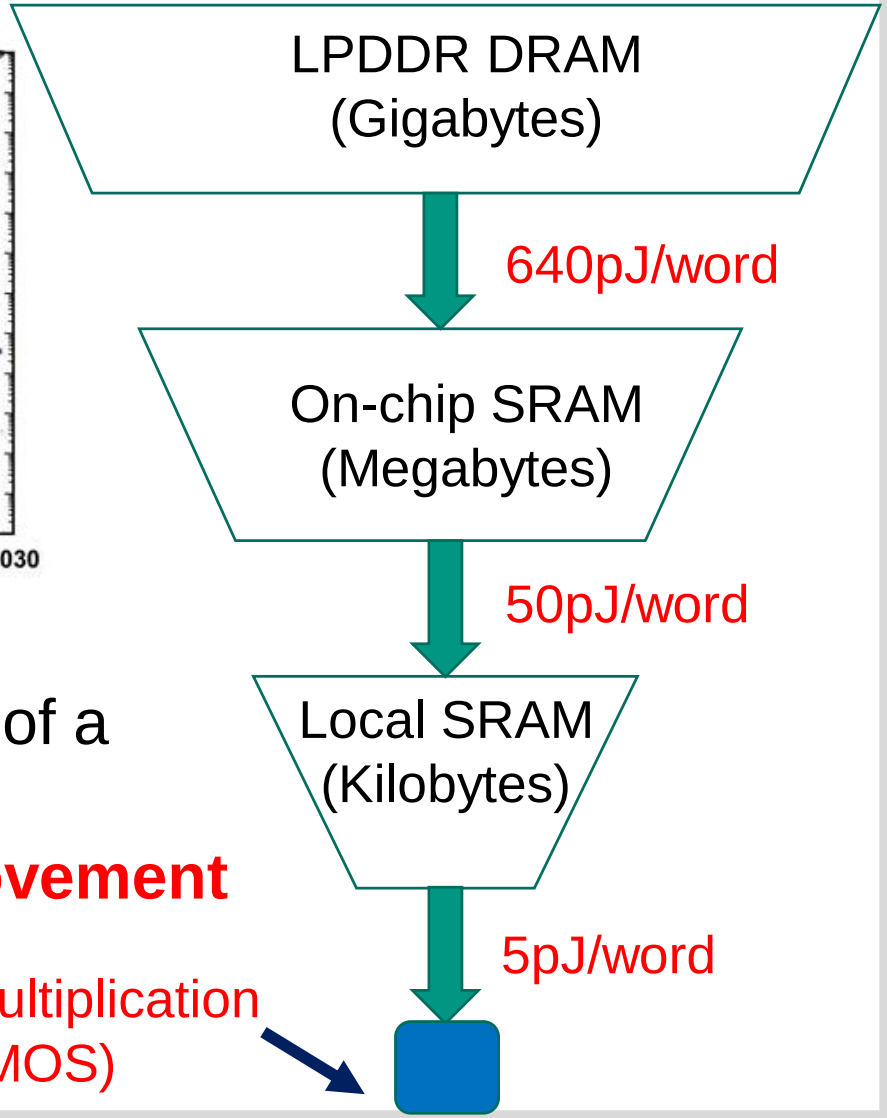


Image Source : Google

Memory and Power Walls



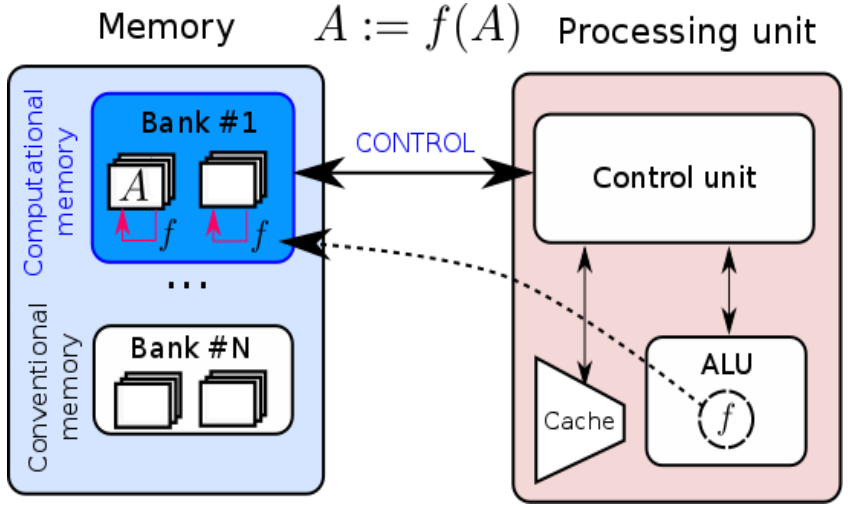




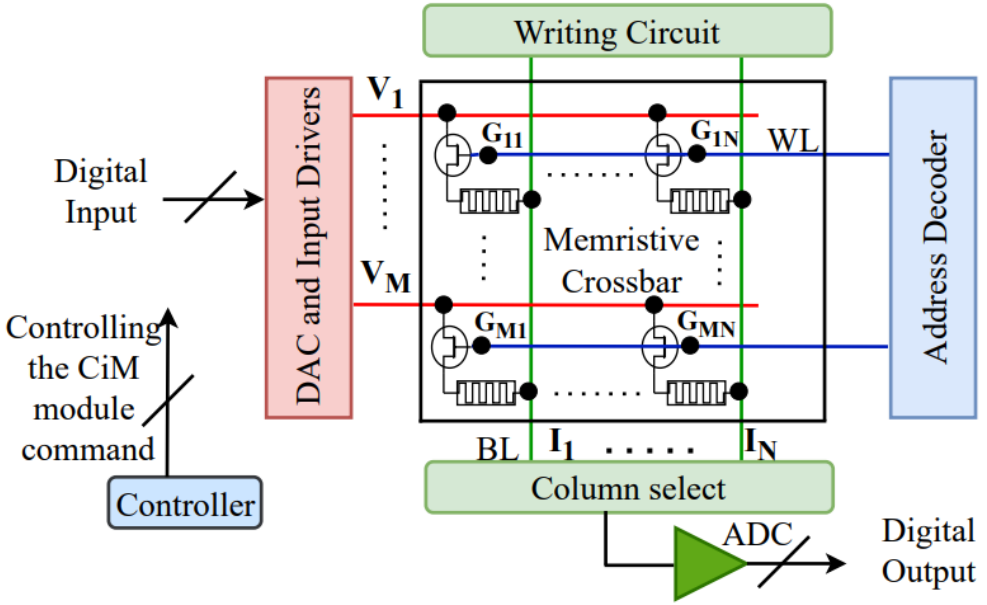
- A memory access consumes **~100-1000X** the energy of a complex addition
- **62.7%** of the total system energy is spent on **data movement**

NN accelerator based on CiM

Processing unit & Computational memory



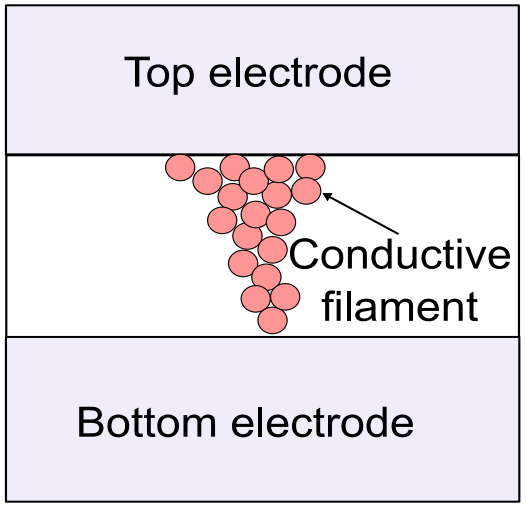
- Perform “certain” computational tasks **in place in memory**
- Achieved by exploiting **the physical attributes of the memory devices**, their **array level organization**, the **peripheral circuitry** as well as the **control logic**



- Neurons and synapses **co-located with memory** cells for seamless processing
- Exploits **parallelism** inherent in memory, speeding up neural network computations

Resistance-based memory devices

ReRAM

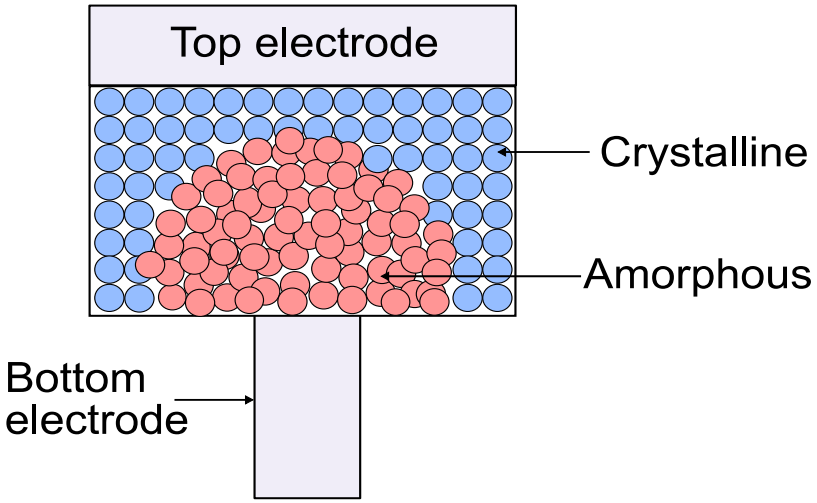


Resistance range = 10^3-10^7

Access time (write) = 10ns - 100ns

Endurance = 10^6-10^9

PCM

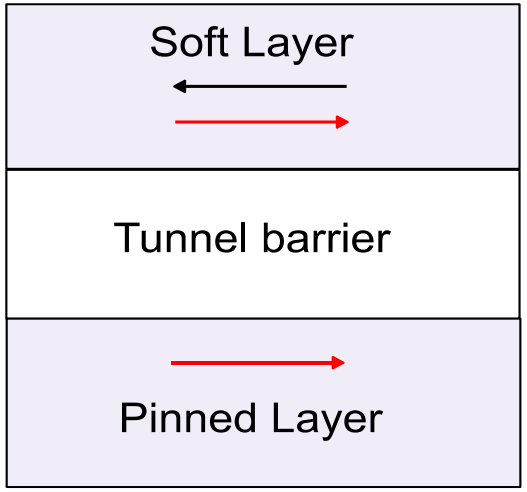


Resistance range = 10^4-10^7

Access time (write) ~ 100ns

Endurance = 10^6-10^9

STT-MRAM



Resistance range = 10^3-10^4

Access time (write) < 10ns

Endurance > 10^{14}

- **ReRAM:** Migration of defects such as oxygen vacancies or metallic ions
- **PCM:** Joule-heating induced reversible phase transition
- **STT-MRAM:** Magnetic polarization of a free layer with respect to a pinned layer
- Resistance-based memory devices also referred to as **memristive devices**

Wong and Salahuddin, Nature Nanotechnology (2015)

Agenda

- Introduction
- Problem Definition and Motivation
- Side Channel Analysis for CiM based NN accelerator
- Countermeasures for CiM tailored NN accelerator
- Results and Discussion
- Conclusion

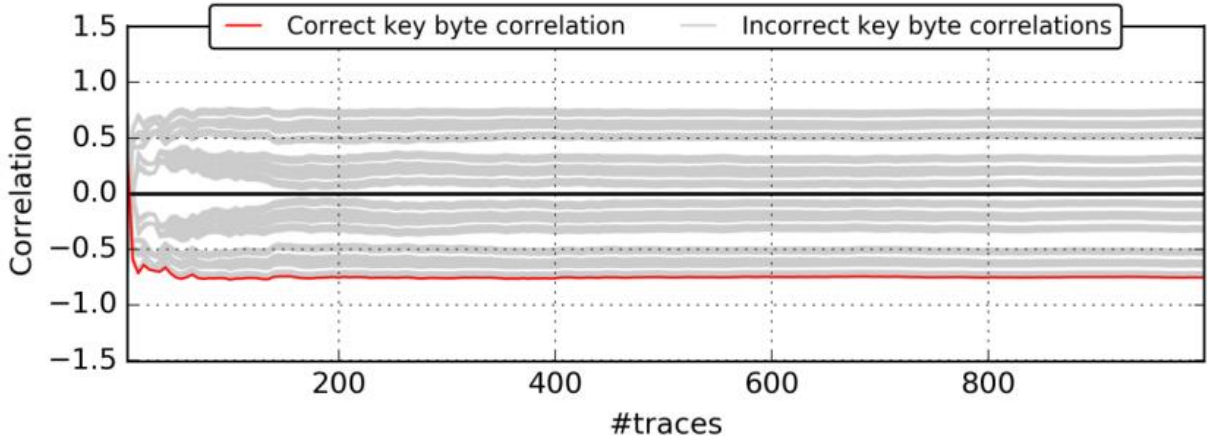
Security Challenges:

- The main industry concern
  - Widespread adaptation of CiM based NN accelerator
  - In analog computing, main channel and side channels are mixing

- Our purpose
  - Accurate yet scalable side channel vulnerability analysis of CiM based NN accelerator
  - Security evaluation of necessary peripheral devices along with the accelerator
  - Effective CiM-tailored countermeasures for NN accelerator

Power Side Channel Analysis (SCA)

- Power usage of a device may vary in a data-dependent manner
- Power consumption measurements
  - Statistical analysis
  - Extraction of information by inspection
- Correlated Power Analysis (CPA)
  - Using hypothesized power model
- Test Vector Leakage Assessment (TVLA)
  - First order statistical leakage
  - t-value < 4.5 shows absence of leakage



Related Work

- CiM realization with NVMs in security applications
  - Encryption/decryption functions with (NVMs) [1] and SRAM [2]
- CiM based NN accelerator
  - Replication attack and protection [3]
  - Power and timing analysis with countermeasures [4]
- Countermeasures result in significant increases (double) in latency and power
- No security evaluation of peripheral devices in CiM based NN accelerator

[1] Dodo et al., “Secure STT-MRAM Bit-Cell Design Resilient to Differential Power Analysis Attacks”, 2019

[2] Xie et al., “Securing Emerging Nonvolatile Main Memory With Fast and Energy-Efficient AES In-Memory Implementation”, 2018

[3] C. Yang et al., “Thwarting replication attack against memristor-basedneuromorphic computing system,” TCAD, 2019

[4] Z. Wang et al., “Side-channel attack analysis on in-memory computing architectures,” IEEE TETC, 2023

Agenda

- Introduction
- Problem Definition and Motivation
- Side Channel Analysis for CiM based NN accelerator
- Countermeasures for CiM tailored NN accelerator
- Results and Discussion
- Conclusion

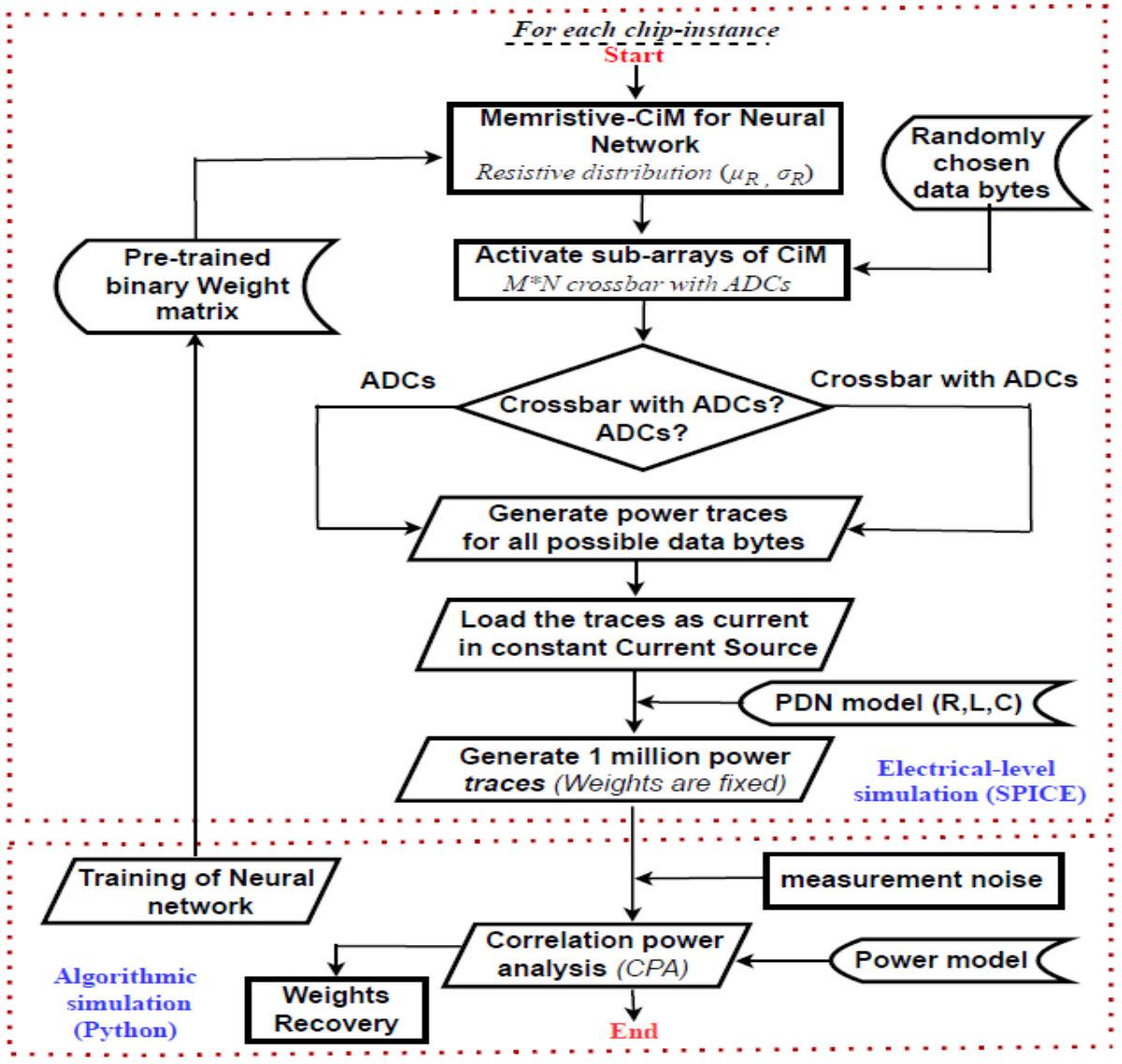
Threat Model

- Attacker’s goal is to query the CiM-implemented NN model
  - Controlling its inputs
  - Observing the side-channel information
- Knowledge of the hardware structure of CiM, crossbar size, and ADCs type
- Physical access: power consumption of the circuit
- Logical access: input/output ports, crossbar, and ADCs
- No access to individual memristive devices
- No knowledge of the NN mapped into the crossbar

Simulation Approach

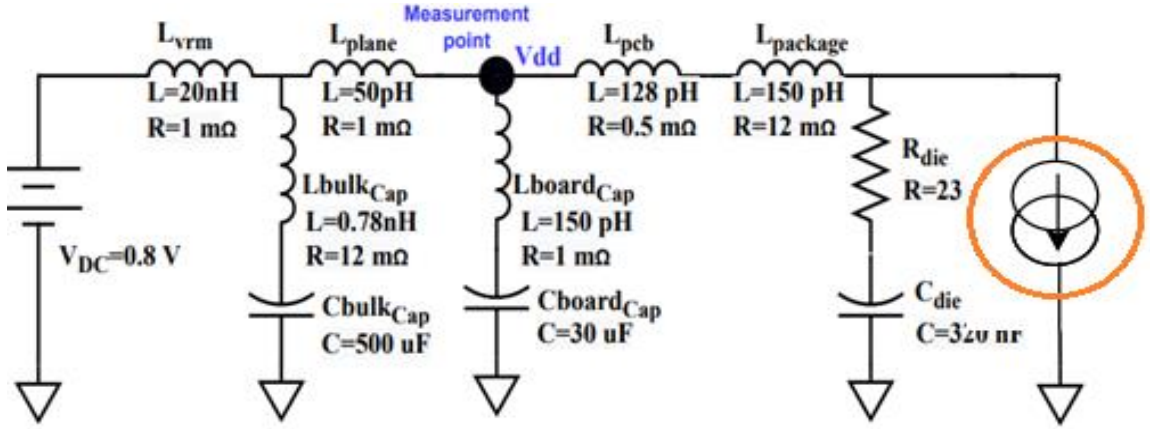
- Challenges and requirements
  - Accurate simulation of entire CiM block with accurate device models
    - Accurate technology models of memristive devices
    - Effect of Power Delivery Network (PDN)
    - Measurement Noise (MN)
  - Generate million of power traces in realistic simulation environment
- Use-case analysis
  - MNIST data stored in CiM based NN accelerator
  - MAC operation
- Source of leakage: device mismatch
  - Need to accurately model mismatch (Monte Carlo on Memristive devices)

Flow Diagram



Effect of PDN and MN

- PDN delivers a stable voltage down to each transistor
  - Consist of R, L, C suitable for specific technology
  - PDN drives CiM based NN accelerator as load



- MN as Gaussian noise with SNR = 20 dB
  - Considered in realistic Oscilloscope error for SCA

Side Channel Analysis Attack

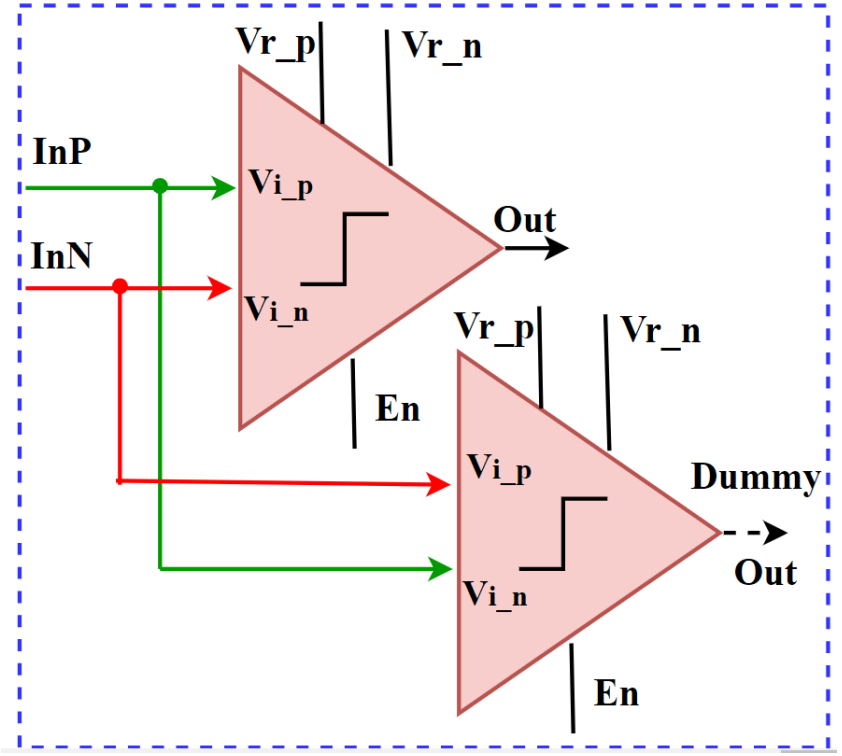
- Attack on CiM implemented NN accelerator consisting MAC operation
  - Vulnerability analysis of entire CiM
  - Vulnerability analysis of peripheral devices of CiM such as ADCs
- SCA exploited by the mismatch in the memory devices
  - Information leaking faster from peripheral devices
- Hamming Weight power model
- Pearson Correlation analysis with data dependent power

Agenda

- Introduction
- Problem Definition and Motivation
- Side Channel Analysis for CiM based NN accelerator
- Countermeasures for CiM tailored NN accelerator
- Results and Discussion
- Conclusion

Countermeasure: Hiding Method

- Hide circuit activity by reducing the signal-to-noise (SNR) ratio
  - Data-dependent power vs power noise
- Power equalization by adding duplicated logic
  - Doubling number of comparators only
  - Interchanging the inputs of duplicated comparator
  - Balancing activated charging and discharging path

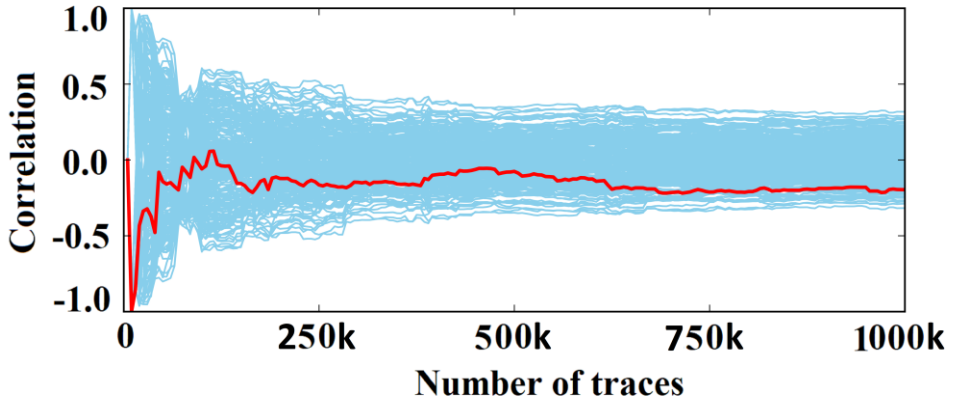


Agenda

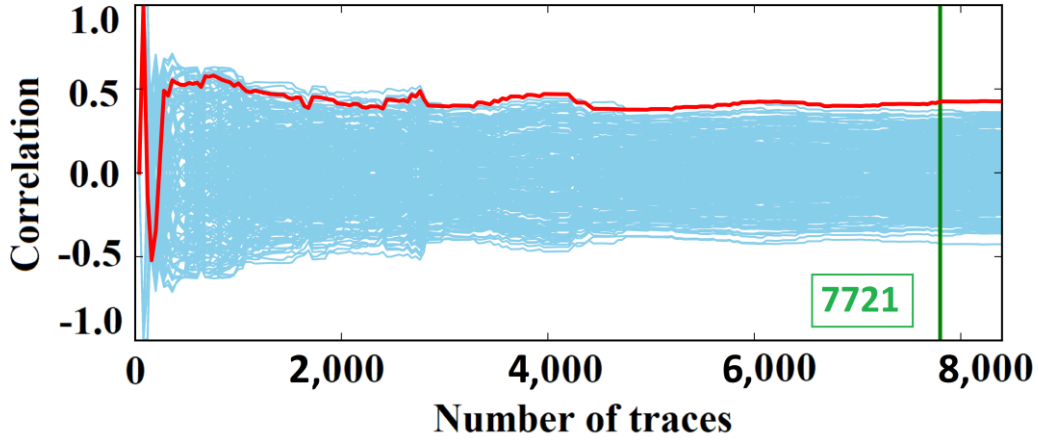
- Introduction
- Problem Definition and Motivation
- Side Channel Analysis for CiM based NN accelerator
- Countermeasures for CiM tailored NN accelerator
- Results and Discussion
- Conclusion

CPA on CiM based NN accelerator

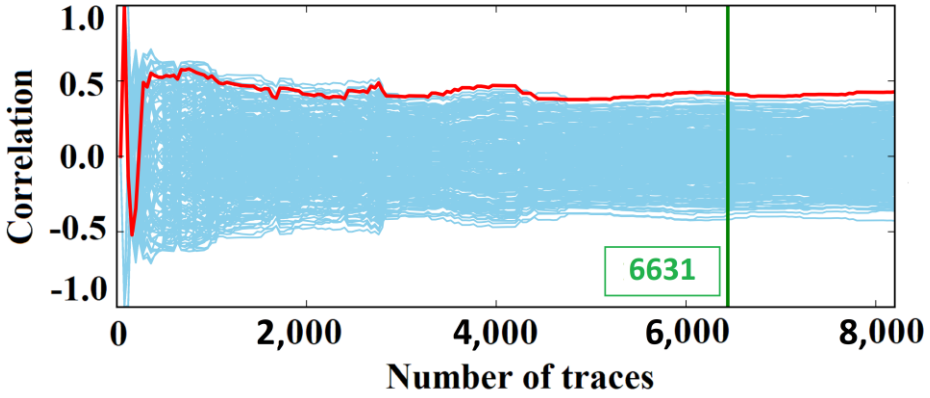
- Crossbar does not leak with 1 million traces
- Entire CiM starts leaking within 8,000 traces
- ADCs start leaking within 7,000 traces



Failed attack on crossbar



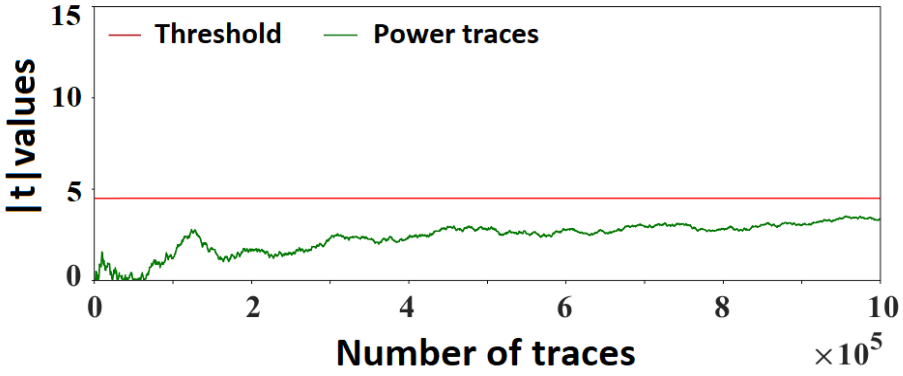
Successful attack on entire CiM



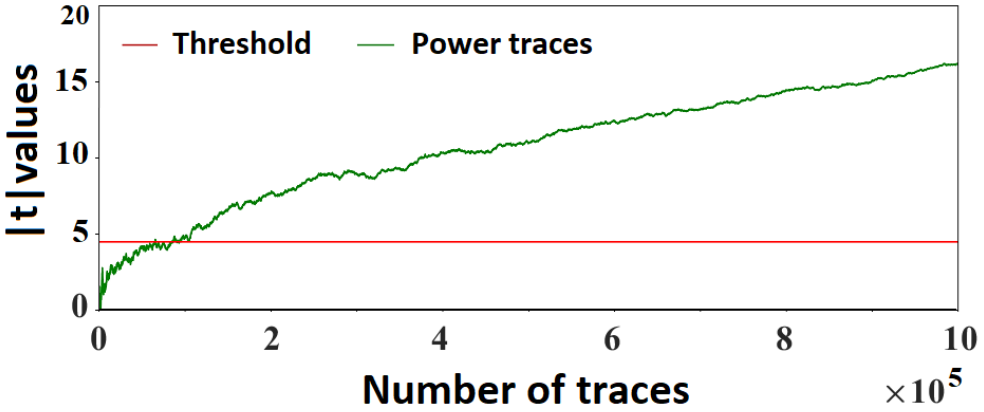
Successful attack on ADCs only

TVLA on CiM based NN accelerator

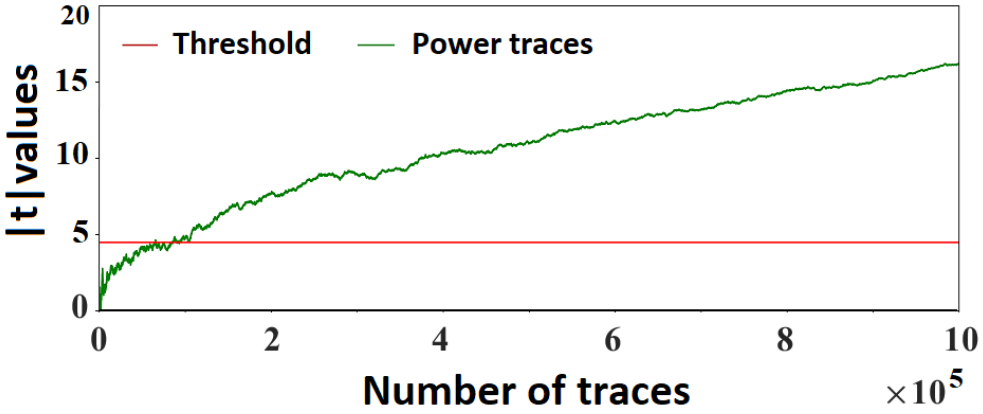
- TVLA does not show any leaking in crossbar
- TVLA leaking after few thousand of traces in entire CiM and ADCs



No leaking from crossbar



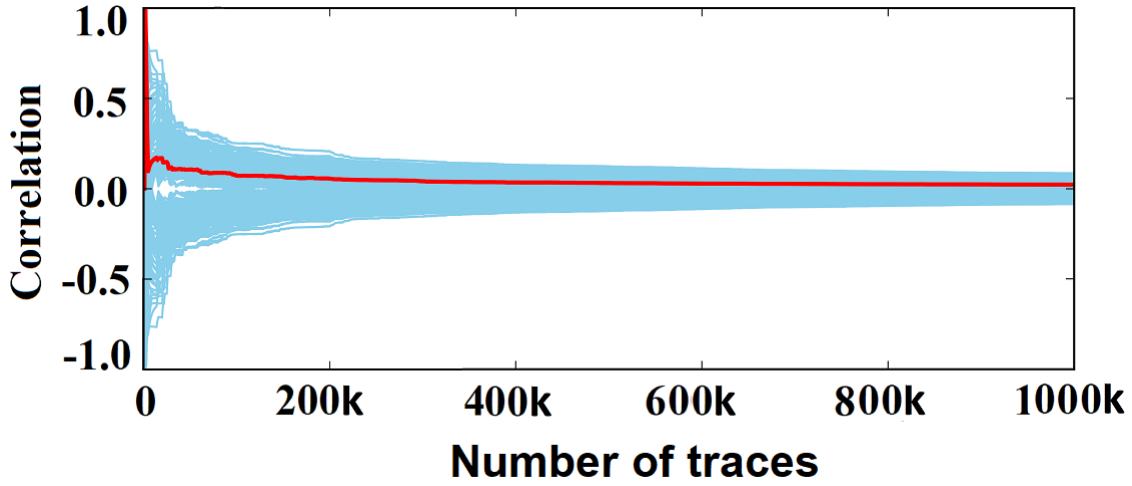
Leaking from entire CiM

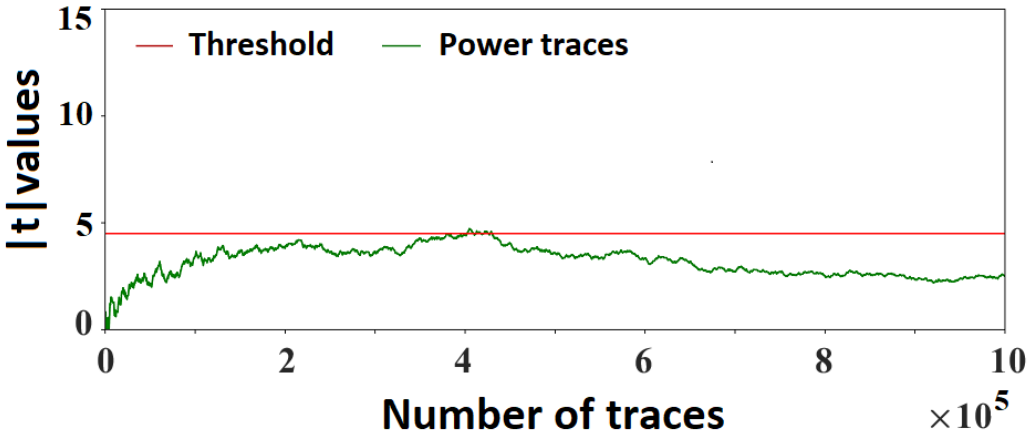


Leaking from ADCs only

Hiding Protected Design

- No CPA attack even with 1 million power traces
- No TVLA information leaking even after 1 million traces





Comparison: Performance Overhead

- Increase in the DNN size does not impact area overhead due to shared ADCs
  - Robustness of the protection mechanism to DNN scaling
- Power overhead of 50.25% compared to the unprotected design
  - More favorable than duplicating the entire CiM (100% power increment)

Metrices	Design		Overhead	
	Unprotected	Protected	Absolute	Percentage (%)
Power	$1.20mW$	$1.803mW$	$0.603mW$	50.25
Area	$5.2um^2$	$8.9um^2$	$3.7um^2$	71.15
Latency	$3.82us$	$3.88us$	$0.06us$	0.01

Agenda

- Introduction
- Problem Definition and Motivation
- Side Channel Analysis for CiM based NN accelerator
- Countermeasures for CiM tailored NN accelerator
- Results and Discussion
- Conclusion

Conclusion

- Computation in Memory (CiM) based Neural Network accelerator
  - Promising solution against conventional NN accelerator
  - Using emerging Non-volatile Resistive Memories
- Analyzing and mitigation security challenges
  - Accurate yet scalable side channel attack analysis flow
    - Vulnerability analysis of CiM crossbar array as well as peripheral devices
    - Effect of variations, power delivery network, measurement noise
    - Million trace analysis with CPA and TVLA
  - Better CiM reliability (higher resistive ratios) means lower SCA security!
- Effective CiM-based countermeasures
  - Hiding: orders of magnitude improvement in SCA

Karlsruhe Institute of Technology

Thank you for your attention!
Any Question?

Please write to:
brojogopal.sapui@kit.edu