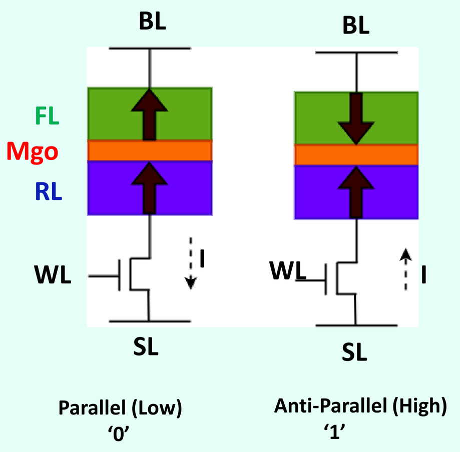
Fault-Tolerant Techniques For Emerging Non-Volatile Memories and Neuromorphic Computing System

DATE 25

Background and Motivation

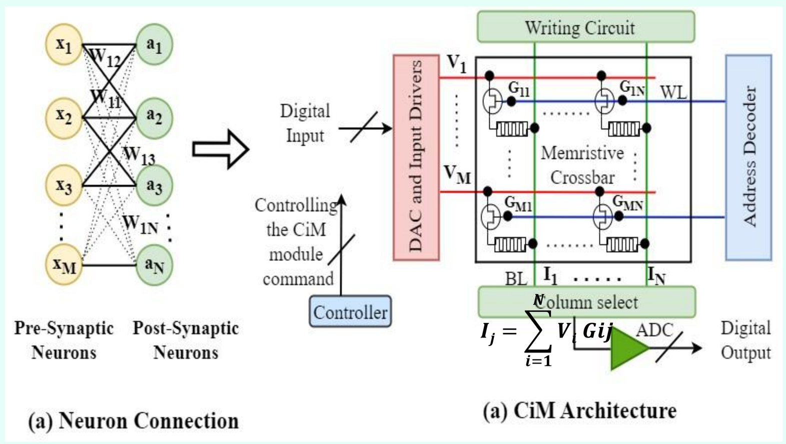
- Emerging non-volatile memories (NVMs) have revolutionized data storage, making them viable alternatives to CMOS memories. Spin-transfer torque magnetic random-access memory (**STT-MRAM**) is the most promising candidate, as shown by several industrial demonstration. However, it has some reliability issues (**Hard and Soft errors**), which need to addressed to improve overall **reliability**.
- Building on the advancements of emerging memories, **Neuromorphic computing systems** have also emerged as a promising approach for accelerating neural network (**NN**) computations. However, ensuring their reliability remains challenging, as memories are prone to errors that can alter stored weight values, degrading **NN accuracy**.



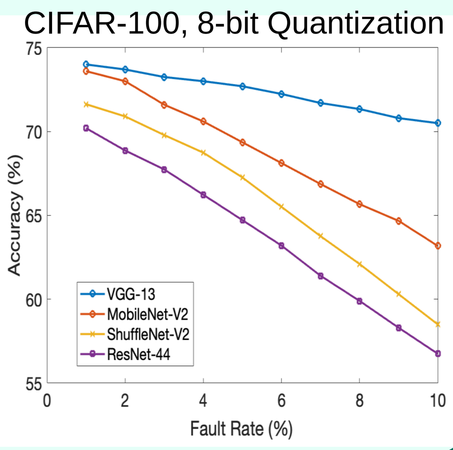
STT-MRAM

- **Soft errors: Momentary bit flip**
  - Asymmetric switching
  - Read failure
  - Retention failure
- **Hard errors: Stuck-at-fault**
  - Oxide breakdown
  - Manufacturing defects

Memristive Crossbar (CiM)

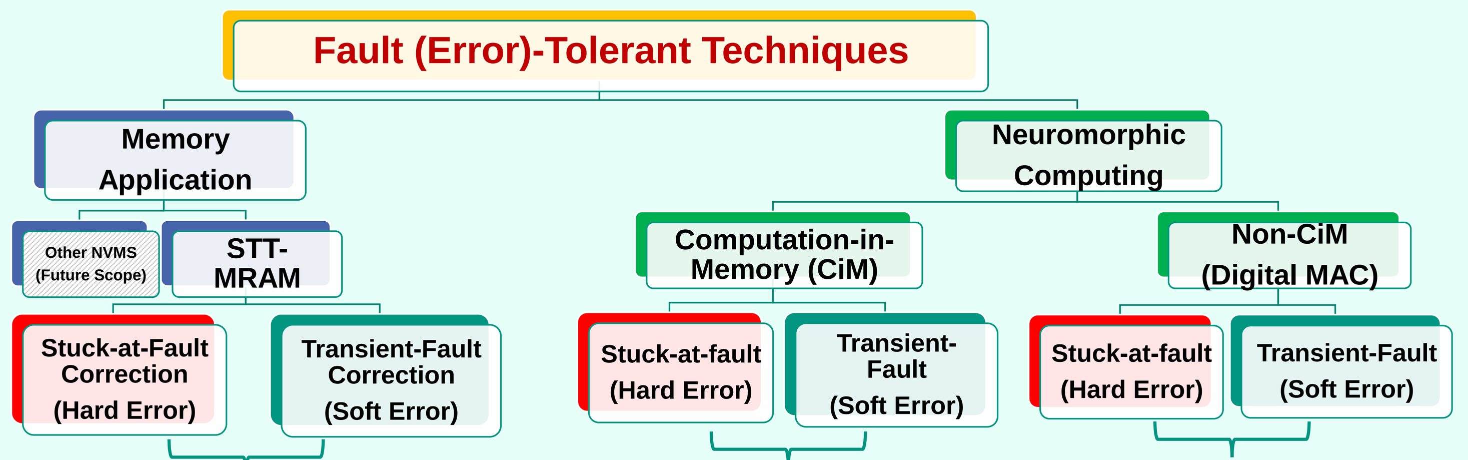


Impact of Errors



Thesis Overview

- This thesis explores an efficient fault-tolerant techniques for non-volatile memories (our focus is on STT-MRAM) and Neuromorphic Computing System, which employ error-correcting codes (ECCs) in conjunction with architectural modifications tailored to specific needs to improve the overall system reliability.



- ❑Study of error correction strategies for STT-MRAM (IEEE Design & Test'24)
- ❑Block error correction pointer for hard error correction (IEEE ASP-DAC'24)
- ❑Mitigating the impact of asymmetric switching (IEEE TCAD'25)
- ❑Mitigating the impact of interconnect on bit-cell reliability (IEEE DATE'25)

- ❑Low overhead Checksum techniques for reliable analog matrix-vector multiplication (MVM) in memristive crossbar for NN application

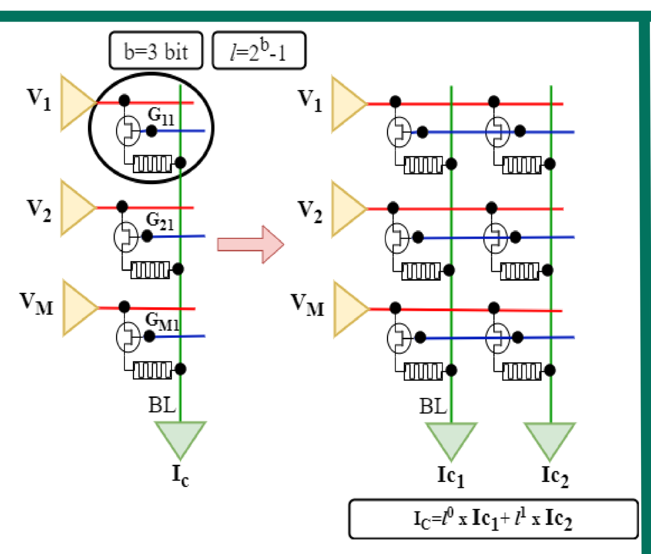
(IEEE IOLTS'22 and VTS'23)

- ❑NN-ECC: Embedding ECC in Neural Network for multi-bit error correction in binary weight memories of digital NN accelerator

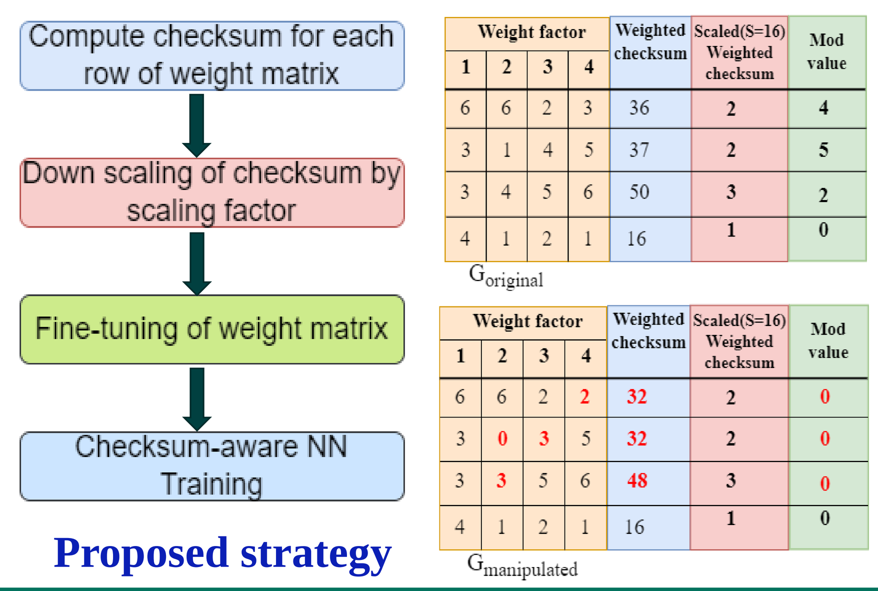
(IEEE VTS'24)

Low Overhead Checksum Techniques for Error Correction in Memristive Crossbar for NN Application

- A memristive cell has a **limited** number of stable states (*l*) to store data bits. This requires **column slicing** to store **checksum** for error correction. This will increase the memory overhead.
- Proposed strategy to **downscale** the **checksum** for storage in a single cell, significantly reducing **memory overhead**.



Column slicing

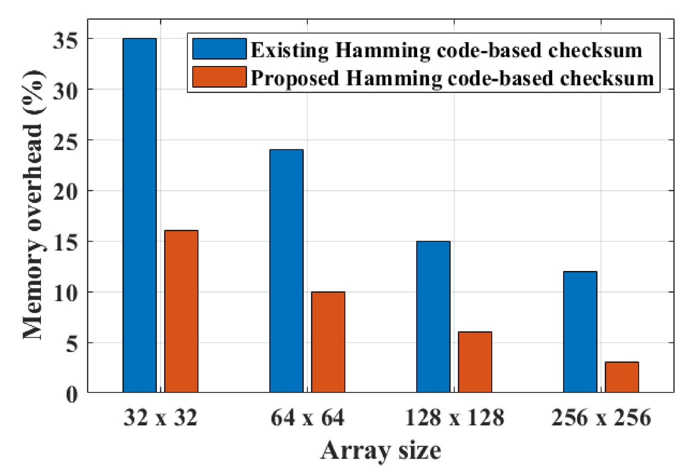


Proposed strategy

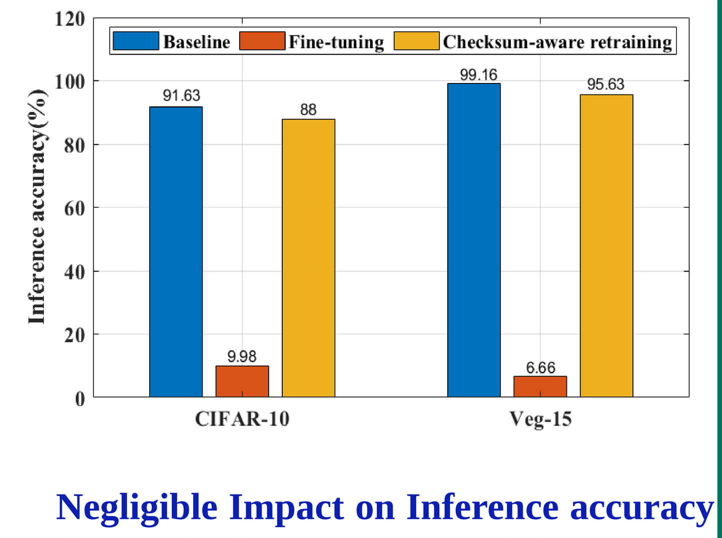
Bit cell precision: 3bit/cell

Model: Resnet-18

Dataset: MLP CIFAR-10



Reduction in checksum Storage for diff array size



Negligible Impact on Inference accuracy due to checksum-aware Training

NN-ECC: Embedding Error Correction Codes in Weight Memories of Neural Network using Multitask Learning

- ECC provide guarantee of a specific number of error correction.
- However, ECC require redundant storage. Given the vast number of weight parameters in NNs, the memory overhead can be significant.
- Proposed strategy to embedding ECC in weight matrix of NN using multi-task learning, while having **zero ECC parity bit overhead** to perform **multi-bit** error correction.

- Consider a scenario:
  - Weight quantization: **8-bit**
  - Single error correction per **64-bit**
- Idea of ECC encoding:
  - Decide the codeword size equals to **64-bit**
  - Size of weight remains 64-bit after encoding

Data: $[W]_{1x57}$ Parity: $[Wp]_{1x7}$

Proposed ECC Encoding Process

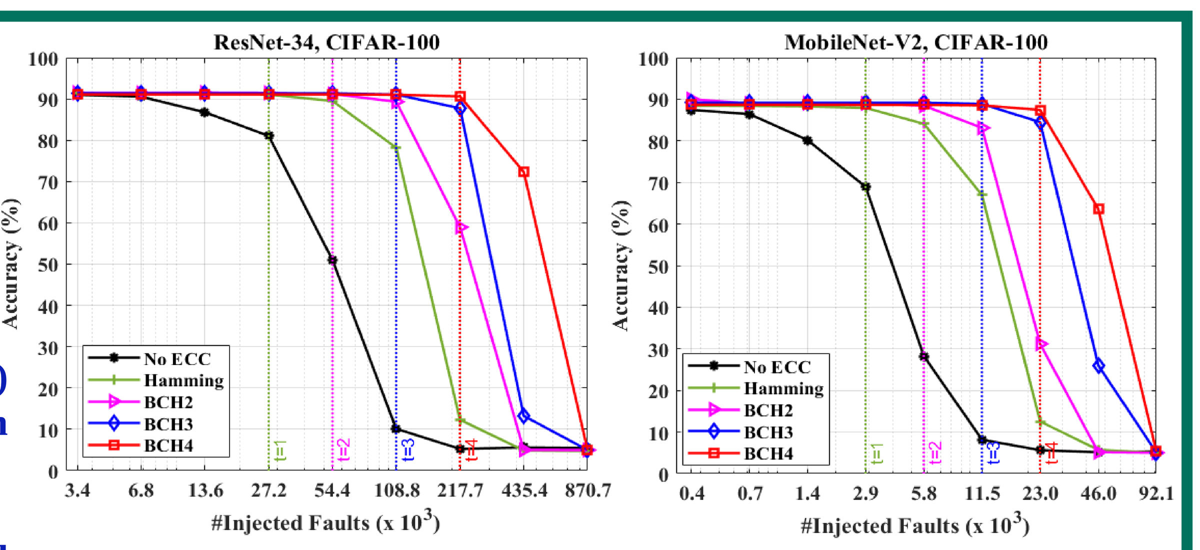
- **During forward-pass**
  - Loss is computed based on both parity and freely learnable weights
- **During backward-pass**
  - Weight matrix is updated based on update rule (Gradient Descent)
  - Parity is calculated based on updated weights
  - Straight through estimator technique is used for forward and backward pass for ECC parity

Proposed Multi-Task Learning

Topology (Dataset: CIFAR-100)	Accuracy (%) Baseline	BCH (t=4)
ResNet-34	91.54	90.42
MobileNet-V2	89.31	88.72

Negligible drop (1-2%) due to encoding even with t=4 BCH code

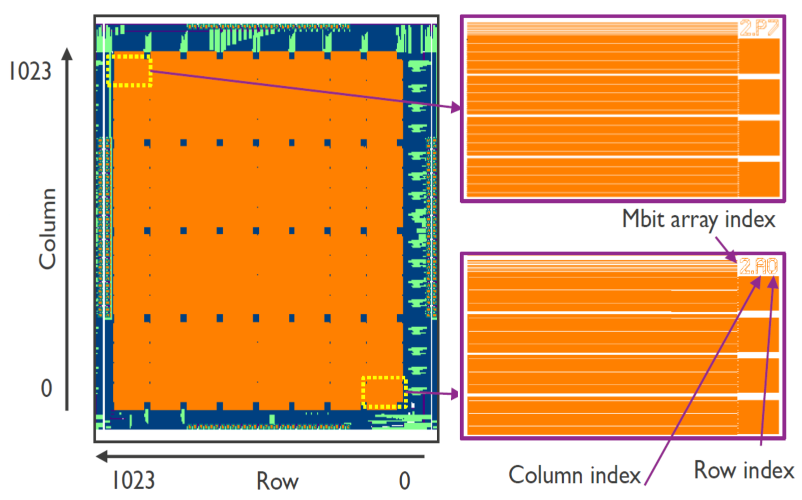
Zero memory overhead for ECC parity bits



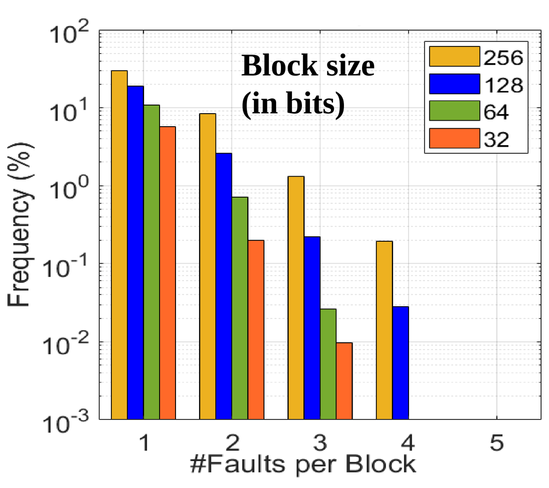
Impacts of error correction on inference accuracy

Block Error Correction Pointer for Hard Error Correction in STT-MRAM

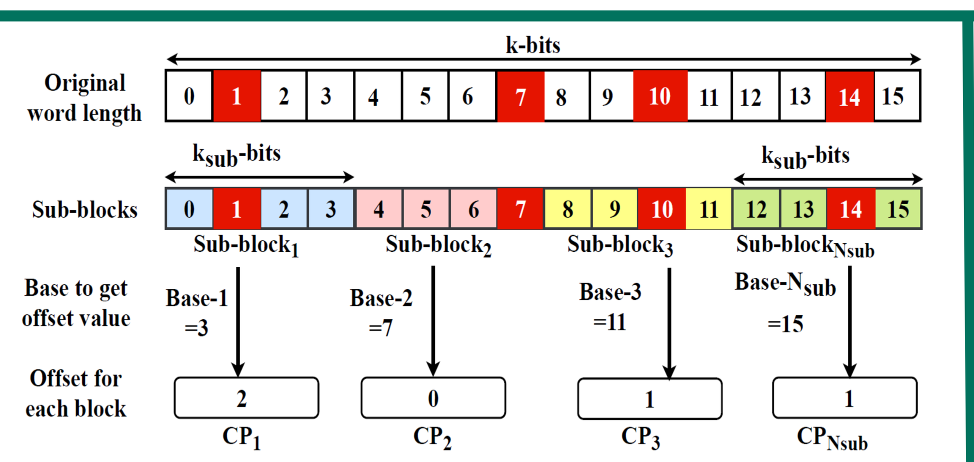
- Proposed an efficient block error correction pointer strategy, dividing memory words into smaller blocks and storing offset instead of the absolute address of the hard error.
- Incorporated experimental measurement data obtained from manufactured STT-MRAM chips to get the hard error distribution.



Physical structure of 1Mb STT-MRAM

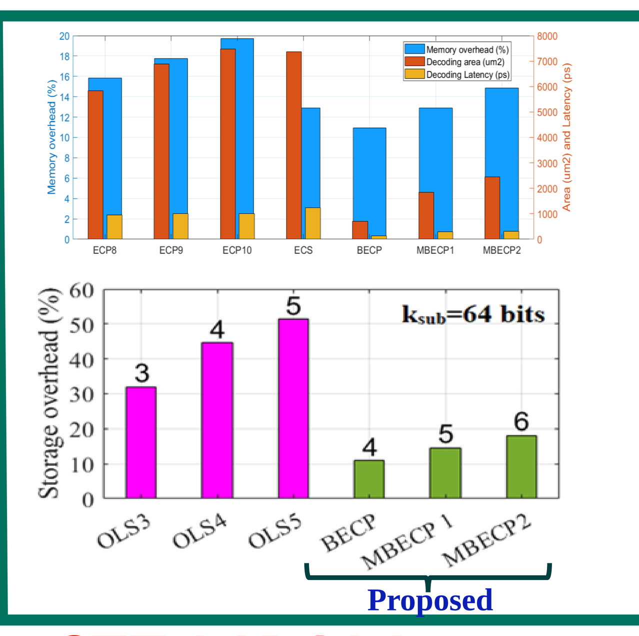


Hard Fault distribution



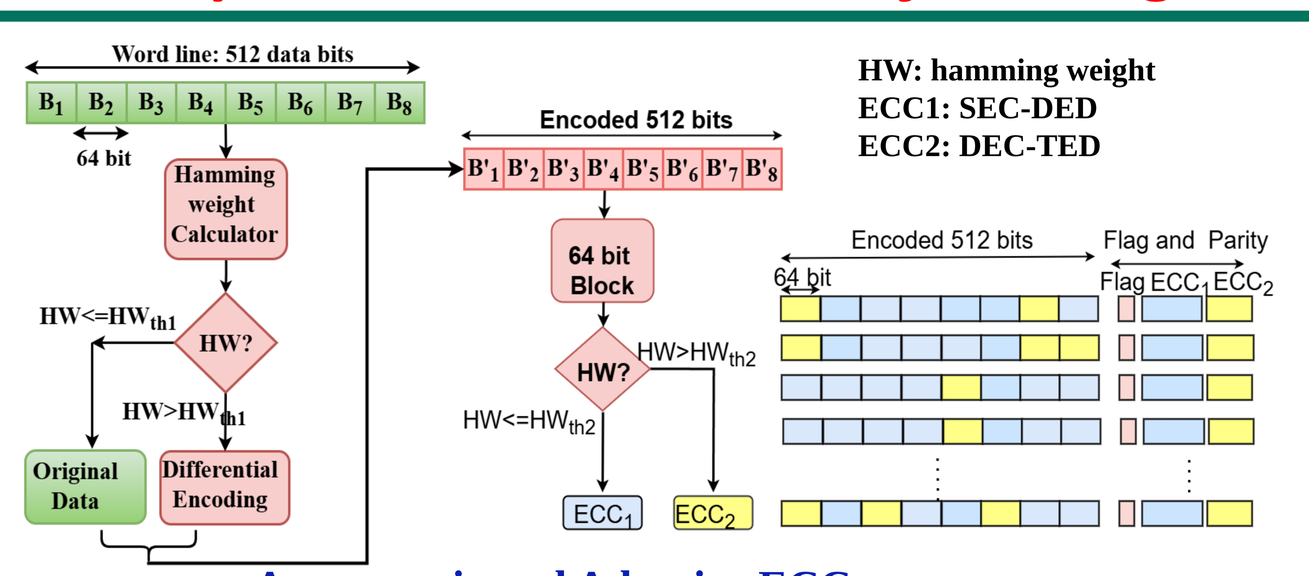
k= word length k_{sub}: sub-block size N_{sub}: #sub-blocks

Block Error Correction Pointer

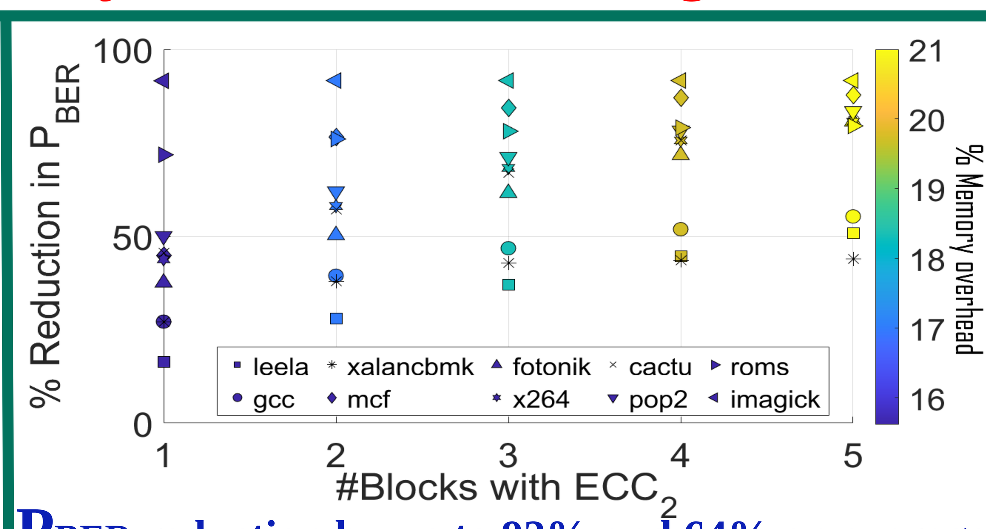


Asymmetric and Adaptive Error Correction for Mitigating Asymmetric Switching Failure in STT-MRAM

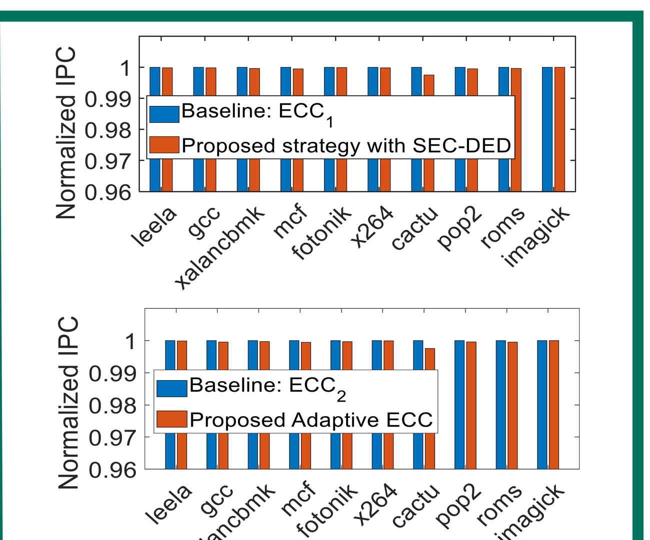
- The switching characteristic of STT-MRAM is **asymmetric**.
- Proposed an efficient asymmetric and adaptive error correction strategy based on the **Hamming weight** of data bits.
- Reliability improvement is measured by L2 cache block error rate (**PBER**)
- System performance is measured in terms of instruction per cycle (**IPC**)



Asymmetric and Adaptive ECC



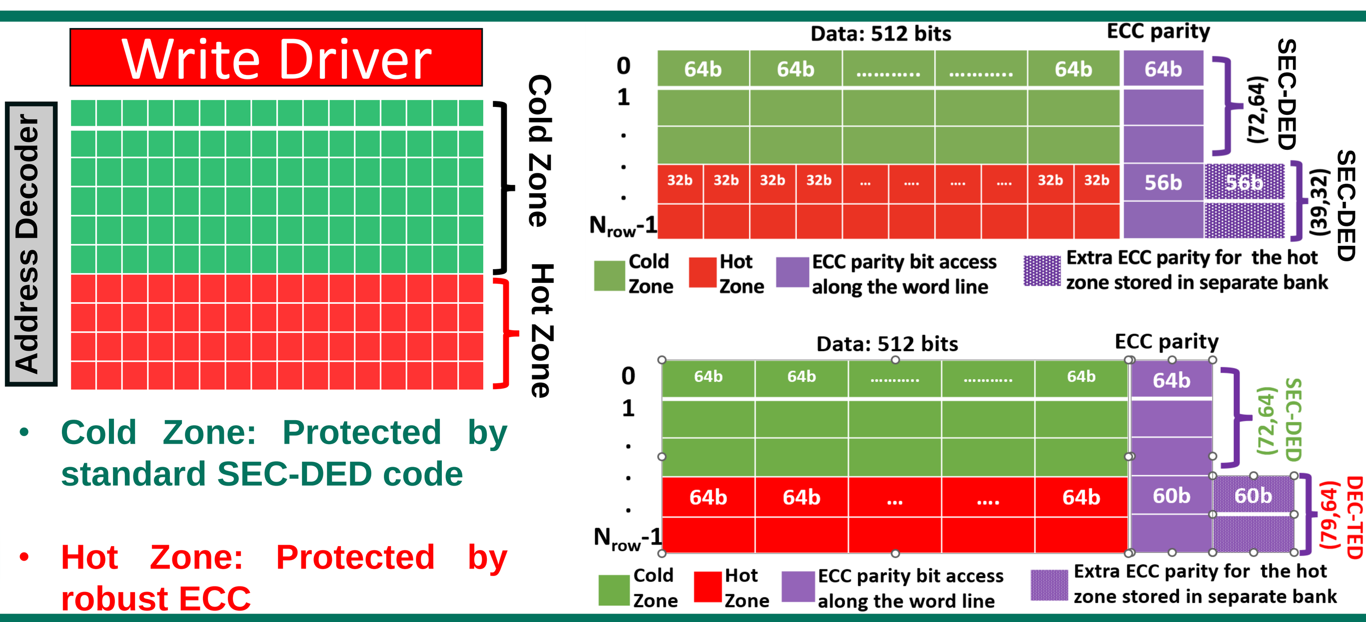
PBER reduction by up to 92% and 64% on average across different SPEC CPU2017 benchmarks



IPC penalty (<0.2%)

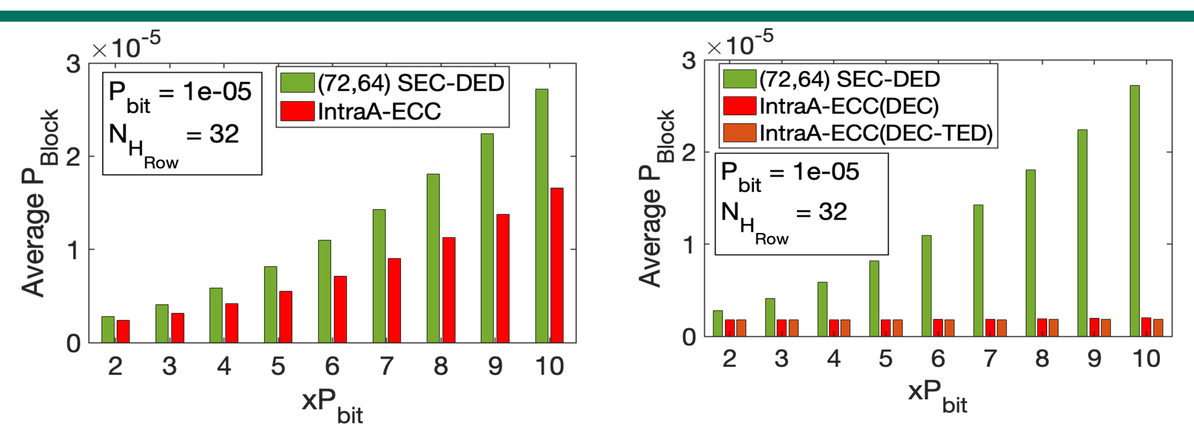
InterA-ECC: Interconnect-Aware Error Correction in STT-MRAM

- As technology scales down, **interconnect parasitics** become more dominant.
- Interconnect parasitic, potentially compromising the reliability of memory cells located far from the **write driver**.
- InterA-ECC strategy selectively applies **robust** error-correction to specific rows within the subarray rather than uniformly across all rows



- Cold Zone: Protected by standard SEC-DED code
- Hot Zone: Protected by robust ECC

Average block error rate reduction (Pblock) while considering last NH=32 rows under hot zone



- Reduction in ECC parity requirements by up to 40% compared to uniform ECC
- Bit error rate (Pbit) resiliency can be enhanced by upto 15× in the hot zone
- Average block error rate reduction by up to 4.5x with 5x Pbit in the hot zone

Surendra Hemaram, Email: surendra.hemaram@kit.edu
Thesis Advisor: Mehdi B. Tahoori
Chair of Dependable Nano Computing
Karlsruher Institut für Technologie