Structural Coding:

A Low-Cost Scheme to Protect CNNs from Large-Granularity Memory Errors

Ali Asgari

Prashant Nair

Karthik Pattabiraman

Florian Geissler

Qutub Syed Sha

Michael Paulitsch





Machine Learning

Machine Learning (ML) Systems work around humans



Home Care



Law Enforcement



Self-Driving Cars

Machine-learning is increasingly used in safety-critical systems

Soft Errors: Effects

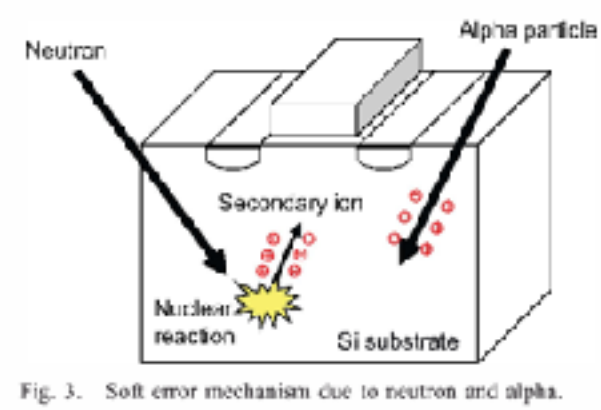
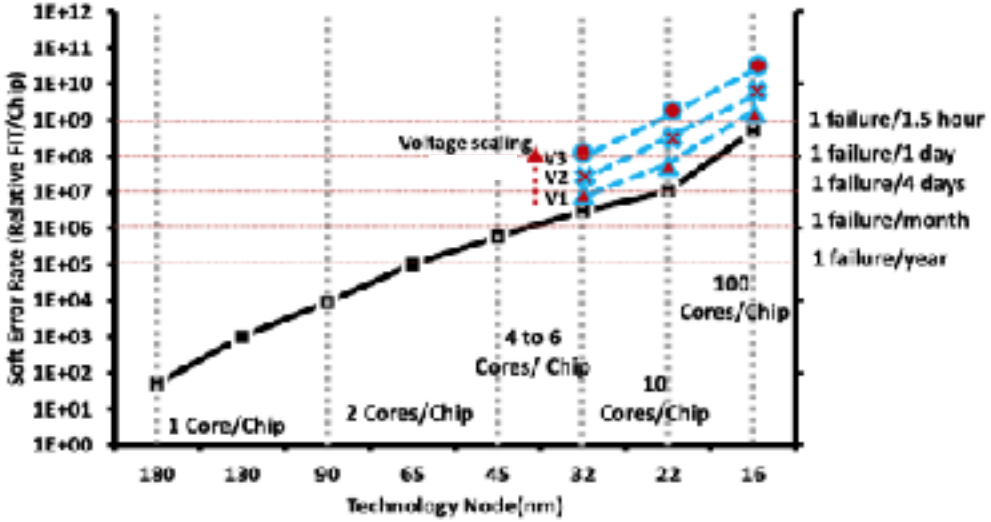


Fig. 3. Soft error mechanism due to neutron and alpha.

1	0	0	1	0	1	1	1
1	0	0	1	1	1	1	1

Image copied from Hashimoto et al, “Soft error immunity of subthreshold SRAM”, ASICON’13

Increasing Soft Error Rate



[Upasani, 2015]



Blue Waters:
1 application crash due to a soft error every 15 mins

[Martino, 2015]

Soft Errors and AVs





ISO 26262

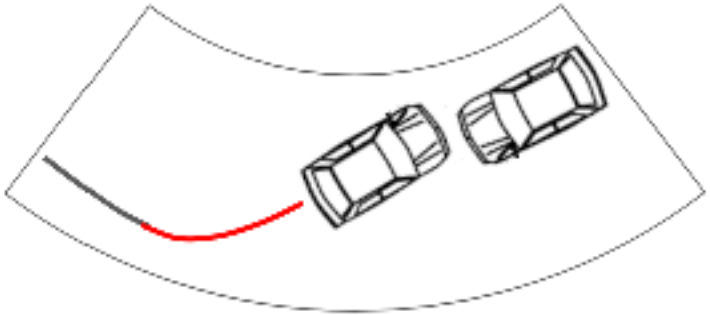
[Our work - SC 2017]

- **Safety standard – Automotive Safety Integrity Level (ASIL-D)**
  - Error rate <10 FIT (per 1 billion hours) – ISO 26262
  - DNN systems do not meet the requirement without protection

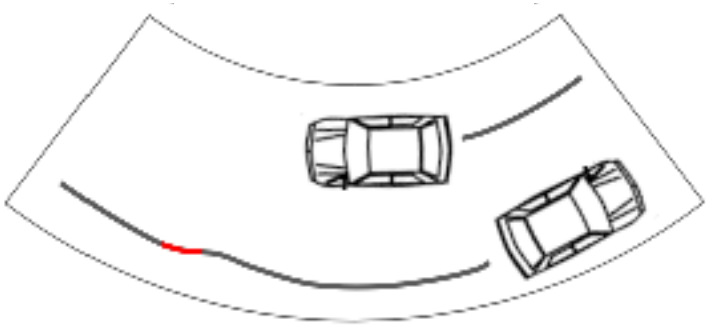
Fault Model: Faults in Parameters

Parameters are mostly comprised of the DNN's weights

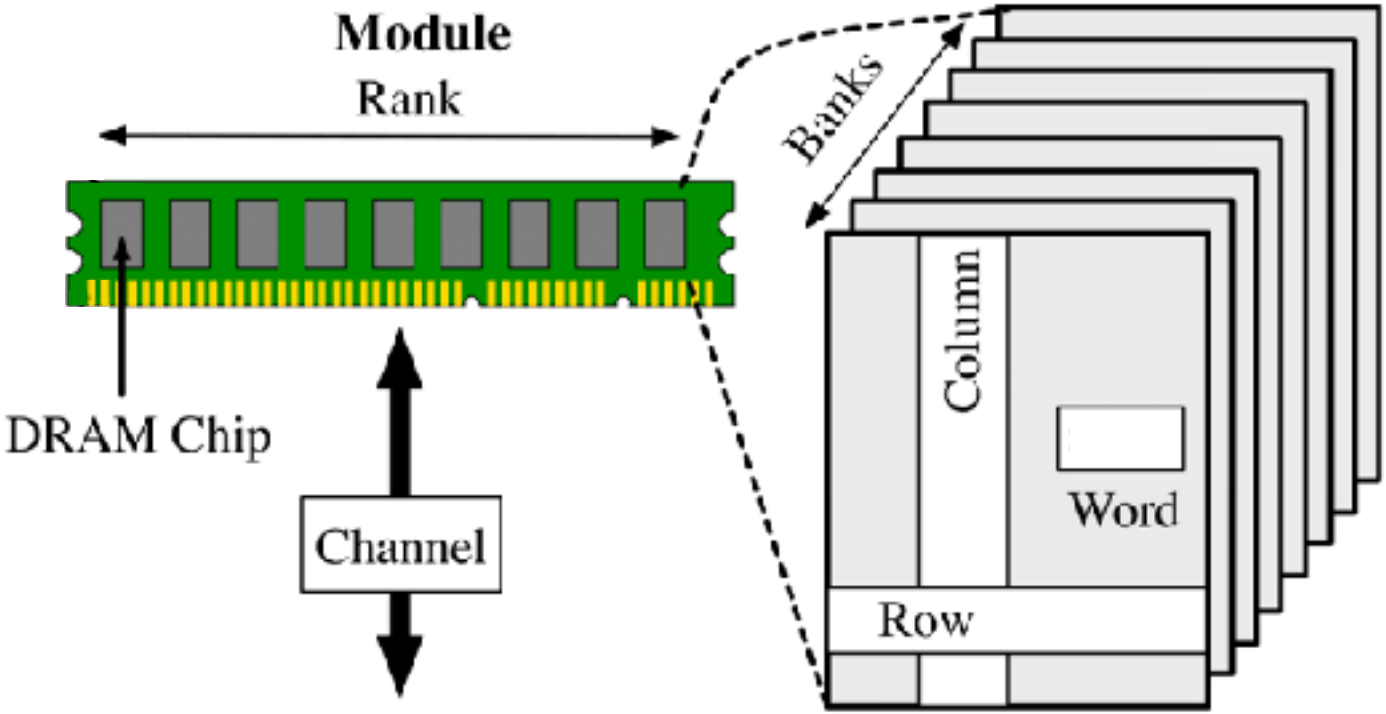
Faults in parameters **remain** across many inferences



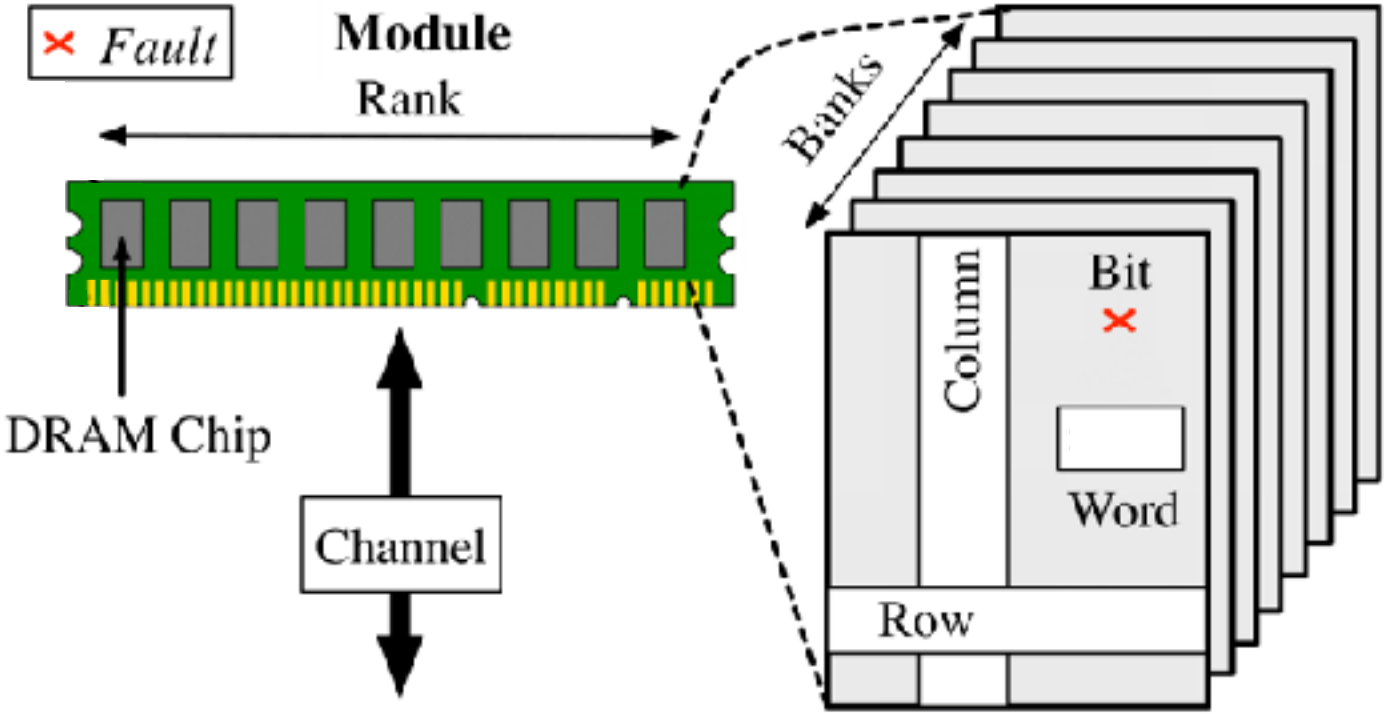
Faults in activations **overwritten** by next inference



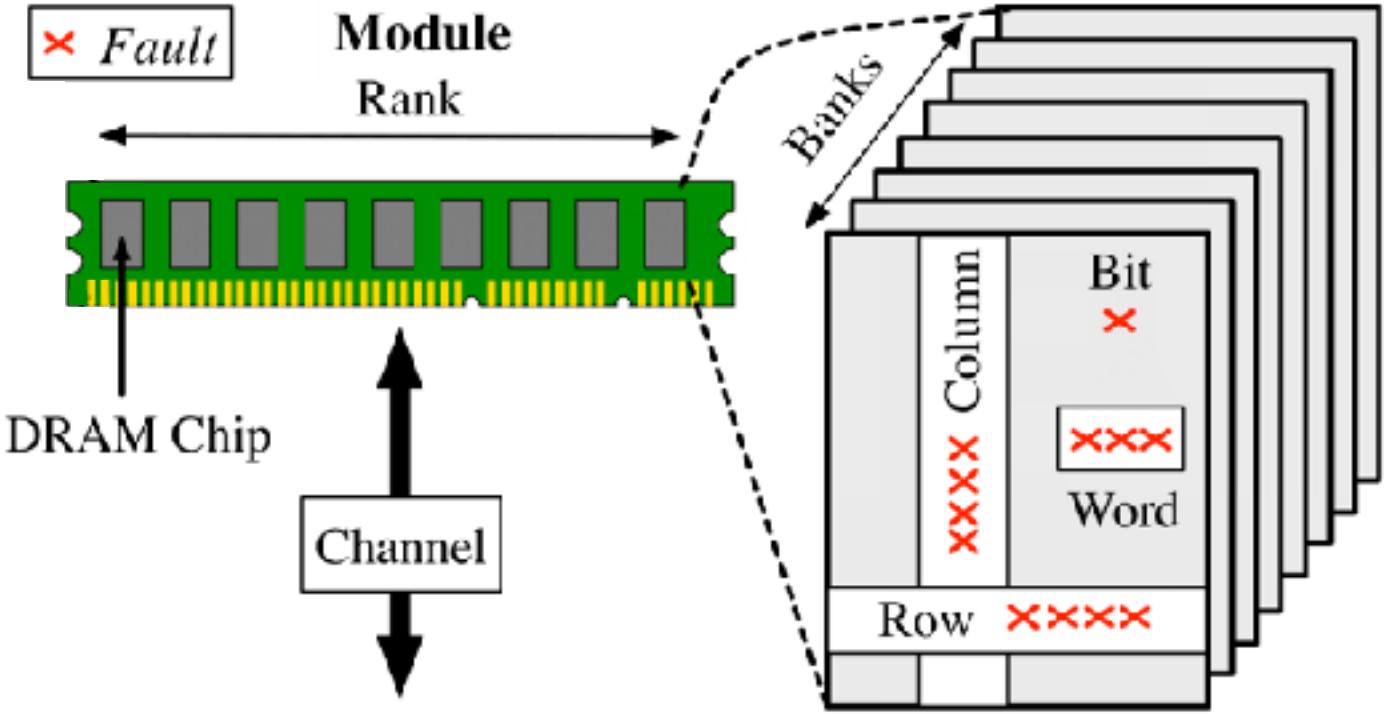
Faults in DRAM



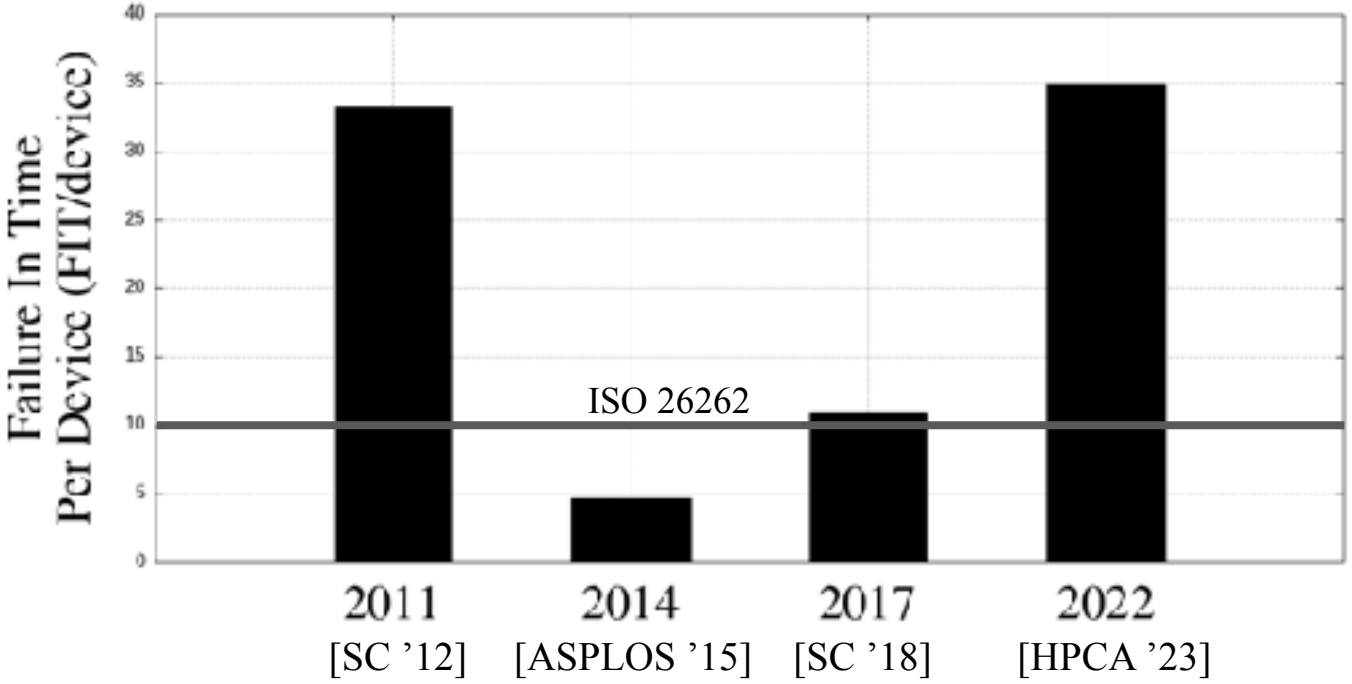
Faults in DRAM



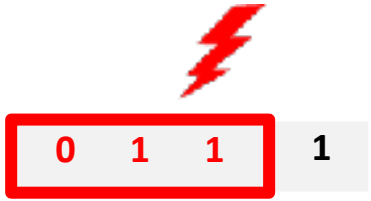
Faults in DRAM: Large-Granularity Faults



Multi-bit DRAM FIT rates (2011-2022)

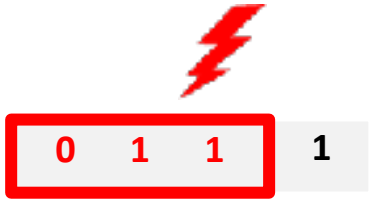


Hardware ECC



SEC-DED cannot detect/correct multi-bit (>2) faults

Hardware ECC

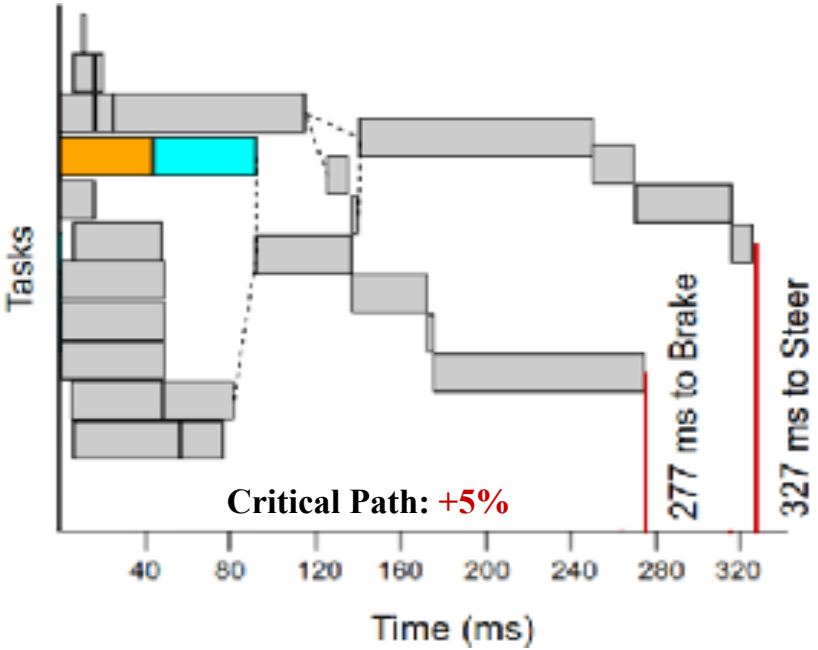


SEC-DED cannot detect/correct multi-bit (>2) faults



Chipkill is energy/bandwidth-inefficient and ~2x as expensive [Nikolaou, 2015]

Importance of Low Overhead



[Rydzewski, 2021]



Outline

Motivation and Background

Structural Coding: Approach

Experimental Evaluation

Conclusion

Outline

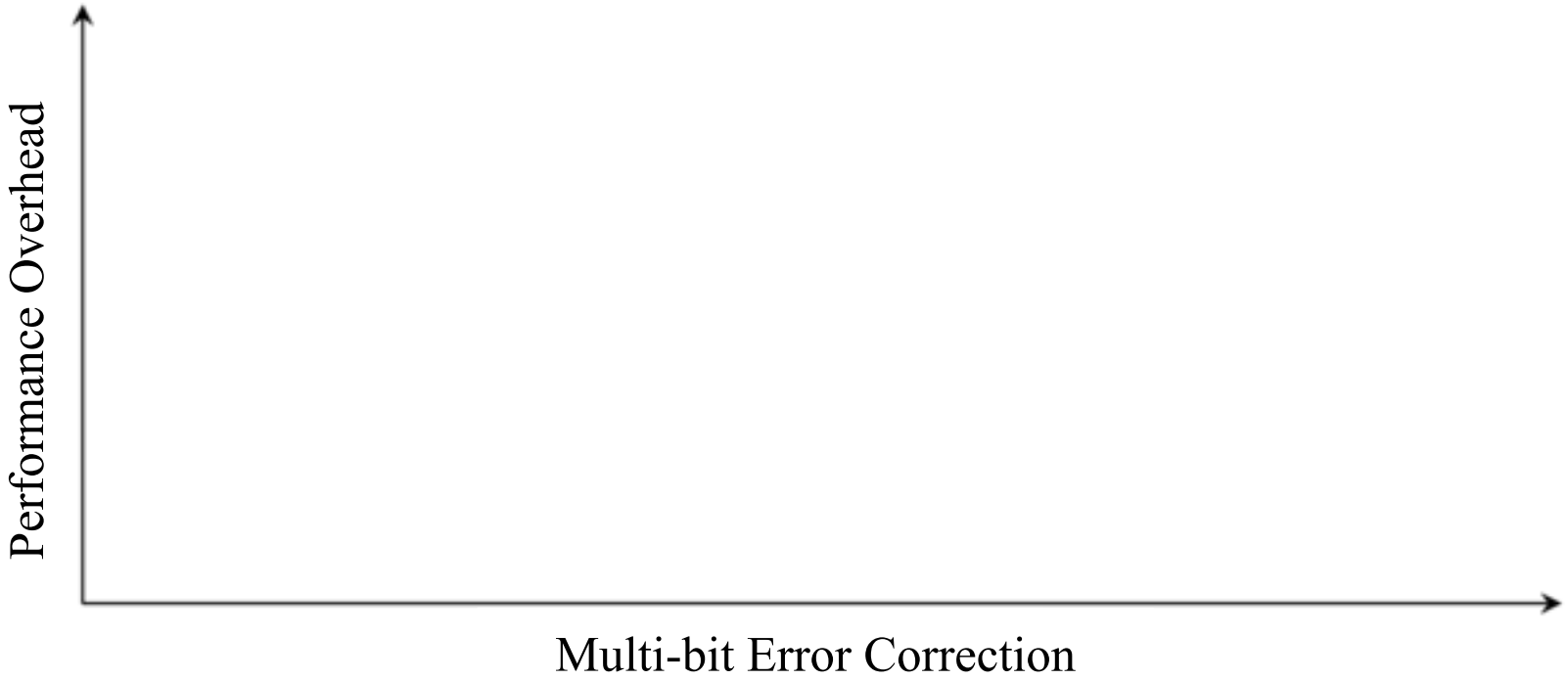
Motivation and Background

Structural Coding: Approach

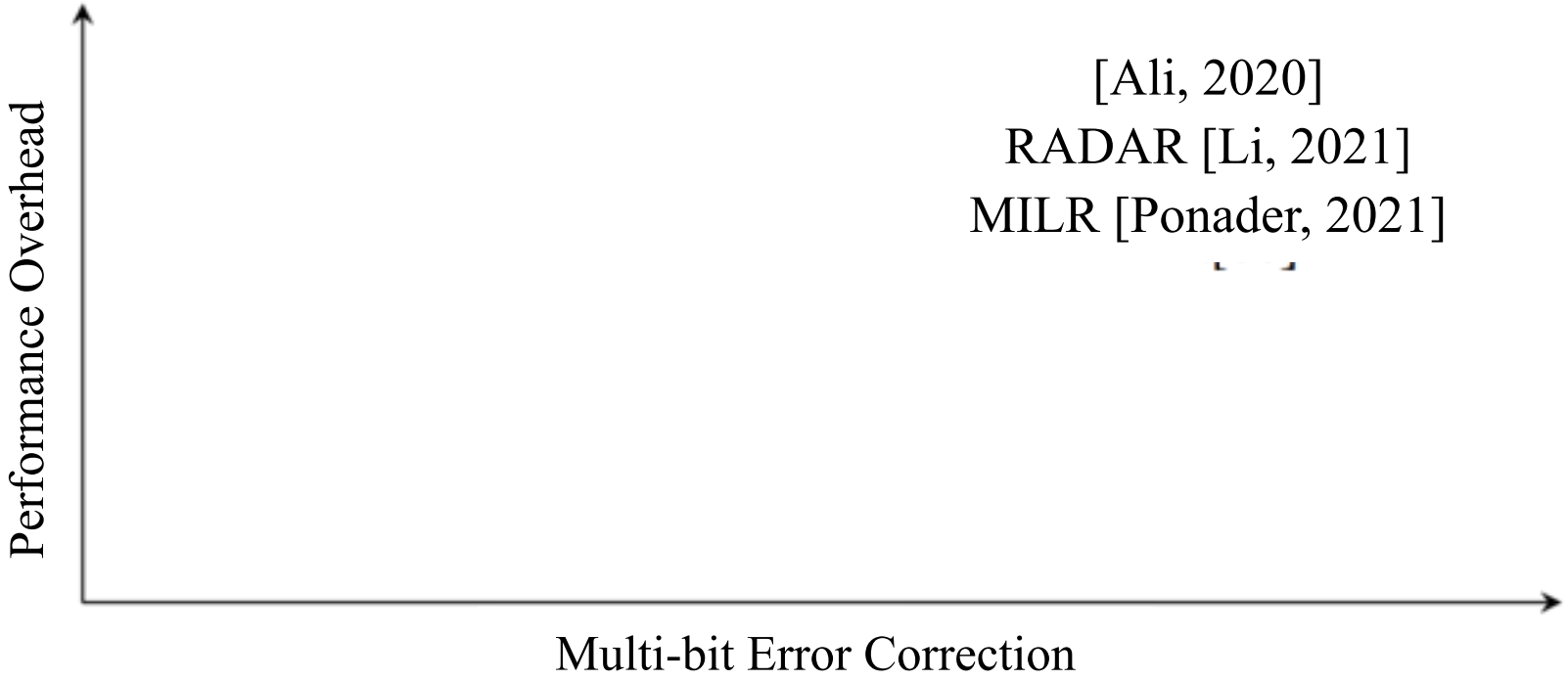
Experimental Evaluation

Conclusion

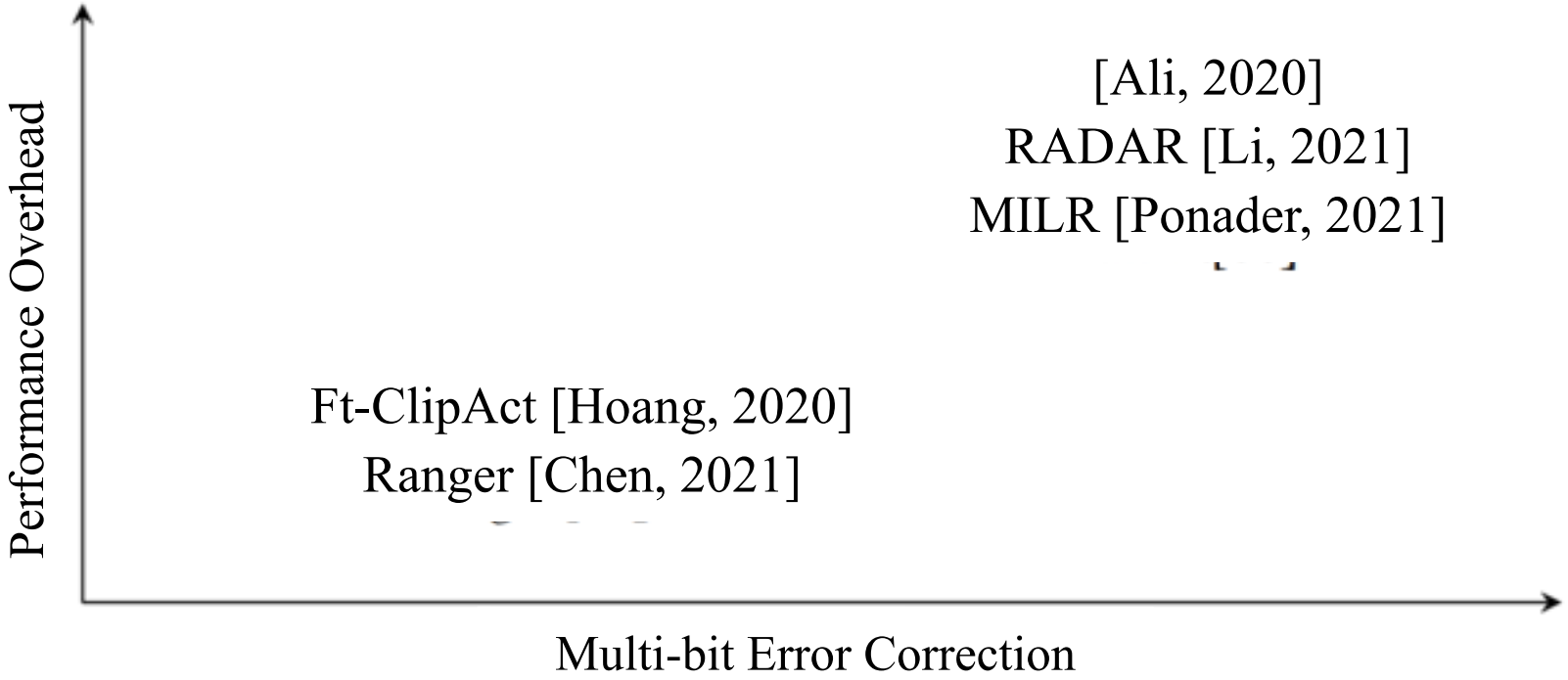
Application-specific (CNN-specific) Software ECC



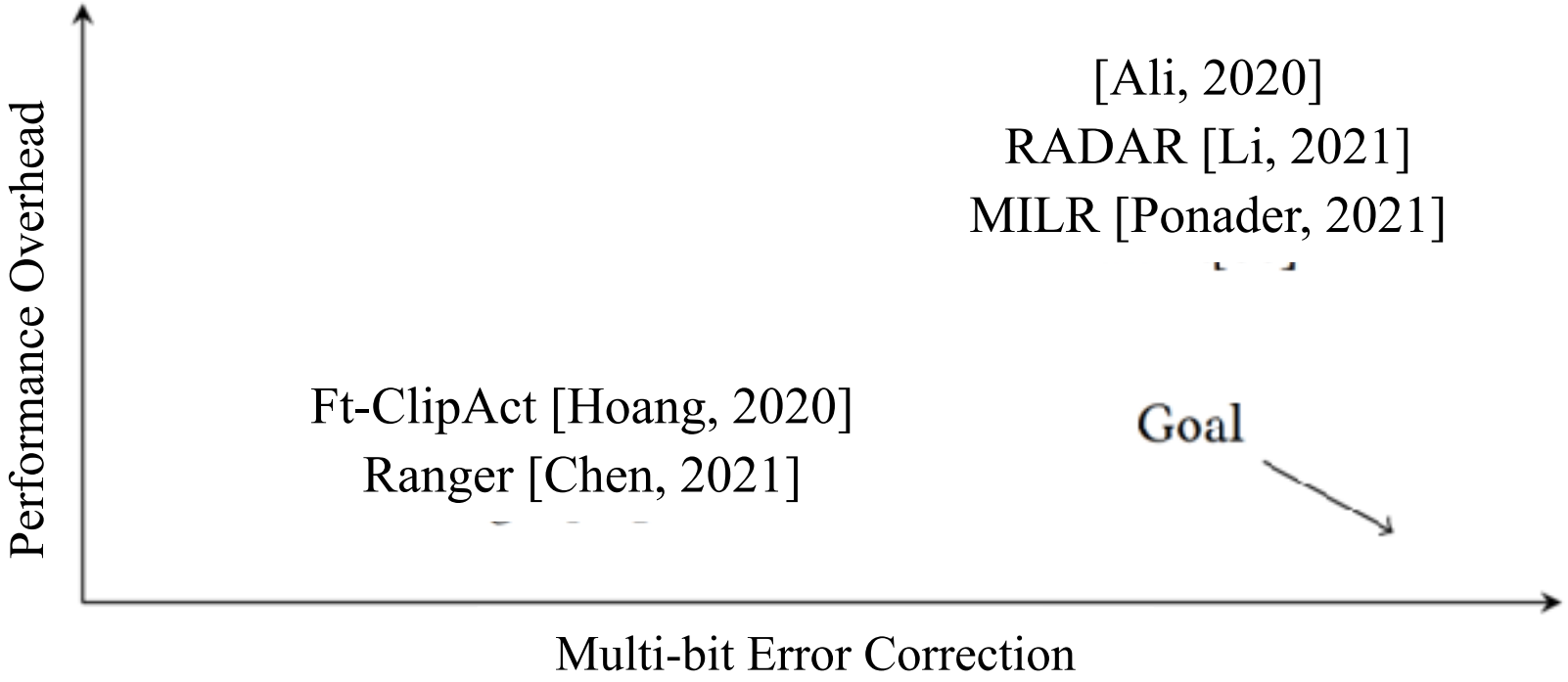
Application-specific (CNN-specific) Software ECC



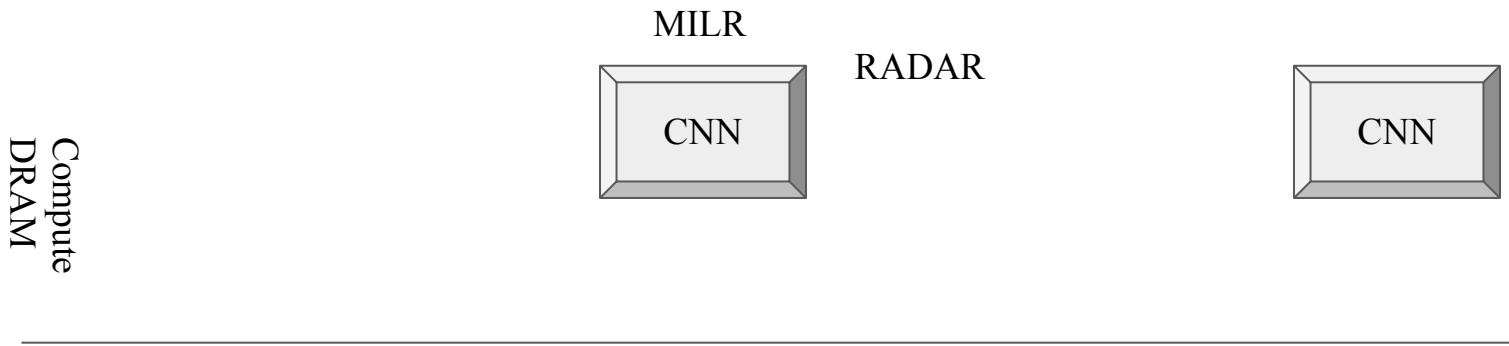
Application-specific (CNN-specific) Software ECC



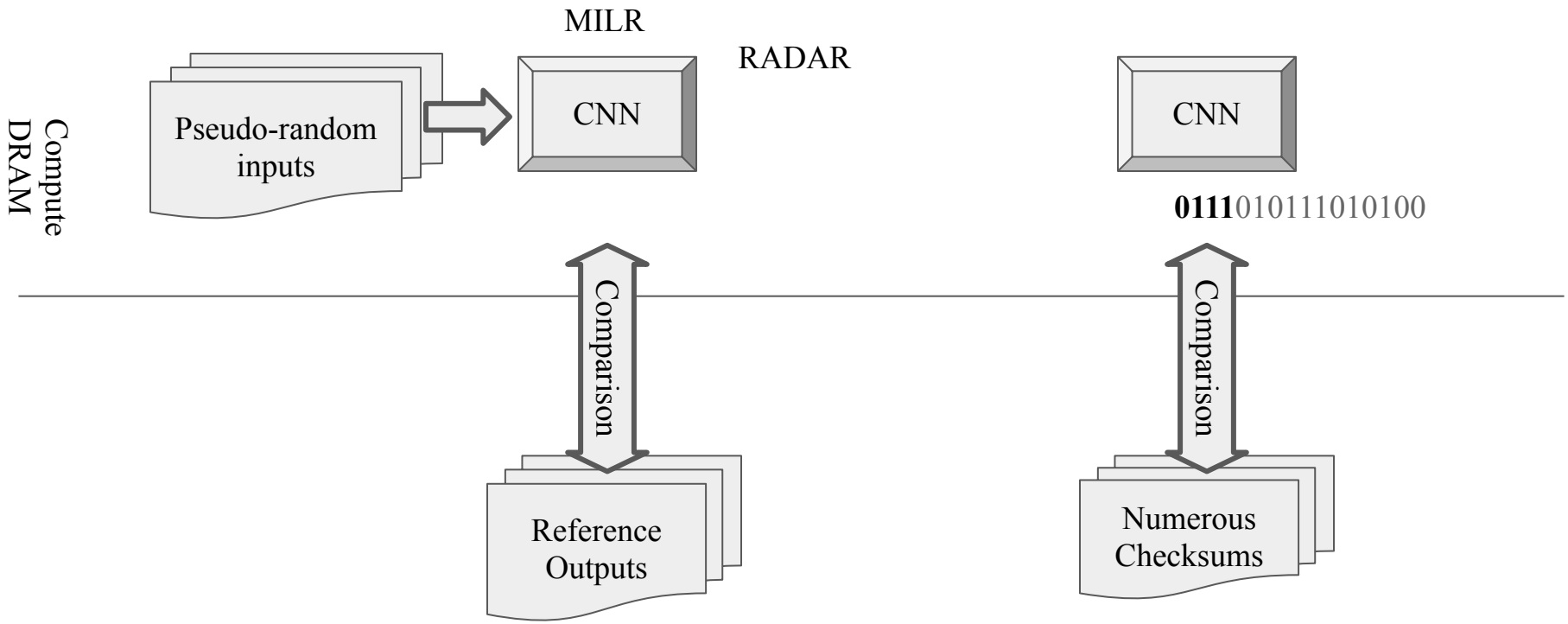
Application-specific (CNN-specific) Software ECC



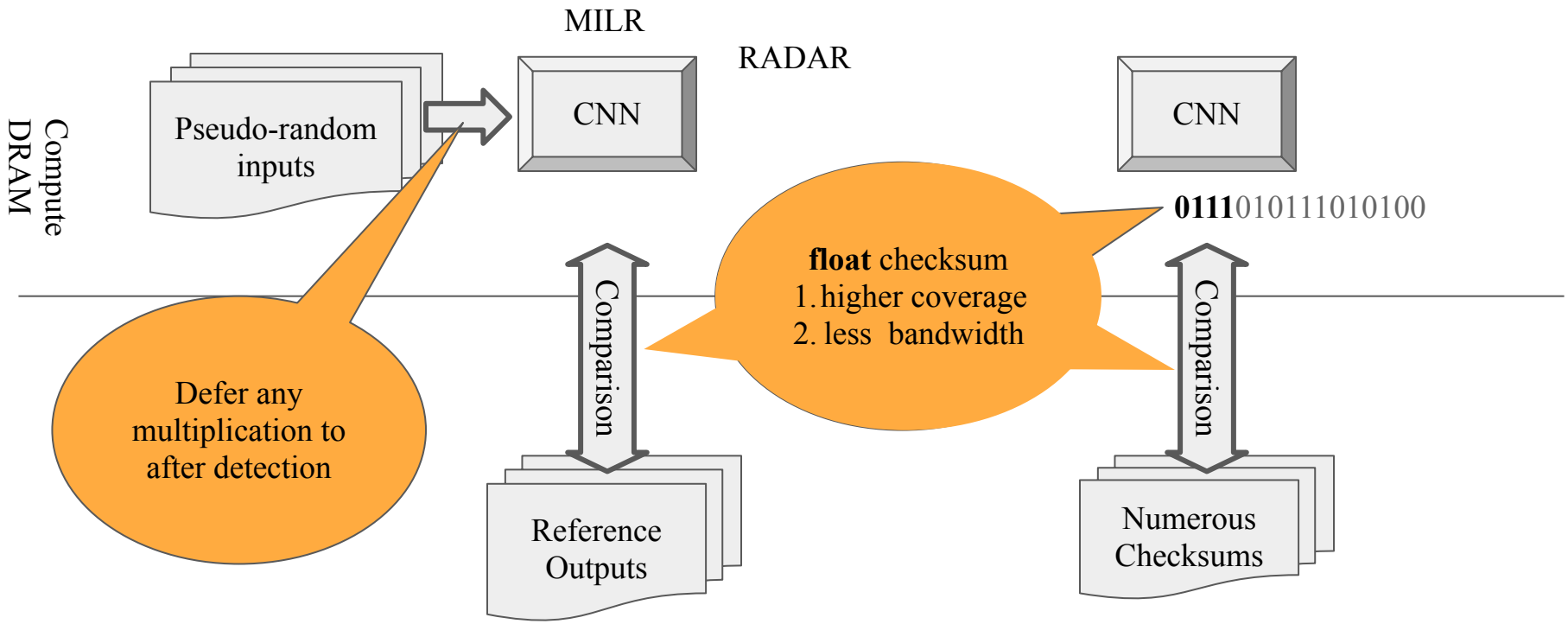
Where is the overhead?



Where is the overhead?



Structural Coding (SC)

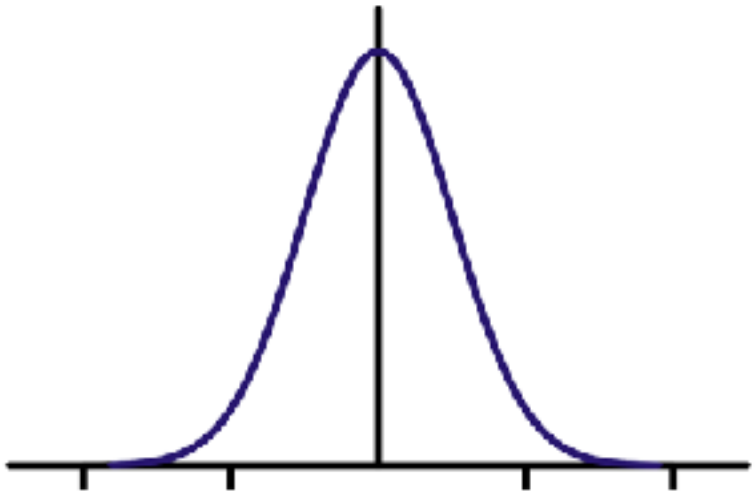


Float sum won’t overflow easily

- DNN parameters are close to zero

One checksum - many parameters

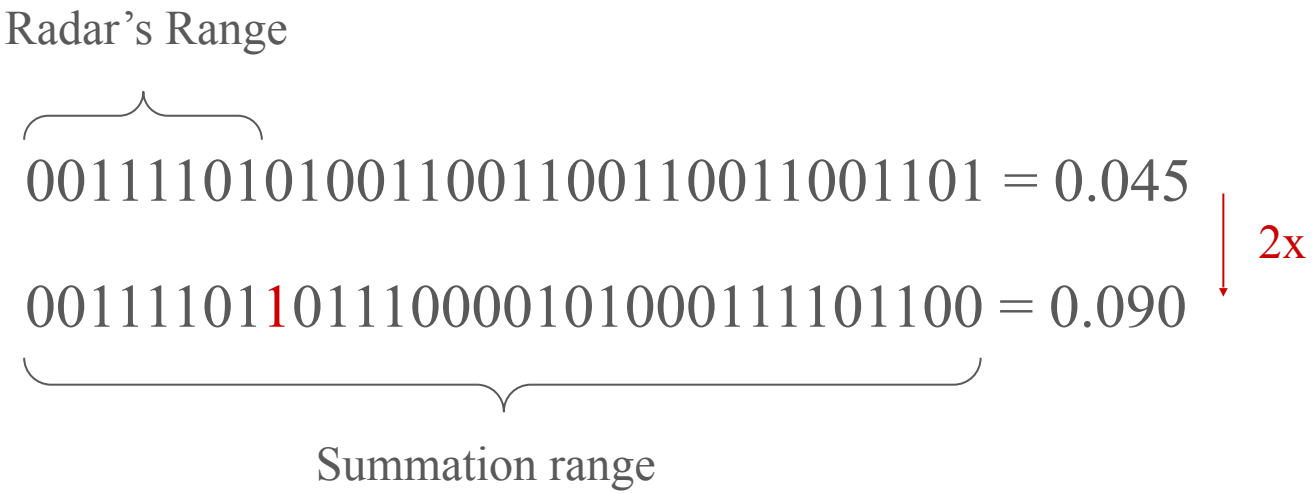
Less checksums → less overhead



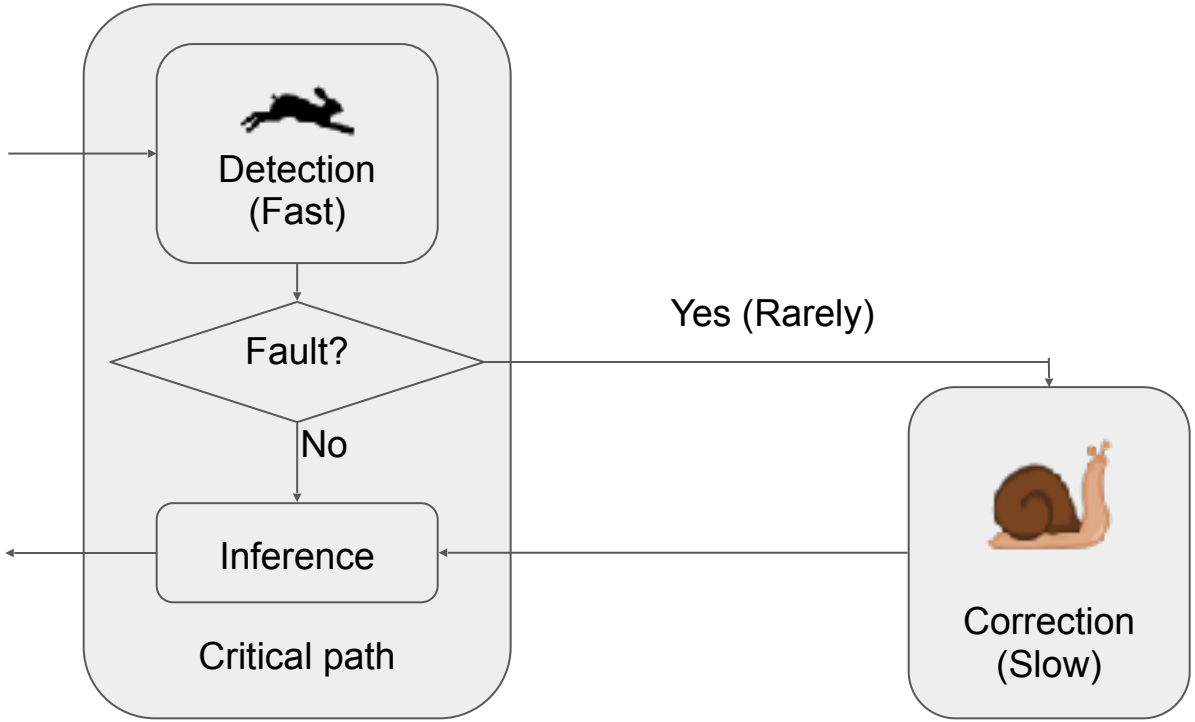
Parameters

Why are Float sums better?

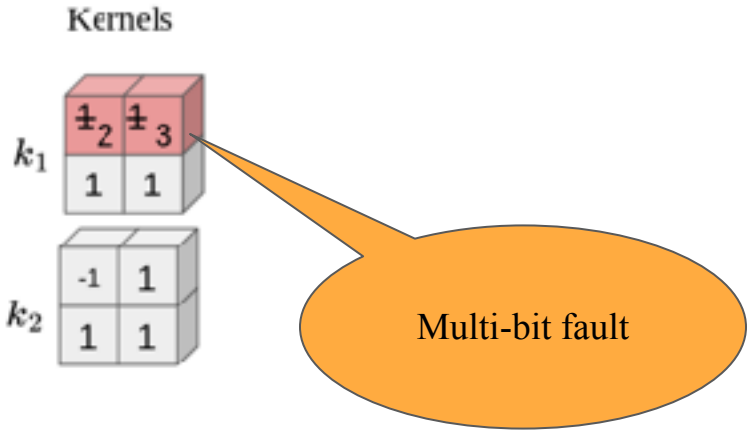
Float sums are better than XOR checksums (used by RADAR)



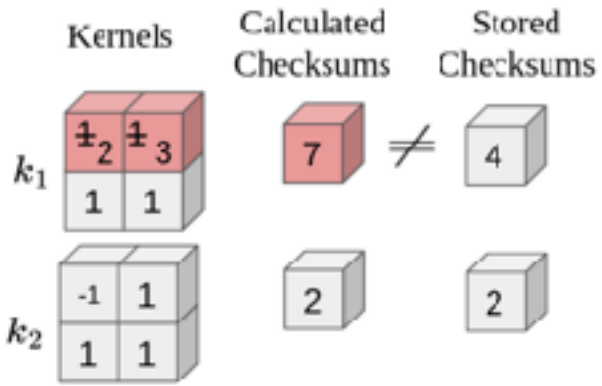
Structural Coding: Operation



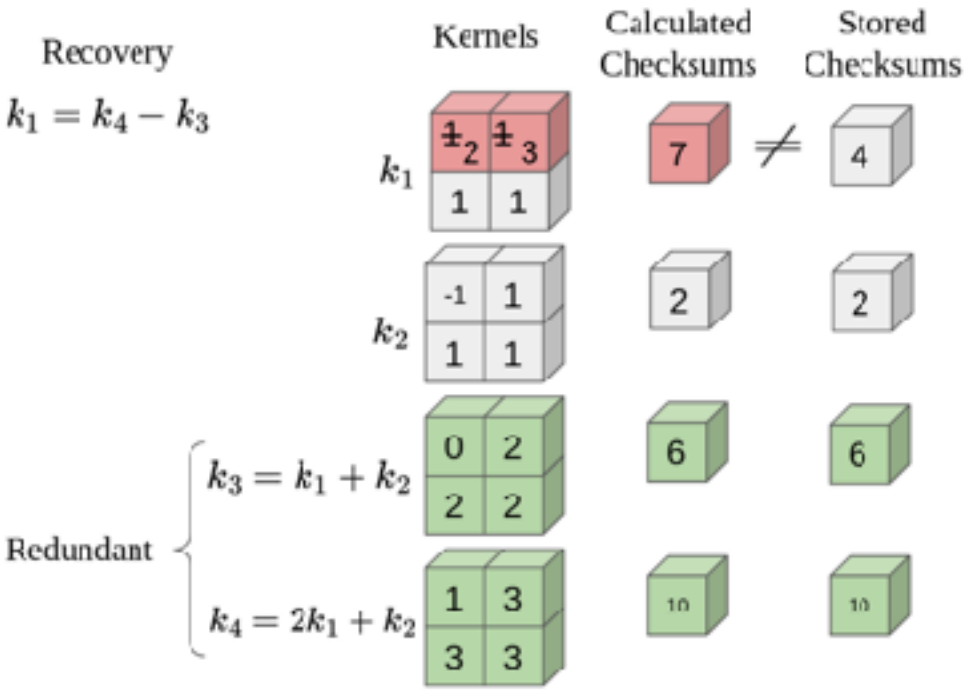
Structural Coding



Structural Coding (Detection)



Structural Coding (Recovery)



Implementation

```
model = apply_sc_automatically(
    model
)
```





https://github.com/DependableSystemsLab/structural-coding

Outline

Motivation and Background

Structural Coding: Approach

Experimental Evaluation

Conclusion

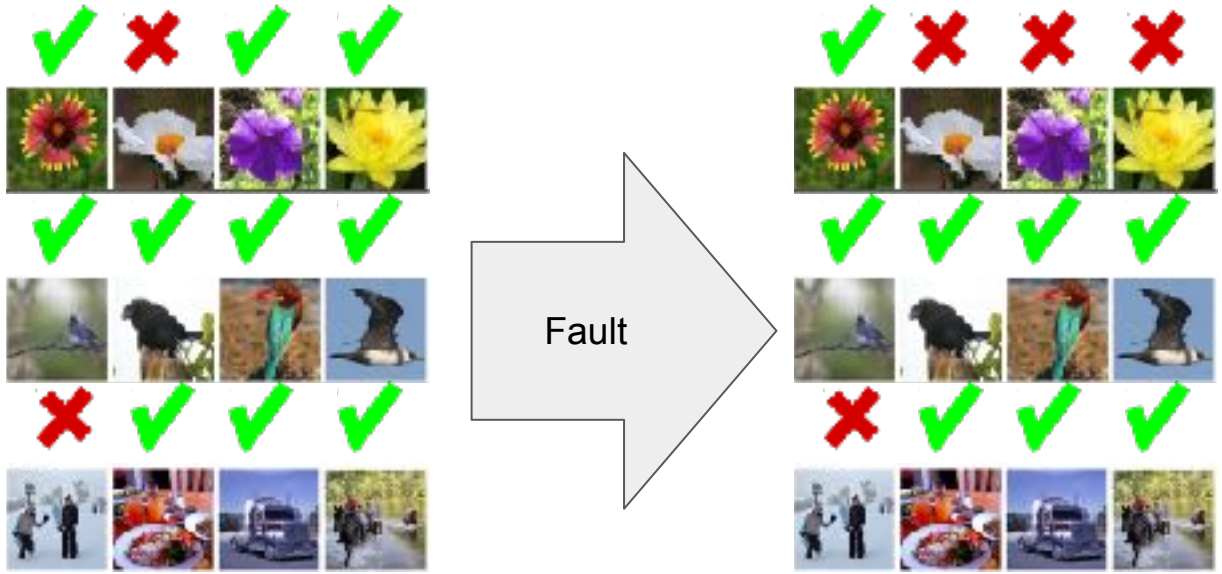
Metrics

SDC: misclassifications due to faults



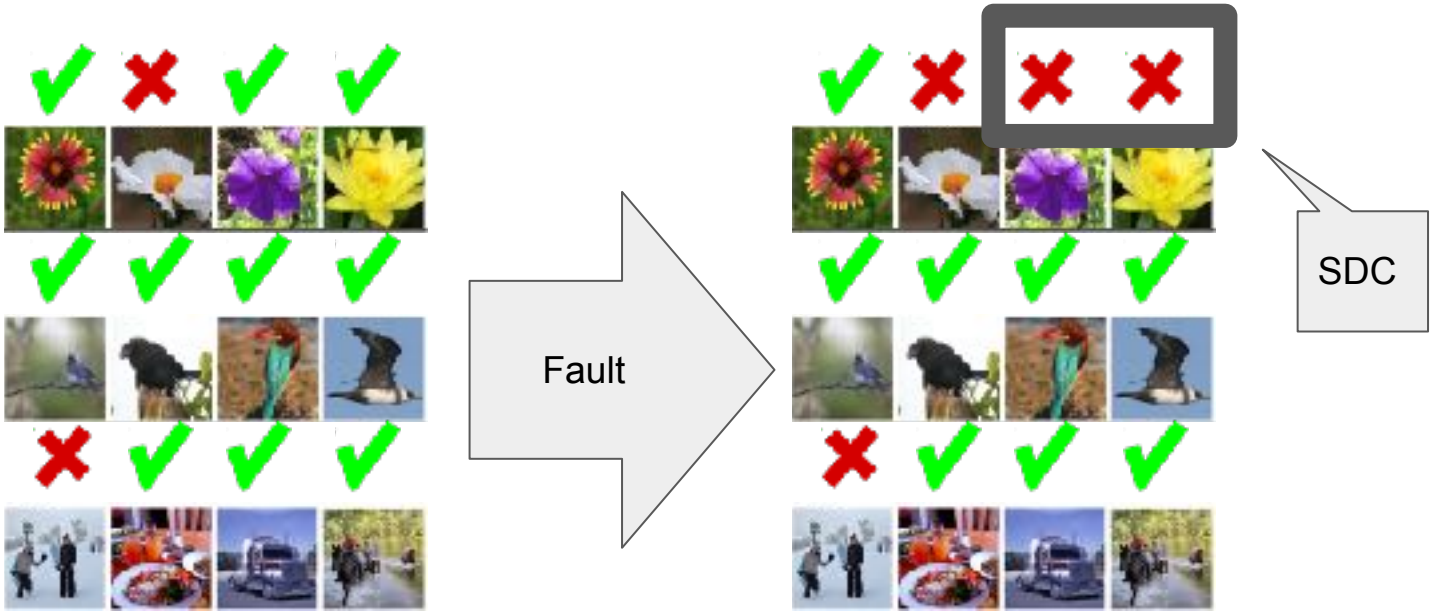
Metrics

SDC: misclassifications due to faults



Metrics

SDC: misclassifications due to faults



Metrics

SDC rate: misclassifications due to faults

Measured via Fault Injection
(1000s of runs - 95% confidence)

CPU time

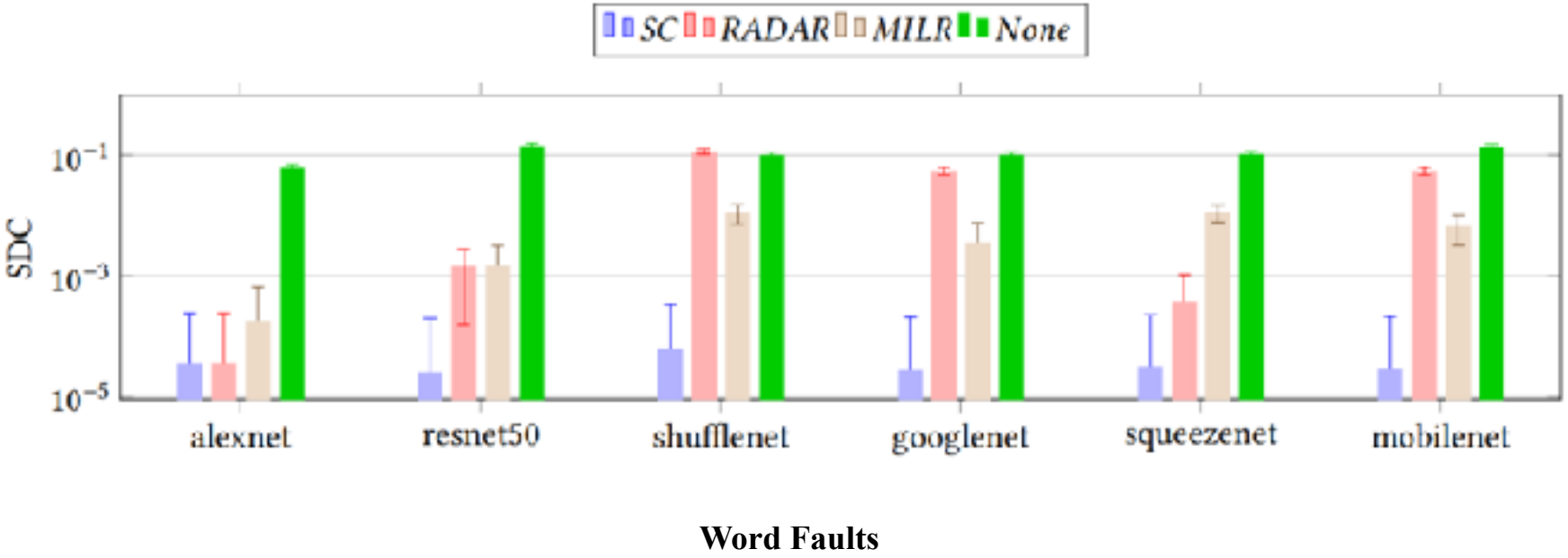
Measured on a real machine

Energy consumption (package + DRAM)

Estimated using CPU counters

SDC Coverage of Structural Coding (SC)

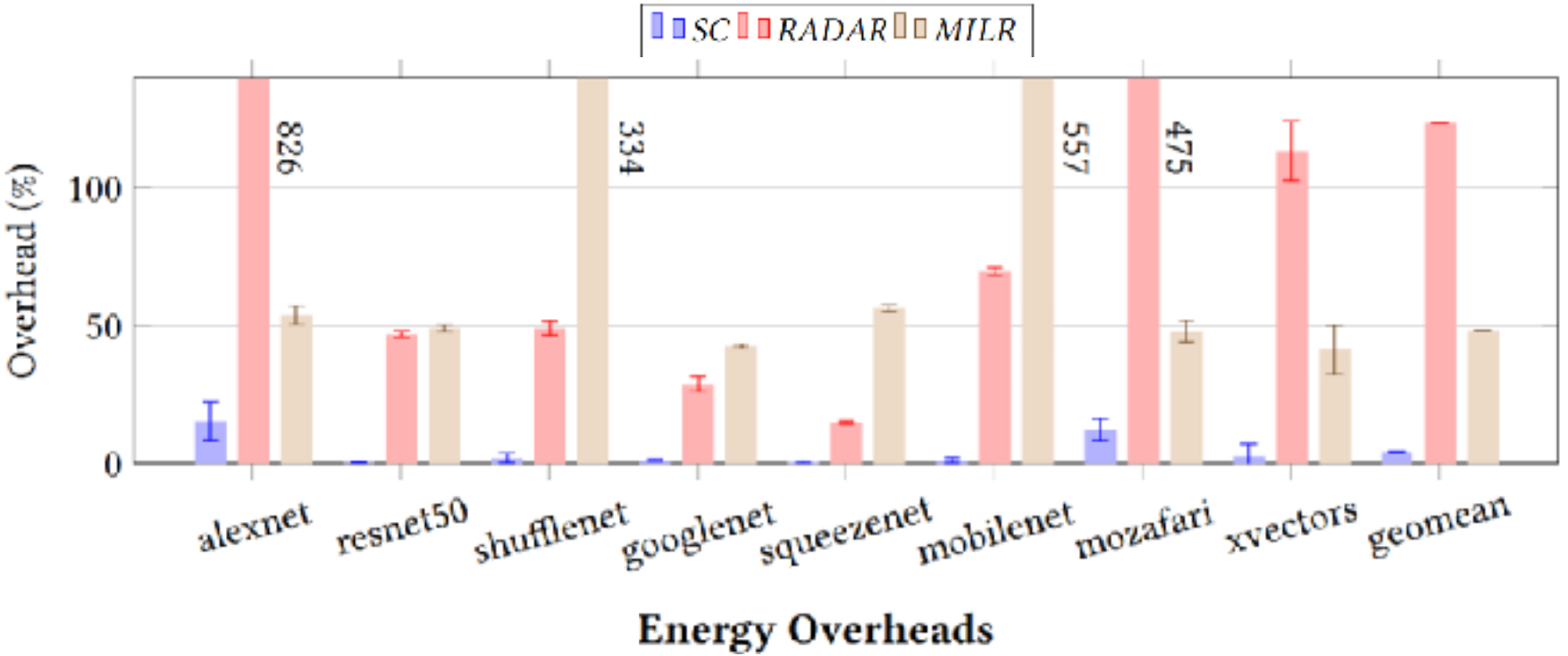
95% Confidence Intervals



Performance overheads of Structural Coding (SC)



Energy overhead of SC (Compared to Software Techniques)



Energy overhead of SC (Compared to Hardware Techniques)

Estimated using Micron's data sheet

Protection	Energy overhead (%)
Commercial Chipkill Dual-channel	32.37
Commercial Chipkill Single-channel	10.41
Structural Coding	**4.38**

Structural Coding has lower energy overheads than even hardware-based Chipkill

Outline

Motivation and Background

Structural Coding: Approach

Experimental Evaluation

Conclusion

Conclusion: Structural Coding

Goals:

Multi-bit error correction

Low overheads, Automated





Results:

Two Orders of Magnitude higher SDC coverage than prior techniques

Performance ~ 2%: Much lower than prior techniques (> 20x)

Energy < 5%: Lower than both s/w and h/w based techniques

Karthik Pattabiraman, karthikp@ece.ubc.ca