TWINRELECT

Cross-Layer Reliability Analysis of Slimmable Neural Networks under Permanent Faults

VTS 2026 – SPECIAL SESSION

N. Zazatis, A. Veronesi, K. Varakliotis, P. Tsoumanis,
C. Sotiriou, L. M. Bolzani Pöhls, M. Andjelkovic, D. Bertozzi

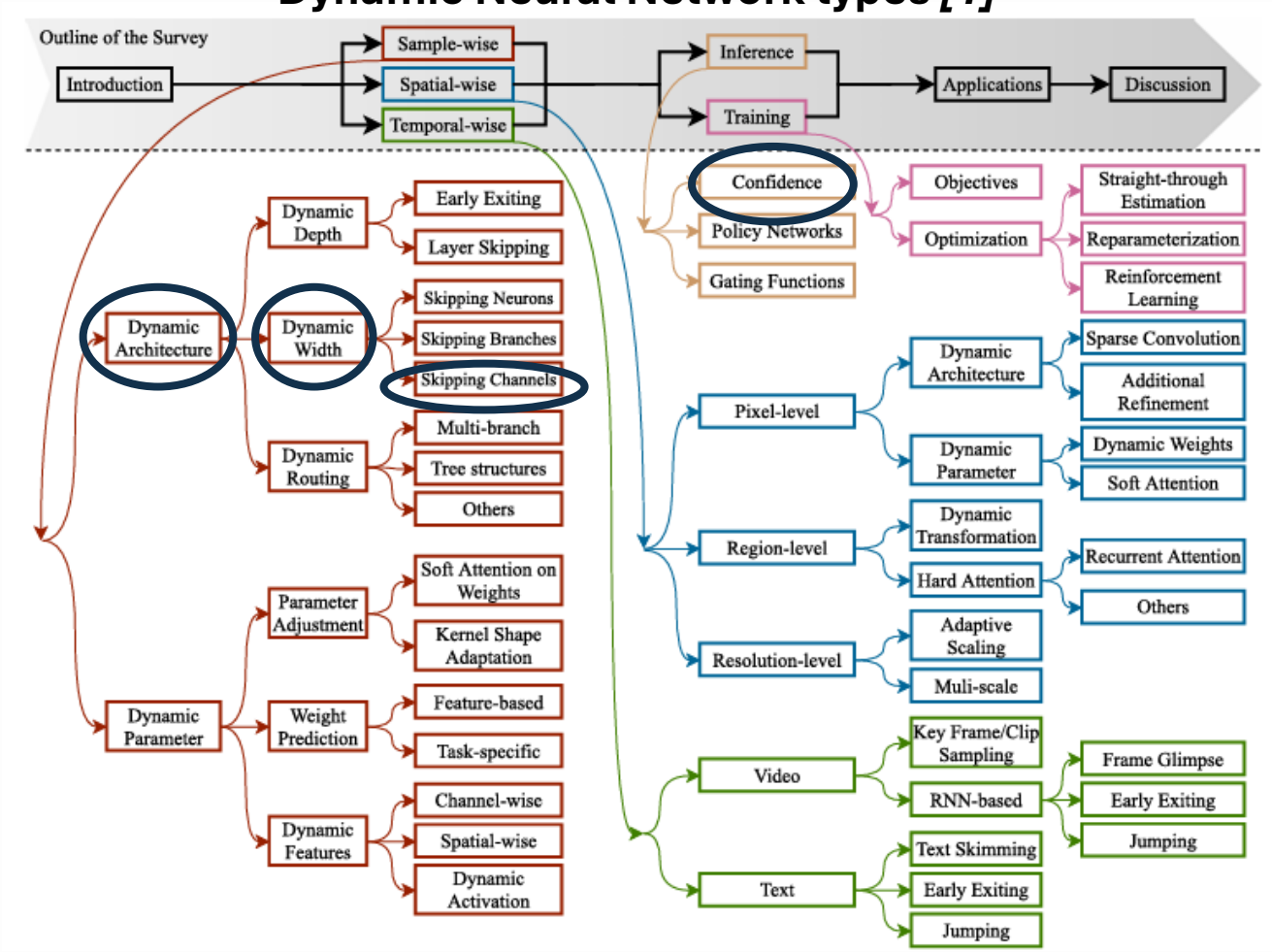
Co-funded by the European Union

UK Research and Innovation

Dynamic Neural Networks

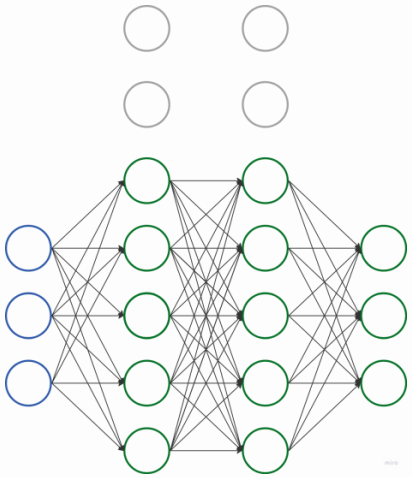
- **DyNNs** adjust their structures or parameters based on the specific input data *[1]*
- Many different flavors! *[1]*
- We focus on Skipping Channels Dynamic Width NN *[2]*
  - Known as **Slimmable NN**
- The width size is based on the **Confidence**

Dynamic Neural Network types *[1]*

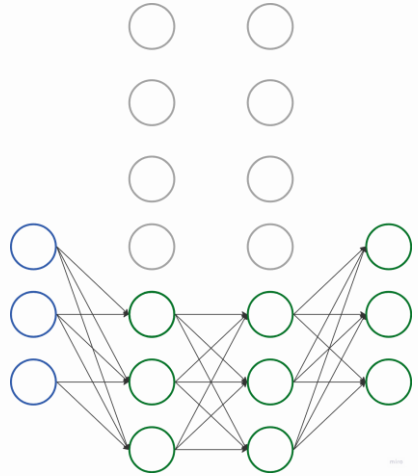


Slimmable Neural Network

- **Slimmable NN** adjust the output channels of the network based on the input
- Advantageous over other Dynamic NN types *[1, 3]*
  - ***Memory footprint***: One model is stored
  - ***Conditional Evaluation***: A simple criteria determines the width of the Network
  - ***Demonstrated Real-World Speedup*** on various hardware platforms



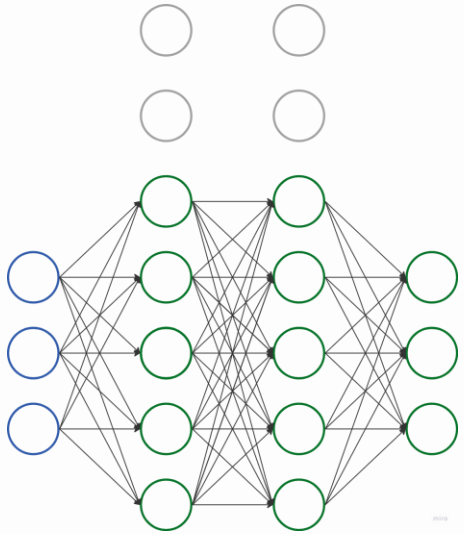
CAT



DOG

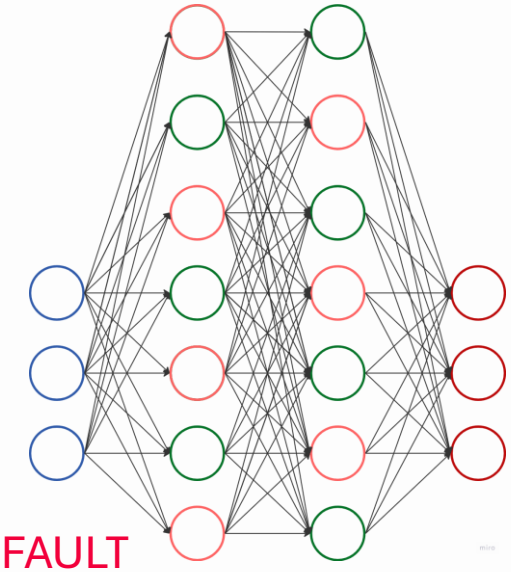
Reliability Analysis on Slimmable NN

- Reliability Analysis on Slimmable NNs is important!
  - **Accuracy Drop**: The predictive power worsens / Functional metric
    Common in Static NNs
  - **Increase in Inference Time**: The network may utilize more increments thar are needed meaning more **unnecessary** computations / Non-functional metric
    Not common in Static NNs
- We must be able to accurately estimate **both** metrics degradation!
  - *Can we quantify the accuracy loss?*
  - *Can we compensate by increasing the inference time?*



CAT

NO FAULT
EXPECTED OUTCOME & UTILISED NEURONS



DOG

HARDWARE FAULT
WRONG OUTPUT **AND / OR**
MORE UTILISED NEURONS

Fault Injection Setup

Test Reliability on NVDLA

- Open-source Architecture by NVDIA

Injecting Stuck-at 0 & 1

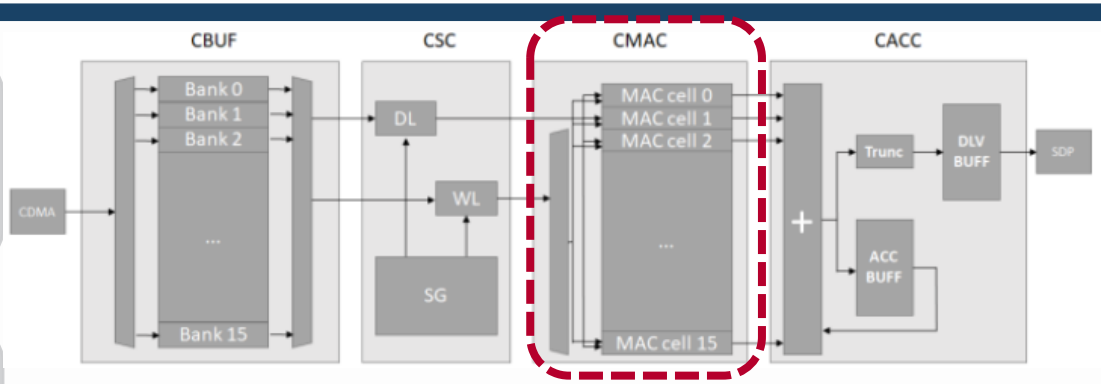
- Ware-out effects: caused by the EOL of the design

Targeting the MAC Array

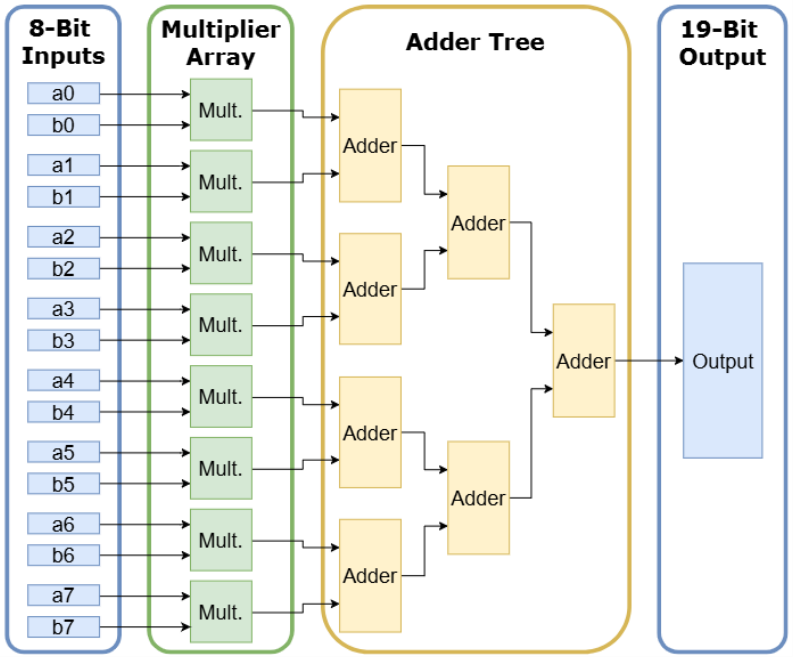
- Computing core of NVDLA

Hybrid Simulation-based FI

- Simulate fauly MAC units
- Utilize EMBER fault simulator
- Requires RTL only for MAC units
- Non faulty components are simulated with pytorch



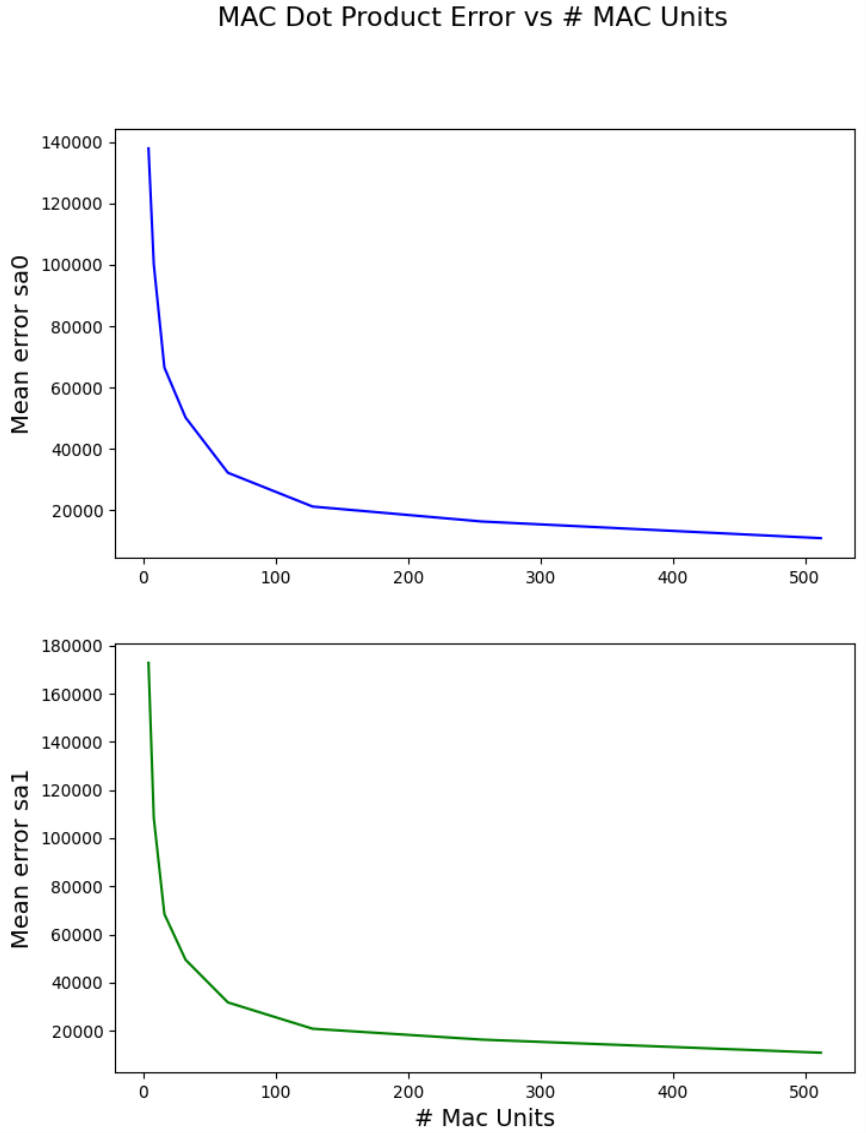
NVDLA MAC Pipeline *[10]*



NVDLA MAC Unit Architecture

Mapping MAC Units to NN Layers

- **N input** vectors -> **multiple MAC Units** are utilized
- **Matrix-Matrix** and **Matrix-Vector** multiplications are just **multiple dot product** operations
  - MAC Units are great at performing those!
  - NN operations are essentially matrix operations
- MAC Units are **NOT infinite** …
  - The available MAC Units must be mapped (tiling) to the needed operations / layers
  - Leads to **hardware reusage**!
  - Less MAC Units -> More reuse
  - A **single fault** in a **single MAC Unit** is present at
    - multiple layers
    - multiple operations in a single layer
  - **A single inference requires multiple RTL simulations**



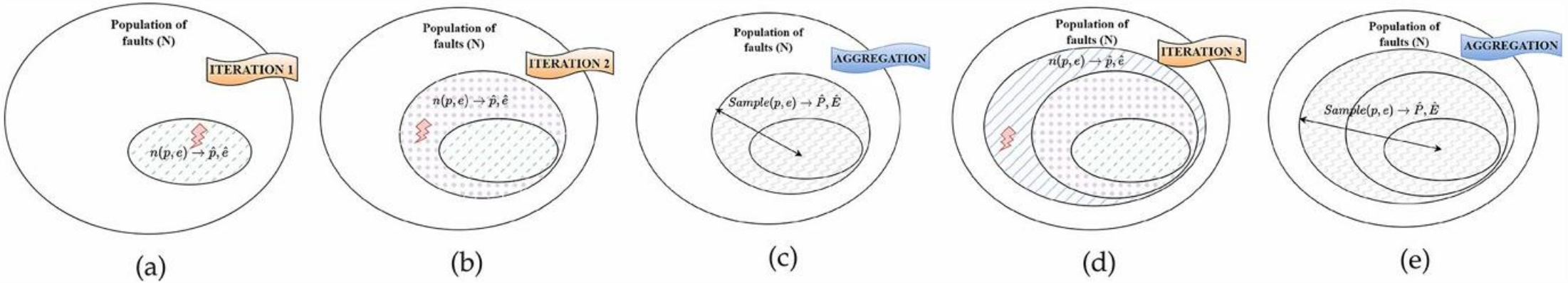
Statistical Fault Injection

MAC Unit has 1642 gates

Fault Injection Candidates: 1642 times the number of MAC units

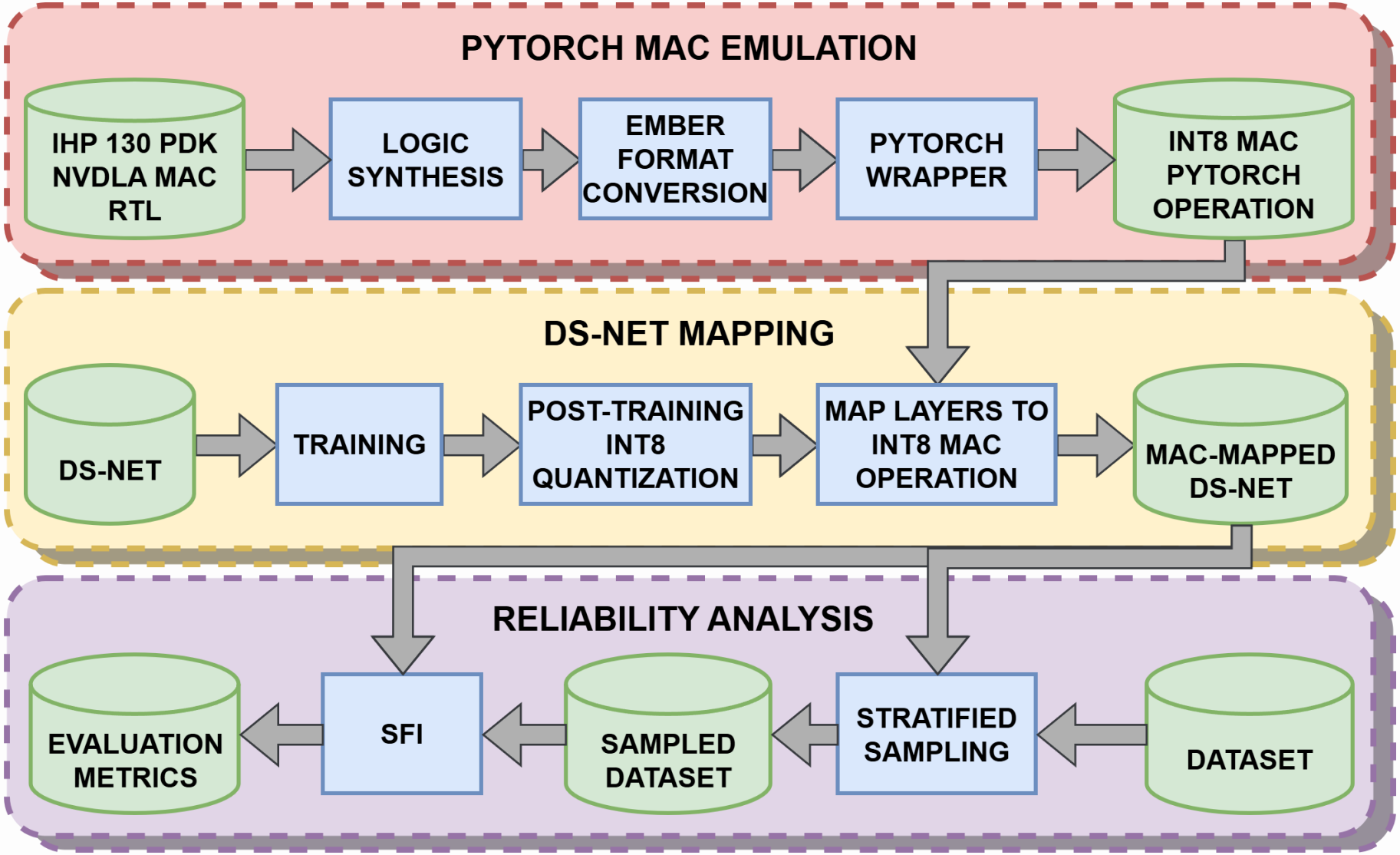
Impossible to inject fault at every candidate

Solution: Statistical Fault Injection *[12]*

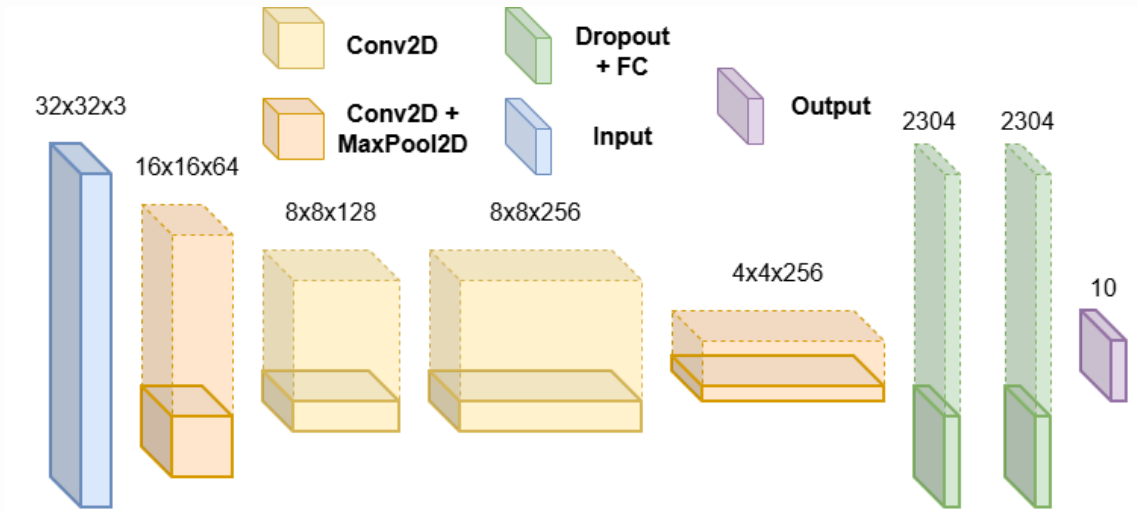


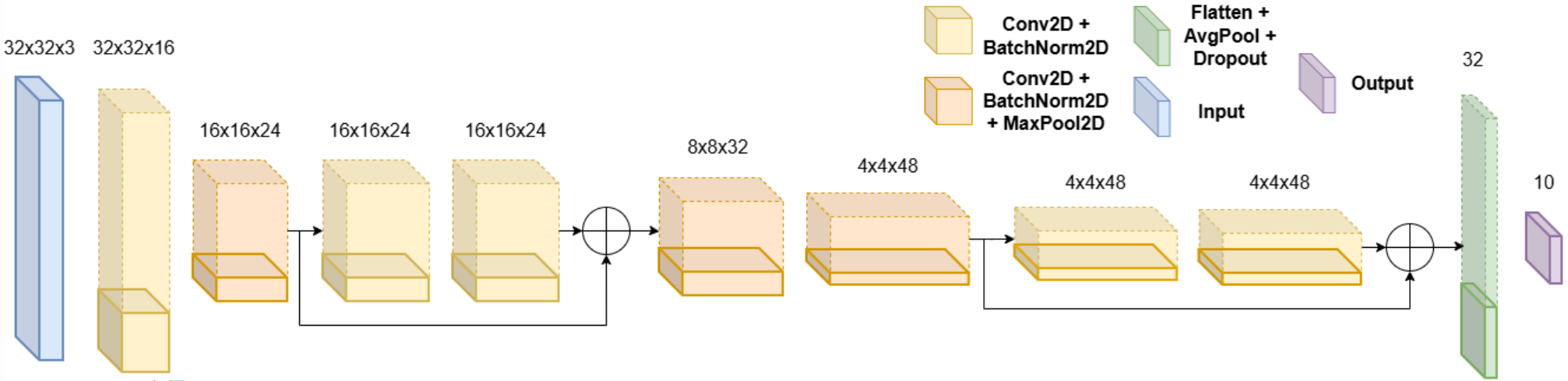
Source: Ruospo, Annachiara, et al. "An effective iterative statistical fault injection methodology for deep neural networks." IEEE Transactions on Computers (2025).

Methodology



Experimental Setup I



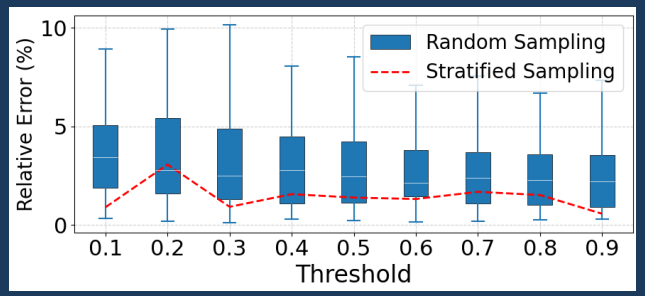


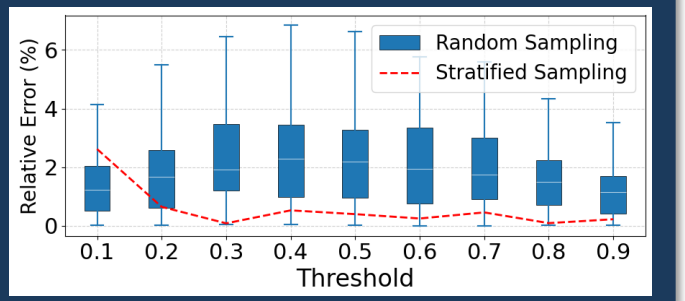
- **Smaller - pruned** version of **AlexNet & ResNet9**
  - DS-AlexNet
  - DS-ResNet
- Deployed **4** different **channels / increments**
  - **0.25x, 0.5x, 0.75x, 1x**
- Configured for **cifar10** dataset
  - 10000 test images
  - 32x32x3 size images
  - 10 output classes

Experimental Setup II

Stratified Dataset Sampling

Size 10000	Dataset Size
N 200	Number of samples
Deviation <5%	Deviation to original dataset





Statistical Fault Injection

Estart 5%	Initial error threshold
Egoal 2%	Target error threshold
Confidence 99%	Statistical confidence

Simulation Cost Reduction

131M → 124,800 inferences (8 MAC)
262M → 127,800 inferences (16 MAC)

≈ 99.9% reduction

Experimental Results I

SDC Rates exceeding 50% for most threshold values

- Much higher than hardware-agnostic approaches report

Different behavior for unique architectures

- DS-Alexnet SDC spikes at T=0.2 and exhibits a plateu for T > 0.2
- DS-Resnet9 SDC spikes at T=0.7 and exhibits a ringing effect across all T values

Similar behavior between fault models

- SA0 and SA1 exhibit almost identical SDC values for a given architecture

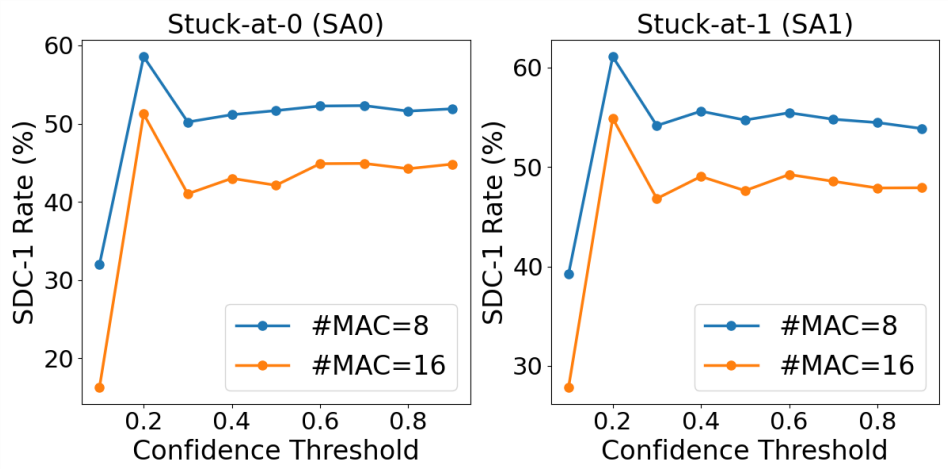
Significantly lower SDC at Threshold == 0.1

- Less utilized neurons
- Less faulty hardware reuse

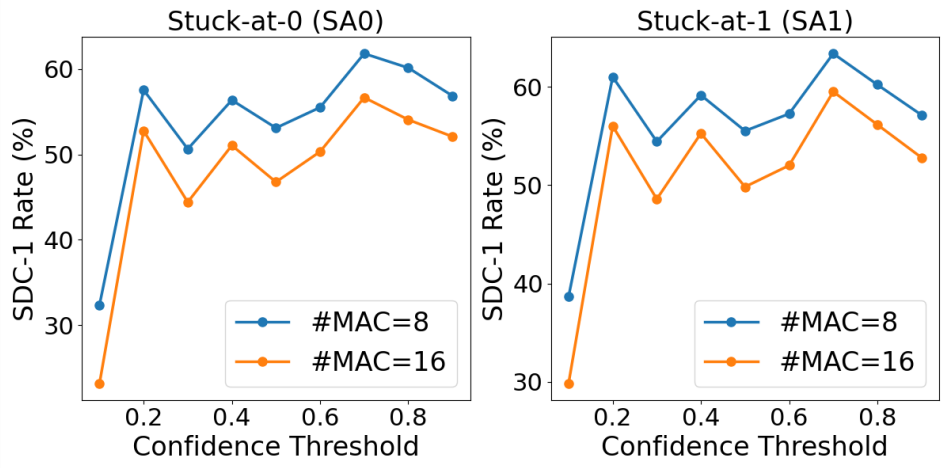
16 MAC Unit configuration is less error prone

- Similar behavior with 8 MAC Unit configuration but at lower level

DS-Alexnet

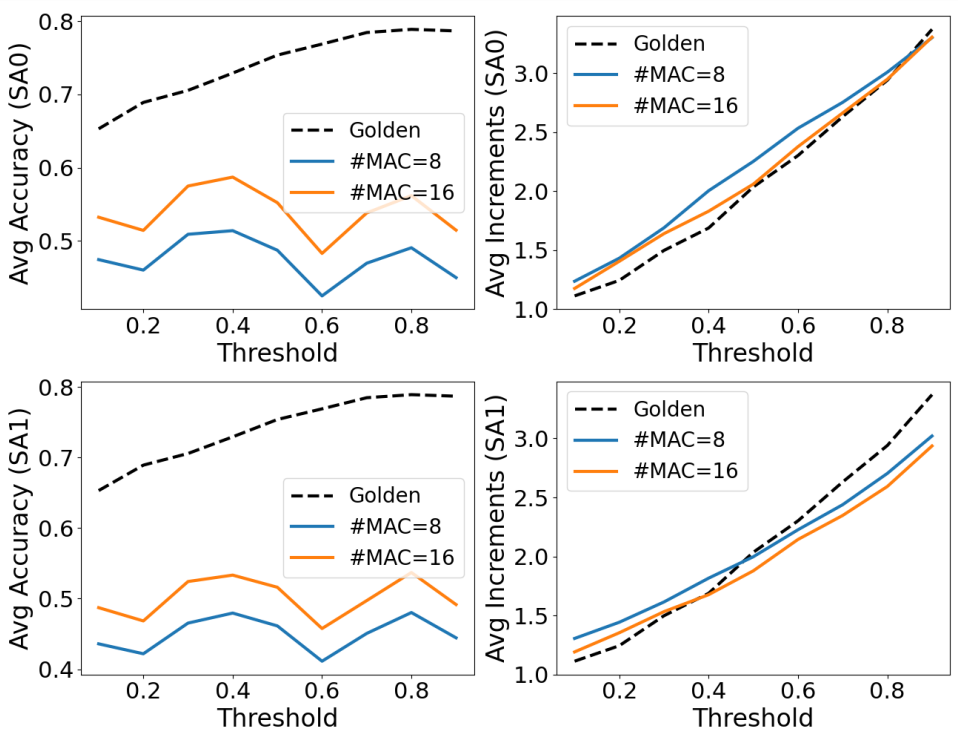


DS-Resnet9

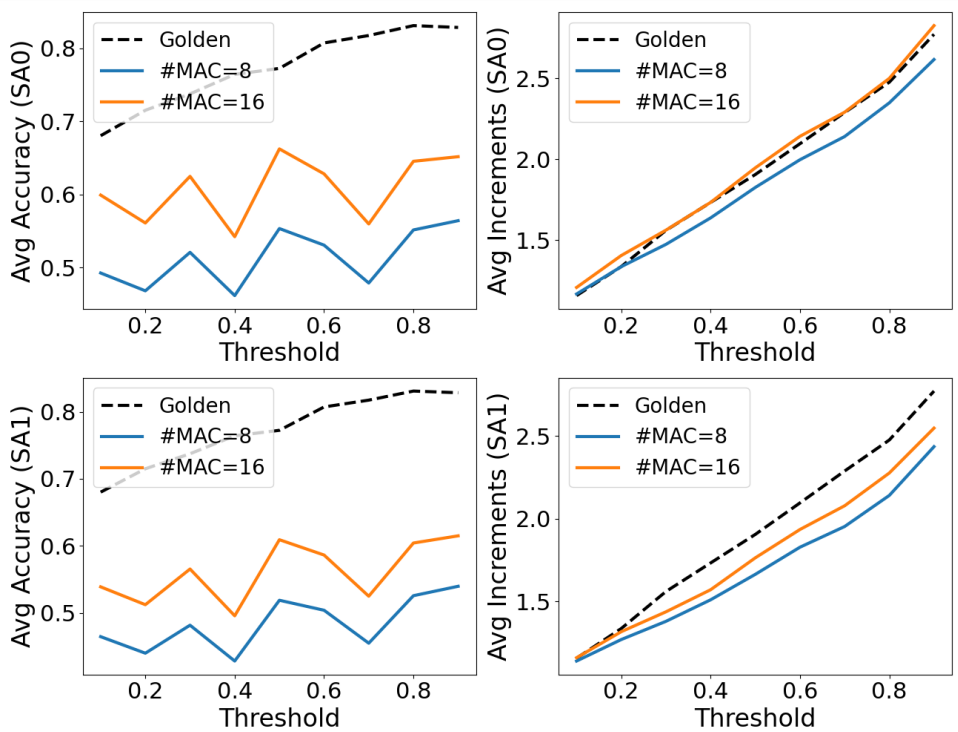


Experimental Results II

DS-Alexnet



DS-Resnet9



Significant Accuracy Drop

Accuracy is not a monotonic function of Threshold

Faulty MAC can both under- and overestimate the utilized neurons

Utilized Increments is a monotonic function of Threshold

More MAC Units lead metrics closer to Golden

Key Takeaways

Hardware Abstraction is Dangerous

References

1. *Y. Han, G. Huang, S. Song, L. Yang, H. Wang and Y. Wang, "Dynamic Neural Networks: A Survey," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 11, pp. 7436-7456, 1 Nov. 2022, doi: 10.1109/TPAMI.2021.3117837.*
2. *H. Tann, S. Hashemi, R. I. Bahar and S. Reda, "Runtime configurable deep neural networks for energy-accuracy trade-off," 2016 International Conference on Hardware/Software Codesign and System Synthesis (CODES+ISSS), Pittsburgh, PA, USA, 2016, pp. 1-10.*
3. *Li, Changlin, et al. "Dynamic slimmable network." Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. 2021.*
4. *Ziade, Haissam, Rafic A. Ayoubi, and Raoul Velazco. "A survey on fault injection techniques." Int. Arab J. Inf. Technol. 1.2 (2004): 171-186.*
5. *Salih, Nadir K., et al. "A survey on software/hardware fault injection tools and techniques." 2022 IEEE Symposium on Industrial Electronics & Applications (ISIEA). IEEE, 2022.*
6. *A. Ruospo, E. Sanchez, L. M. Luza, L. Dilillo, M. Traiola and A. Bosio, "A Survey on Deep Learning Resilience Assessment Methodologies," in Computer, vol. 56, no. 2, pp. 57-66, Feb. 2023, doi: 10.1109/MC.2022.3217841.*
7. *Mohammad Hasan Ahmadilivani, Mahdi Taheri, Jaan Raik, Masoud Daneshtalab, and Maksim Jenihhin. 2024. A Systematic Literature Review on Hardware Reliability Assessment Methods for Deep Neural Networks. ACM Comput. Surv. 56, 6, Article 141 (June 2024), 39 pages. https://doi.org/10.1145/3638242*
8. *D. Xu et al., "Reliability Evaluation and Analysis of FPGA-Based Neural Network Acceleration System," in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 29, no. 3, pp. 472-484, March 2021, doi: 10.1109/TVLSI.2020.3046075.*
9. *A. Ruospo et al., "Assessing Convolutional Neural Networks Reliability through Statistical Fault Injections," 2023 Design, Automation & Test in Europe Conference & Exhibition (DATE), Antwerp, Belgium, 2023, pp. 1-6, doi: 10.23919/DATE56975.2023.10136998.*
10. *A. Veronesi, F. Dall'Occo, D. Bertozzi, M. Favalli and M. Krstic, "Exploring Software Models for the Resilience Analysis of Deep Learning Accelerators: the NVDLA Case Study," 2022 25th International Symposium on Design and Diagnostics of Electronic Circuits and Systems (DDECS), Prague, Czech Republic, 2022, pp. 142-147, doi: 10.1109/DDECS54261.2022.9770169.*
11. *Veronesi, Alessandro, et al. "Ember: A cycle-based framework for early-stage reliability assessment in parametric rtl designs." 2025 IEEE 34th Asian Test Symposium (ATS). IEEE, 2025.*
12. *Ruospo, Annachiara, et al. "An effective iterative statistical fault injection methodology for deep neural networks." IEEE Transactions on Computers (2025).s*

Thank you!

Contact Details:
znikolaos-g@uth.gr

SCAN ME

TWINRELECT
TWINNING FOR EXCELLENCE IN RELIABLE ELECTRONICS

UNIVERSITY OF THESSALY FOUNDED 1984

CAS lab
University of Thessaly
Greece

Co-funded by the European Union

UK Research and Innovation

Grant Agreement Nº 101160314