Leibniz Institute for High Performance Microelectronics

AI and Microelectronics – Reliability View

Volos, Twin-Relect

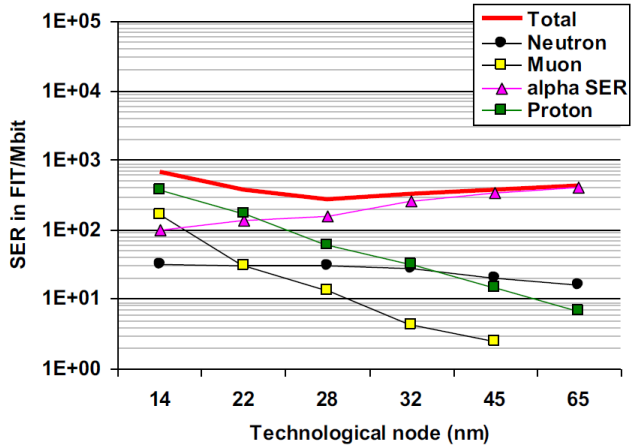
Prof. Dr.-Ing. Milos Krstic

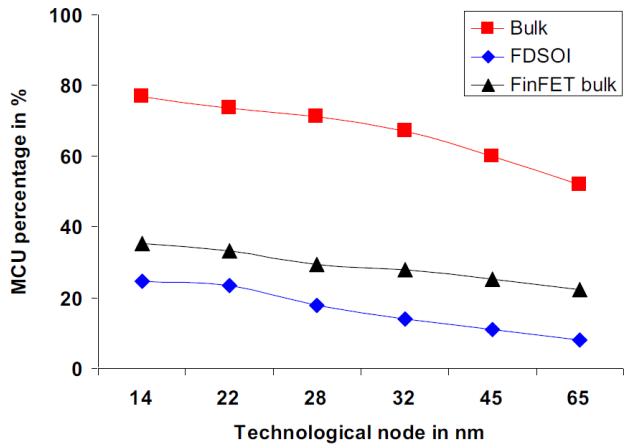
July-25

Introduction

- Soft errors are increasingly becoming a problem with technology scaling of CMOS and enhanced system integration
  - The effects that have not been so common in the past (mutliple cell upsets) are now almost dominant

- Emerging technologies are introducing new reliability and variability issues (example RRAM)





Source: G. Hubert, L. Artola, D. Regis, Impact of scaling on the soft error sensitivity of bulk, FDSOI and FinFET technologies due to atmospheric radiation, *INTEGRATION, the VLSI journal,* http://dx.doi.org/10.1016/j.vlsi.2015.01.003

Deep Learning in Safety Critical Applications

Autonomous Control Requires Intelligence

- Automotive Industry
  - Heavily working on DL based autonomous systems
  - Use cases: Path planning, object detection, lane detection, depth estimation, etc.
- Space Industry
  - Increasing the level of autonomy
    - Recently MARS rover name ‚Perseverance‘ Landed on Mars autonomously
    - **11-minute** communication delay between Earth and Mars
  - Space missions become increasingly frequent and complex.
  - Growing demand for fast and self-adjusting DL based autonomous handling and navigation capabilities.

Automated Driving Scenario [2]

Space Mission [3]

[2] https://becominghuman.ai/computer-vision-applications-in-self-driving-cars-610561e14118
[3] http://www.sci-news.com/space/deep-space-radiation-learning-memory-07464.html

Deep-Learning Reliability as Technology Enabler

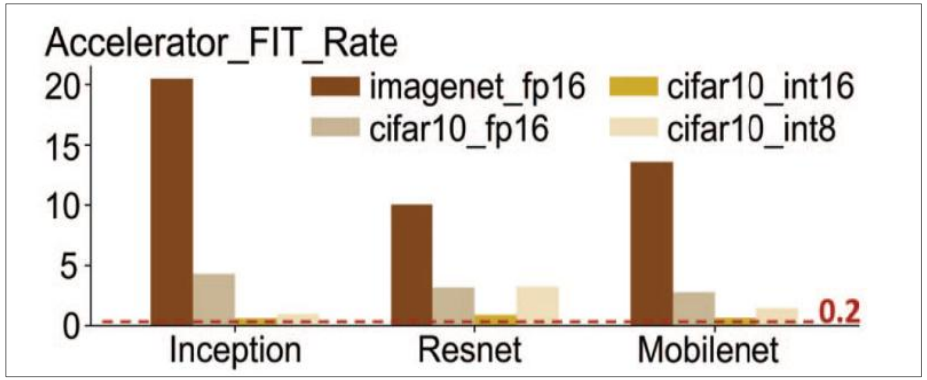
With energy efficiency, **reliability** is the **key topic** for future **deployment of AI in mission-critical application domains.**

Previous studies pointed to **Neural Networks** (NNs) as a **very robust computing paradigm**.

- Stochastic Nature
- Naturally resilient to internal states fluctuations

The presence of **faults** within the NNs' data **still produces an accuracy drop**.

- **Not suitable for every application scenario.**

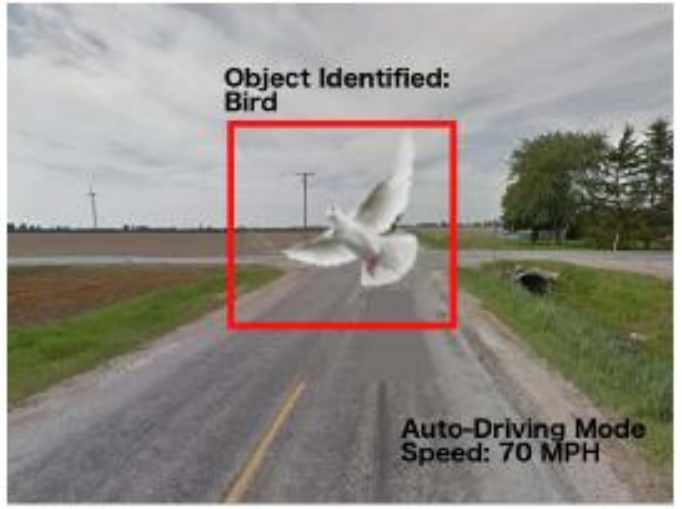


Yi He, Prasanna Balaprakash and Yanjing Li
«FIdelity: Efficient Resilience Analysis Framework for Deep Learning Accelerators», *In Proceeding of MICRO 2020.*

Fault Effects in AI based Safety Critical Applications

- Deep Neural Networks Accelerator
  - Hard/Soft errors in:
    - Computational Elements
    - Storage Elements
- High Energy Particles / Radiation Strikes
  - Hard and Soft Errors:
    - Hard and soft errors compromise system functionality by causing permanent damage or bit-flips in memory or computational elements
    - As a result we can have erroneous calculation or even functional interrupts

Fault Free Execution [4]



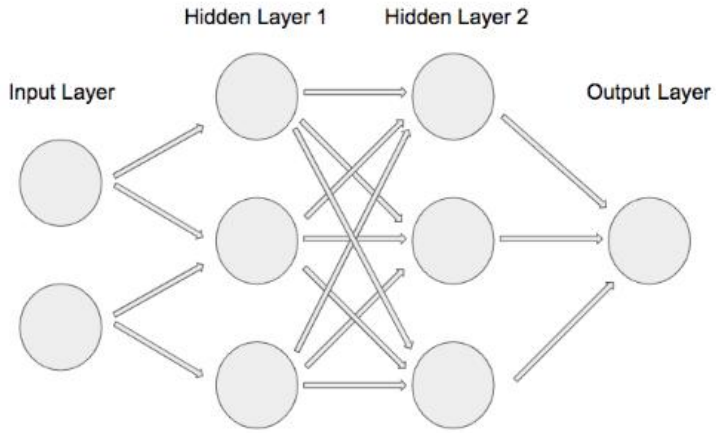
Erronous Execution [4]

- **Reliability validation is a must!**

[4] G. Li, S. K. S. Hari, M. Sullivan, T. Tsai, K. Pattabiraman, J. Emer, and S. W. Keckler, "Understanding error propagation in deep learning neural network (dnn) accelerators and applications,"

Factors Affecting Resiliency of Neural Networks

- Network architecture
  - Fully Connected NN, CNN (VGG, Resnet , Lenet, Inception)
  - Use of batch normalization layers (average the output values of layers)
  - Depth of the Network (deeper NN – more resilient)
- Layer Type
  - Fully Connected Layers
  - Convolution Layers (Conv layers are less resilient)
- Model Compression: (good for edge devices)
  - Pruning (removes the unnecessary connections)
  - Quantization (reducing precision)
- Data type used
  - The vulnerability is proportional to the dynamic value range of the data type can represent. i.e. FP32, FP16, FxP
- Bit position of the fault
  - Generally, high-order bits are more vulnerable(especially that goes from 0 to 1)
  - MSBs causes large deviation in magnitude.
- Data Reuse buffers
  - Data reuse is not a property of the DNN itself but of its hardware implementation



Neural Networks

State-of-the-Art in DLA Reliability Analysis

The **DLA Reliability Analysis frameworks** currently reported in literature follow **three major approaches:**

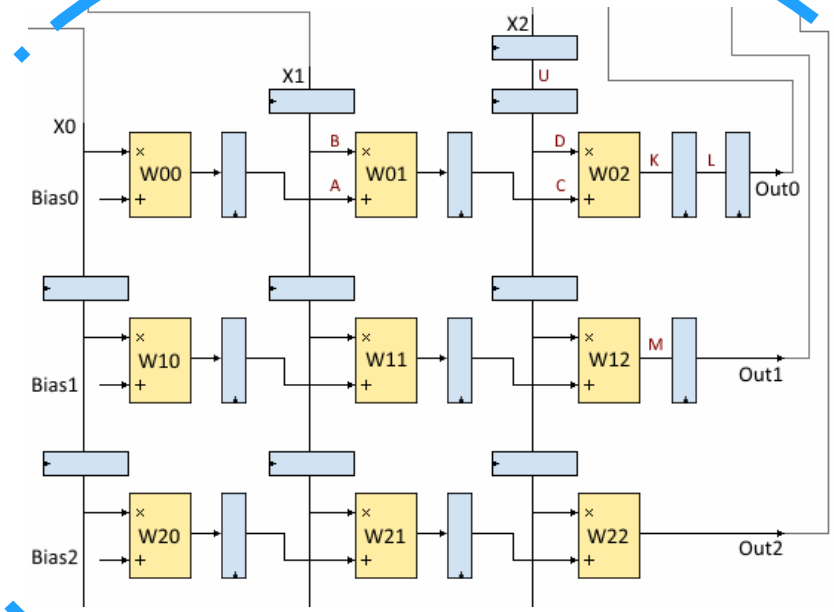
- **RTL simulation:**
  - Highly Accurate, Slow Simulation Speed;
- **FPGA Emulation:**
  - Faster than RTL, Need Synthesizable Design
- **Software Simulation:**
  - Low Accuracy, High Simulation Speed

We can define **evaluation accuracy** as the **number of injectable fault locations** (both in space and time domains)

State-of-the-Art in DLA Reliability Analysis

The **DLA Reliability Analysis frameworks** currently reported in literature follow **three major approaches:**

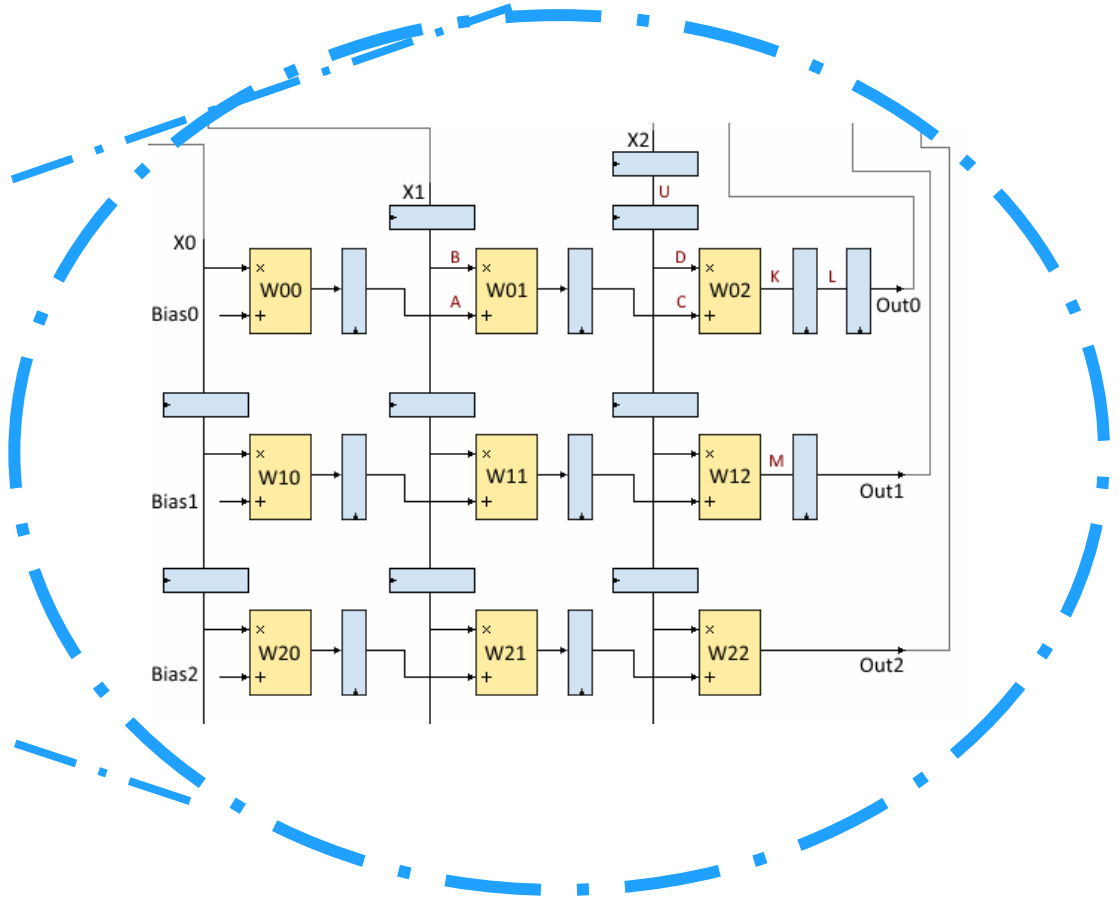
- **RTL simulation:**
  - Highly Accurate, **Slow Simulation Speed**;
- **FPGA Emulation:**
  - Faster than RTL, Need Synthesizable Design
- **Software Simulation:**
  - Low Accuracy, High Simulation Speed



State-of-the-Art in DLA Reliability Analysis

The **DLA Reliability Analysis frameworks** currently reported in literature follow **three major approaches:**

- **RTL simulation:**
  - Highly Accurate, Slow Simulation Speed;
- **FPGA Emulation:**
  - Faster than RTL, **Need Synthesizable Design**
- **Software Simulation:**
  - Low Accuracy, High Simulation Speed



State-of-the-Art in DLA Reliability Analysis

The **DLA Reliability Analysis frameworks** currently reported in literature follow **three major approaches:**

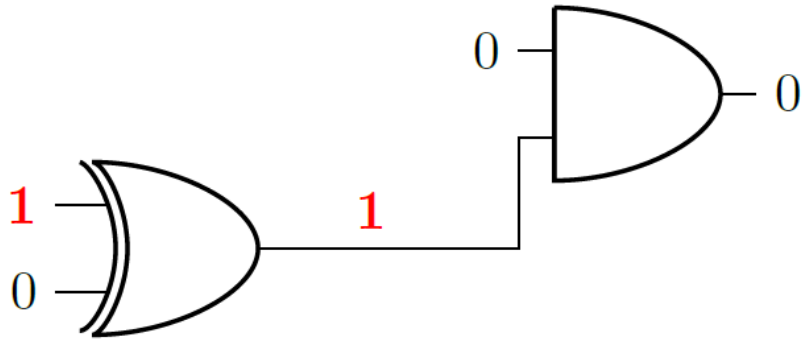
- **RTL simulation:**
  - Highly Accurate, Slow Simulation Speed;
- **FPGA Emulation:**
  - Faster than RTL, Need Synthesizable Design
- **Software Simulation:**
  - **Low Accuracy**, High Simulation Speed

Identification of Critical Flip-Flops in digital Systems

- Fault injection
  - Very high effort
- Error propagation probability calculation
  - High effort
  - Doesn‘t scale well for large circuits
  - Limited consideration of reconvergent paths
- Search for methods with lower complexity by higher level of abstraction
- Structural netlist analysis
  - Identification of critical flip-flops by structural properties directly derived from netlist
  - Reduced complexity
  - Sufficient accuracy?

Counting Gates with Controlling Values

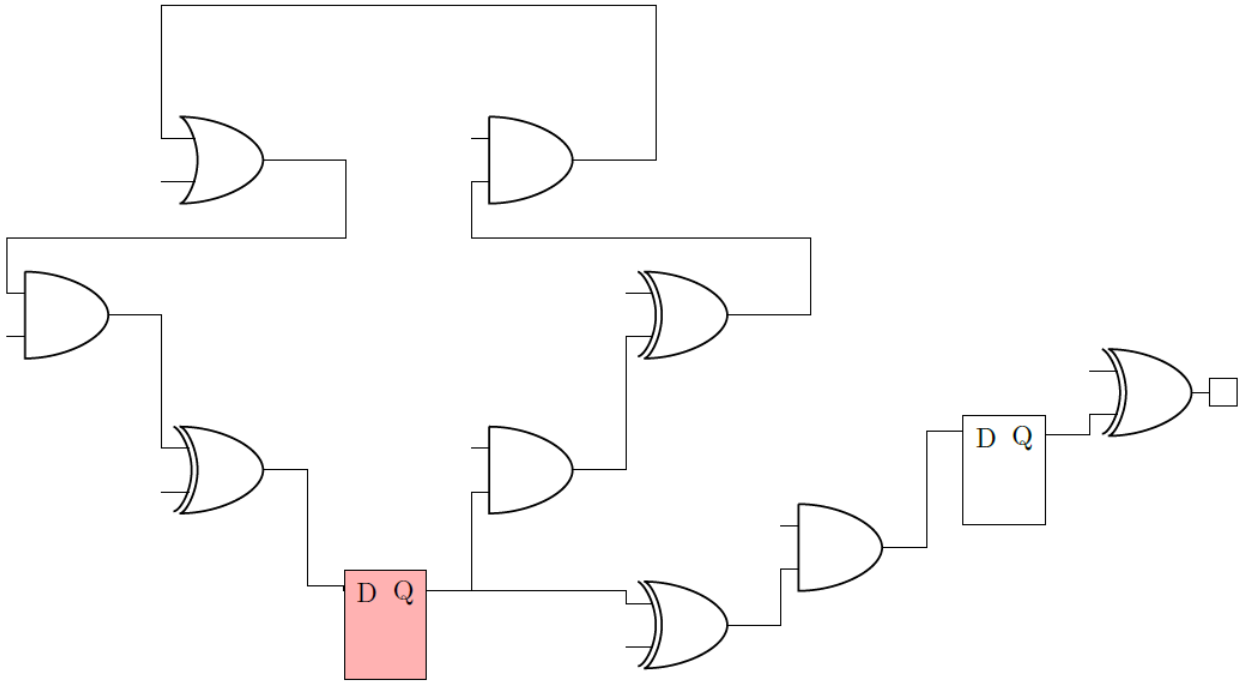
- Two effects of faults regarded as critical
  - Propagation to primary outputs
    - Output paths with low probability of masking
  - Permanent error in system state
    - Feedback paths with low probability of masking
- Logical masking by gates with controlling value
  - AND
    - Controlling value ‘0’
    - One input ‘0’ → output ‘0’ independent of the other input
  - XOR
    - No controlling value
    - Every transition at the inputs causes a transition at the output



- Idea:
  - Paths with low number of gates with controlling value have low probability of masking
  - Select Flip-Flop for protection if
    - on cycle with a low number of gates with controlling value
    - Path to output with a low number of gates with controlling value

Breitenreiter, Anselm, et al. "Selective Fault Tolerance by Counting Gates with Controlling Value." 2019 IEEE 25th International Symposium on On-Line Testing and Robust System Design (IOLTS). IEEE, 2019.

Counting Gates with Controlling Values – Illustration of the criteria

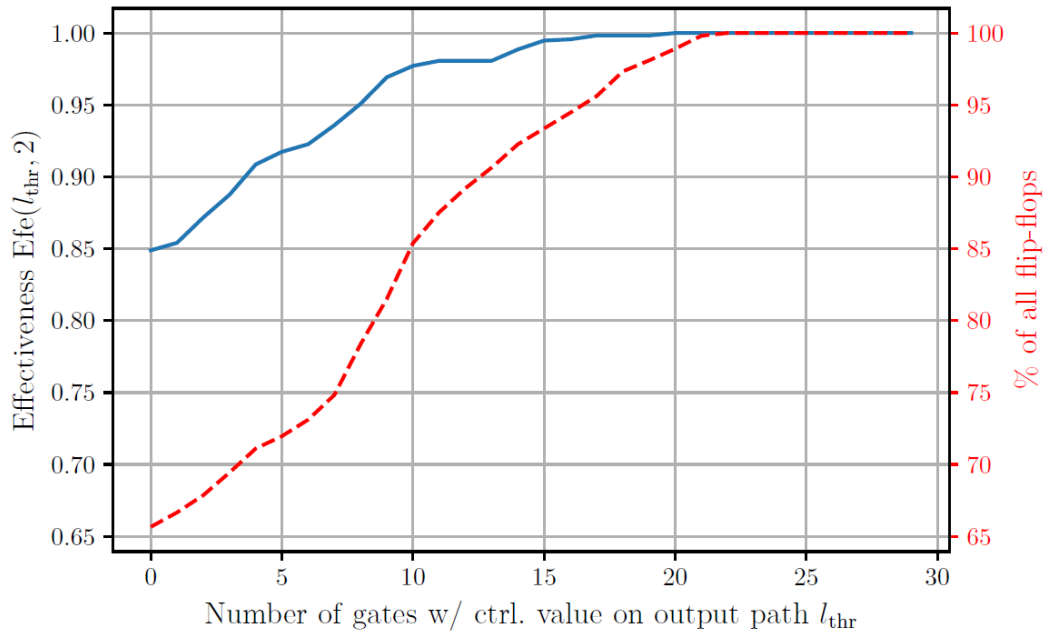


- The red flip-flop
  - is on a cycle with c = 4 gates with controlling value
  - has an output path with l = 1 gates with controlling value

Breitenreiter, Anselm, et al. "Selective Fault Tolerance by Counting Gates with Controlling Value." 2019 IEEE 25th International Symposium on On-Line Testing and Robust System Design (IOLTS). IEEE, 2019.

Counting Gates with Controlling Values - Evaluation Results

- Effectiveness of flip-flop selection based on the number of gates with controlling value on the output paths and number of gates with controlling value on cycles
- Threshold for the number of gates with controlling value on cycles c_"thr"

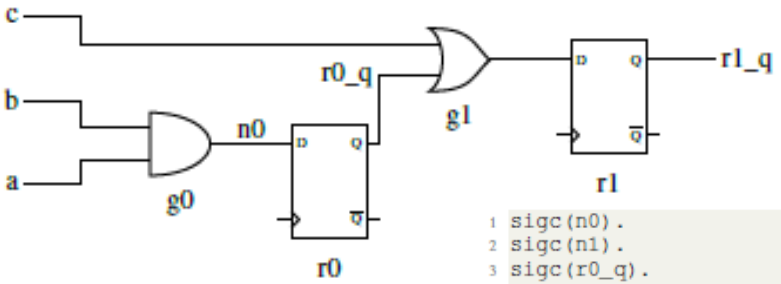


- Protecting about 75% of all flip-flops can avoid 95% of the errors

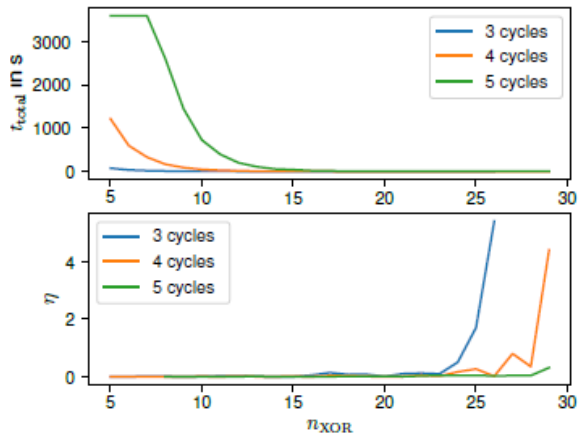
Breitenreiter, Anselm, et al. "Selective Fault Tolerance by Counting Gates with Controlling Value." 2019 IEEE 25th International Symposium on On-Line Testing and Robust System Design (IOLTS). IEEE, 2019.

Fault Susceptibility Analysis by Combinatorial Search

- Error Propagation Probability
  - How likely is it for a fault in a certain circuit element to propagate and manifests as an error at the outputs?
- Answer Set Programming
  - Combining Logic Programming with Stable Model Semantics
- Approximate Model Counting
  - Number of possible fault in a complex circuit too high to count
  - Split search spaces in subspaces, count solutions in one subspace and multiply by number of subspaces
- Features
  - Input data format similar to Verilog netlist
  - Adaptable error definition
  - Reduction of search space by integration of application information



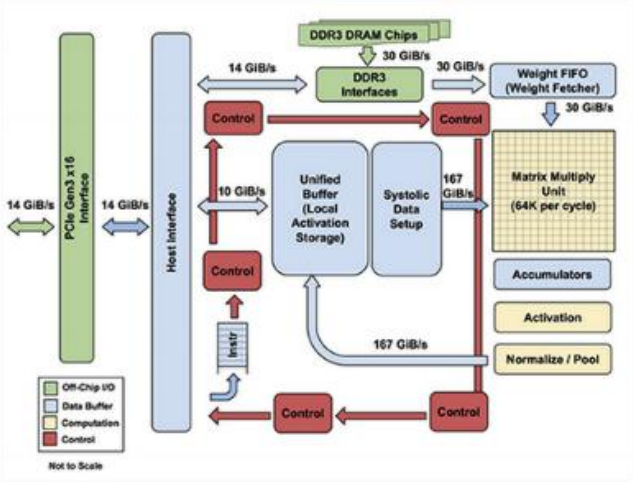

```
sigc(n0).
sigc(n1).
sigc(r0_q).
sigc(r1_q).
gatec(g0, and, n0, a, b).
gatec(g1, or, n1, r0_q, c).
regc(r0,r0_q,n0).
regc(r1,r1_q,n1).
input(a;b;c).
output(r1_q).
```



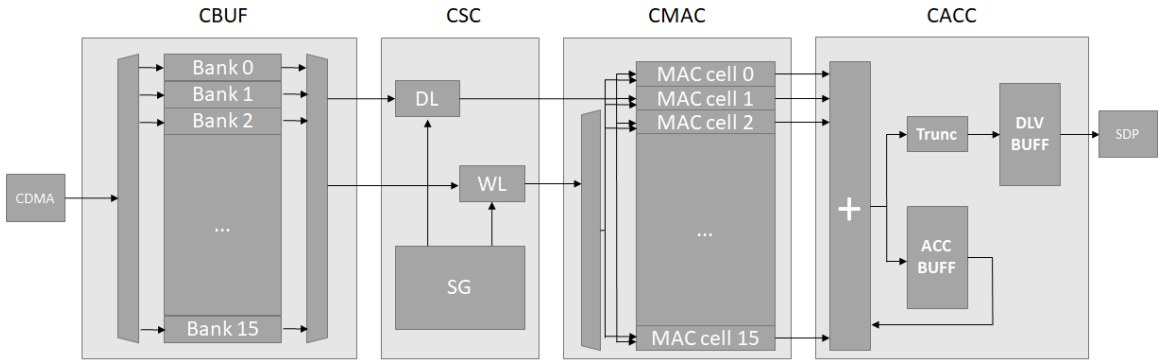
A. Breitenreiter, M. Andjelković, O. Schrape and M. Krstić, "Fast Error Propagation Probability Estimates by Answer Set Programming and Approximate Model Counting," in *IEEE Access*, vol. 10, pp. 51814-51825, 2022.

Deep-Learning Hardware Acceleration

Former studies[1] demonstrated how **Deep-Learning Accelerators (DLAs) may bring to effective performance gain while cutting energy consumptions.**



In 2016, **Google's TPU** was granting a **15-30x performance boost** and an **80x performance/Watt** against (at the time) **state-of-the-art GPUs.**



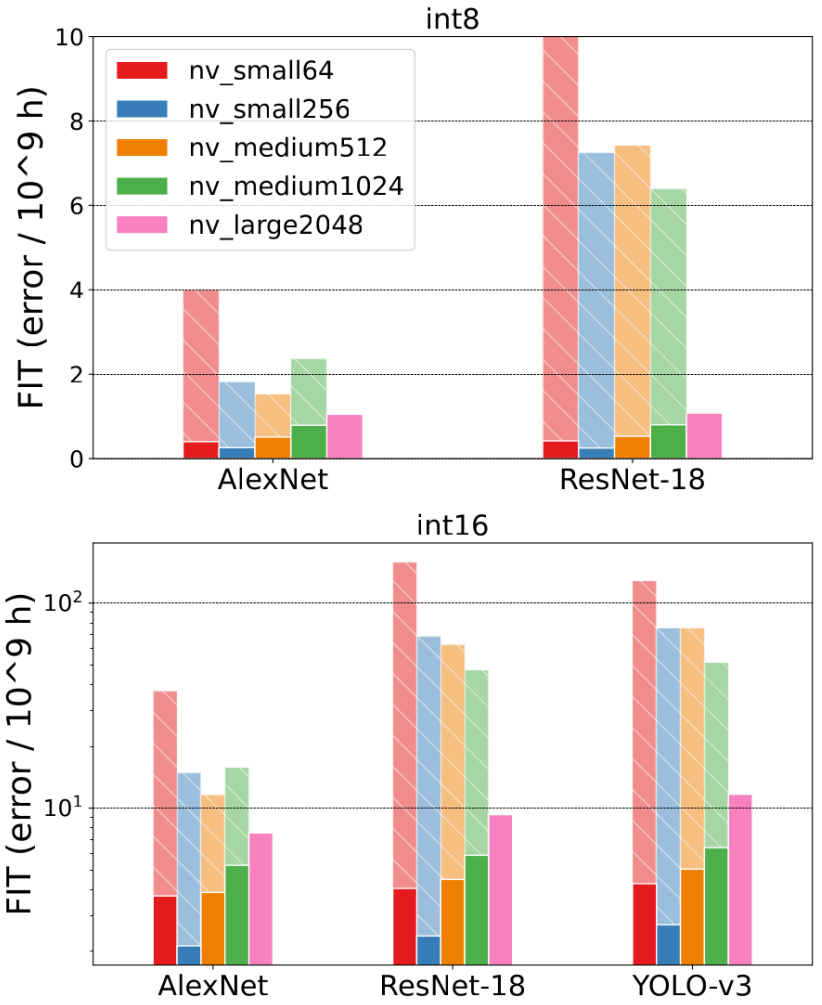
Our DLA research focusses onto the **open-source NVIDIA DLA** (NVDLA).

- Industry-grade design (implemented into Jetson boards)
- Both HW & SW completely open-sourced

[1]: *N . P. Jouppi et al.* «in-Datacenter Performance Analysis of a Tensor Processing Unit», *In Proceeding of ISCA 2017.*

DLA Error Analysis and Architecture Exploration

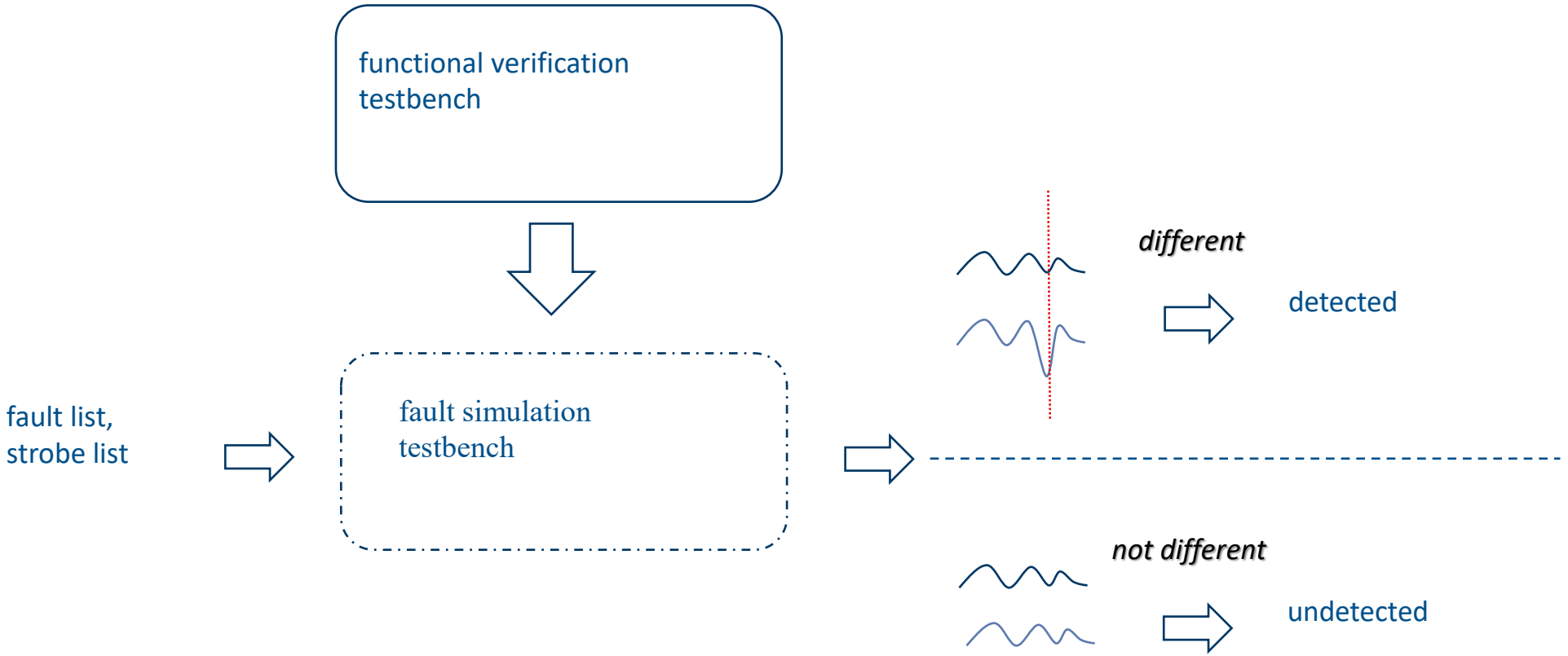
- Correlation of the application errors with the choice of hardware architecture
- Data reuse techniques applied in hardware have a strong impact on error propagation and ultimately on Silent Data Corruption (SDC).
- To effectively correlate hardware errors with application errors, high-speed simulation techniques are required.
- For this purpose, a proprietary full-cycle RTL simulation tool has been developed
- We currently have a case study (NVDLA accelerator) and a fault model (SEU), further fault models are in the program, which are also to be integrated (MBU, SAF)



A. Veronesi, A. Nazzari, D. Passarello, M. Krstic, M. Favalli, L. Cassano, A. Miele, D. Bertozzi, C. Bolchini, Cross-Layer Reliability Analysis of NVDLA Accelerators: Exploring the Configuration Space, 29th IEEE European Test Symposium 2024, Outstanding Student Presentation Award.

Fault injection approach

(Example Fault injection with Xcelium fault simulator)

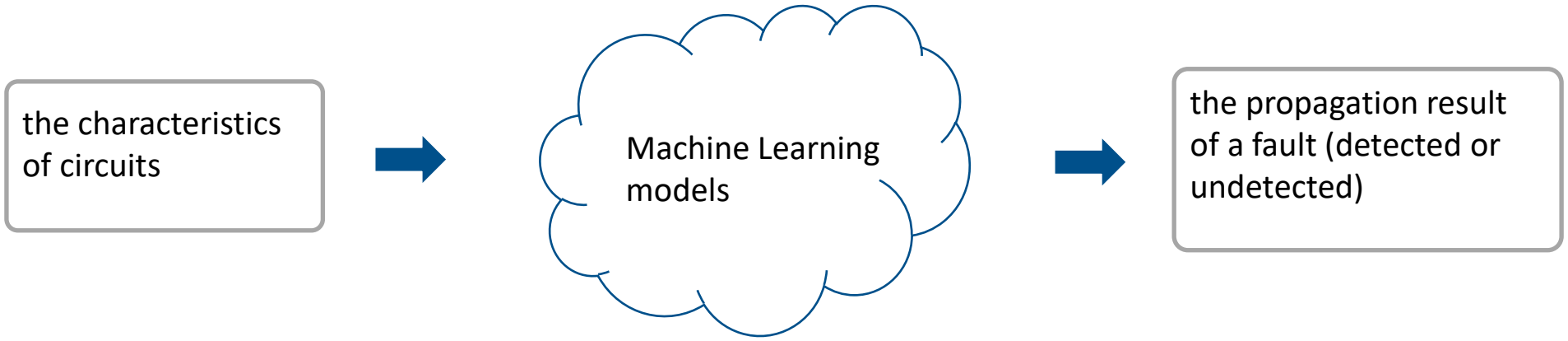


Accelerating Fault Injection

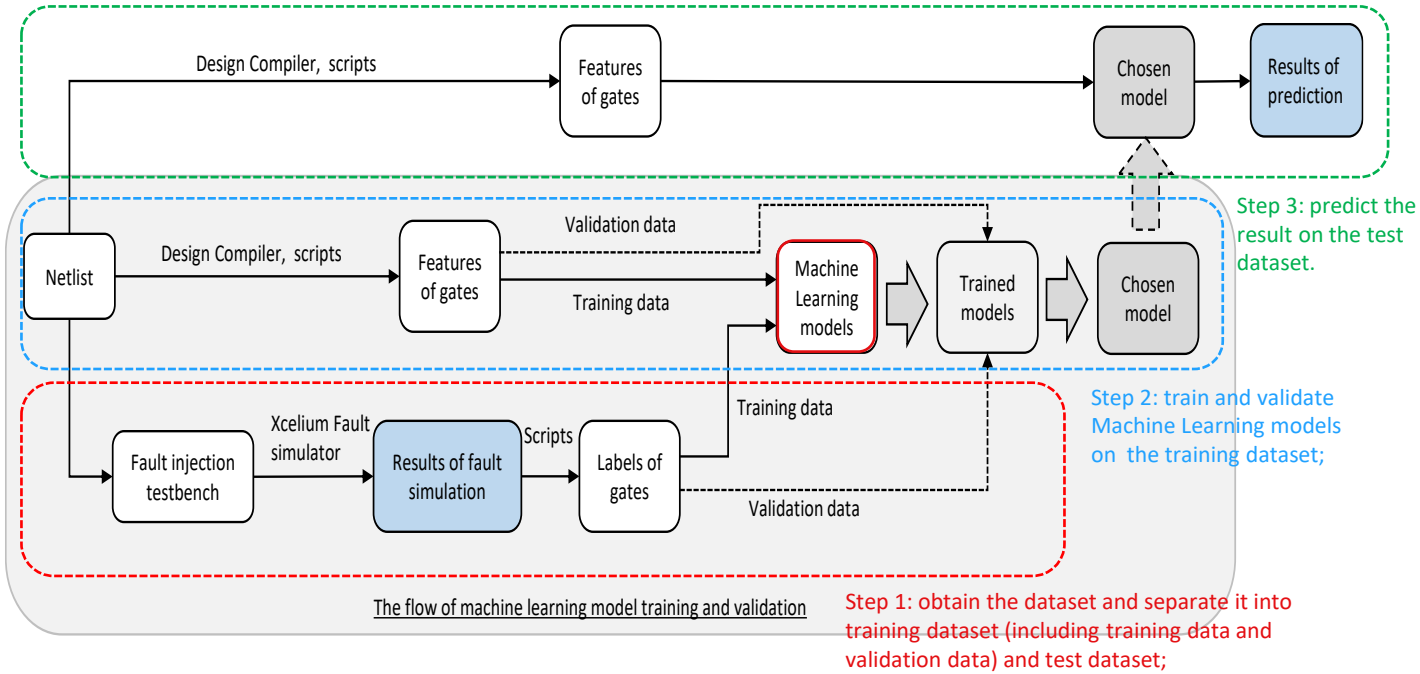
- Simulation-based fault injection is time-consuming.

  Fault models & the number of gates & the time to inject fault

- Machine learning algorithms could build a model based on experience data to get the prediction capability.
- Apply machine learning in fault injection to reduce the time involved.



Accelerate fault simulation with Machine Learning - NNs



- With Neural Networks (NNs), we can use the features of gates to predict their fault simulation results;
- This method is validated on an ADC core designed for space application and the prediction accuracy achieves up to 95%.

Reference: This parameter indicate library cell of the gate.

Input_pins_num: This parameter indicates the number of the input pins of the gate, also the number of driver pins of the gate.

Output_pins_num: This parameter indicates the number of the output pins of the gate.

Loads_num: This parameter to indicate the number of pins connected with the output pins of the gate.

Vector_of_driver_pins_type: This vector describes the types of cells connected to the input pins of this gate, also the cell types of driver pins.

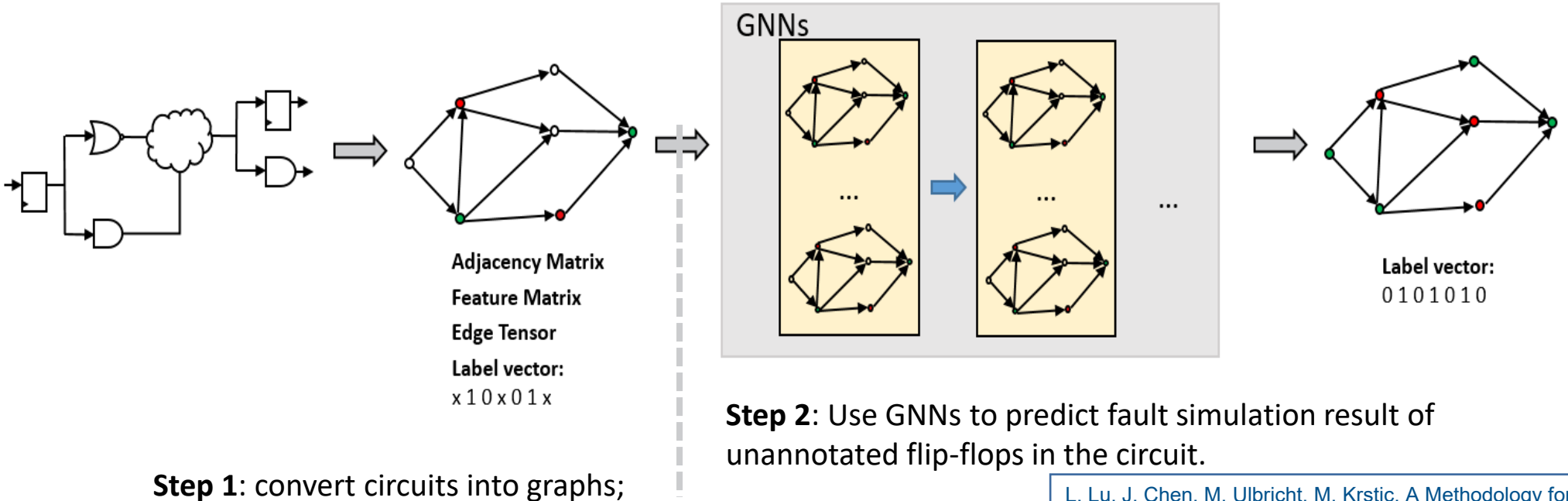
Vector_of_load_pins_types: This vector describes the types of cells connected to the output pins of this gate, also the cell types of load pins.

...

L. Lu, J. Chen, A. Breitenreiter, O. Schrape, M. Ulbricht, M. Krstic, Machine Learning Approach for Accelerating Simulation-based Fault Injection, IEEE NorCAS 2021.

Accelerate fault simulation with Machine Learning - GNNs

- NNs are only able to utilize the features of individual components.
- Graph Neural Networks (GNNs) can learn from the features of neighbours and correspondingly get the capability to learn from structural features.



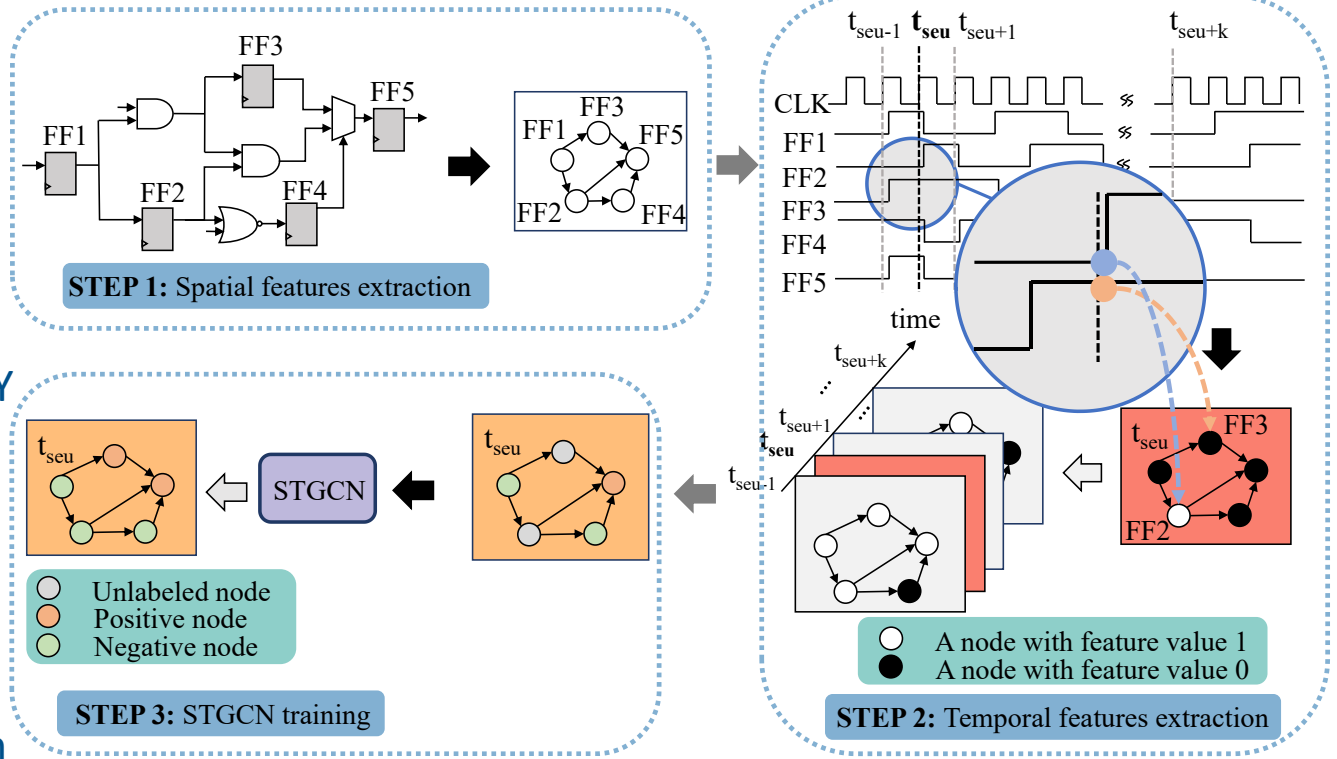
Step 1: convert circuits into graphs;

Step 2: Use GNNs to predict fault simulation result of unannotated flip-flops in the circuit.

L. Lu, J. Chen, M. Ulbricht, M. Krstic, A Methodology for Identifying Critical Sequential Circuits with Graph Convolutional Networks, 2022 IEEE Computer Society Annual Symposium on VLSI (ISVLSI), 2022

AI for Fault Analysis – achieved Results

- Simulation-based fault injection for reliability analysis is time-consuming.
  - Machine learning could be the alternative
- GNNs were able to identify critical flip-flops in terms of SEU tolerance.
  - The accuracy achieved with Ibex and RI5CY is up to 98%.
- Spatio-temporal Graph Convolutional Network (STGCN) could predict the results of SEU error simulation.
  - We validated the proposed method on four test circuits and achieved a prediction accuracy of 94-96%.

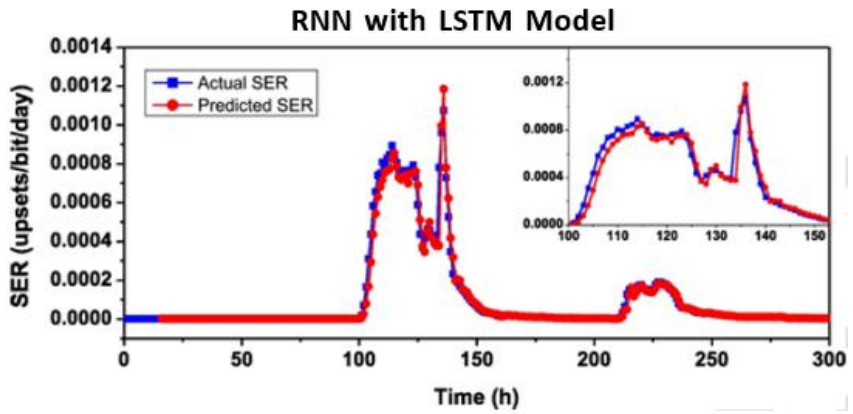


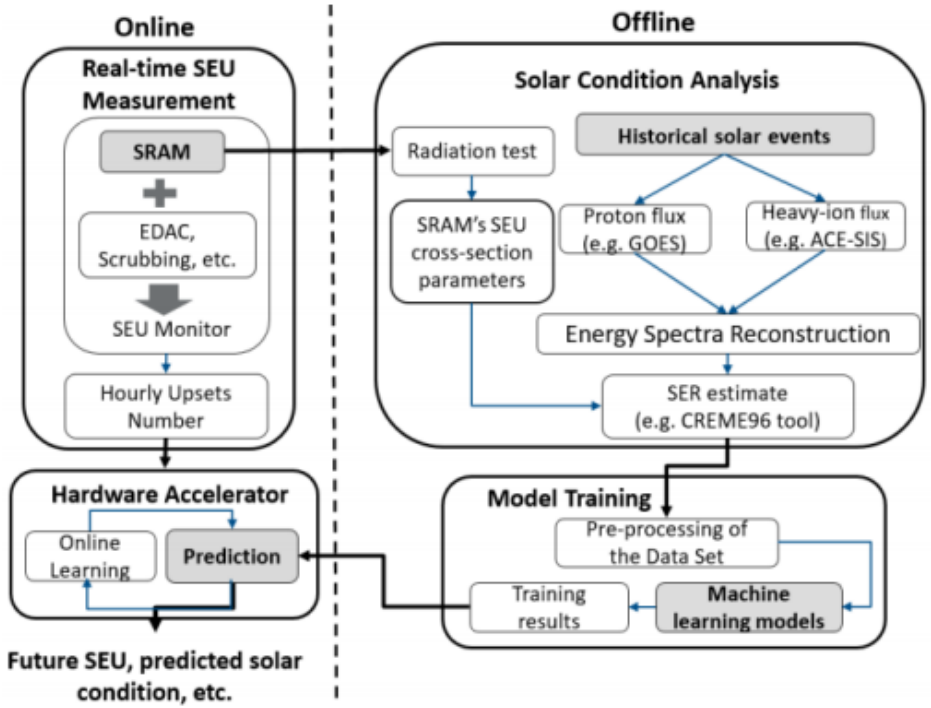
Circuit	Number of Samples	Percentage of Positive Samples (%)	Precision (%)	Recall (%)	Accuracy (%)
I2C	7650	70.56	97.25	97.71	96.52
SPI	6550	62.60	96.19	97.63	96.23
XGE-MAC	48888	52.12	96.54	96.33	96.26
Ibex	63780	16.51	83.32	78.55	93.96

L. Lu, J. Chen, M. Ulbricht, M. Krstic, Machine learning methodologies to predict the results of simulation-based fault injection, **IEEE TCAS I**, 2024.
L. Lu, J. Chen, M. Ulbricht, M. Krstic, Towards Critical Flip-Flop Identification for Soft-Error Tolerance with Graph Neural Networks, **IEEE TCAD**, 2024
L. Lu, J. Chen, M. Ulbricht, M. Krstic, Towards SEU Fault Propagation Prediction with Spatio-temporal Graph Convolutional Networks, **DATE**, 2024.

AI for Resilience
Solar Particle Events Prediction with Machine Learning

- Solar Particle Events (SPEs) dominate the space radiation environment increasing Soft Error Upset (SEU) rate several orders of magnitude for certain time
- How to predict SPEs?
- Analyze a data from large number of historical SPEs and Supervised-machine learning for data and model training
- Collect the suitable SPE&SEU prediction model
- Novel SRAM-based SEU monitor for in-flight real-time hourly SEU number detection
- Customized low-cost hardware accelerator for SPE & SEU prediction
- Achieved: Prediction of SPEs 1 hour in advance



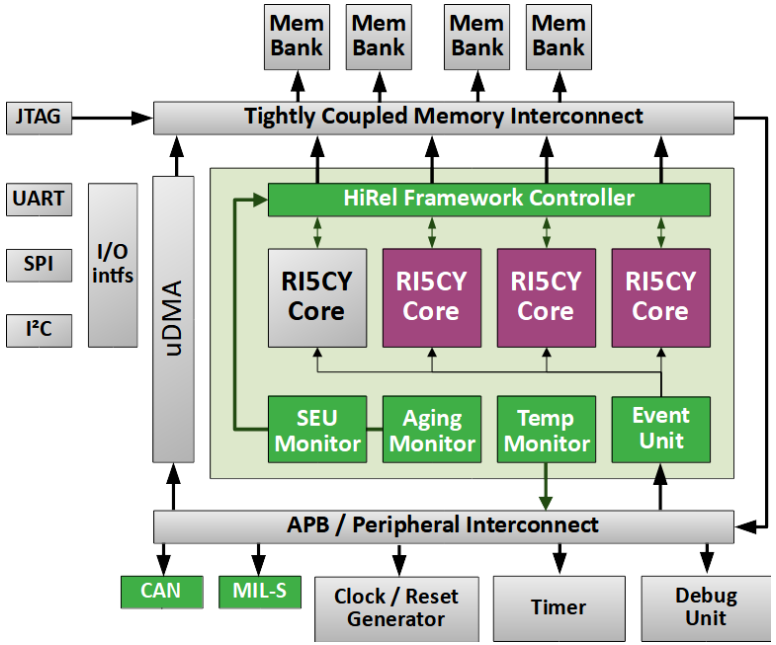


J. Chen, T. Lange, M. Andjelkovic, A. Simevski, L. Lu and M. Krstic, "Solar Particle Event and Single Event Upset Prediction from SRAM-based Monitor and Supervised Machine Learning," in IEEE Transactions on Emerging Topics in Computing, 2022

Application: TETRISC SoC

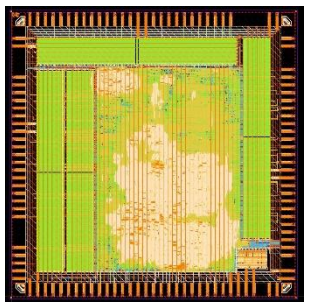
The TETRISC SoC - a **smart highly reliable RISC-V based multiprocessor platform**, suitable for harsh environments (such as aviation, space, etc.)

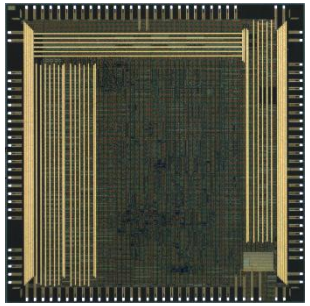
- RISC-V based adaptive quad-core SoC
- Based on the open-source single-core microcontroller architecture PULPissimo
- Multiple on-chip monitors (SEU, Temperature, Aging)
- Smart framework controller for on-line system adaption
- Dynamic trade-off between reliability, performance and power consumption

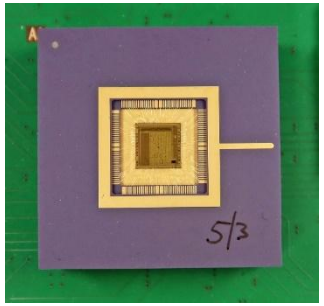


Chip Specs

- **Area** - 43,56 mm2 (6.6*6.6mm), **Nominal clock frequency** - 30MHz, **Power consumption** - <1W (estimated), **Memory** - 4*8192*40 Bit SRAM, **Pads** - 81 Signal pads, 35 others





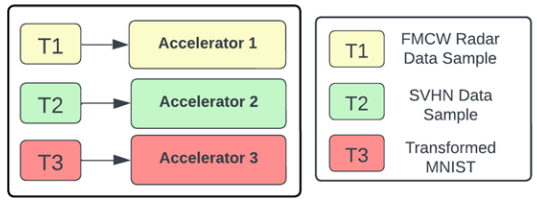


M. Ulbricht, J. Chen, L. Lu, M. Krstic, The TETRISC SoC - A resilient Quad-Core System based on the ResiliCell approach, Microelectronics Reliability, Vol. 148. 2023.

Reliable AI Processing in FPGA

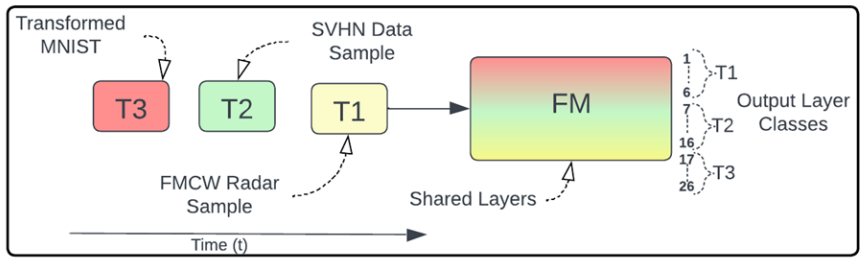
1 Shared layers methodology to efficiently map CNN models on hardware

- Shared-layers methodology leverages fundamental working principles of CNNs to learn patterns
- Fused and Branched CNN architectures are evaluated for three distinct tasks based on accuracy, quantization, pruning, hardware resource utilization, power, and latency.



The traditional way of multiple-task execution

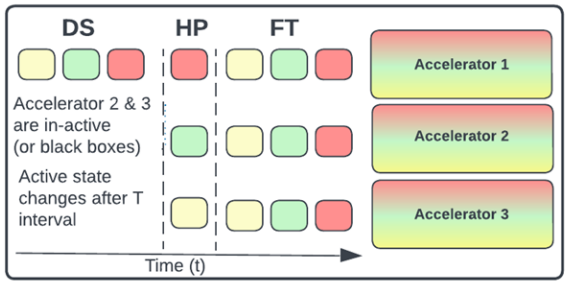
Three tasks/datasets two modalities



Shared-layers-based fused and branched model capable of executing multiple tasks from multiple modalities

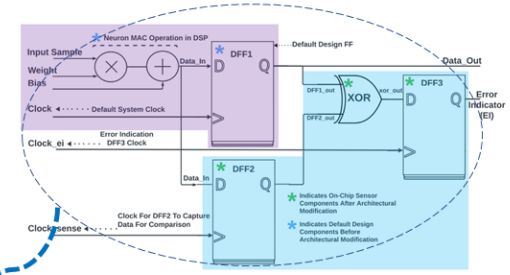
2 Reconfigurable CNN accelerators based on Shared-layers methodology

- Extend the shared-layers methodology and proposed fault-tolerant CNN accelerators with reconfigurable capabilities
- **FT:** DMR/TMR mode for reliable AI processing
- **HP:** Parallel execution of multiple tasks
- **DS:** Aging aware mode to reduce aging and power consumption.

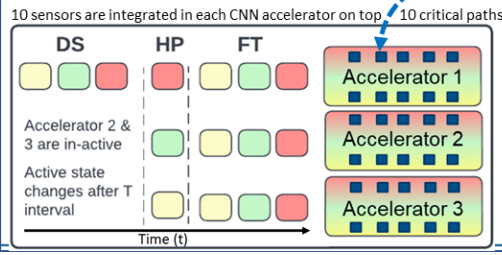


3 Aging and Soft Error Aware Reconfiguration in CNN Accelerator

- Integration of the on-chip sensor in a CNN hardware accelerator to detect aging and soft errors.
- Reconfigure the CNN accelerator by utilizing the intelligence provided by the on-chip aging and soft error sensor.



Architectural modification in a single neuron computation to integrate sensor



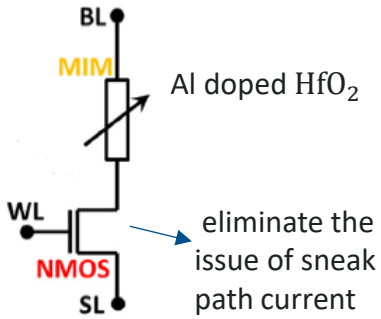
R. T. Syed, Y. Zhao, J. Chen, M. Andjelkovic, M. Ulbricht, M. Krstic, FPGA Implementation of a Fault-Tolerant Fused and Branched CNN Accelerator with Reconfigurable Capabilities, IEEE Access, 2024.

R.T. Syed, M. Andjelkovic, M. Ulbricht, M. Krstic, Towards Reconfigurable CNN Accelerator for FPGA Implementation, IEEE Transactions on Circuits and Systems II: Express Briefs, 2023

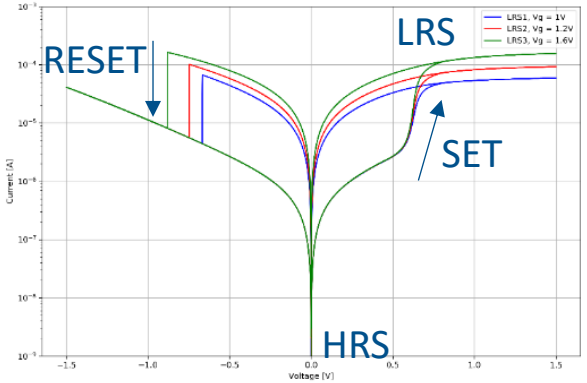
Aleternative Technologies - Memristor and 1T1R Structure

Memristor:

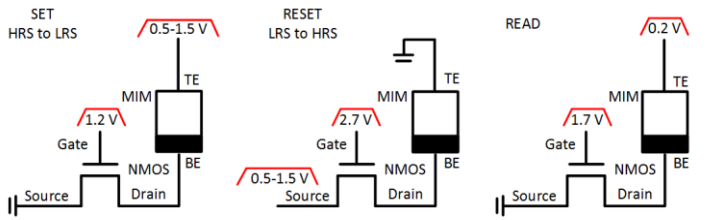
- Two-terminal device, the resistance state is retained based on the history of applied voltages
- Resistive RAM
- Non-volatile memory
- In-memory computing
- Fully compatible with CMOS



Structure of 1T1R

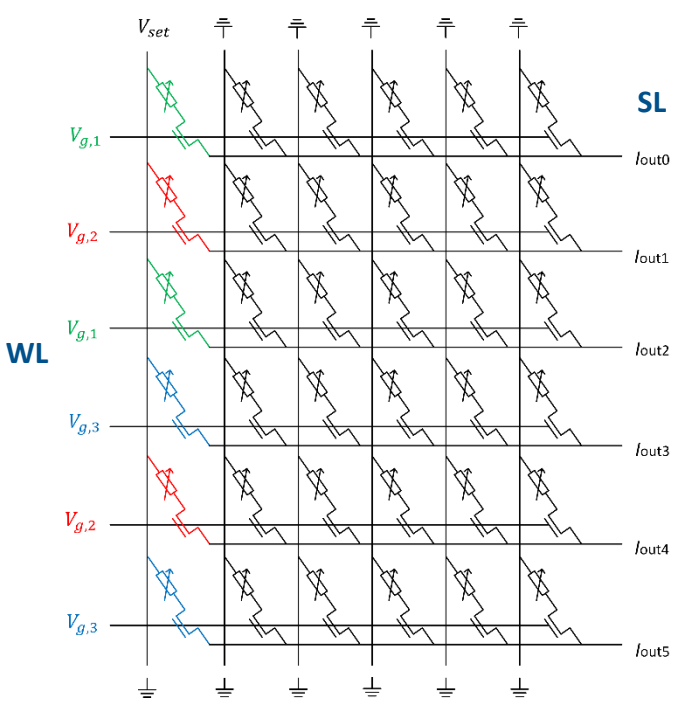


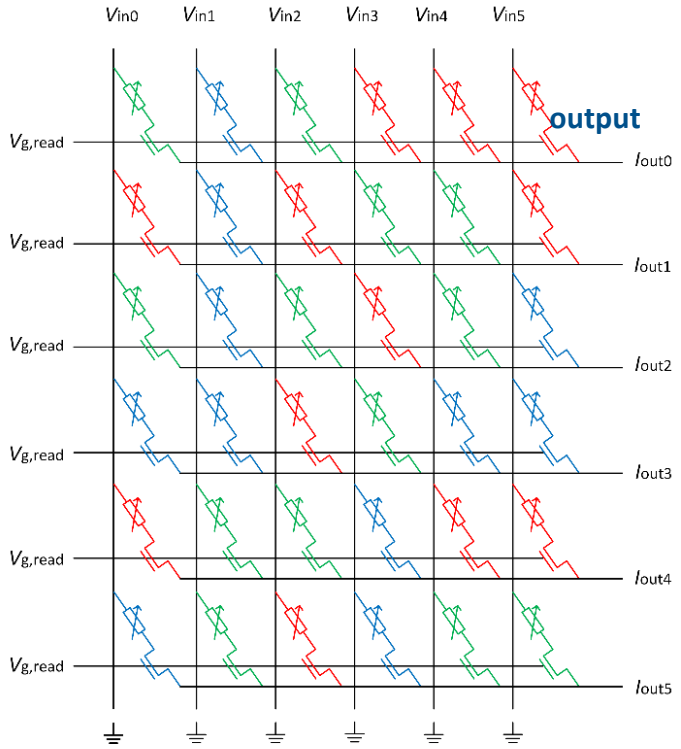
***IV* relationship of 1T1R**



Operations of the 1T1R structrue

Memristive Crossbar





- VMM in memory
- High parallelism and throughput
- Reuse of matrix

$$I_{\text{out},i} = \sum_{j=0}^{N} V_{\text{in},j} \times G_{i,j}$$

Mapping matrix to the crossbar:

Column-wise SET/RESET

Currents received at SLs to perform write-verify

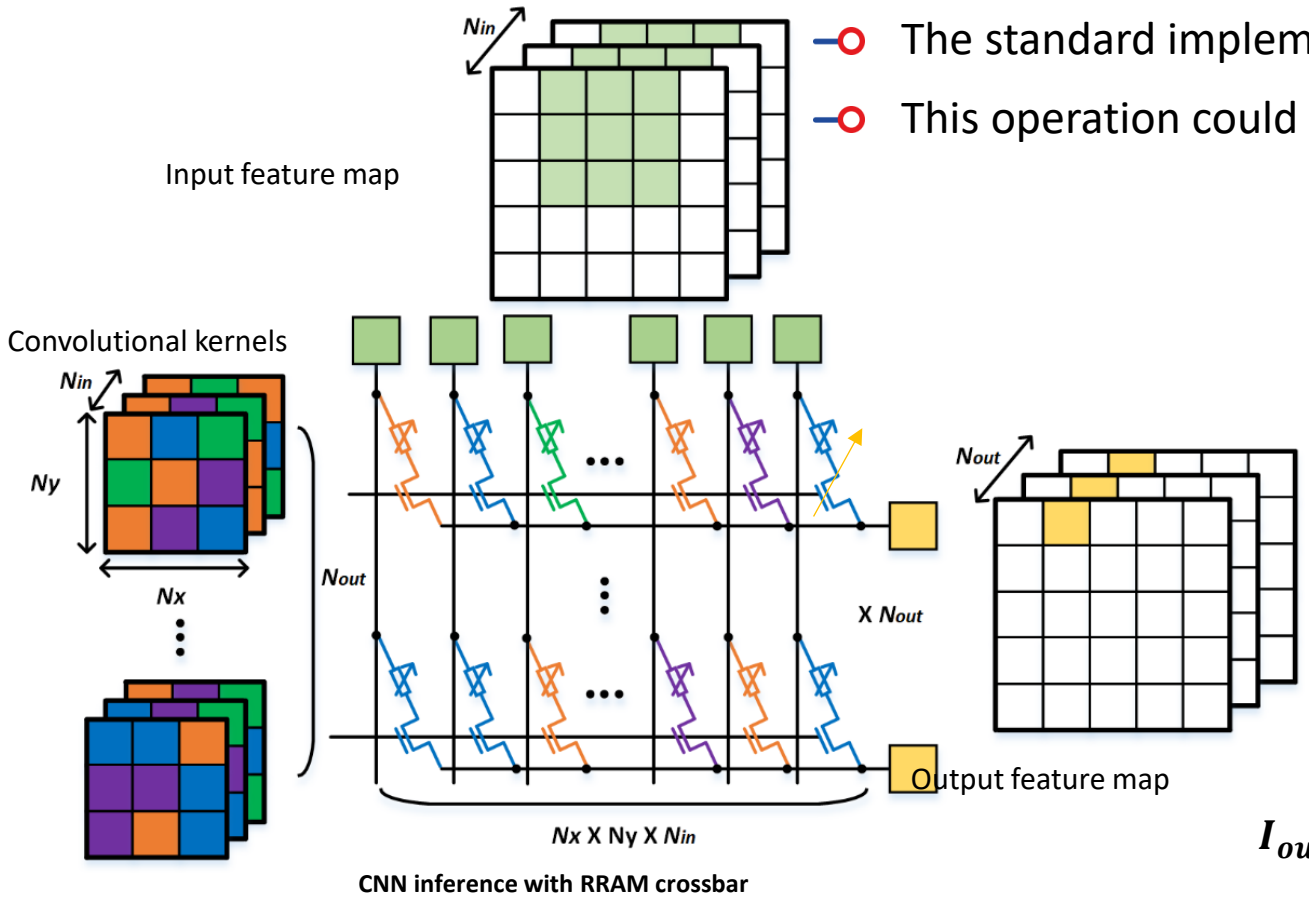
Multiplication:

Assigning input voltage vector on BLs (2 bits)

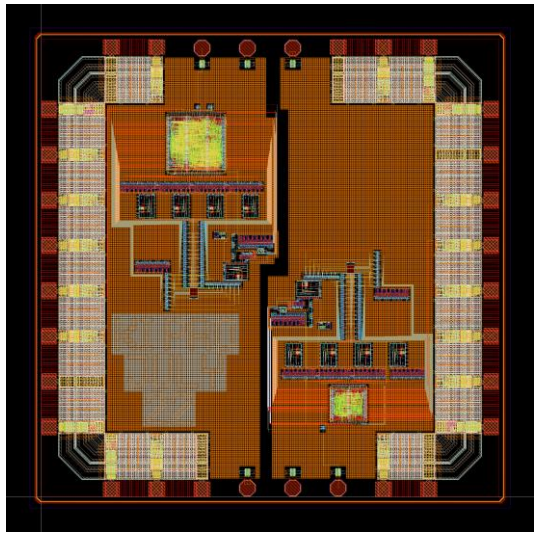
Reading output current as results of VMM

RRAM for AI acceleration - In-memory computing

- Machine learning based on **vector matrix multiplication**
- The standard implementation of this process requires a lot of time and energy.
- This operation could be implemented more optimally with **RRAM technology**



$$I_{out,i} = \sum_{j=0}^{N-1} V_{in,j} \times G_{i,j}$$



BMBF project **KI-PRO**
RRAM crossbar AI accelerator
(Design with TUM)
~ 6 mm2, TO Jul 2022

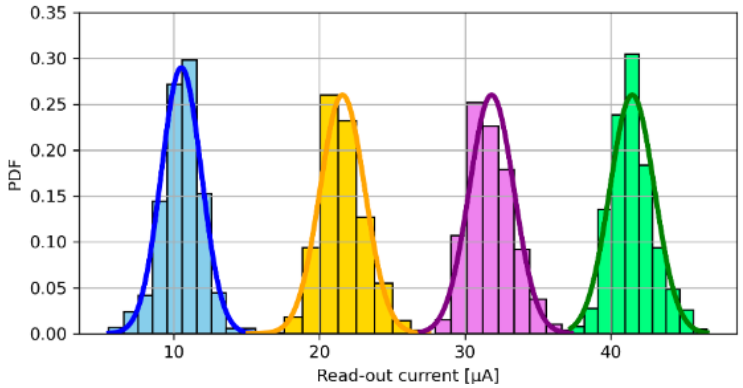
MLP for Breast Cancer Detection

Objective:

System-level evaluation of RRAM-based neural network
Investigating effects of limited word length and device variabilities

Benchmark:

Wisconsin Breast Cancer Dataset (WBCD)



Statistical model of device-to-device variability from measurements

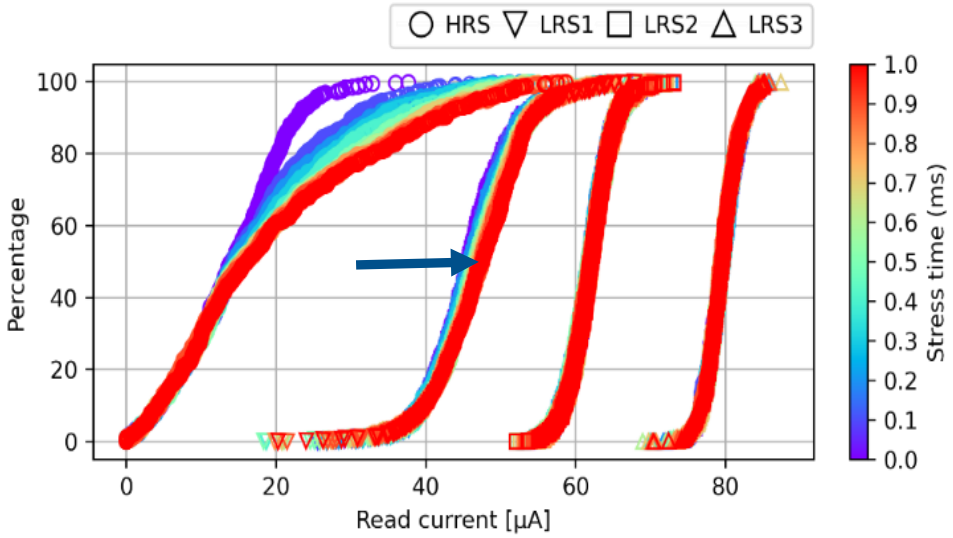
Design		Simulated MLP		
Architecture		9-4-1		
Description		Baseline	Quantization	Variability
Train.	Acc.	97.81%	96.16%	96.23%
	Std. Dev	1.03	0.76	0.76
Val.	Acc.	96.3%	97.81%	97.92%
	Std. Dev	0.67	1.03	1.12

*Models are evaluated with 5-fold cross validation.
*Quantization: 2-bit weights and activations

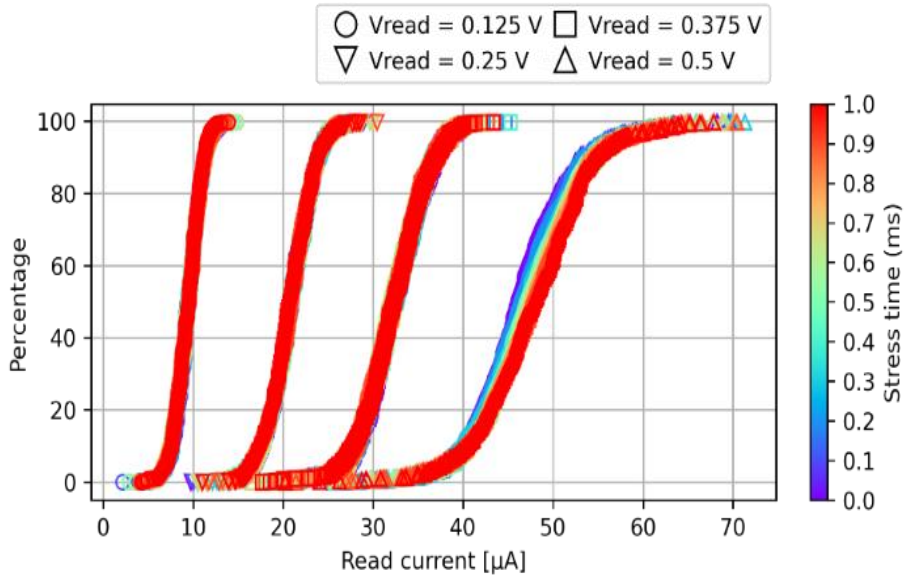
- Accuracy is maintained after the quantization. Standard deviation is increased
- Small-scale MLP is capable of tolerating variabilities
- Some other non-ideal effects should be considered (stuck devices)

Read Disturb in RRAM

The stored internal RRAM conductance is unintentionally modified during reading



$V_{read} = 0.5$ V, 512 devices, 1000 pulses with $T = 1$ µs



LRS1, 512 devices, 1000 pulses with $T = 1$ µs

- States with low conductance suffer from read disturb (HRS, LRS1)

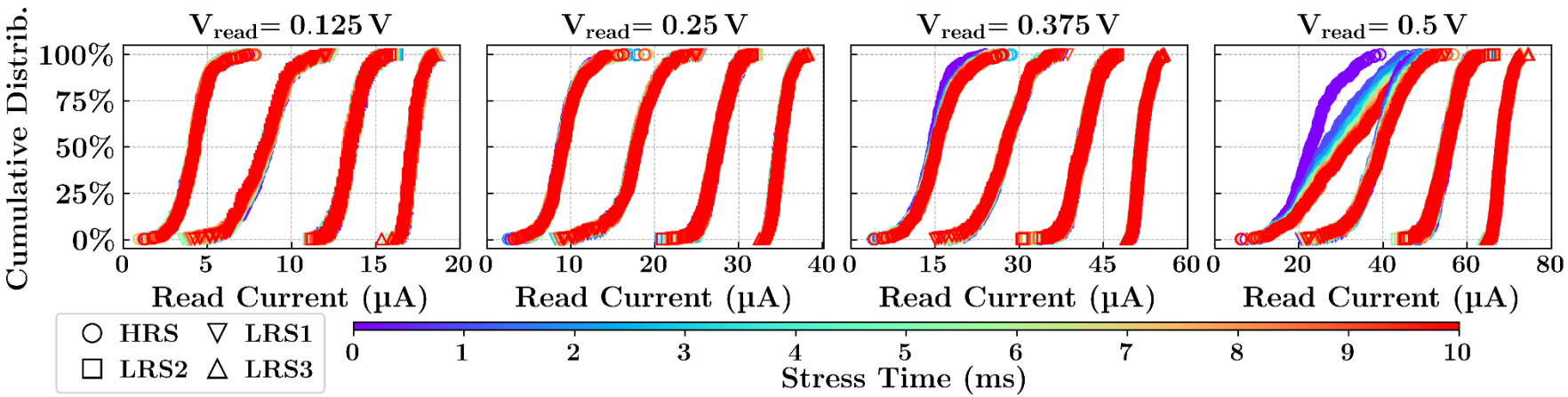
- Study the multi-bit input => Higher requirement for read disturb
- Incorporating device-to-device variability
- Evaluating the impact of read disturb with different mapping strategies

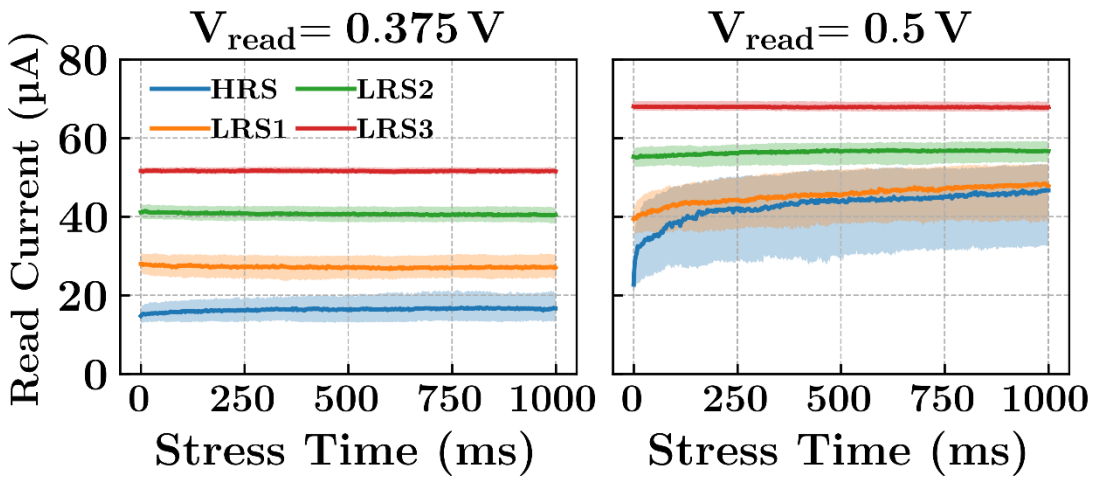
J. Wen, A. Baroni, E. Perez, M. Ulbricht, C. Wenger and M. Krstic, "Evaluating Read Disturb Effect on RRAM based AI Accelerator with Multilevel States and Input Voltages," *IEEE International Symposium on Defect and Fault Tolerance in VLSI and Nanotechnology Systems (DFT)*, Austin, TX, USA, 2022, pp. 1-6,

Read Disturb Characterization

- 100k read pulses with 10 us each → **1 second**
- 4 states × 4 read voltages
- 256 devices

10 ms → **1s**





Short time period (**10 ms**):

- Devices at HRS slightly drift @ V_{read} = 0.375 V
- Devices at HRS and LRS1 drift @ V_{read} = 0.5 V, **overlap**
- **Overall drift + increased device-to-device variability**

Long time period (**1000 ms**):

- HRS @ V_{read} = 0.5 V**, decreased drift speed, no saturation**
- **Significant overlap** between HRS and LRS1, and also LRS2
- **Weak conductive fiaments for HRS and LRS1**

J. Wen, at all, ReDiM: An Efficient Strategy for Read Disturb Mitigation in RRAM-Based Accelerators, IOLTS 2025

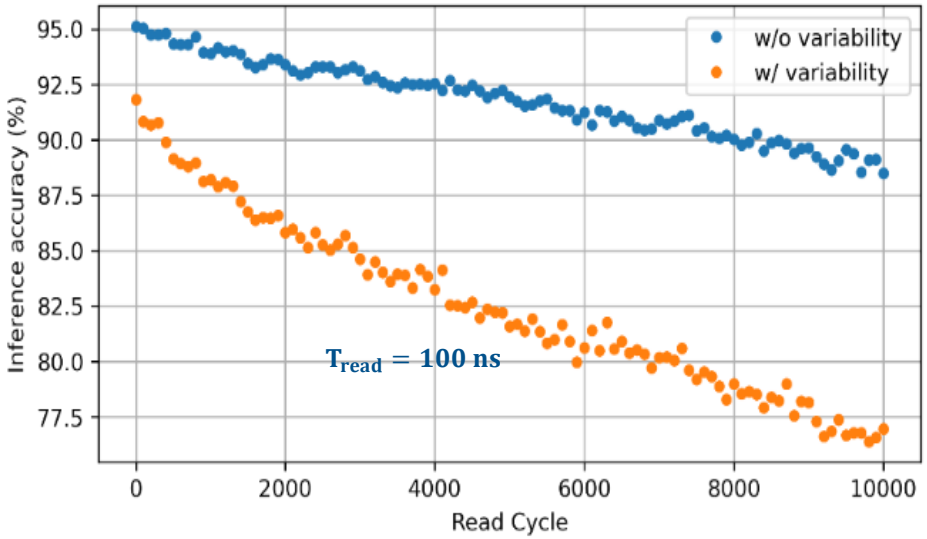
Results: LeNet-5

LeNet-5 with MNIST:

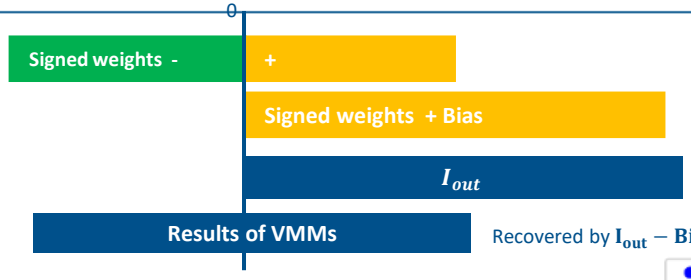
2-bit weights (1 device ⟷ 1 weight)

Signed computation:

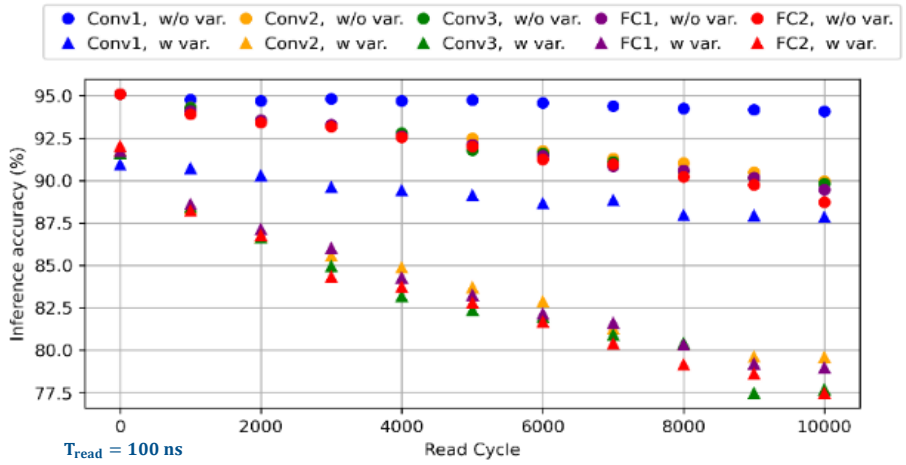
Weight biasing



Inference accuracy of LeNet-5 with the consideration of read disturb and variability



Weight biasing to achieve signed computation



Inference accuracy of LeNet-5 with a specific hardened layer
Hardened layer => assumed as fault-free against read disturb

- Accuracy degradation caused by read disturb
- Larger margin between models with and without variability for the increased stressing time
- The first convolutional layer is the critical layer that mostly degrades the inference accuracy

J. Wen, A. Baroni, E. Perez, M. Ulbricht, C. Wenger and M. Krstic, "Evaluating Read Disturb Effect on RRAM based AI Accelerator with Multilevel States and Input Voltages," *IEEE International Symposium on Defect and Fault Tolerance in VLSI and Nanotechnology Systems (DFT)*, Austin, TX, USA, 2022, pp. 1-6,

Results: VGG-7

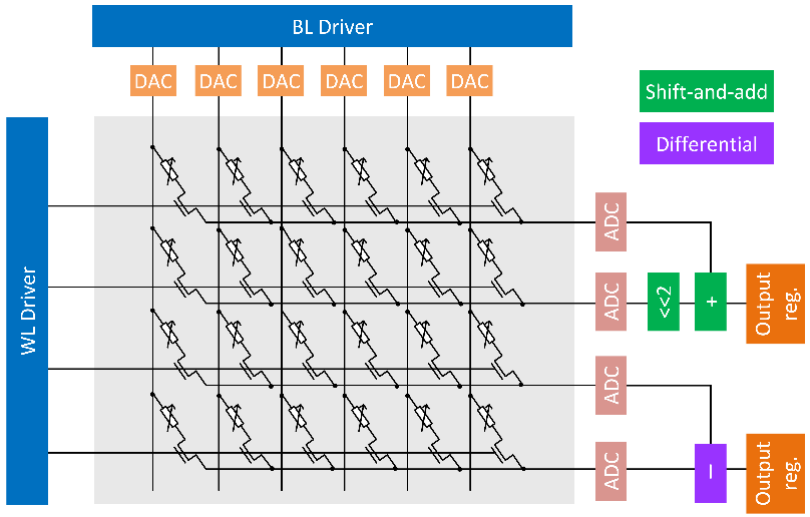
VGG-7 with CIFAR-10:

2 devices 1 weight

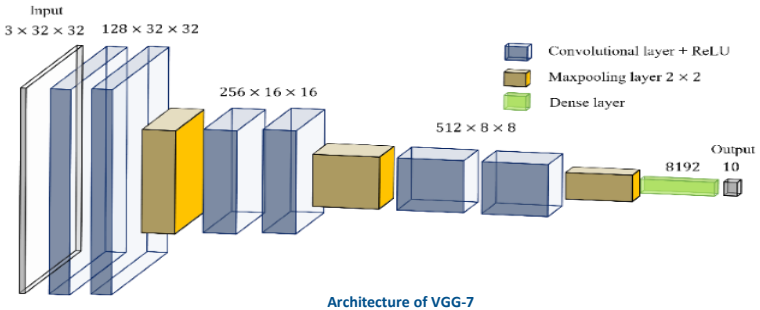
Two implemented methods to cluster devices and achieve the signed computation:
(1)Shift-and-add + Weight biasing

(2) Differential pair

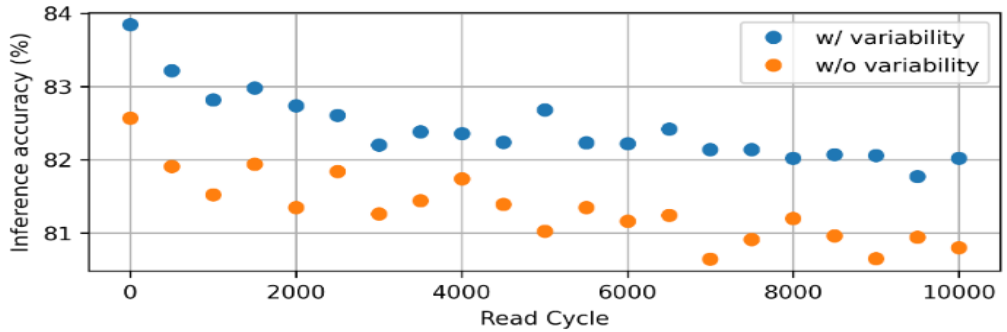
$$I_{out} = (G_+ - G_-) \cdot V_{in}$$



RRAM crossbar array with shift-and-add and differential pair



Architecture of VGG-7



Inference accuracy of VGG-7 with (2) differential pair

- (1) Shift-and-add with weight biasing failed with an accuracy of ~20% due to the amplified errors after shift operations
- Weights mapping with (2) differential pair is helpful to tolerate the impact of read disturb and device-to-device variability

J. Wen, A. Baroni, E. Perez, M. Ulbricht, C. Wenger and M. Krstic, "Evaluating Read Disturb Effect on RRAM based AI Accelerator with Multilevel States and Input Voltages," *IEEE International Symposium on Defect and Fault Tolerance in VLSI and Nanotechnology Systems (DFT)*, Austin, TX, USA, 2022, pp. 1-6,

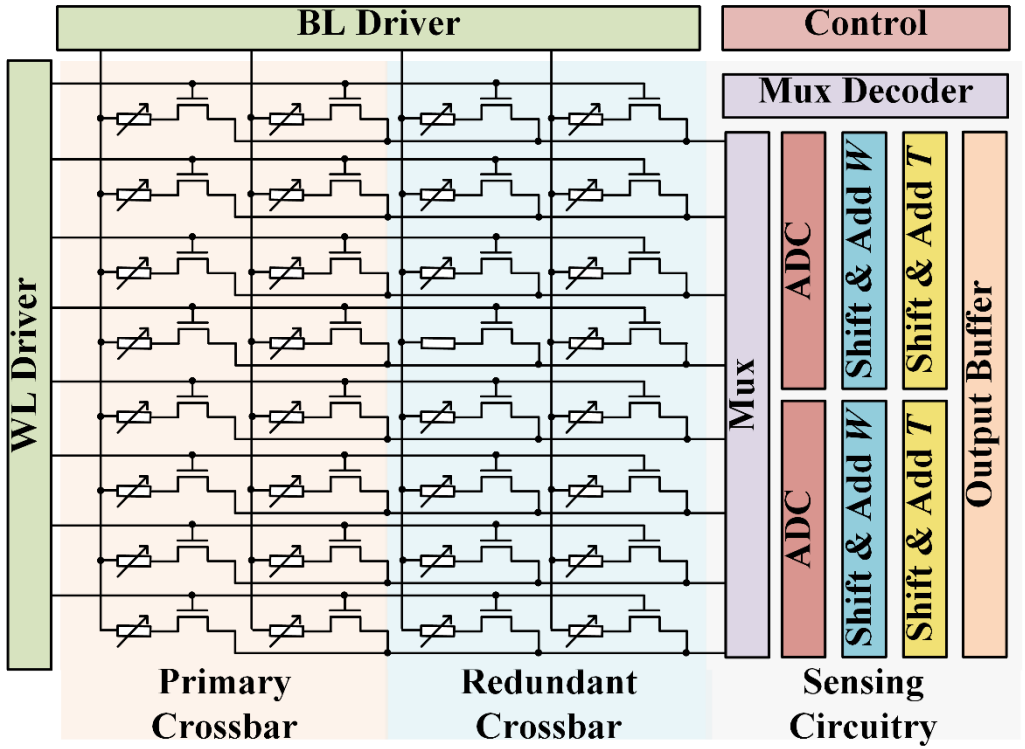
ReDiM

$$\mathbf{I}_{\mathbf{out},i} = \sum_{j=0}^{N-1} \mathbf{V}_{\mathbf{in},j} \cdot \mathbf{G}_{i,j}$$

- **Replicating** the RRAM crossbar
- **Decomposing** the large input voltages into smaller ones

$$\mathbf{I}_{\mathbf{out},i} = \sum_{j=0}^{N-1} \mathbf{V}_{\mathbf{in},j,\mathbf{primary}} \cdot \mathbf{G}_{i,j} + \sum_{j=0}^{N-1} \mathbf{V}_{\mathbf{in},j,\mathbf{reundant}} \cdot \mathbf{G}_{i,j}$$

$$\mathbf{V}_{\mathbf{in},j} = \mathbf{V}_{\mathbf{in},j,\mathbf{primary}} + \mathbf{V}_{\mathbf{in},j,\mathbf{reundant}}$$

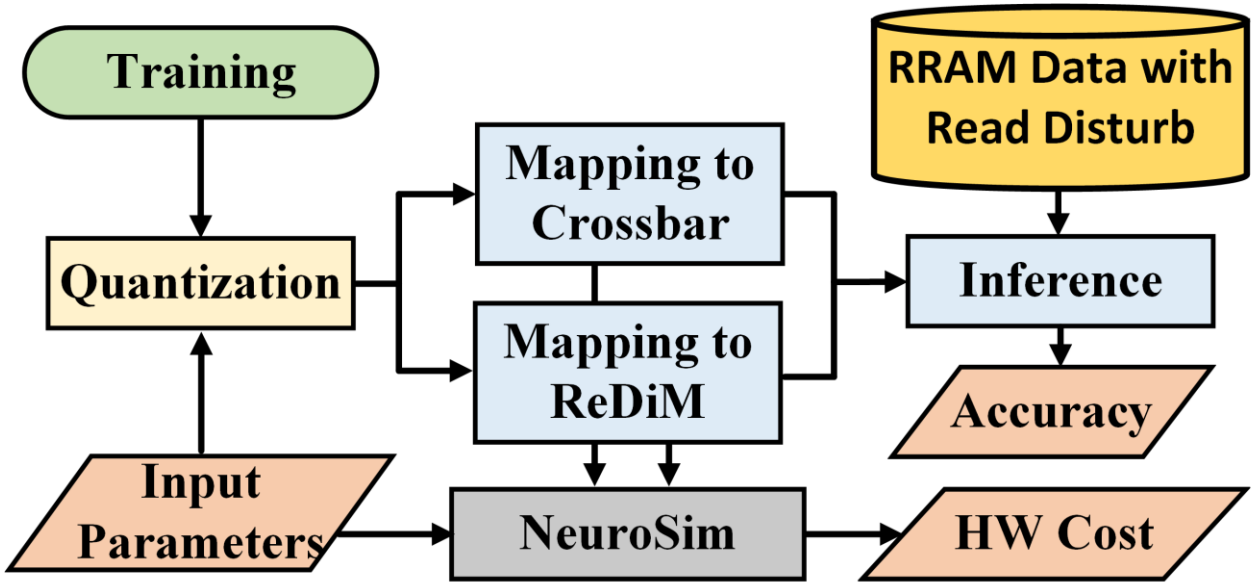


Architecture of ReDiM

- ✓ **Not decreasing** the **sense margin** of the sensing circuitry
- ✓ **No** extra **hardware overhead** on the sensing circuitry
- ✓ The hardware cost of crossbar is **much smaller** than the peripheral circuitry

J. Wen, at all, ReDiM: An Efficient Strategy for Read Disturb Mitigation in RRAM-Based Accelerators, IOLTS 2025

Simulation Framework



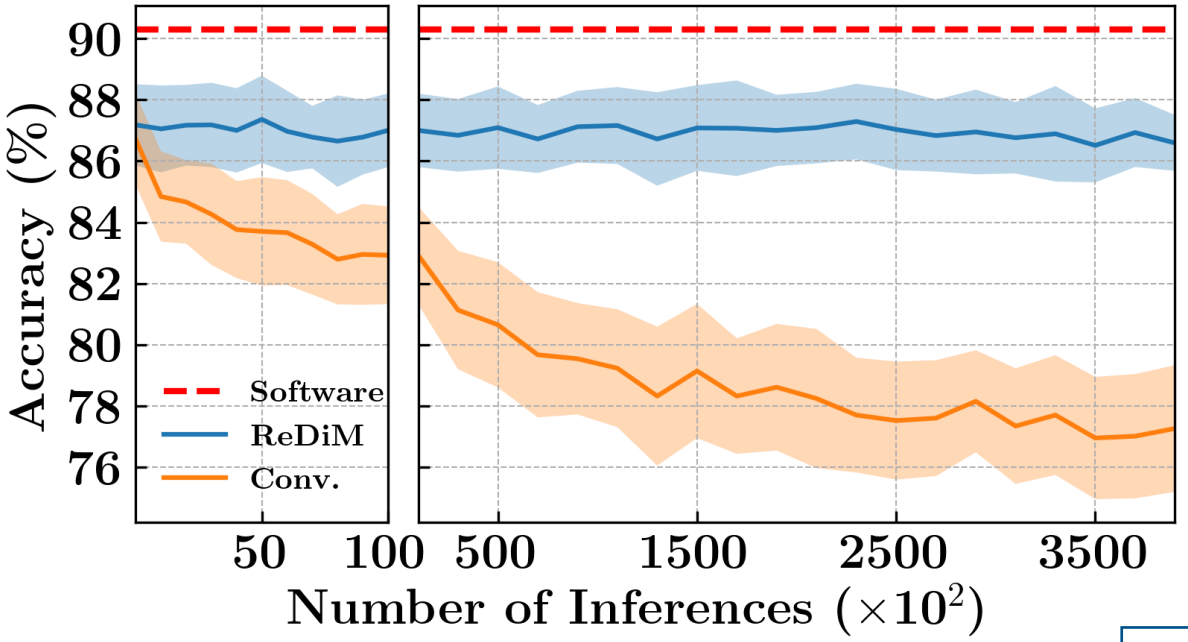
Developed simulation framework

- Evaluating two key aspects: **accuracy** and **hardware cost**
- Utilizing **measured RRAM data with read disturb** to model the device behavior at the system level
- Energy estimation based on **extracted mapped states** in the actual workload

J. Wen, at all, ReDiM: An Efficient Strategy for Read Disturb Mitigation in RRAM-Based Accelerators, IOLTS 2025

Experimental Result: Effectiveness

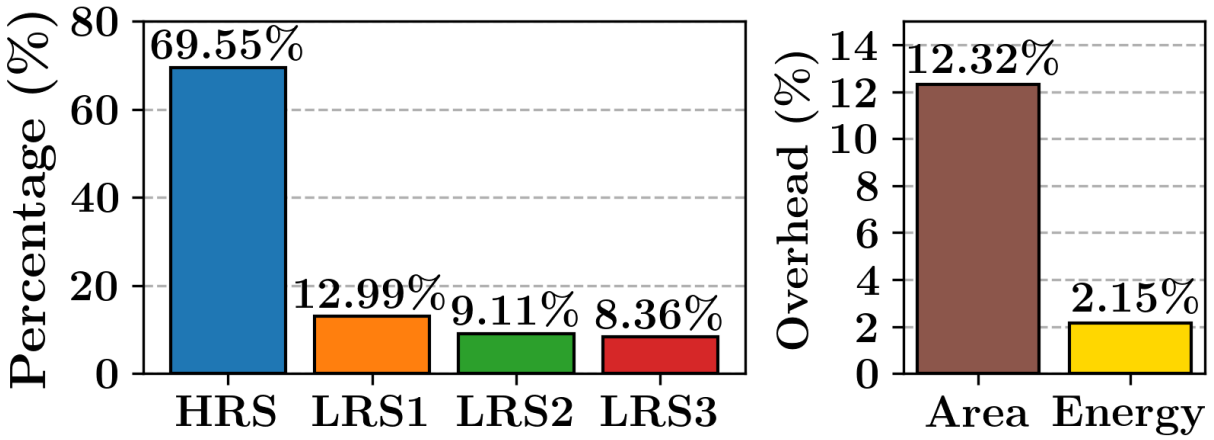
- **VGG-8** on **CIFAR10**
- Validation: **25 batches**, **mean values** and **mean ± 1std**
- Weight width = 8, ADC resolution = 6, num. of input pulses = 3
- Differential pairs for signed computations



- Both ReDiM and Conv. architectures suffer from quantization and device variability
- The **Conv.** architecture is **heavily impacted** by read disturb since **the beginning, continuous degradation**
- **ReDiM** effectively mitigates the read disturb
- **ReDiM** shows a **better stability** ➡ **less sensitive to the inputs**

J. Wen, at all, ReDiM: An Efficient Strategy for Read Disturb Mitigation in RRAM-Based Accelerators, IOLTS 2025

Experimental Result: Hardware Overhead



ReDiM: **Lightweight** design with **strong** mitigation **effectiveness** on **read disturb**

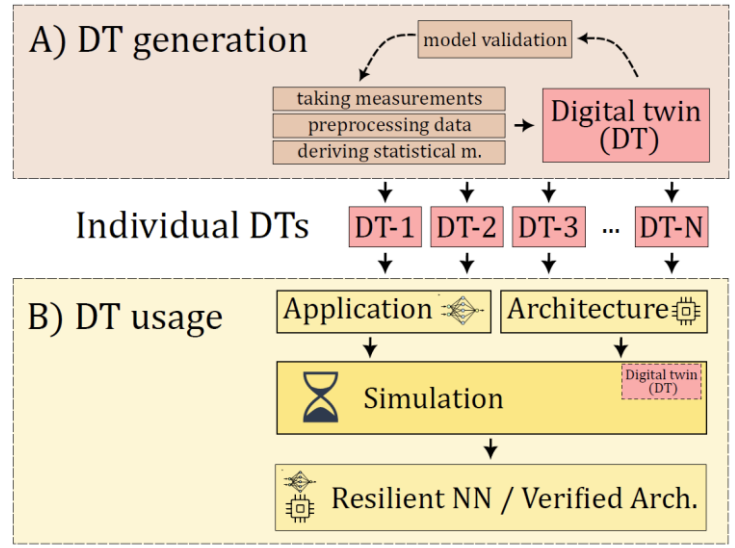
- HRS dominates the conductance states ➡ allign with observed accuracy decrease
- Array-level evaluation for inference ➡ RRAM crossbar, peripheral circuitary
- ReDiM introduces **slight hardware increase** compared to the conventional design: RRAM crossbar, BL driver

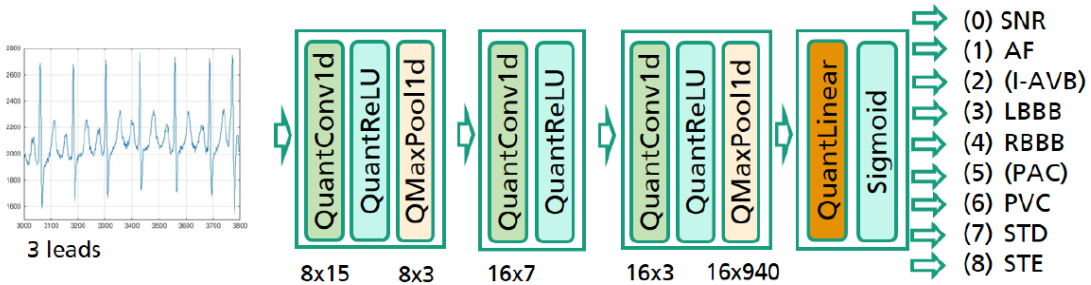
J. Wen, at all, ReDiM: An Efficient Strategy for Read Disturb Mitigation in RRAM-Based Accelerators, IOLTS 2025

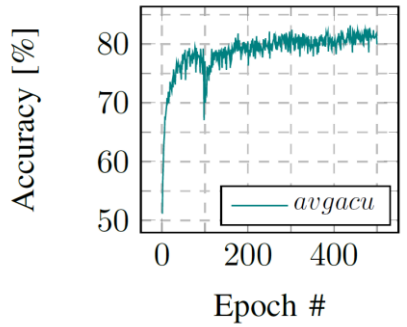
Using Digital Twins to model RRAM variation effects

From raw measurements to application

- Joint work with IHP, BTU, Fraunhofer, RWTH, PGI (7,10)
  - We created an approach for the automated creation of “digital twins”, integrated them into a NN framework & applied FAT
  - At the push of a button this can cover many variation types
    - D2D, C2C, Endurance, …







DT: Fritscher et al.: “Prototyping reconfigurable RRAM-based AI accelerators using the RISC-V ecosystem and Digital Twins”. In: ISC High Performance 2023

Summary

- Reliability validation of AI processing systems is extremely challenging
  - Emerging technologies bringing new challenges
  - Exhausting efforts required for this

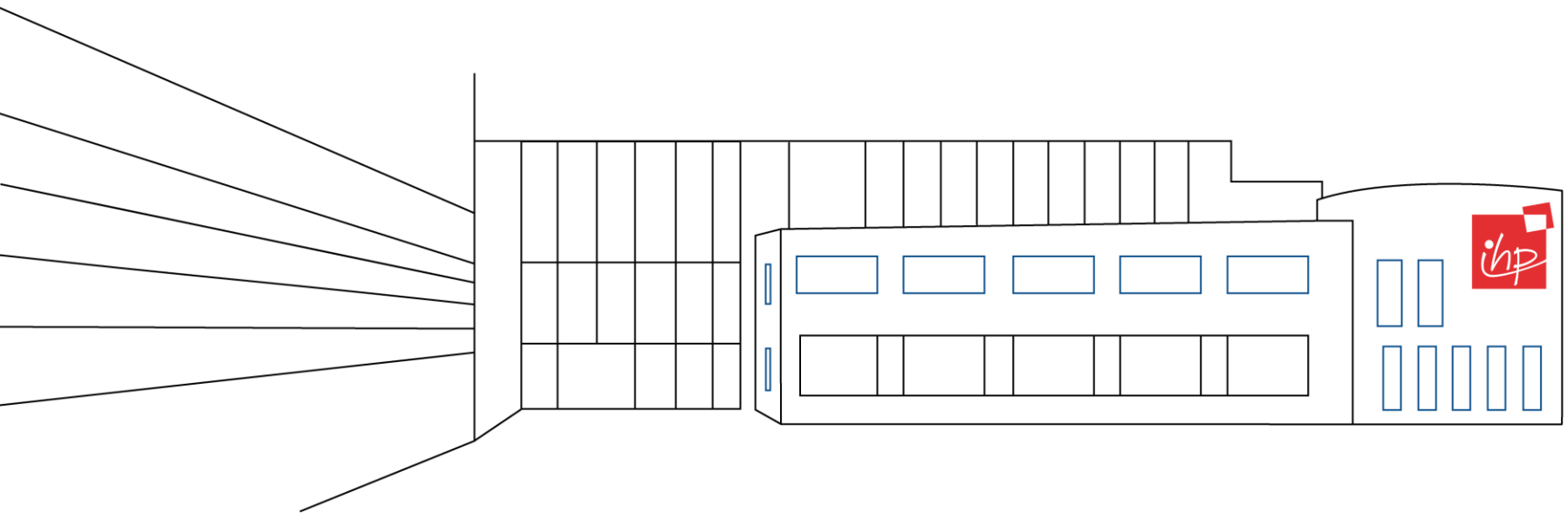
- Innovative approaches in validation are required
  - New modelling approaches
  - Combination of hardware and high level verification
  - New fault injection strategies

Plenty of scope for further research

Thanks to my team members: Anselm Breitenreiter, Junchao Chen, Alessandro Veronesi, Jianan Wen, Li Lu, Rizwan Tariq Syed, Markus Fritscher

Thank you for your attention!

IHP GmbH – Leibniz Institute for High Performance Microelectronics
Im Technologiepark 25
15236 Frankfurt (Oder)
Tel.: +49 (0) 335 5625 729
E-Mail: krstic@ihp-microelectronics.com



Leibniz Association