

Introduction

- **Challenge:** Deep learning advancements improve accuracy and efficiency, but reliability is crucial for safety-critical applications (e.g., autonomous systems, avionics).
- **Limitations of Existing methods:** Redundancy-based fault mitigation is effective but costly in computation and energy.
- **Focus of This Research:** Developing a reliability-aware hardware-software co-design for Systolic Array accelerators to enhance efficiency and fault tolerance.

Methodology

- **RunSAFER: a Runtime SA Fault Detection mechanism [1]**
A **resource-efficient** fault detection mechanism combining **Algorithm-based Fault Tolerance (ABFT)** and **scan chain** methodology
 - Applying **test patterns** to SA to compute **complemented checksums** on the current DNN weight matrix.
 - Comparing SA checksums to reference values exploiting the Accumulators.
 - Producing **Reference values by the Accumulators** working in asymmetric **SIMD**, simultaneously processing weight streams and partial products.
 - Implementation of the method as **custom matmul instruction**, requiring 2 additional clock cycles to process the test vectors.

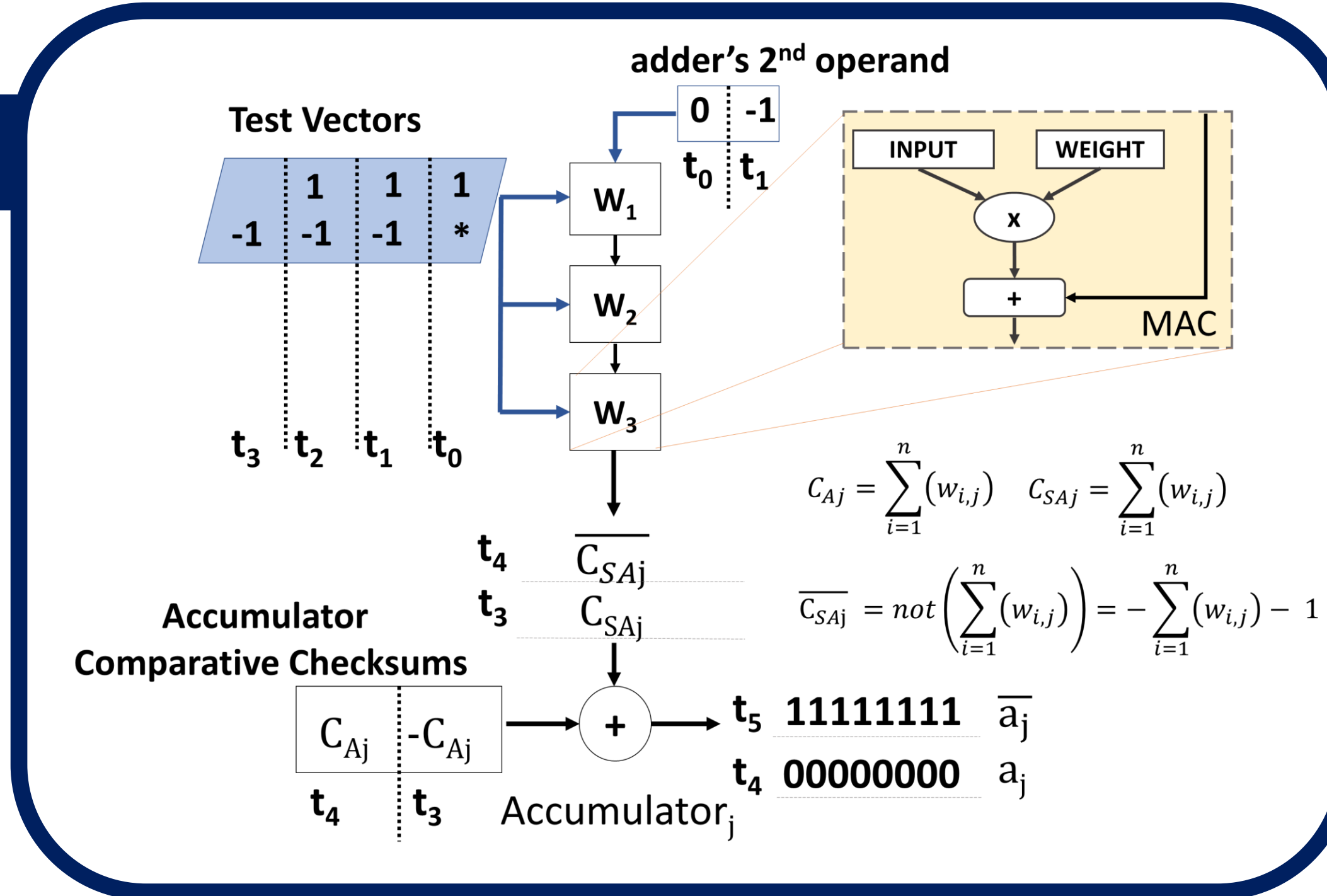


Fig. 1. The RunSAFER method.

- **RePAIR: Reconfigurable Platform for AI Resilience within RISC-V Ecosystem [2]**

The methodology is **integrated into the ISA** of an open-source **TPU** and the **NEORV RISC-V** core to implement **RePAIR**, a reliability-driven RISC-V ecosystem for accelerating DNNs on FPGA. It leverages **Dynamic Partial Reconfiguration** for **fast detection, correction, and recovery** from SEE-induced faults.

- **ZOR: Zero Overhead Reliability Strategies [3]**

Analyzing **fault impact** in SA through **datapath analysis, fault propagation modeling, and hardware fault injection**, showing that data mapping strategies respond differently to permanent faults.

- **Weight Stationary** is **more sensitive** than Input Stationary.
- Proposing a **new mapping reliability-oriented strategy** based on data-driven resource rotation.

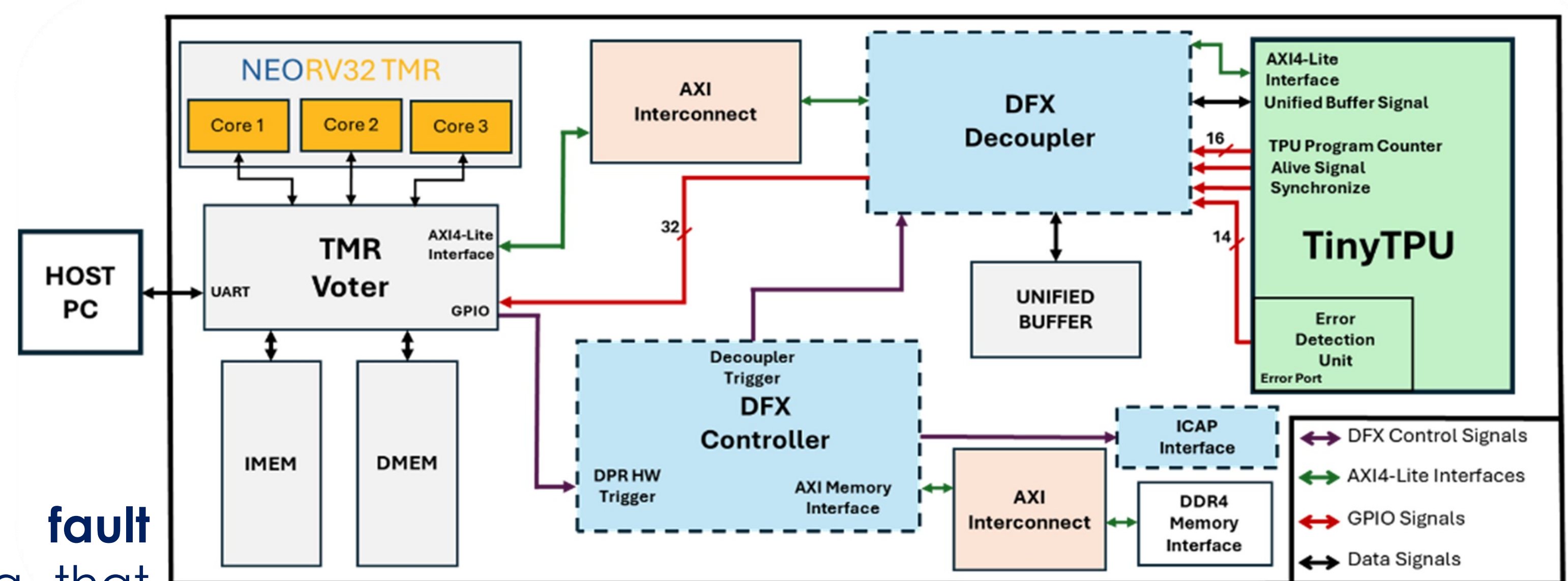


Fig. 2 The RePAIR platform.

Experimental Results

- The proposed methodologies were tested on the tinyTPU core, an open-source accelerator, implemented on the KCU105 board with an AMD Ultrascale SRAM-based FPGA.
- **Fault Model:** bitflip in configuration memory to emulate SEU-induced faults in accelerator datapath
- **Benchmark Application:** CNNs models on MNIST-digit and CIFAR10

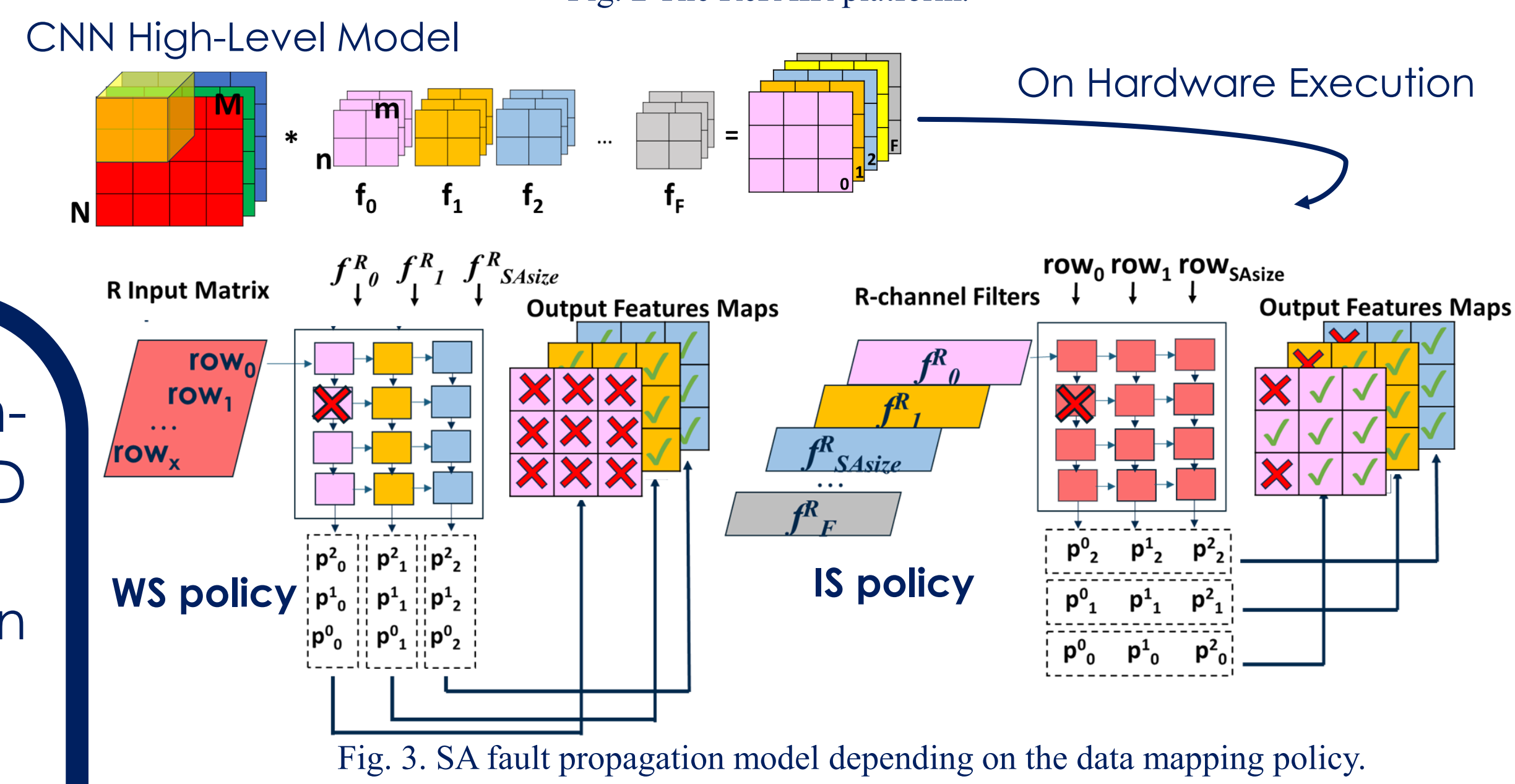


Fig. 3. SA fault propagation model depending on the data mapping policy.

Table 1. RunSAFER characteristics.

COMPARISON WITH ABFT METHODS FOR SYSTOLIC ARRAY N X N			
	[4]	[5]	RunSAFER
PEs	N	0	0
Checksums	N adders	2N+1 adders, 1 MAC	0
Detection Comparators	N	N+2	N
COMPARISON WITH SCAN METHODS FOR SYSTOLIC ARRAY 256x256			
	[6]	[7]	RunSAFER
Required Test Patterns	11	12	2 (+1 for LSB checking)
Area Overhead	NA	5.25%	0.31%
Fault Coverage SA	100%	100%	95% (on FPGA device)
Runtime	X	X	✓

Table 2. RePAIR platform post-implementation details.

Platform Modules	LUTs	FFs	BRAMs	DSPs
TinyTPU	4,294	7,211	181	210
TMR NEORV32	3,219	3,180	3	0
DPR Logic	1,185	989	0	0
Glue Logic Resources	13,874	17,670	95.5	3
Total [%]	9.31%	5.99%	46.58%	11.09%

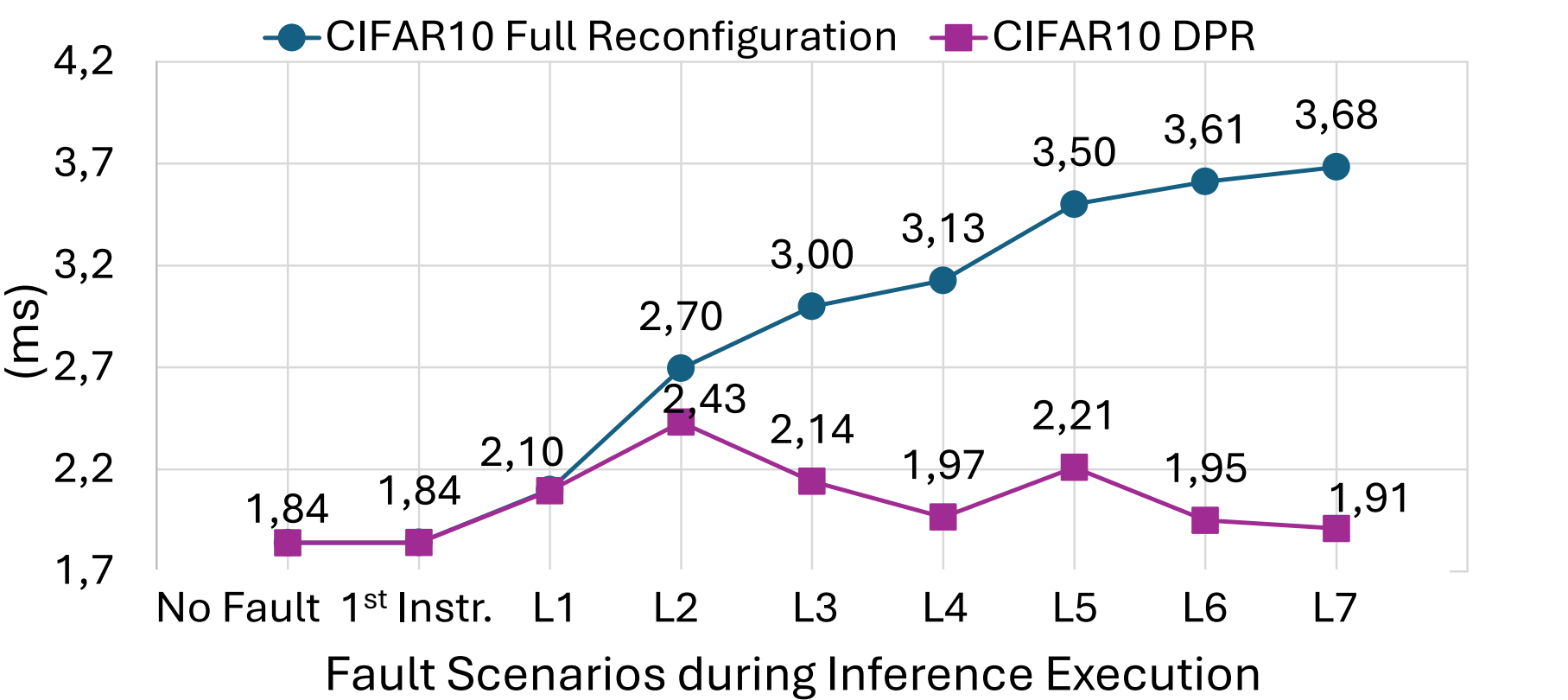
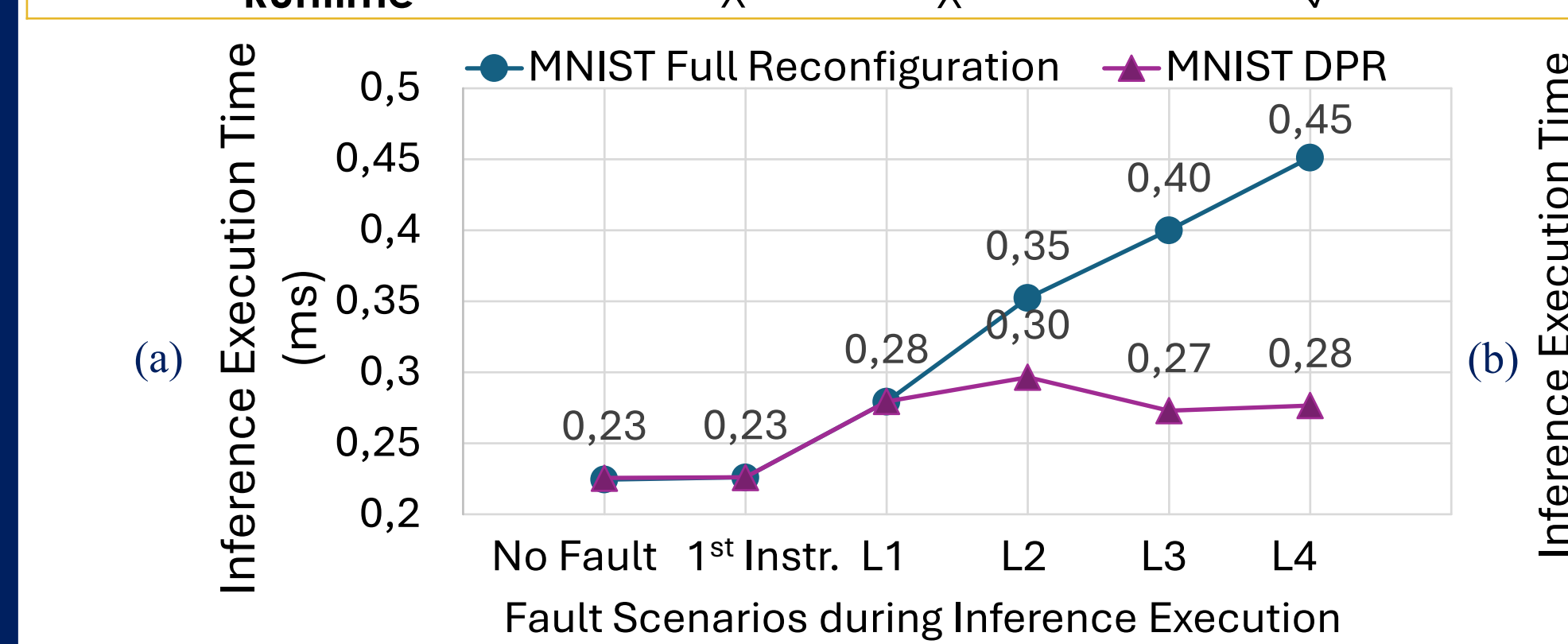


Fig. 5. Total inference execution time considering fault occurrences at different stages of the inference process with and without DPR for MNIST(a) and CIFAR10 (b).

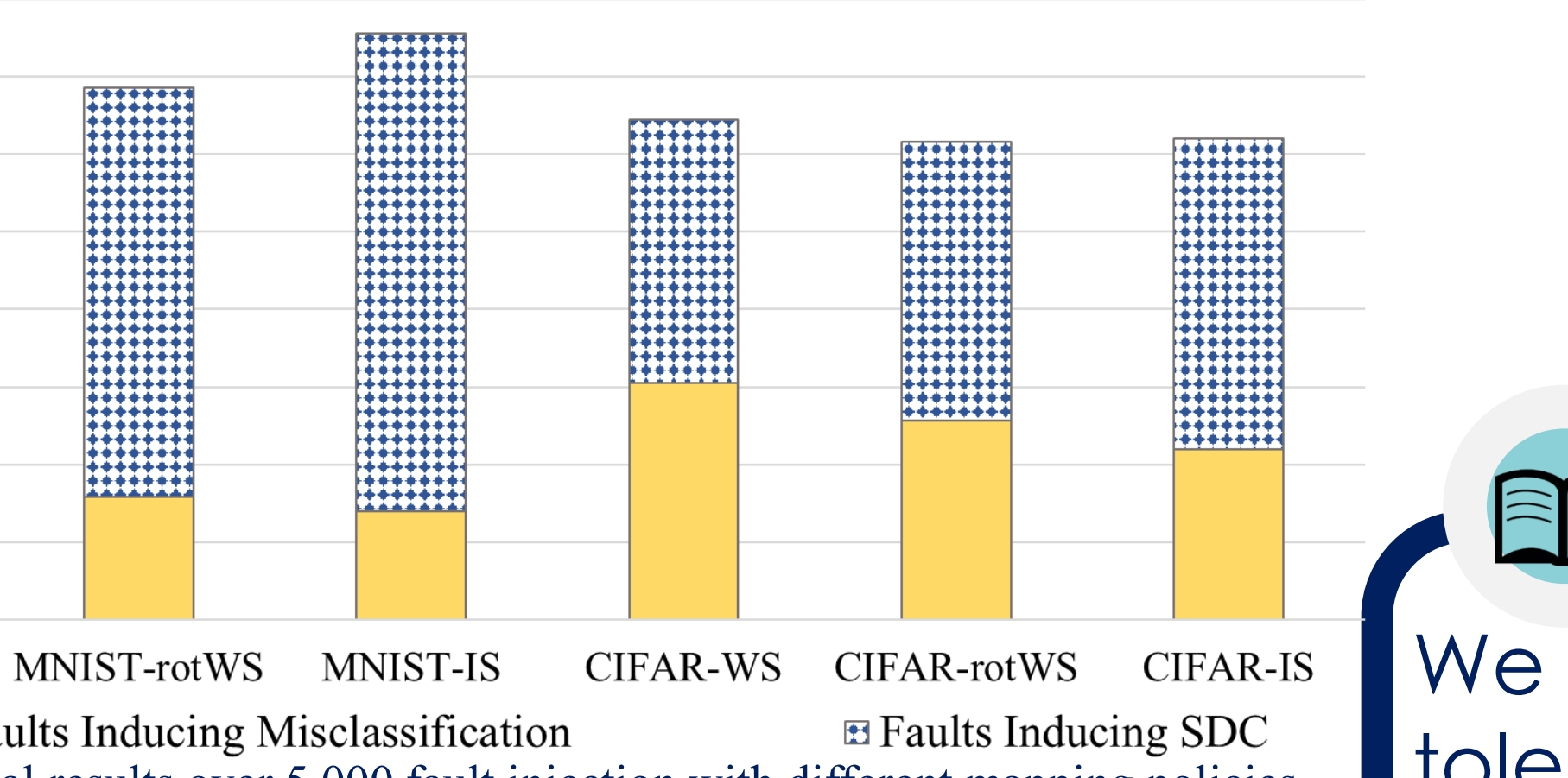
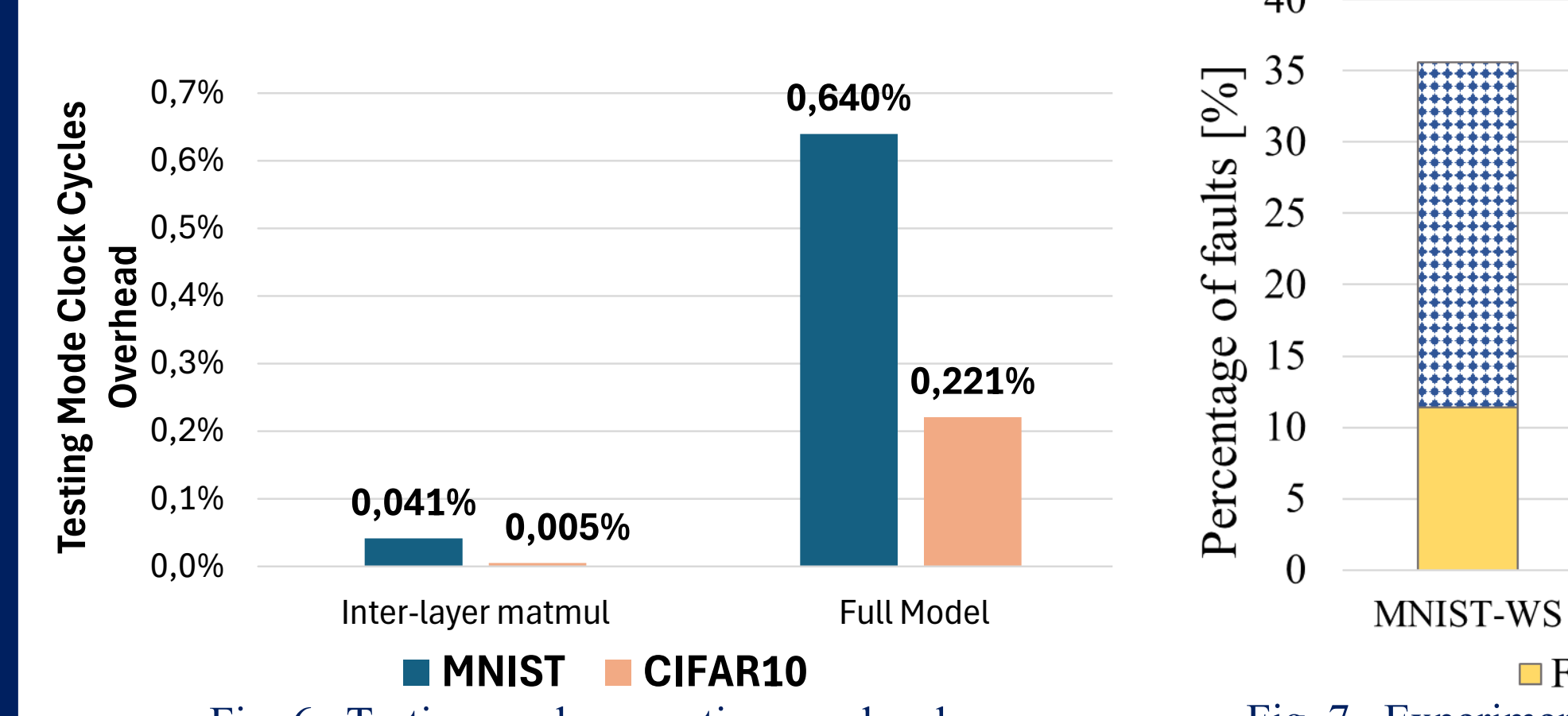


Fig. 6. Testing mode execution overhead. Fig. 7. Experimental results over 5,000 fault injection with different mapping policies.

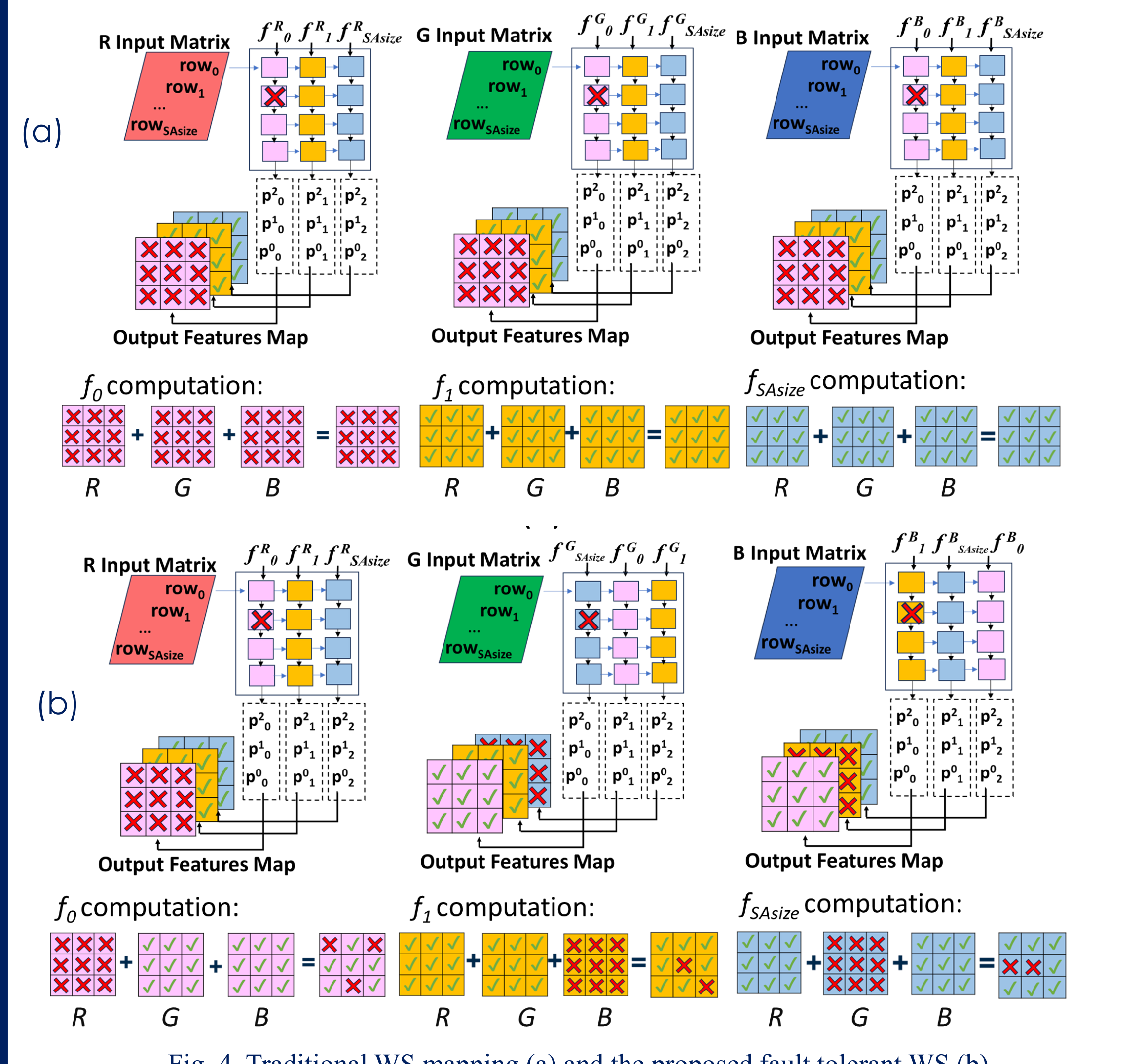


Fig. 4. Traditional WS mapping (a) and the proposed fault tolerant WS (b).

Conclusions and Future Works

We proposed reliability strategies for SA, from testing to tolerance, leveraging fault propagation insights and accelerator datapath analysis for cost-effective solutions. Future work will focus on lightweight, reliability-oriented algorithms.

1. E. Vacca et al., "RunSAFER: A Novel Runtime Fault Detection Approach for Systolic Array Accelerators," IEEE 41st International Conference on Computer Design (ICCD), 2023.
 2. E. Vacca et al., "ZOR: Zero Overhead Reliability Strategies for AI Accelerators," 22nd IEEE Interregional NEWCAS Conference (NEWCAS), 2024.
 3. G. Cora, E. Vacca, et al., "RePAIR: Reconfigurable Platform for AI Resilience within RISC-V Ecosystem," 21st International Symposium on Applied Reconfigurable Computing, 2025.
 4. M. Safarpour et al., "Algorithm Level Error Detection in Low Voltage Systolic Array" in IEEE Transactions on Circuits and Systems II: Express Briefs, Feb. 2022.
 5. F. Libano, et al., "Efficient Error Detection for Matrix Multiplication with Systolic Arrays on FPGAs," in IEEE Transactions on Computers, Aug. 2023.
 6. J. Kim, et al., "ZOS: Zero Overhead Scan for Systolic Array-based AI accelerator", 2022 19th International SoC Design Conference (ISOCC), 2022
 7. H. Lee, et al., "STRAIT: Self-Test and Self-Recovery for AI Accelerator", in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2023.