Enhancing RISC-V Ecosystem for SEU-Resilient Inference on FPGA-Based Implementations

Giorgio Cora, Eleonora Vacca, Corrado De Sio, Sarah Azimi, Luca Sterpone

Department of Control and Computer Engineering -
Aerospace, Safety and Computing (ASaC) Lab

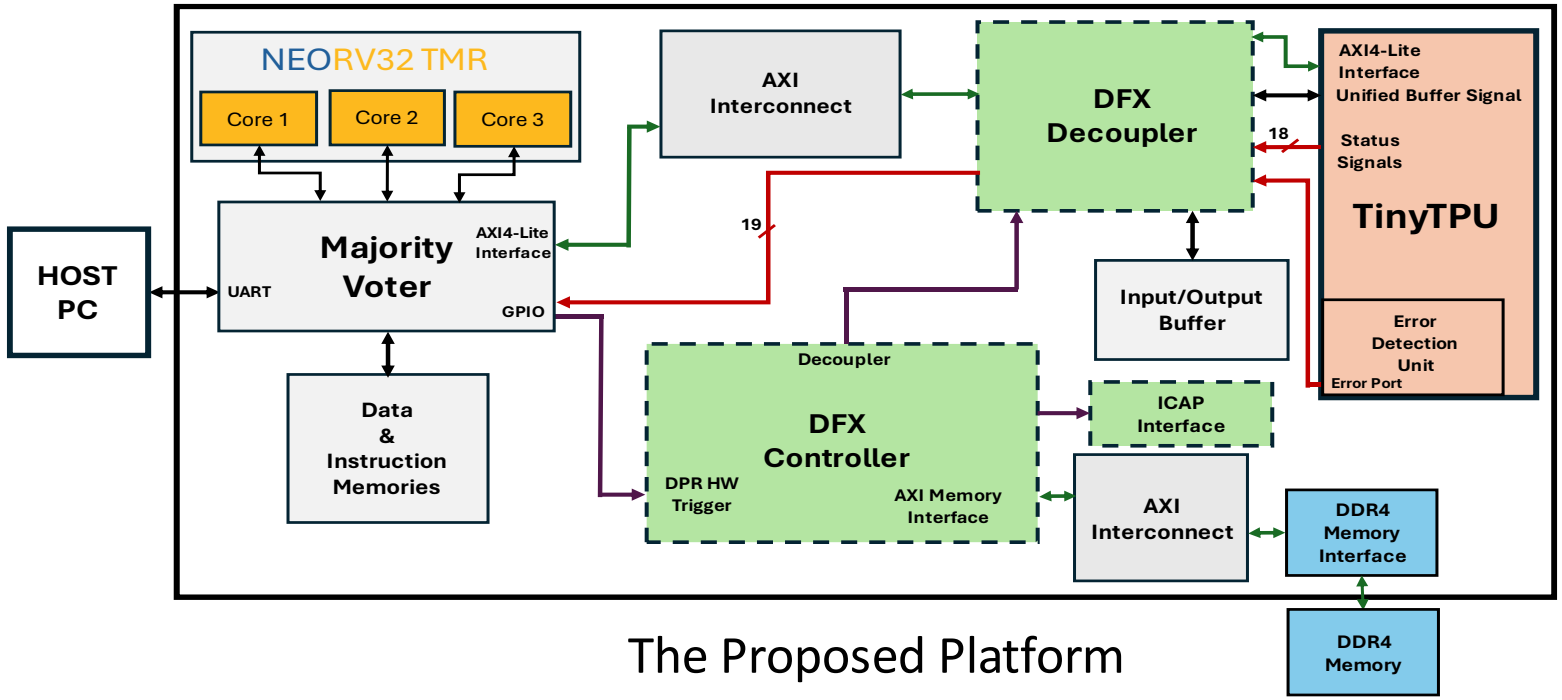
Politecnico di Torino

Motivations

- The increasing adoption of **RISC-V** architectures, combined with **AI accelerators for safety-critical applications**, has led to the development of various platforms, with their **main focus being performance** optimization.
- Adoption of reliability features are often **limited to traditional approaches**, causing significant losses in terms of system's availability time.

- The proposed platform embeds:

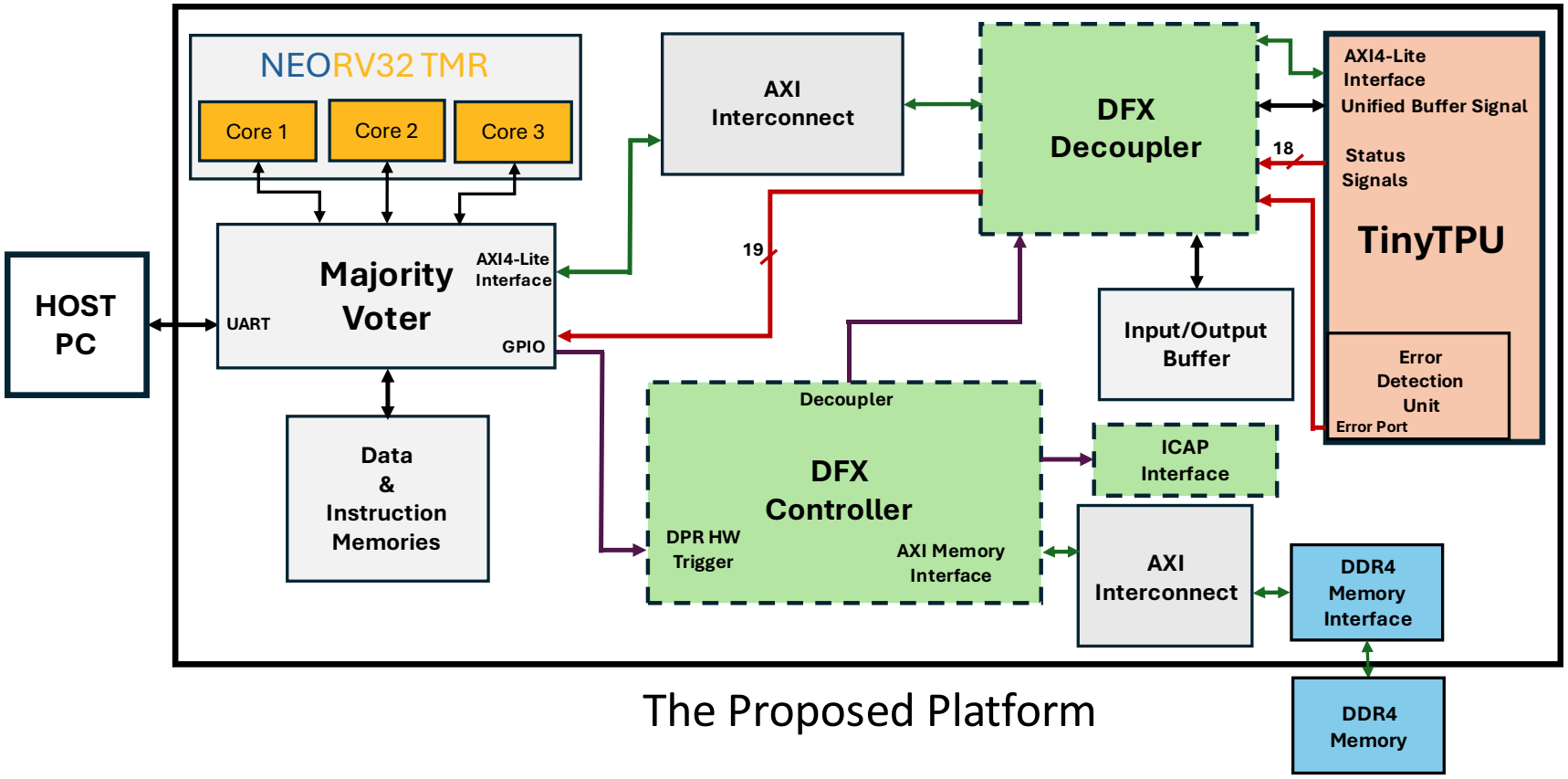
1. **Minimal time losses** in case of errors.
2. **Low-latency error detection** mechanism.
3. A **RISC-V architecture** to reduce area occupation.
4. **Partial reconfiguration** for fast error correction.



The Proposed Platform

Outline

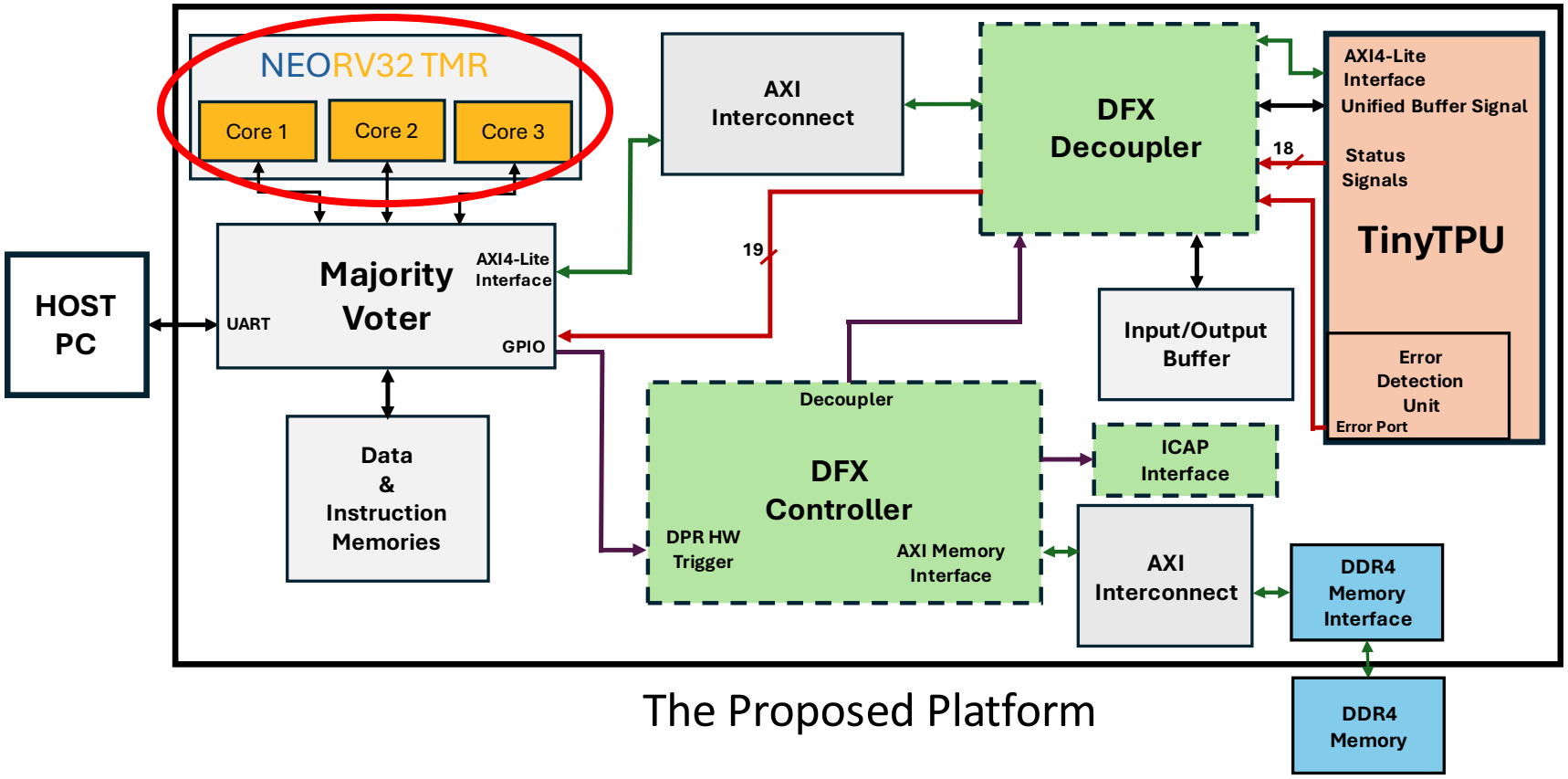
- The Proposed Platform
- TPU Architecture and Error Detection Mechanism
- Hardware and Software Setup
- Result Analysis
- Conclusions and Future Works



The Proposed Platform

Neorv32:

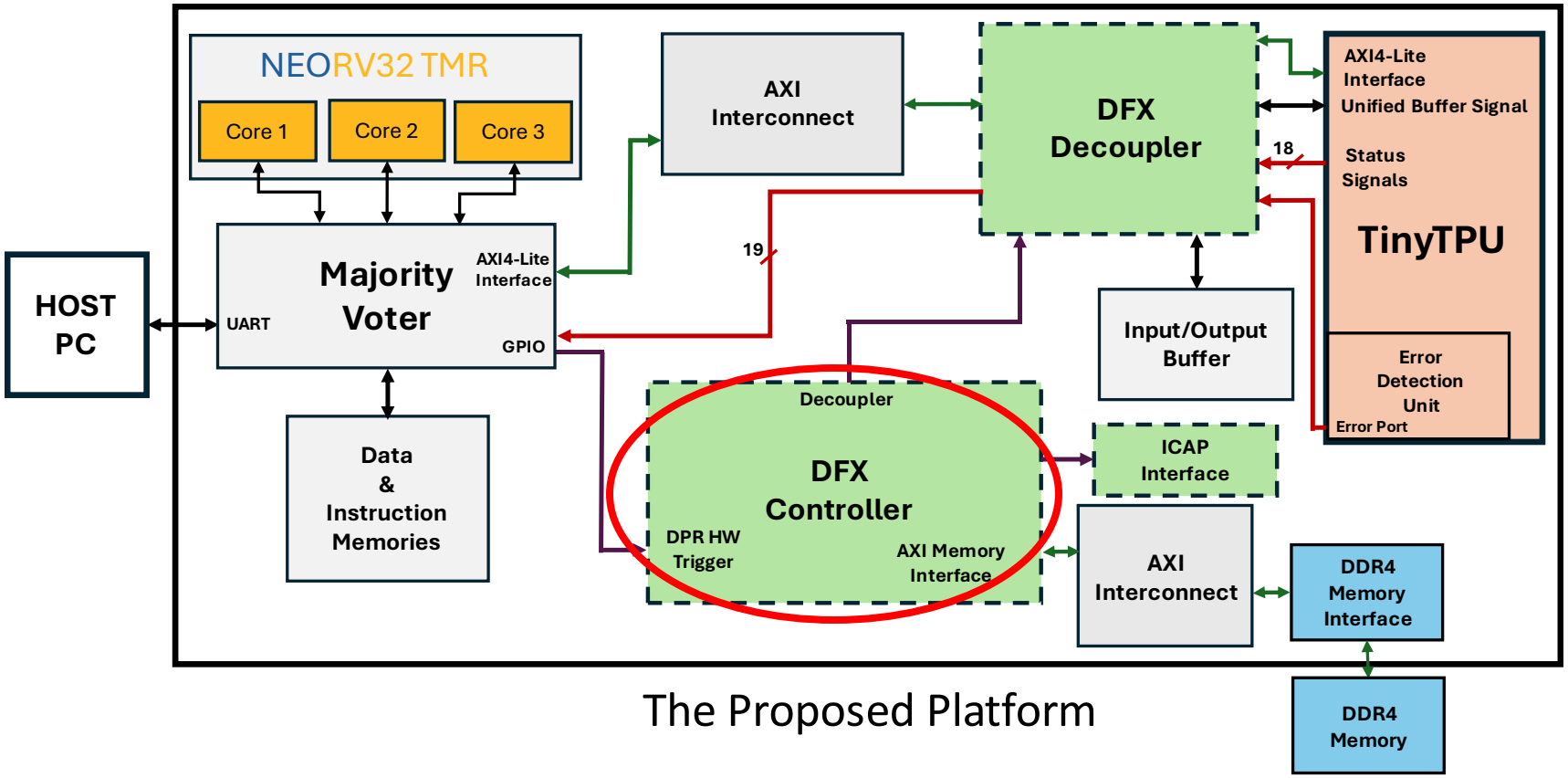
- 32-bit VHDL based Architecture implementing **RV32I ISA**.
- **AXI4-LITE** Interface for TPU and DDR memory communication.
- **UART** Communication with the Host.
- **GPIO Interfaces** for Error Detection and Partial Reconfiguration management.
- **TMR** for Improved Reliability.



The Proposed Platform

Partial Reconfiguration Support:

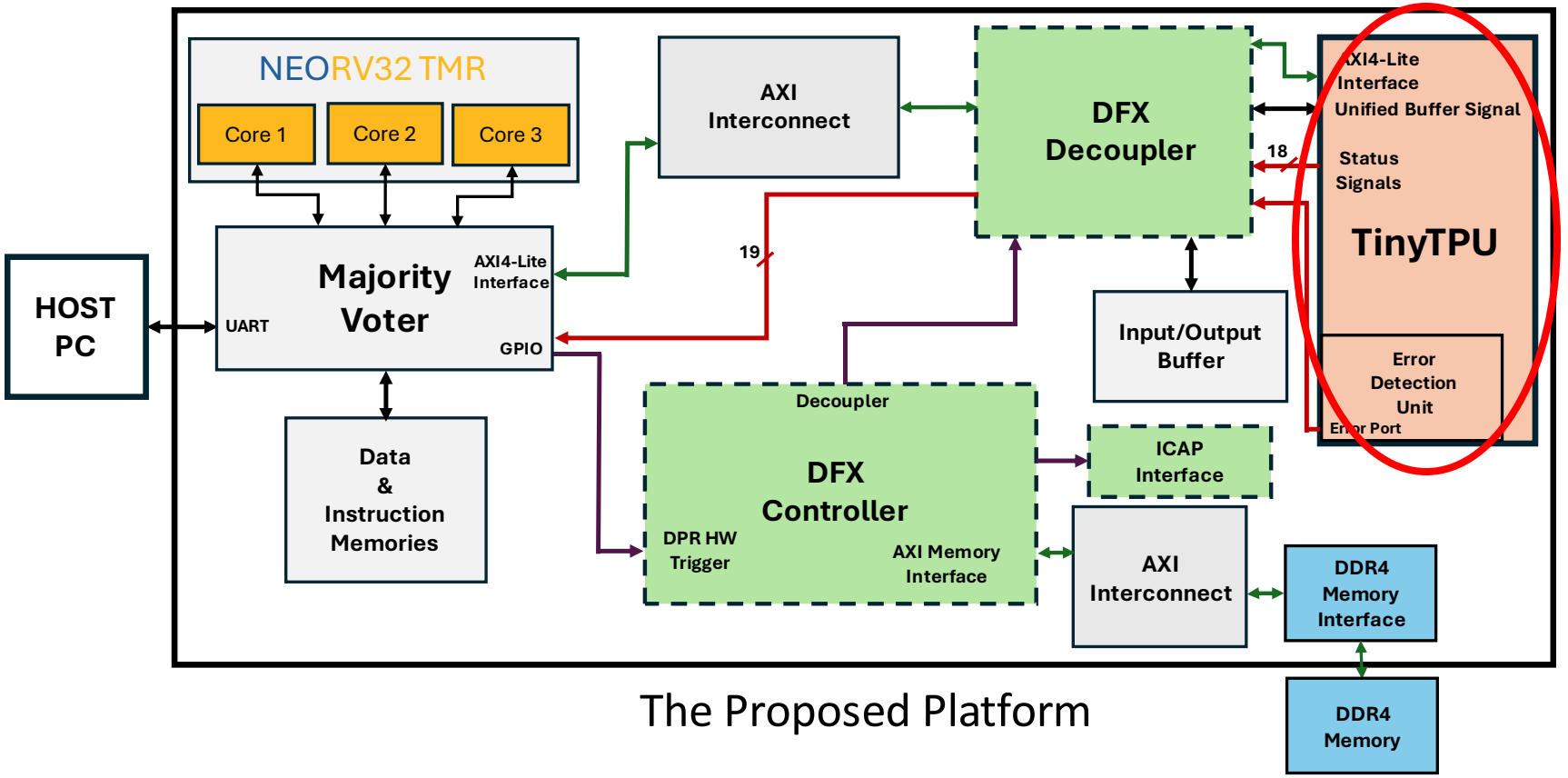
- **Minimal Recovery** Time.
- **Autonomous Recovery** in case of Errors.
- **Execution resumption** from last correct state.
- **Minimal System downtime** in case of errors.



The Proposed Platform

TinyTPU:

- **Configurable Systolic Array size**, from 6x6 to 14x14 MAC units.
- **80-bits ISA**.
- Designed for DNN execution.
- Support for **ReLU and Sigmoid** activation.
- **Custom ISA** for Error Detection.
- **Minimal** execution time **overhead**.



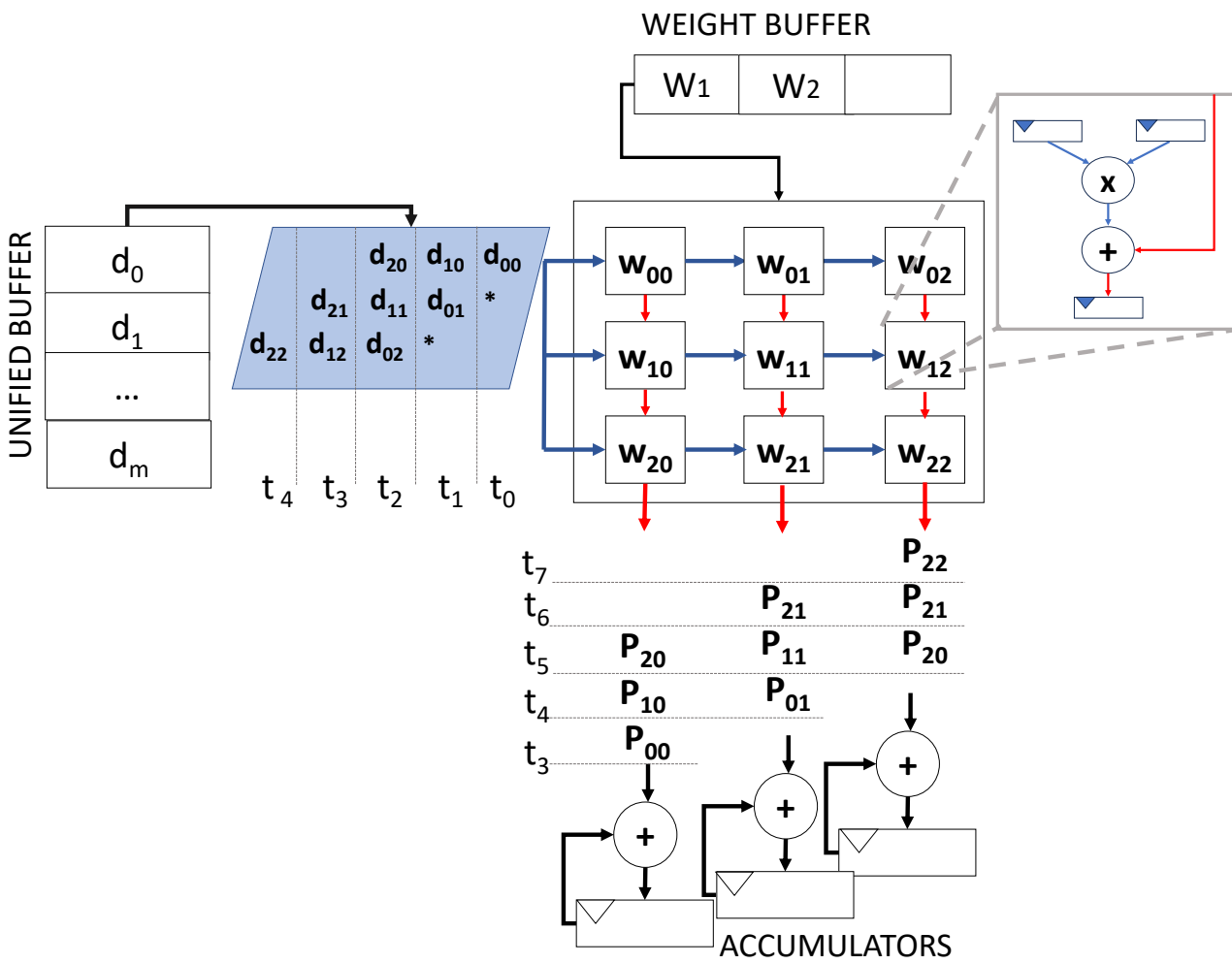
The Proposed Platform

The TPU consist of:

- Systolic-based accelerators consist of **2D arrays of Multiply and Accumulate (MAC) units**.
- **Custom fault detection** architecture.

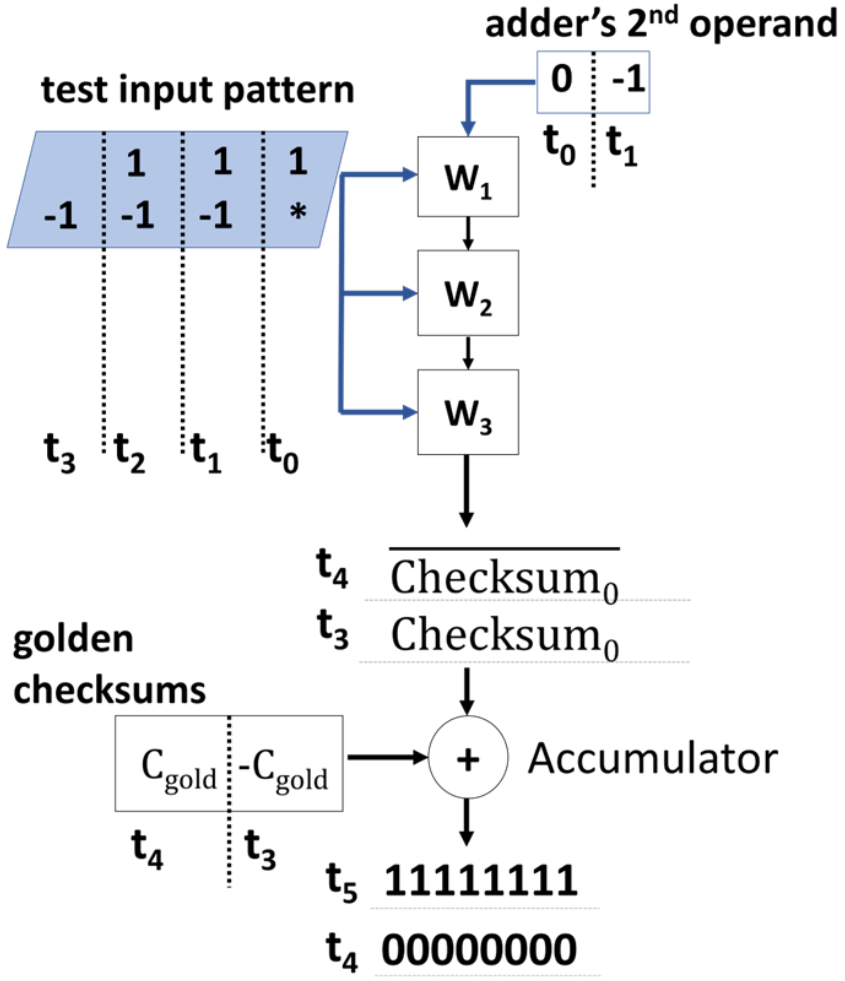
Conventional fault mitigation techniques face significant limitations:

1. High hardware overhead.
2. Lack of runtime testing.
3. Dependence on datapath resources, assuming they are fault-free.



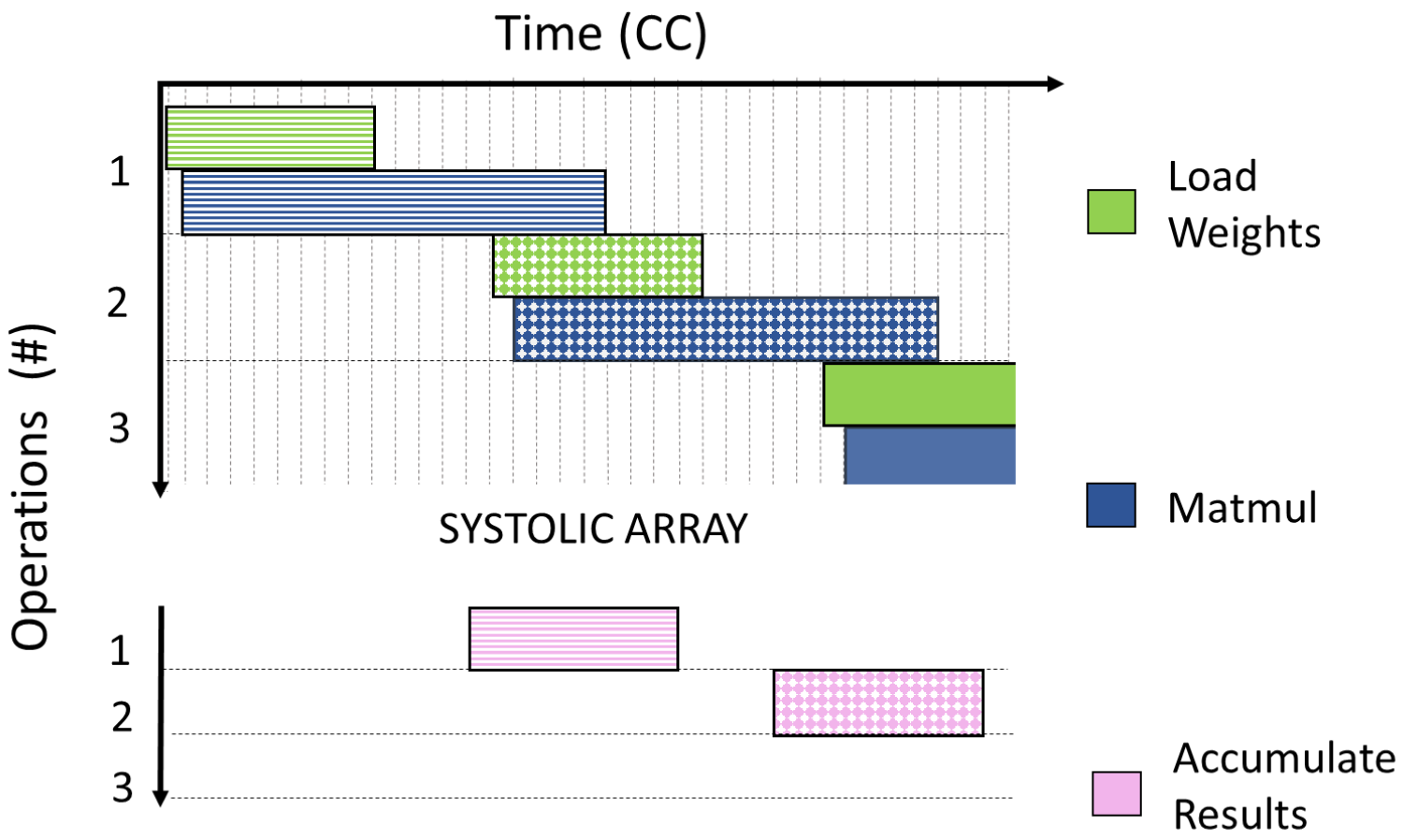
The Systolic Array Datapath

- A **custom fault detection mechanism** is integrated within the accelerator datapath.
- Internal resources are exploited to **minimize hardware overhead**.
- Error detection is achieved using **checksum computations**.
- Checksums computations ensure **minimal time overhead**.

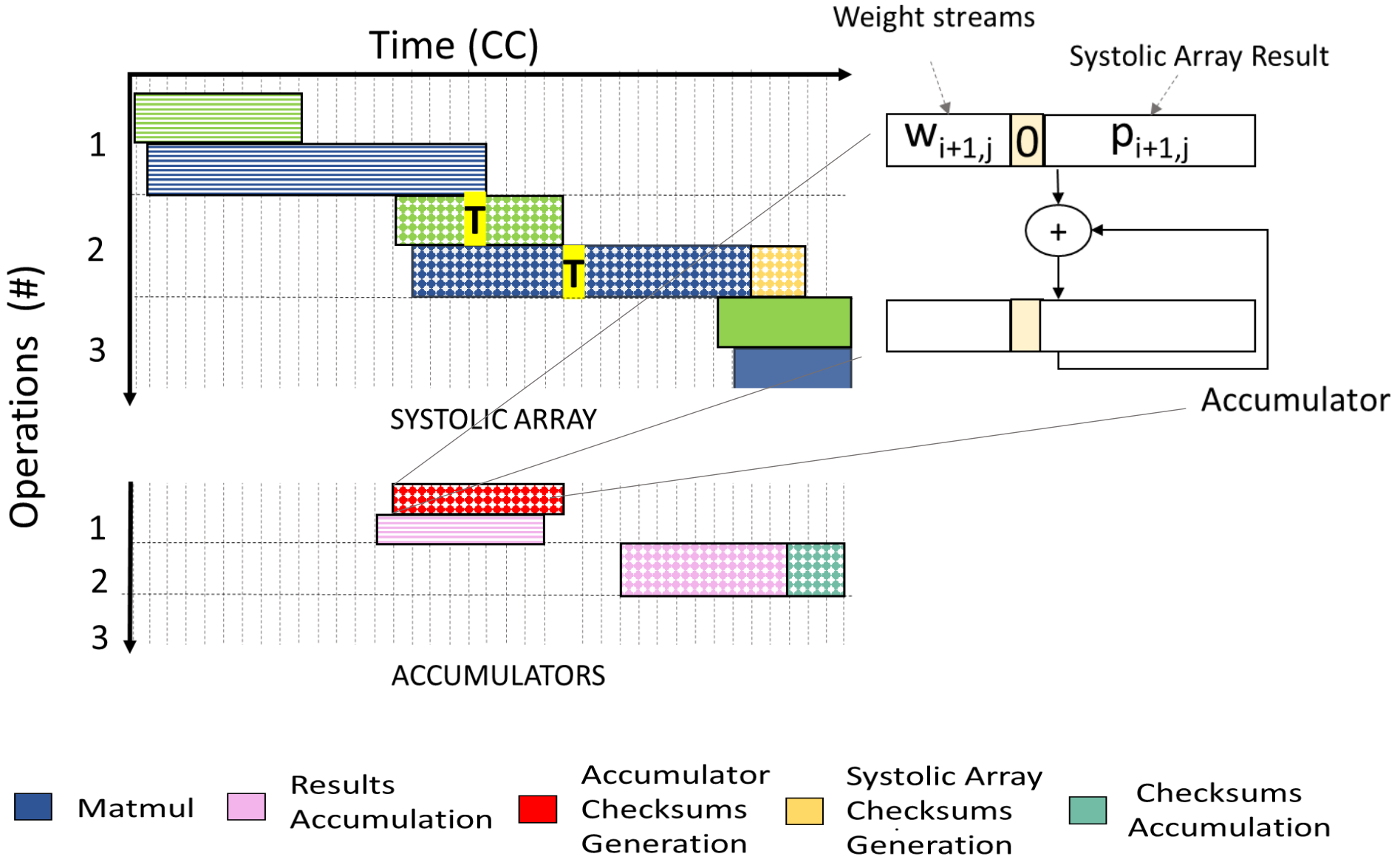


The Systolic Array Datapath

- In both the **original and modified pipelines**, matrix multiplication is executed as a **sequence of load weight** instruction followed by a **matmul** instruction.
- Once the **results are** generated and **processed by the accumulators**, a new set of instructions can be fetched and begins execution.



- In the modified pipeline, **ISA has been extended** to support **normal and testing operation modes**.
- In testing mode, each instruction requires a **maximum overhead of 3 clock cycles**.



KCU105 Development Board:

- High Number of logic Resources and Memory Elements.
- Better Place&Route Flexibility.
- Direct access to DDR storage.

Adopted CNN-Datasets:

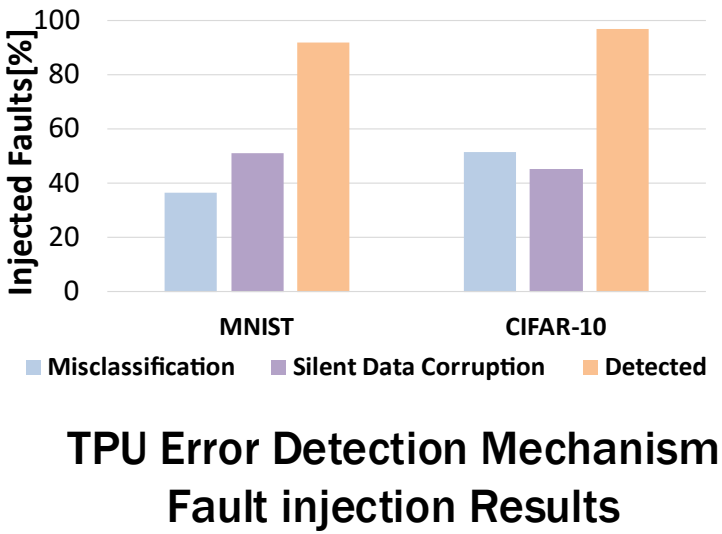
- CIFAR-10
- MNIST

Proposed Platform Resources utilization

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%

Result Analysis: Error Detection Mechanism Performances

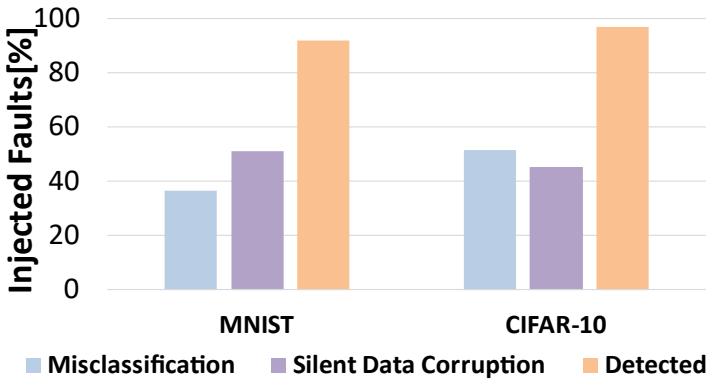
- A total of 20 thousand faults injected in the TPU Datapath.
- **Up to 94%** of faults (misclassification and SDC) **detected**.



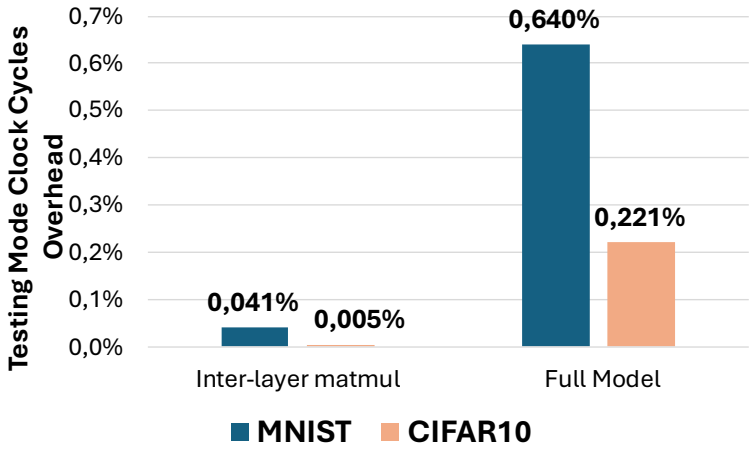
TPU Error Detection Mechanism Fault injection Results

Result Analysis: Error Detection Mechanism Performances

- A total of 20 thousand faults injected in the TPU Datapath.
- **Up to 94%** of faults (misclassification and SDC) **detected**.
- **Maximum overhead of 0,64%** when full model is executed in testing mode.
- **Minimum overhead (0,041%/0,005%)** when inter-layer matmul are executed in testing mode.

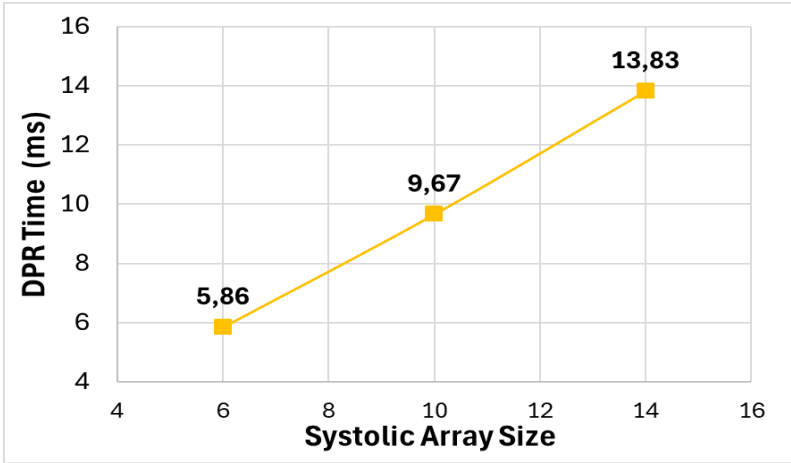


TPU Error Detection Mechanism Fault injection Results



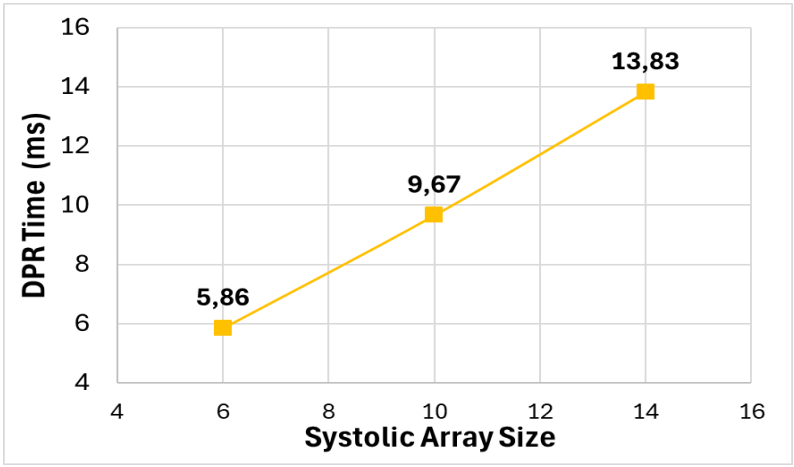
Error Detection mechanism Time Overhead

- The **DPR time scales** linearly with the **size of the Systolic Array**, from less than 6ms in the smallest case to around 14ms for the larges SA size.

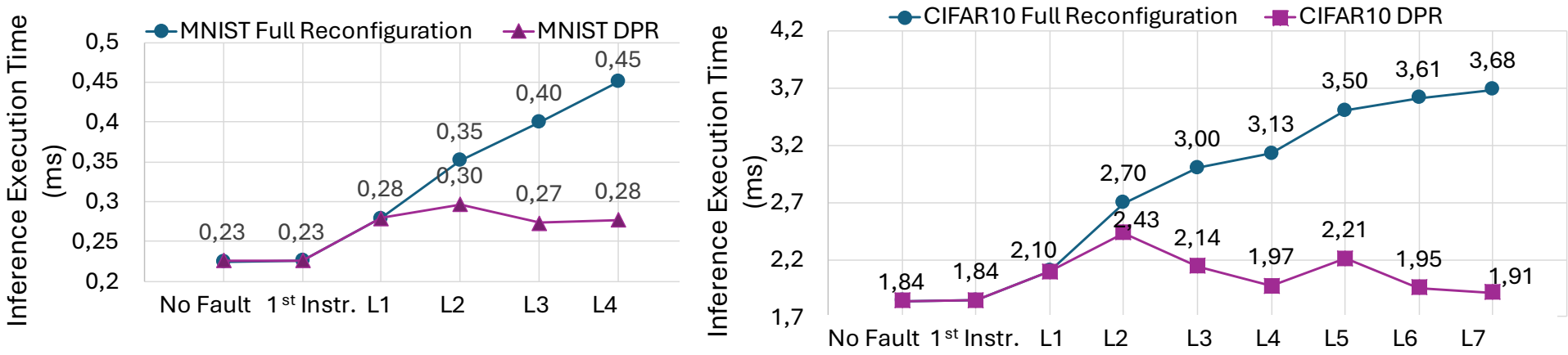


TPU DPR Time

- The **DPR time scales** linearly with the **size of the Systolic Array**, from less than 6ms in the smallest case to around 14ms for the larges SA size.
- The overall **inference execution time is reduced in the DPR case**, allowing operation recover from last correctly executed operation.



TPU DPR Time



Inference Execution Time in Full Reconfiguration and DPR cases

Conclusions and Future Works

- A **reliable platform for DNN execution** in safety-critical environment has been proposed.
- **Error detection capabilites** have been implemented into a Systolic Array.
- The Accelerator has been paired with the **NEORV32 and Partial Reconfiguration to ensure error recovery** and **reduced system downtime**.
- A **fault inejction campaign** took place to validate the effectiveness of the proposed error detection mechanism.
- A detailed **analysis on the efficiency** of the proposed platfrom has been carried out.
- Future works include further improvement to the Error Detection mechanism and more complex coupling between TPU and RISC-V Processor.

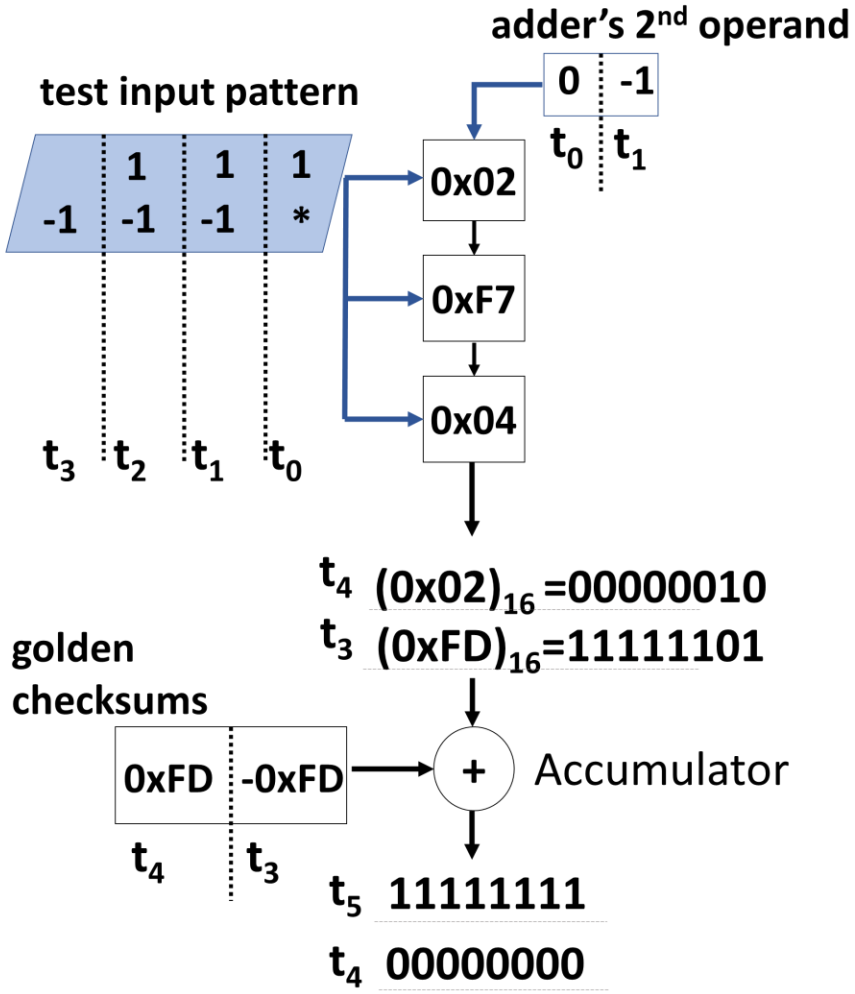
Thanks for the attention!

For further information:
- giorgio.cora@polito.it
- eleonora.vacca@polito.it
- corrado.desio@polito.it
- sarah.azimi@polito.it
- luca.sterpone@polito.it

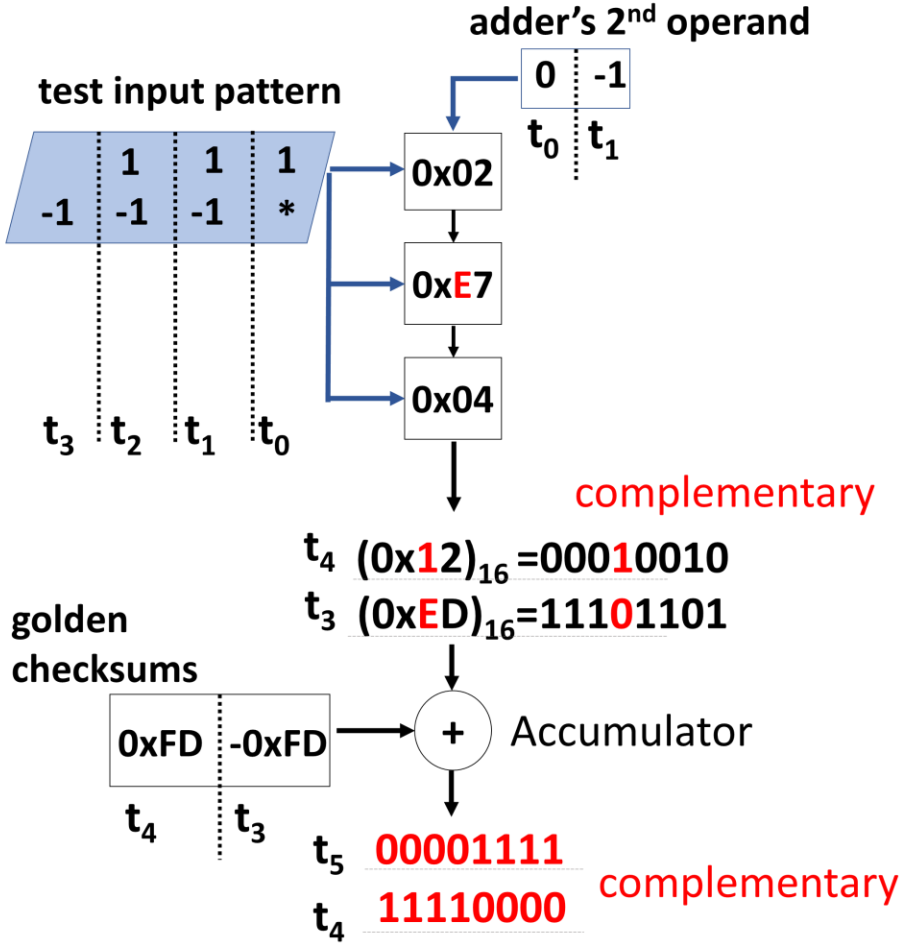
ASAC LAB

AEROSPACE SAFETY AND COMPUTING LAB

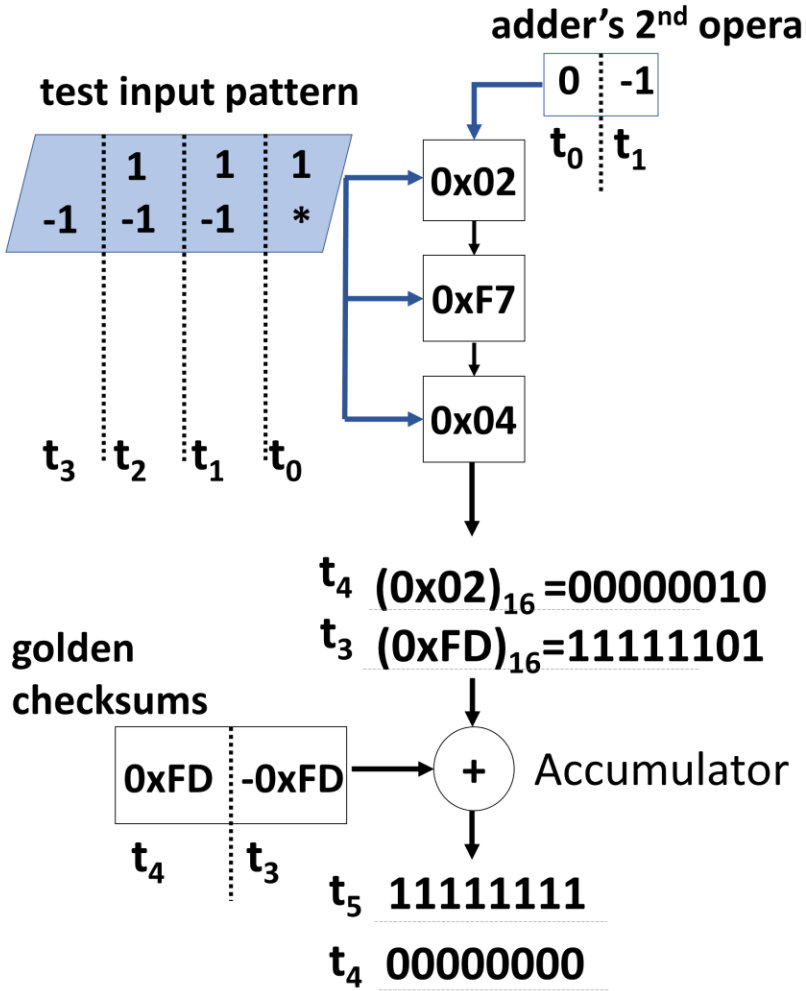
asaclab.polito.it



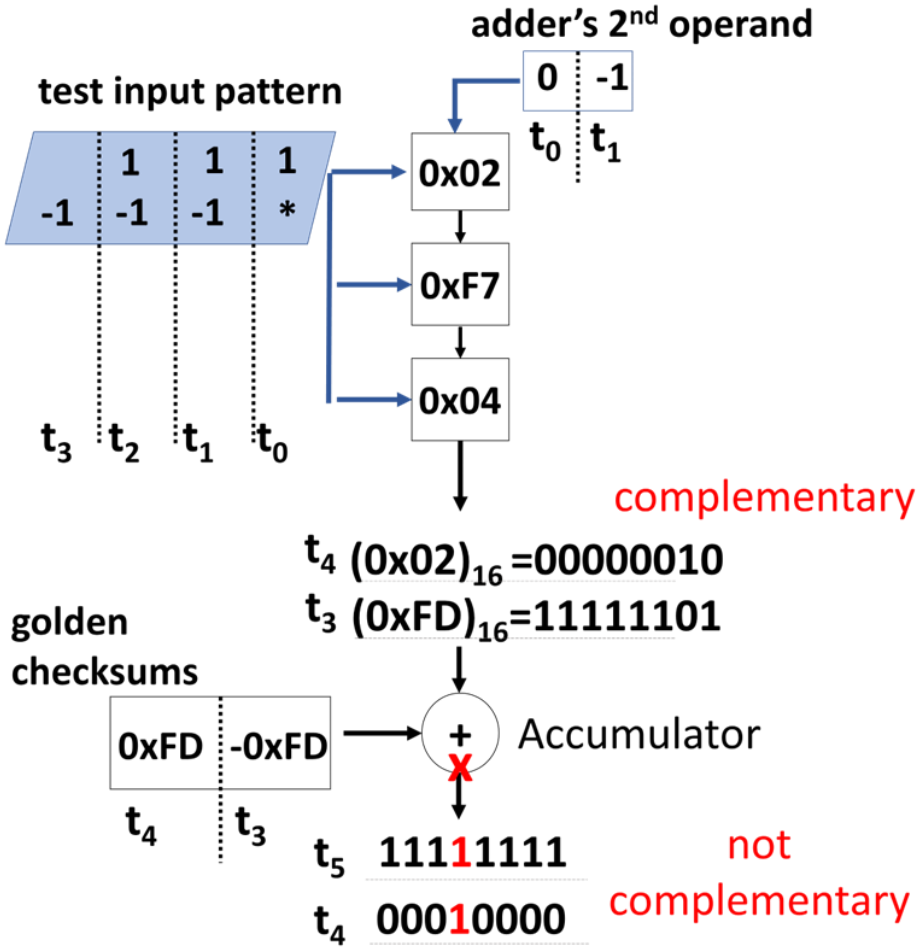
Fault Free Scenario



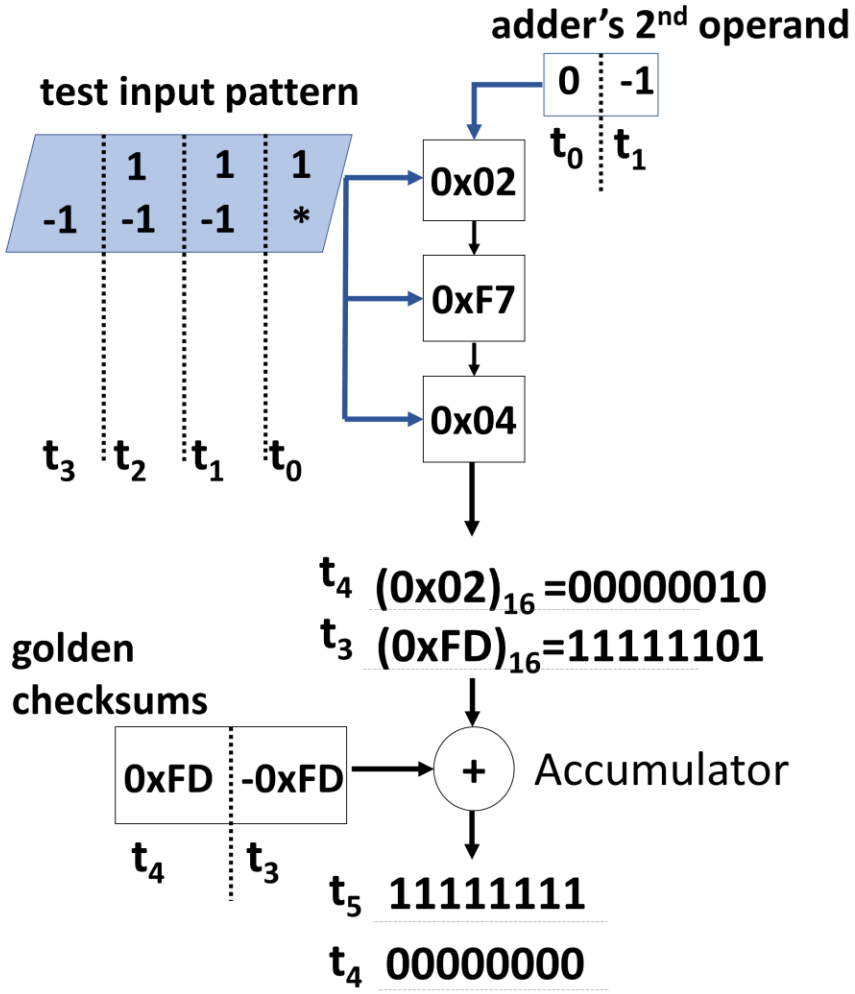
Error in the Weight Register



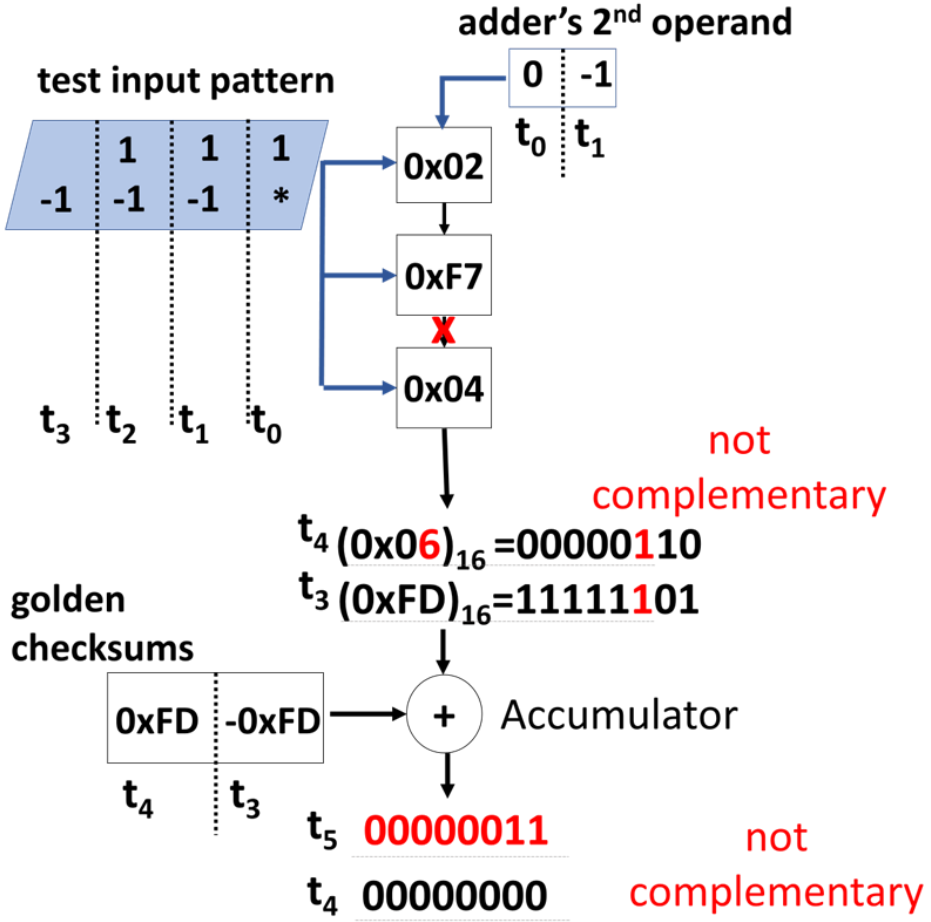
Fault Free Scenario



Error in the Accumulators



Fault Free Scenario



Error in the Output Propagation