New Challenges for the Reliability Evaluation of FPGA’s Accelerators in Artificial Intelligence Platforms

Sarah Azimi
Giorgio Cora
Corrado De Sio
Andrea Portaluri
Eleonora Vacca
Luca Sterpone

Politecnico di Torino

Outline

- Motivations
- Background
- Neural Networks Reliability
- FPGA Accelerators
- Experimental Results
- Conclusions and Future works

Reliability of FPGA Systems

- It matters for mission-critical applications


Avionics


Control

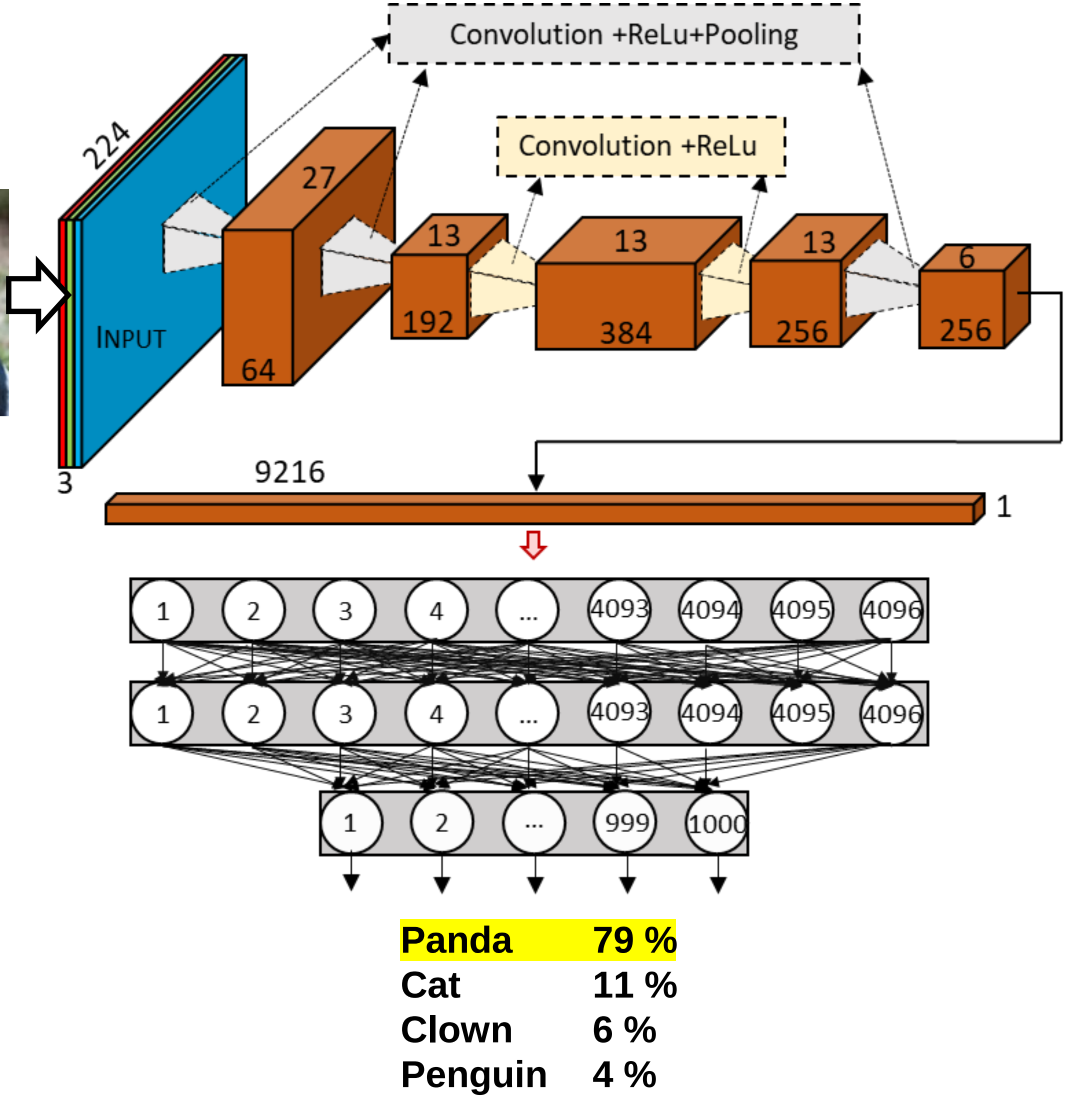

Space


Automotive


Transport

Artificial Intelligence

- Increasing advent of vision-oriented elaboration algorithms
- Adoption of deep learning techniques
- Autonomous computing systems will enable to take autonomous decisions

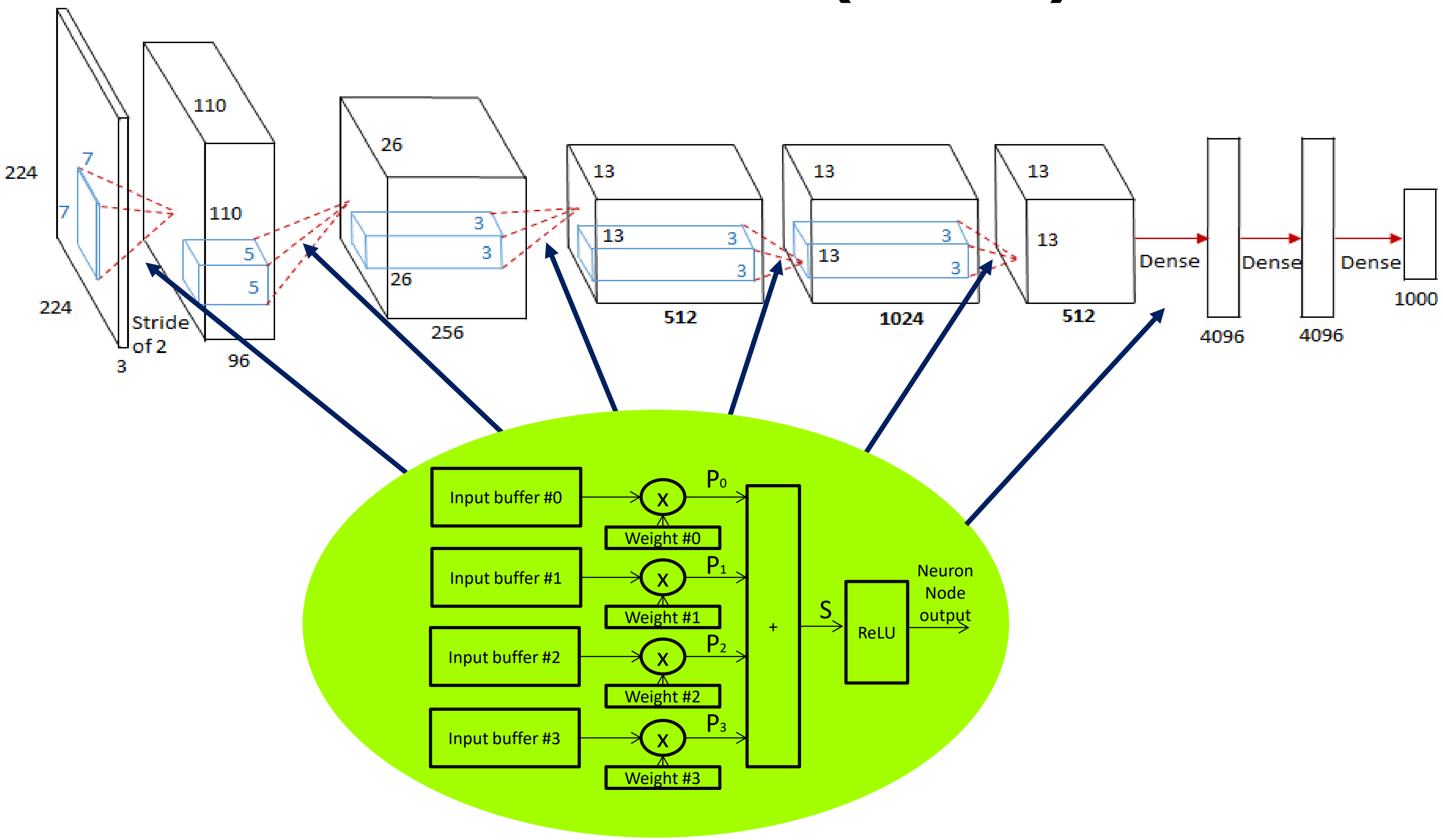


High-performance Computing on FPGAs

- **The high performance and reprogrammable capability lead FPGA as an appealing solution for high-performance demanding algorithms**
  - limited power consumption
  - high efficiency

Neural Networks and Artificial Intelligence

- **The usage of hardware devices capable of supporting Convolutional Neural Networks (CNNs) become strategic**

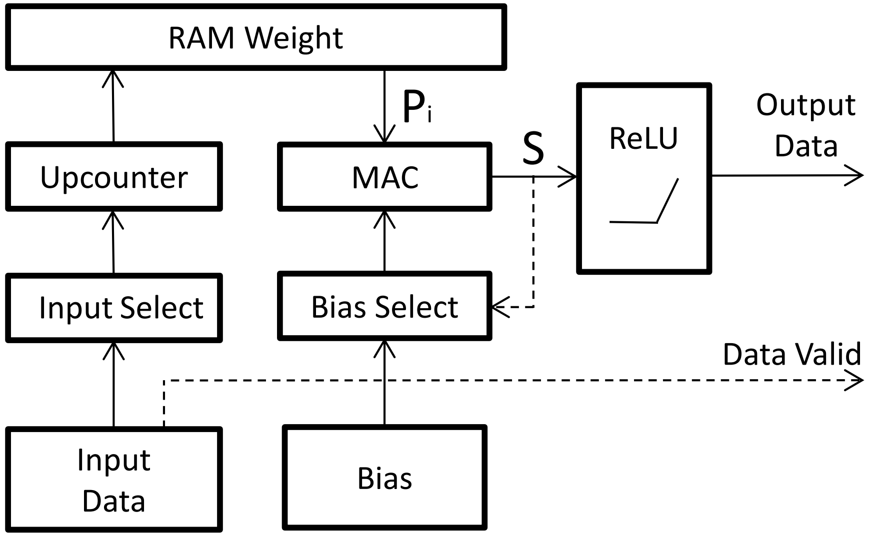


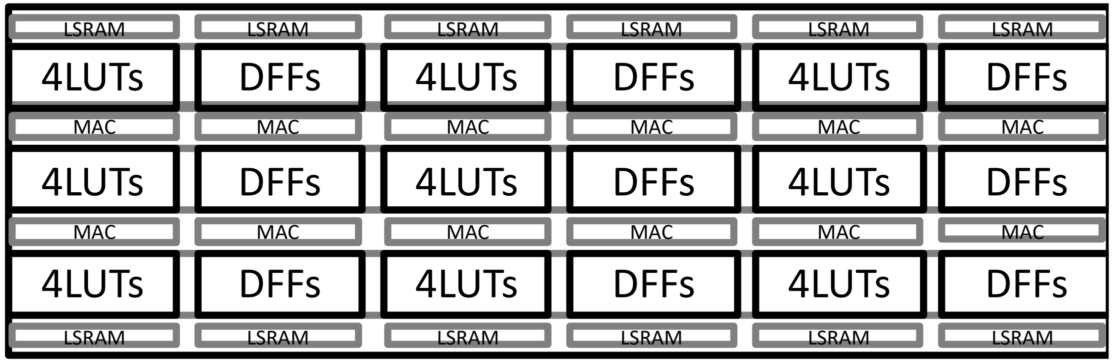
Implementation of CNNs on FPGA (i)

- **Several parallel neurons must be instantiated to implement a complete CNN**
- **All the data flow traversing structure from the synapse inputs to the post-rectified linear output required is limited to a resolution**
- **The product requires higher resolution for the multiplication and extra range for the accumulation to avoid overflow conditions of any arithmetic process**
- **Fully parallel neurons is not optimized for FPGA**

Hardware Synthesizable Neurons

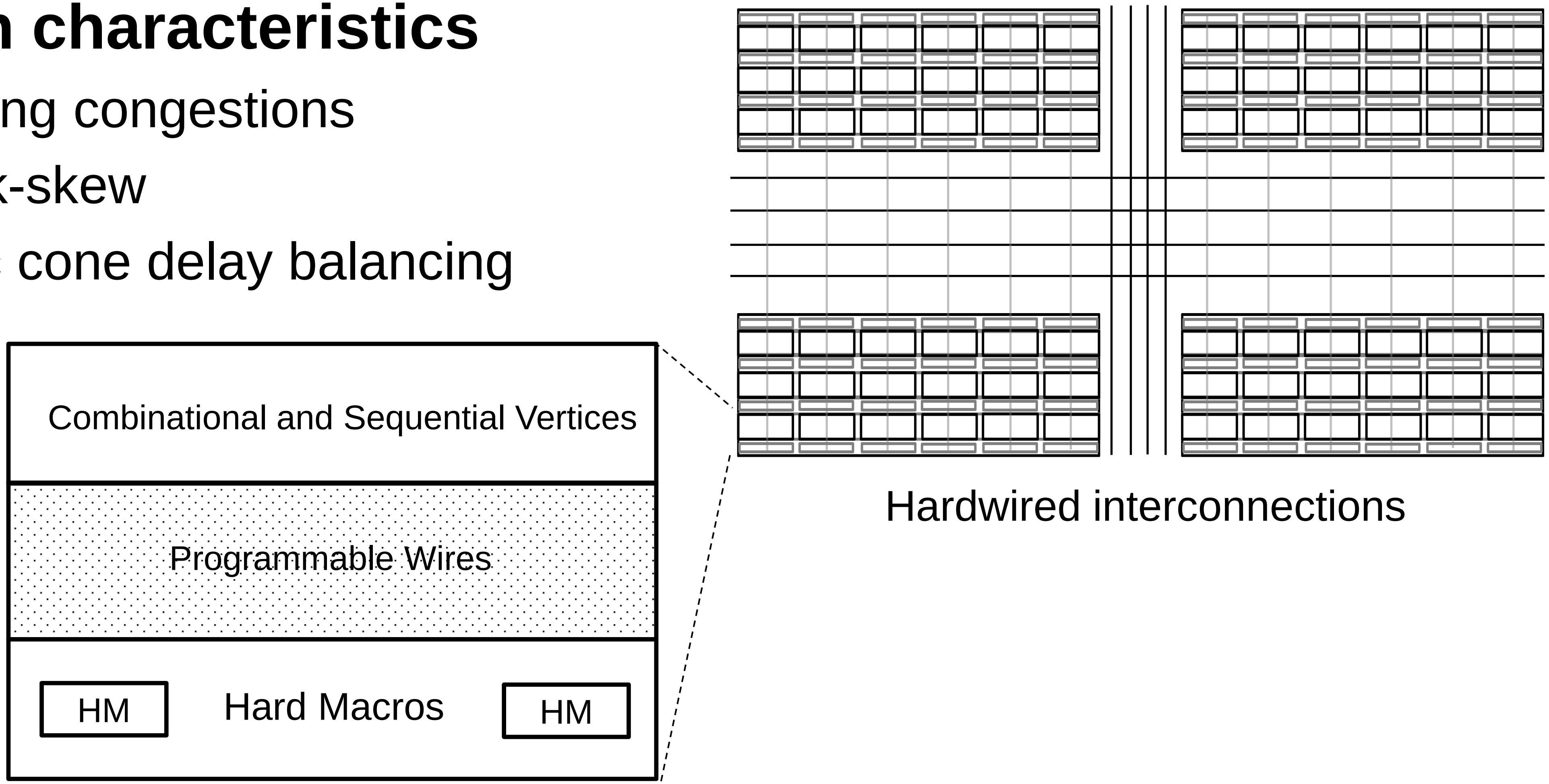
- **Customization of MAC and Hardwired units depending on**
  - architectural organization of the Neural Network
  - physical implementation tailoring depending on the FPGA resources availability





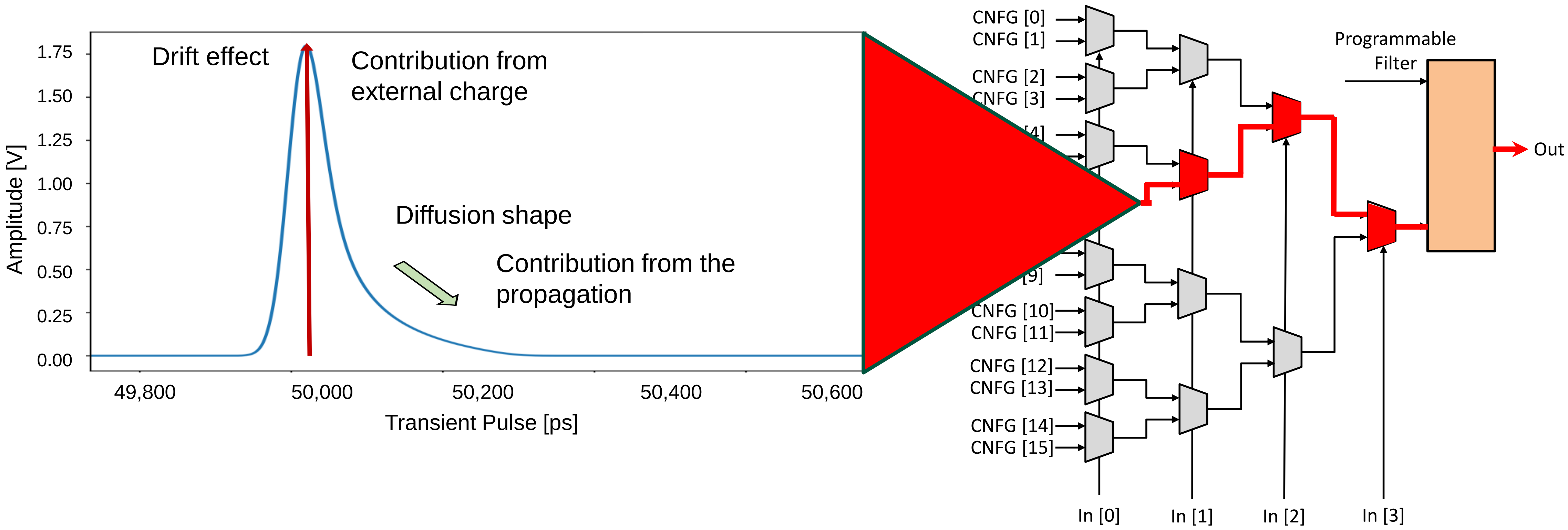
Customizable Placement

- **The placement can be parametrized to manage the Neural Neuron characteristics**
  - Routing congestions
  - Clock-skew
  - Logic cone delay balancing

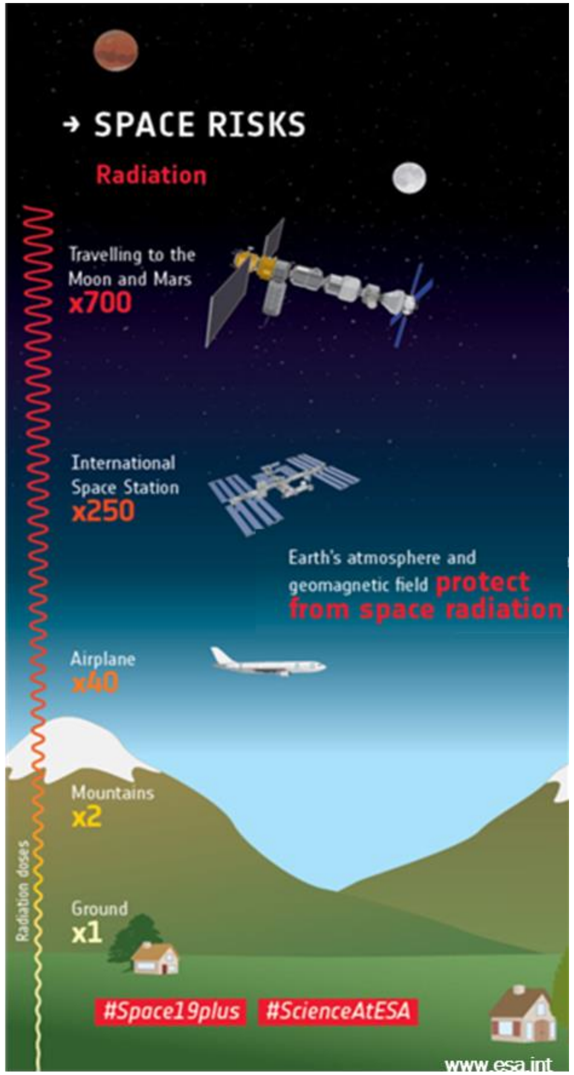


Neural Network Reliability

- **Transient errors have been demonstrated to be dominant effects on the reliability degradation**



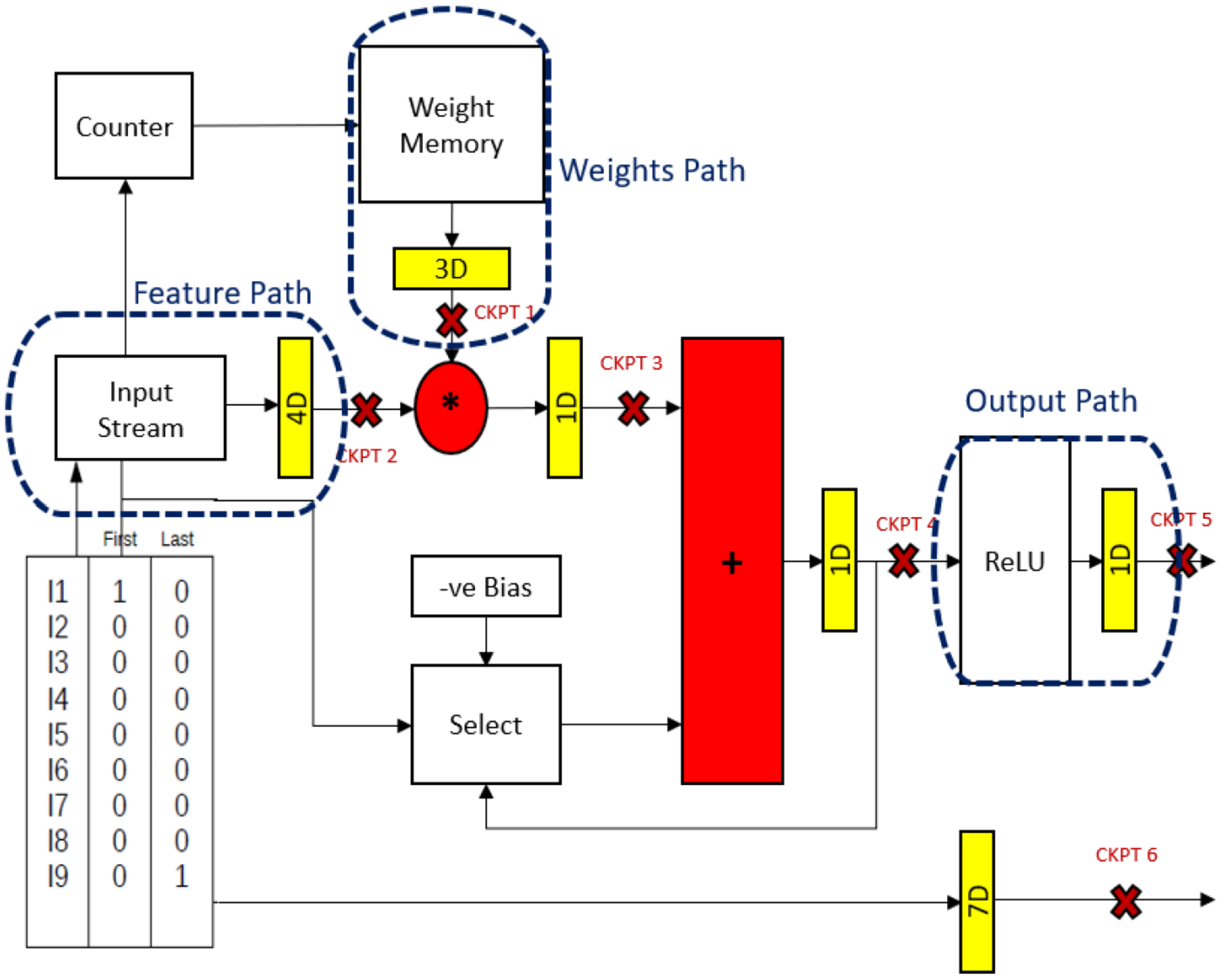
Radiation-induced Transient Errors



Location	Elevation [ft]	Relative Neutron Flux
Seattle, WA	160	1,05
Moscow, Russia	490	1,14
Chicago, IL	590	1,19
Denver, CO	5,280	3,76
Leadville, CO	10,170	10,79
White Mountain	12,500	15,07

Keller and Wirtlin, “Impact of Soft Errors on Large-Scale FPGA Cloud Computing” FPGA 2019

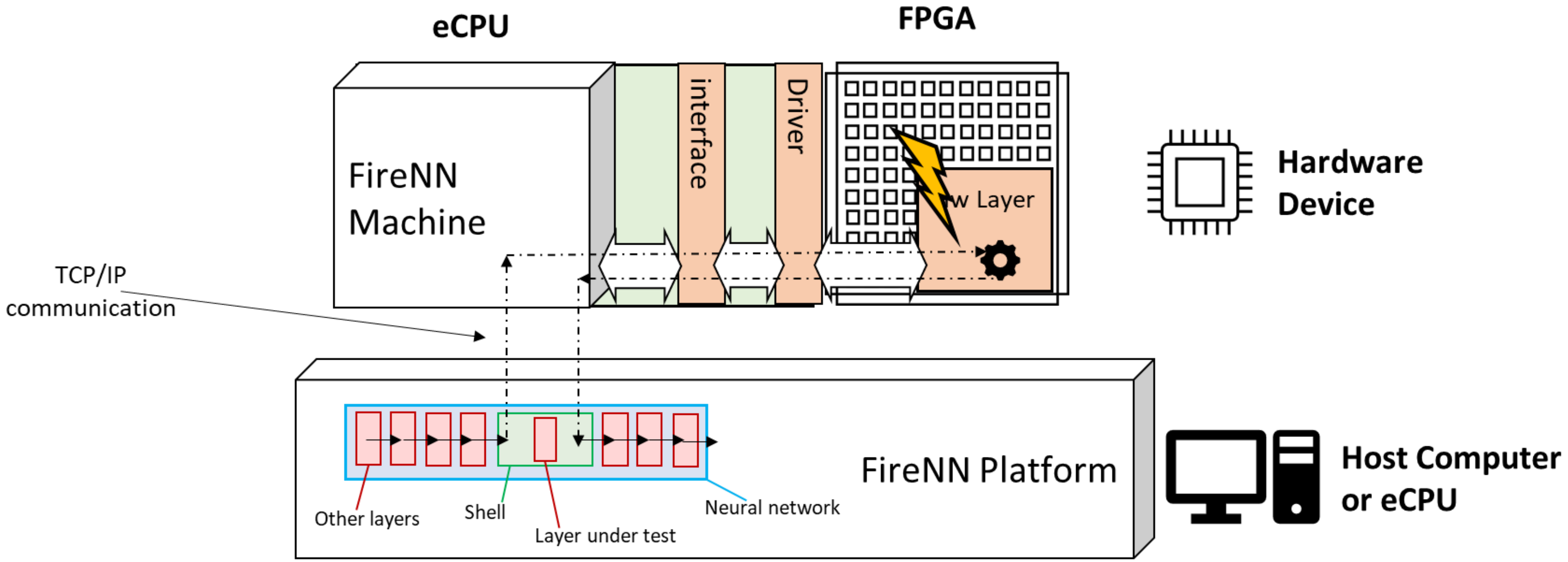
Neuron Architecture



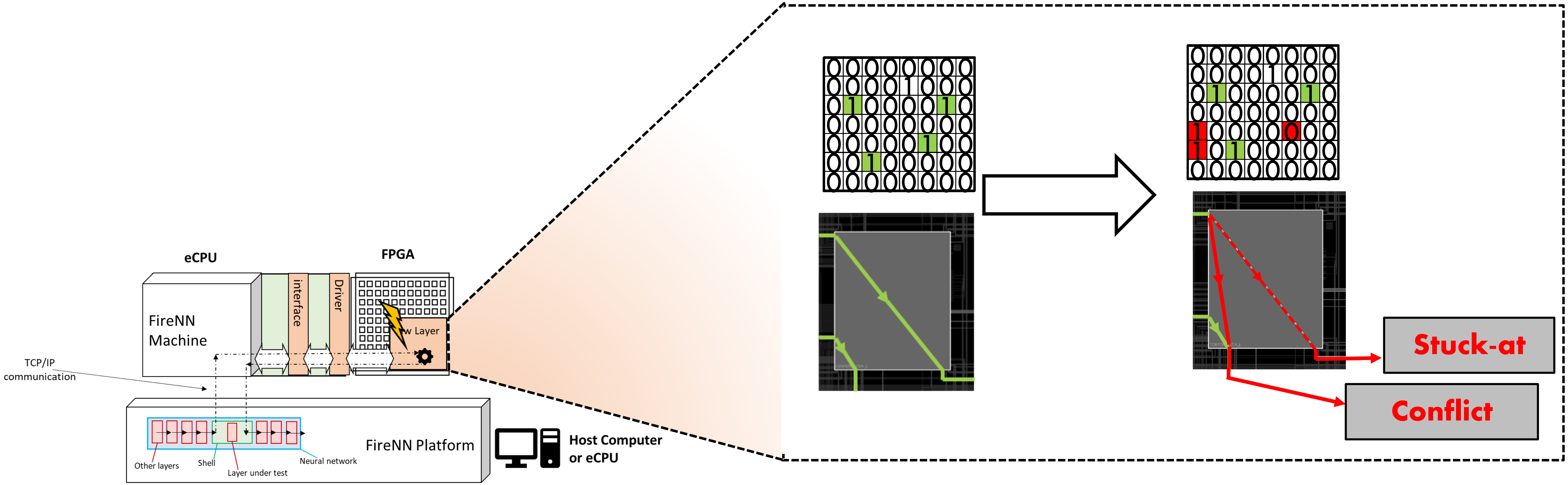
12

Neural Network Reliability Analysis

- **Embedded Fault Injection Platform**
- **Inject faults in the implementing hardware via configuration memory corruption**
- **Structural modification of the Neural Network**



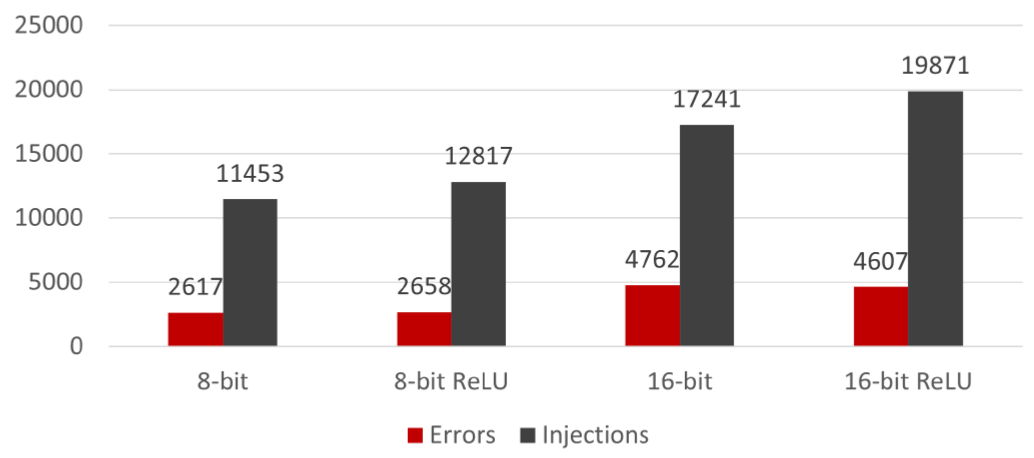
Neural Network Structural Modification Methodology



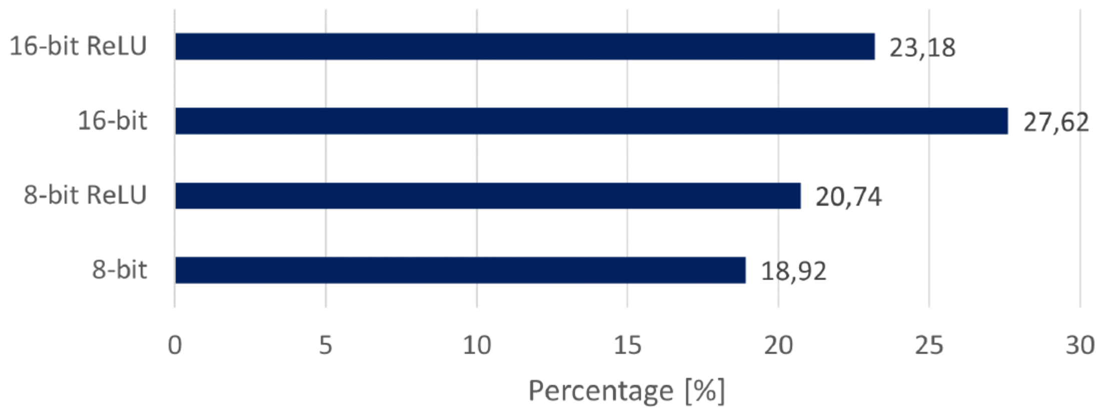
More configuration memory manipulation: *“PyXEL: An Integrated Environment for the Analysis of Fault Effects in SRAM-Based FPGA Routing“*
2018 International Symposium on Rapid System Prototyping (RSP), Torino, Italy, 2018

Experimental Analysis (i)

- **Single bit-flip injection within the Neural Network FPGA configuration memory**

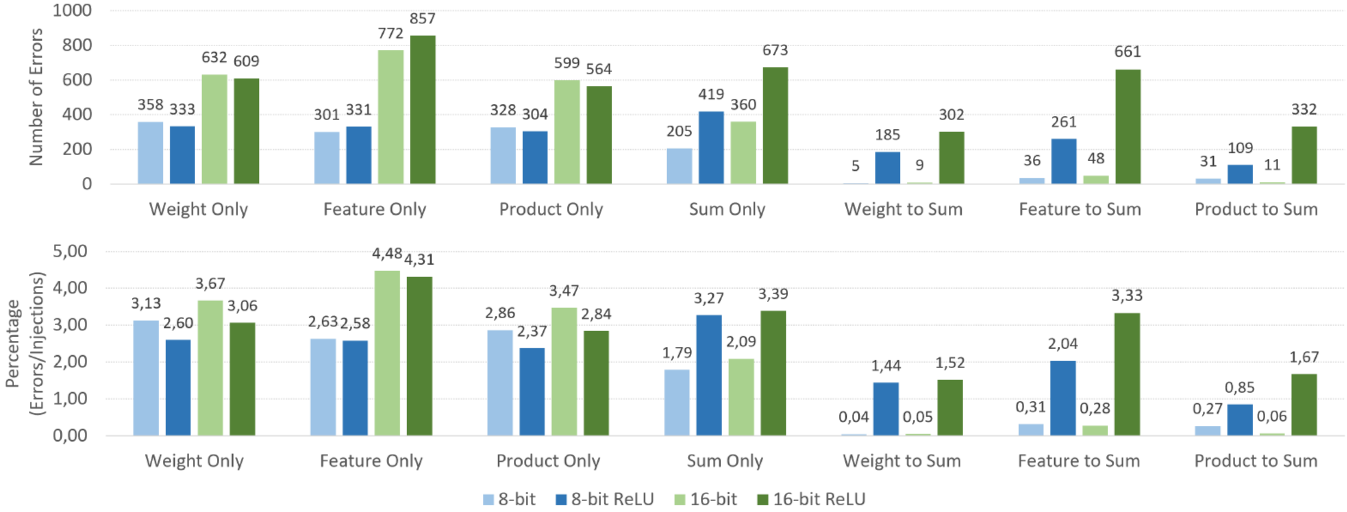


- **Error rate of a single neuron with different resolutions**



Experimental Analysis – Error Propagation (ii)

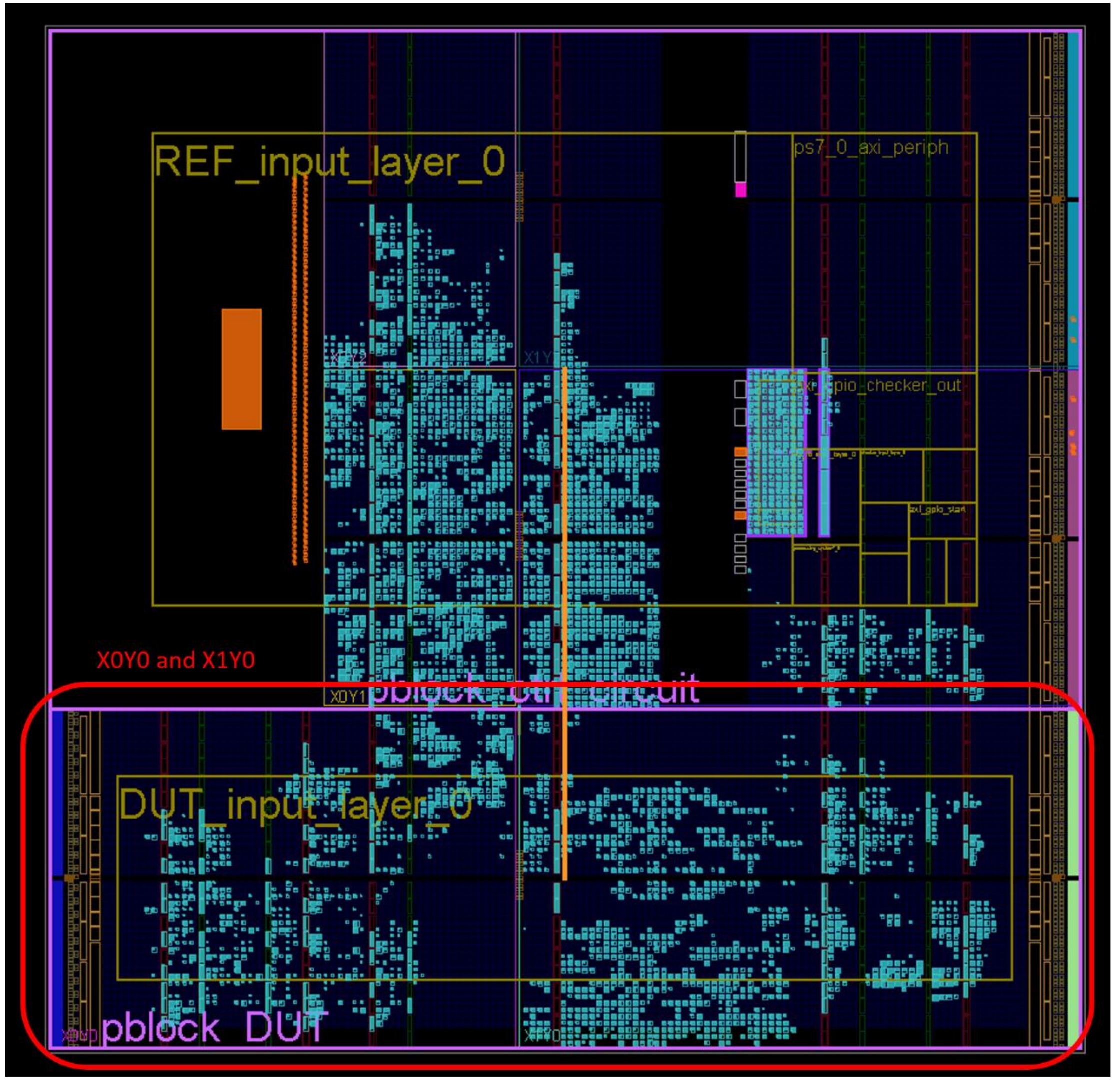
- **Errors not propagated to the Neural Network output**



- **Errors propagated to the Neural Network output**

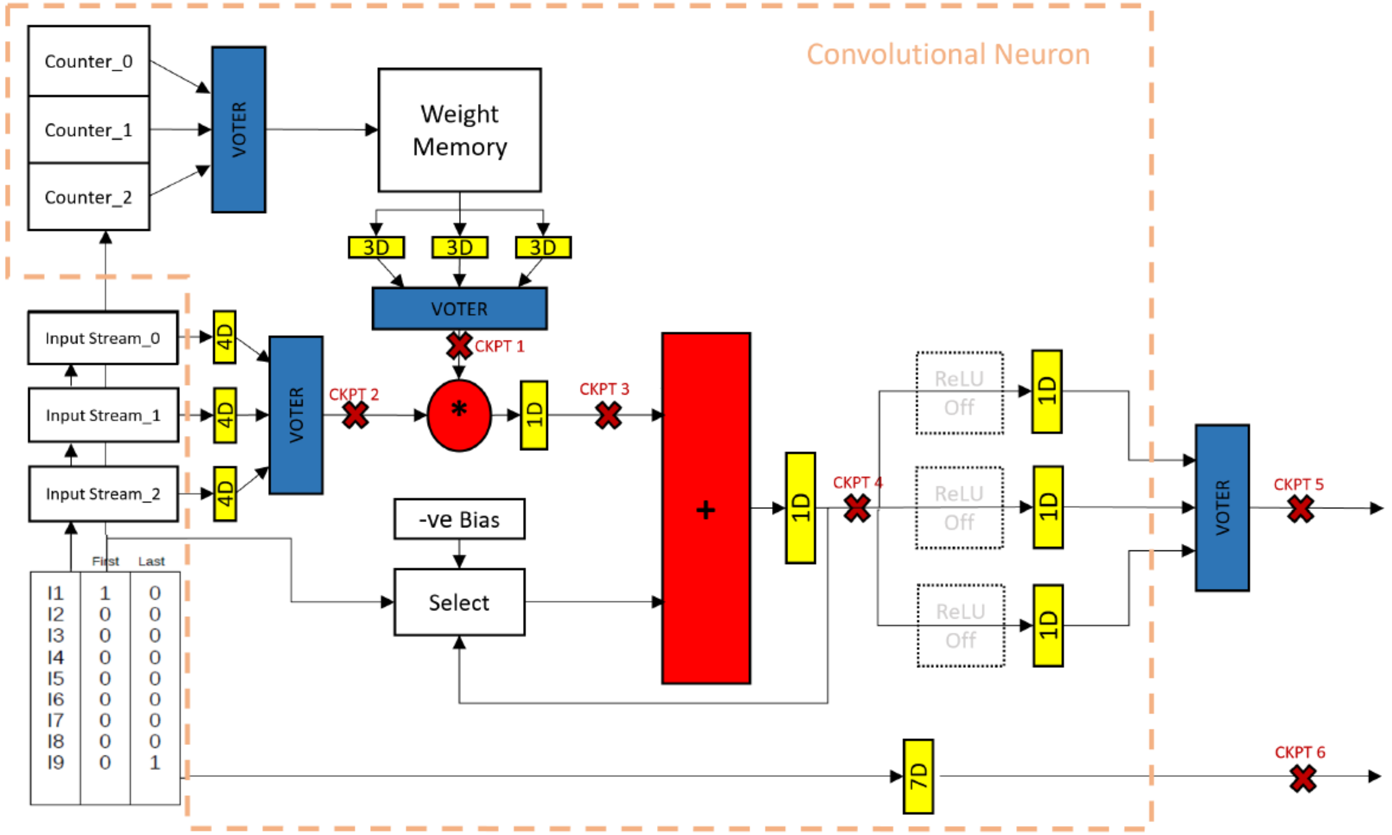
Experimental Analysis (iii)

- **Placement view of the ZFNet Neural Network implementation on a Zynq 7020 SRAM-based FPGA**

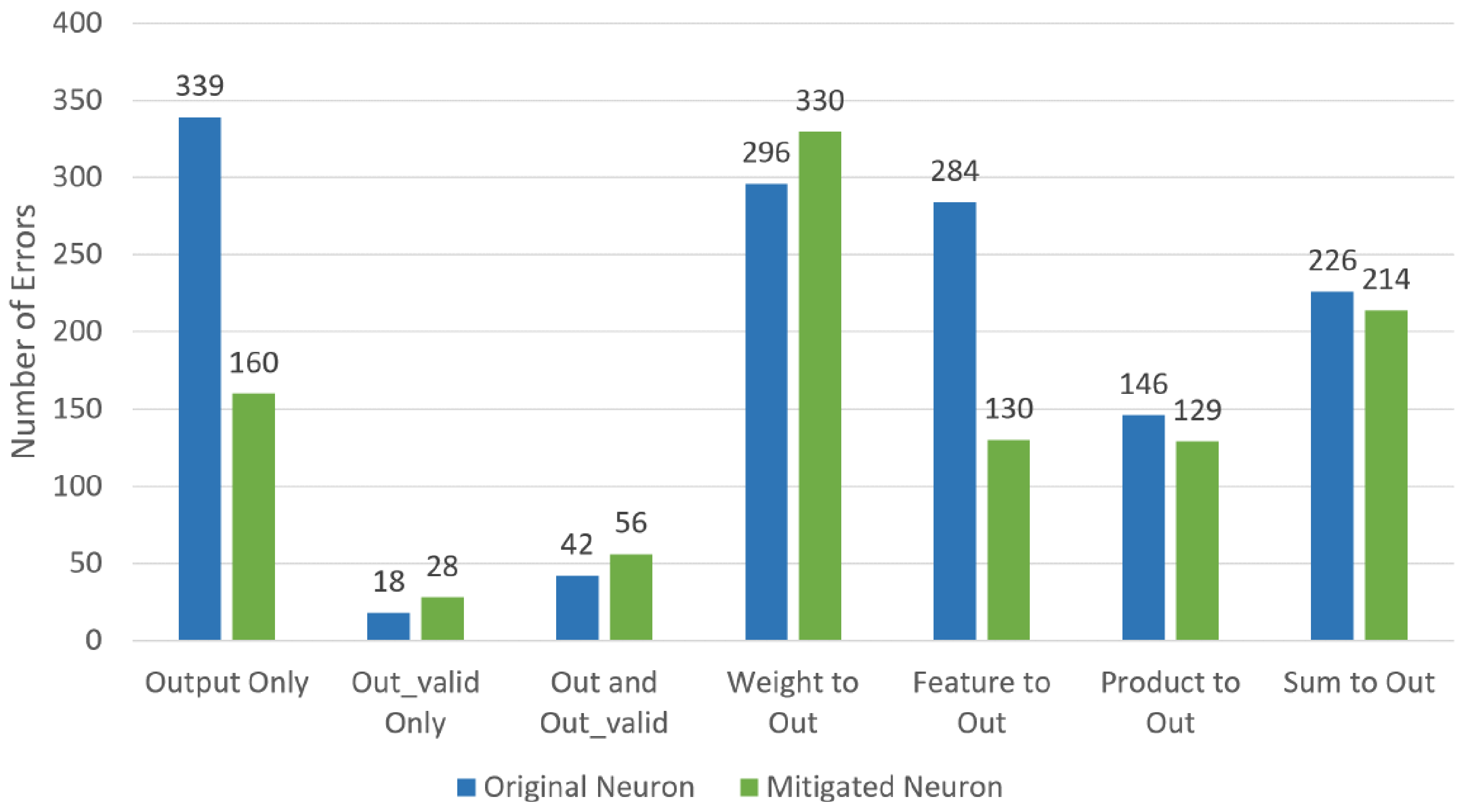


TMR Mitigation Approach

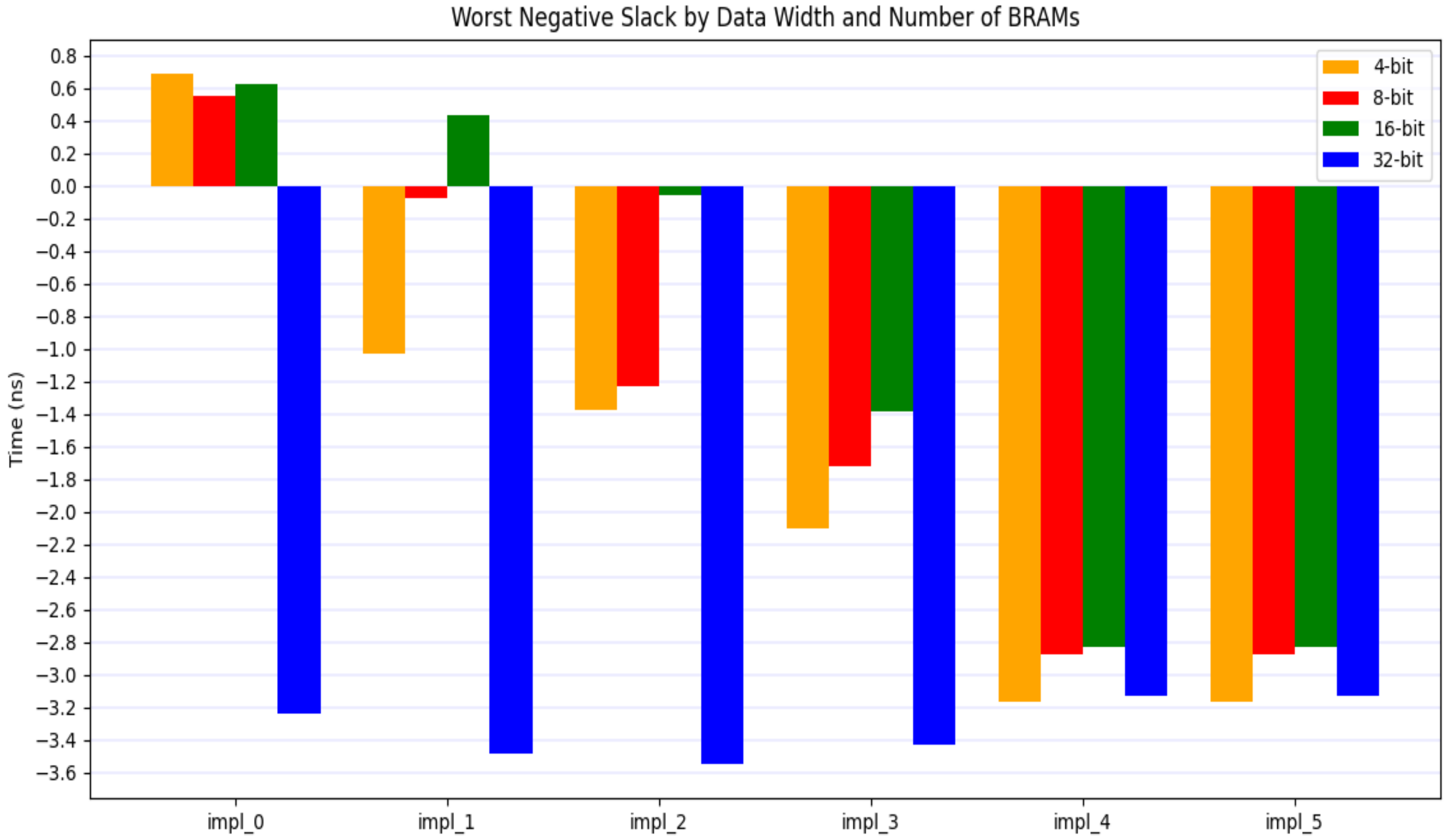
- **Application of Triple Modular Redundancy (TMR) techniques on selective resources**
  - Counters
  - Input Streams
  - Weight Memories
  - ReLU



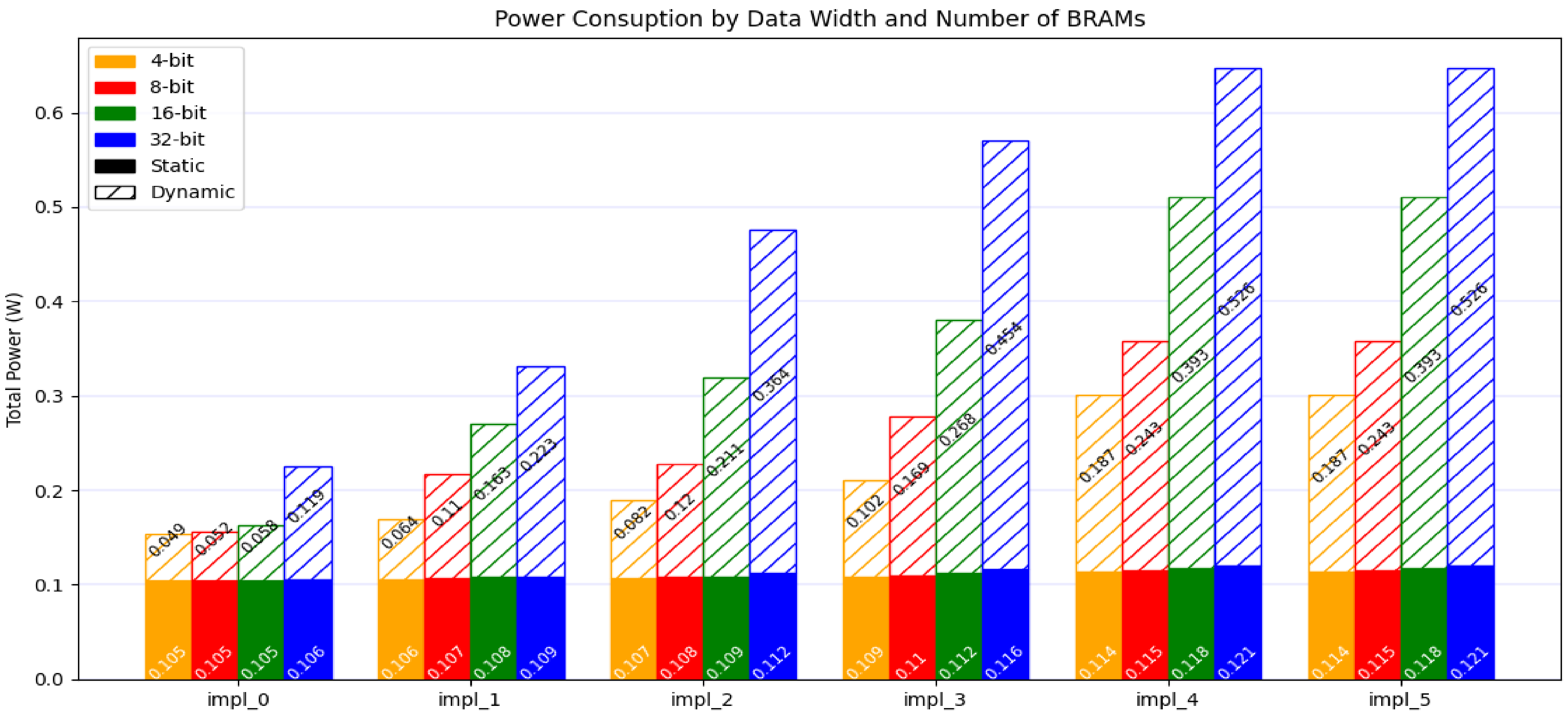
Experimental Analysis (iv)



Timing Analysis

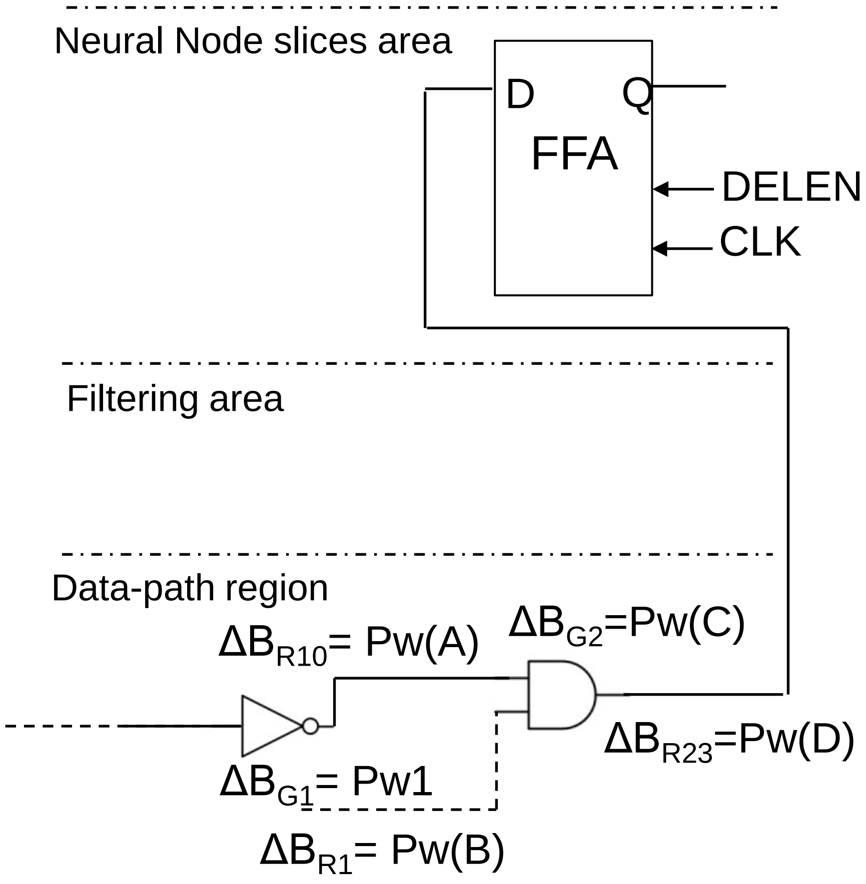


Power Analysis



Transient Errors Mitigation

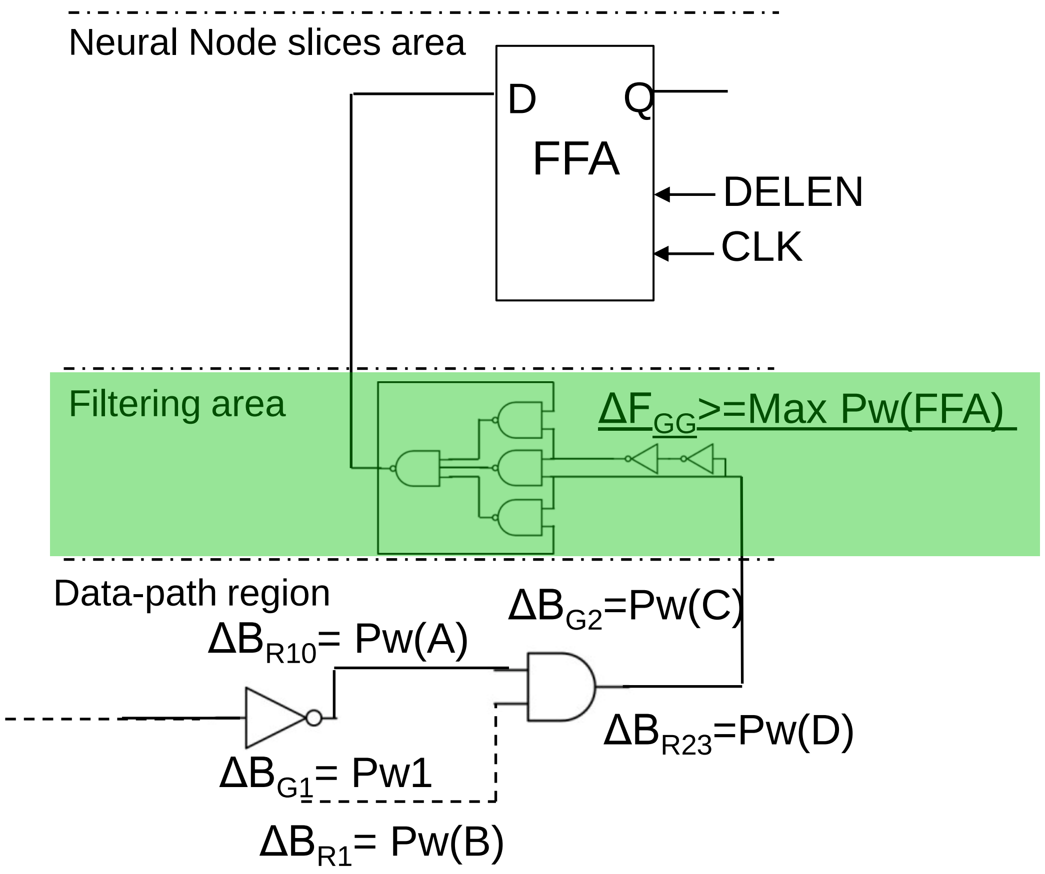
- **Selective activation Transient filtering performs massive SET filtering**
- **Avoid drastic performance degradation**



Transient Errors Mitigation

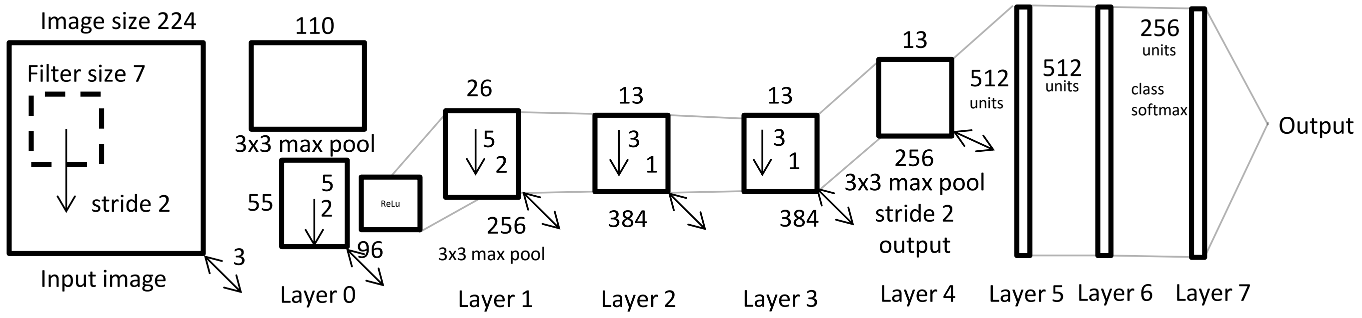
- **Selective activation Transient filtering performs massive SET filtering**

- **Avoid drastic performance degradation**



ZFNet Complete Mitigation on RTG4 FPGA

- **Pruned ZFNet implemented on RTG4 Microchip FPGA**



- **Original**

  Timing performance optimization and without any mitigation

- **Commercial**

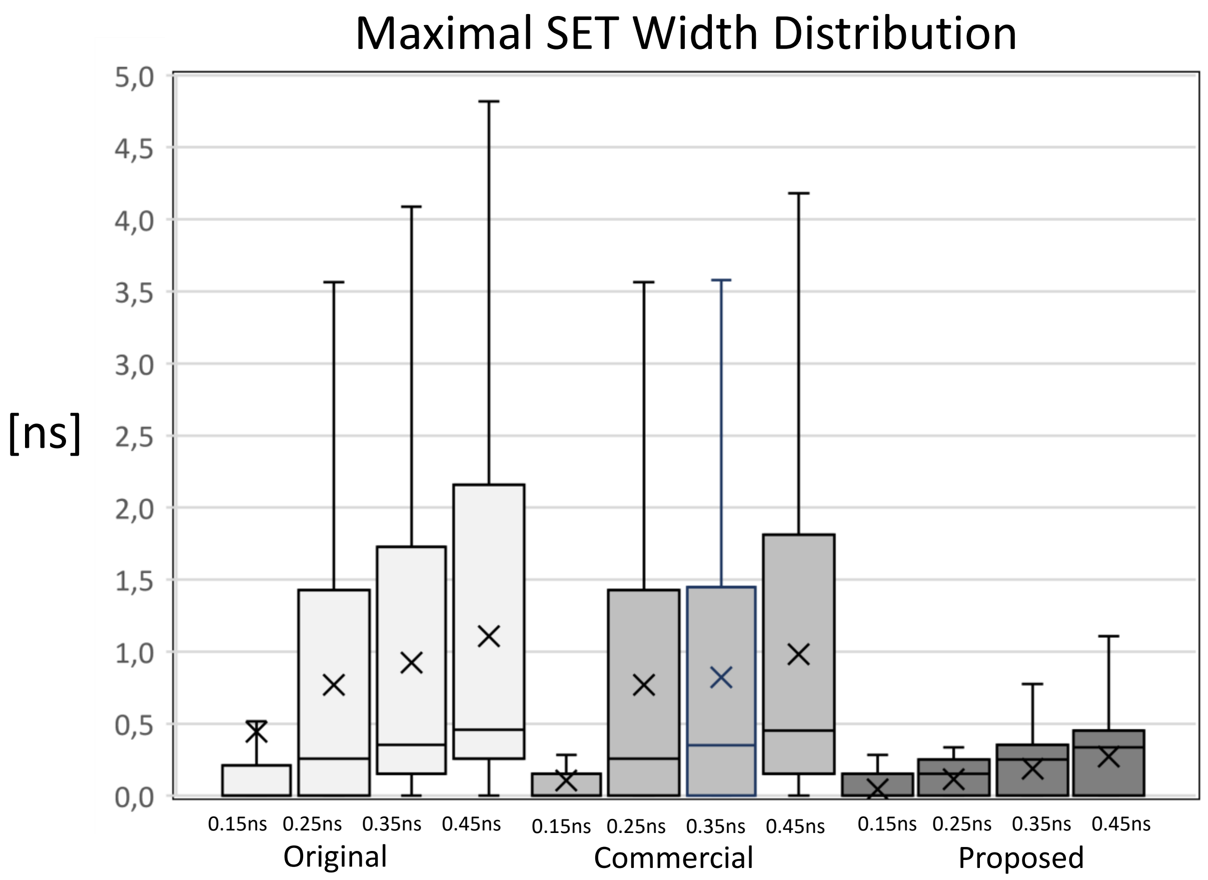
  Implemented with the SET filtering feature up to 600 ps

- **Proposed**

  Implemented with placement constraints targeting DSP performance and LSRAM resources and adopting selective SET filtering

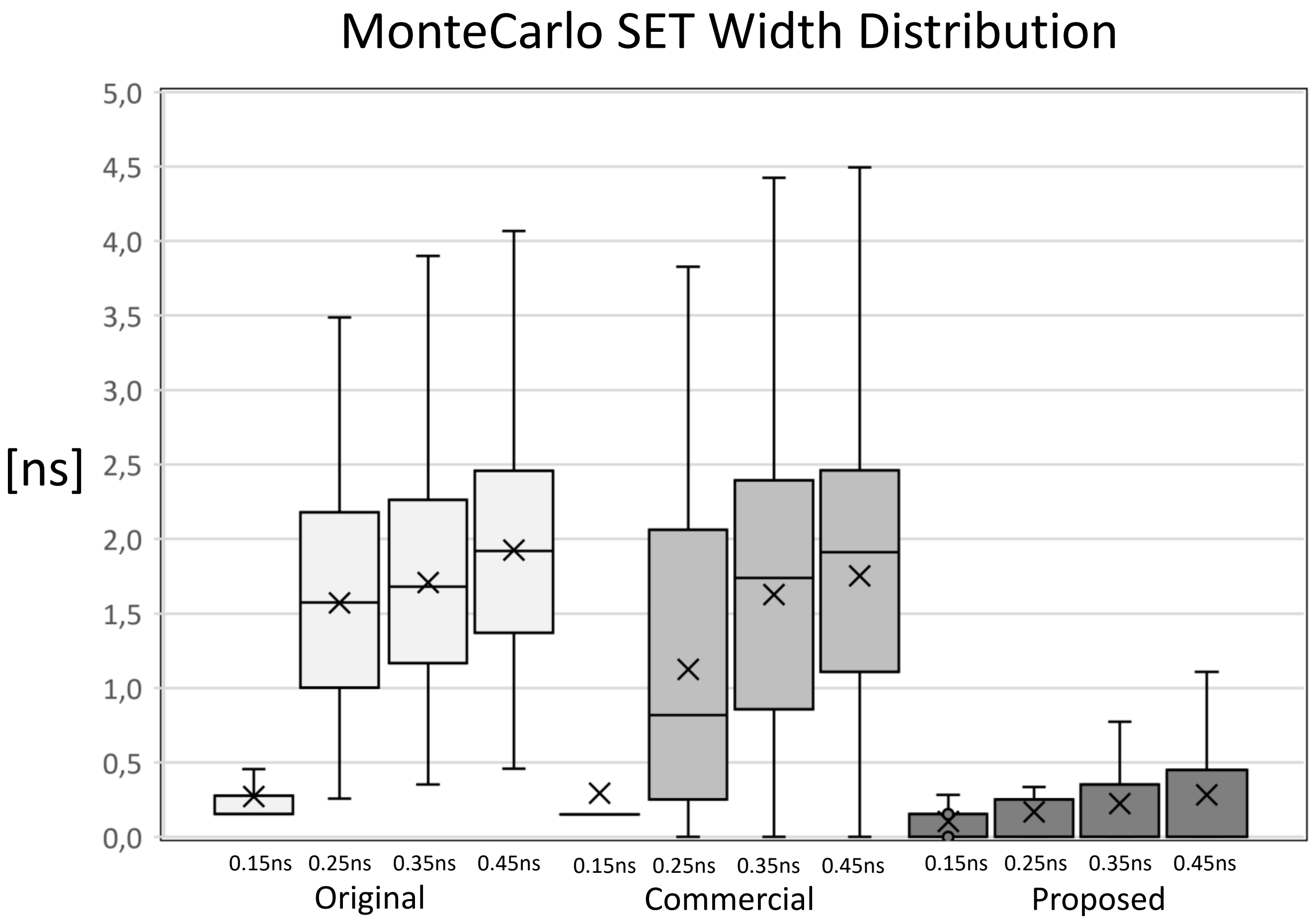
Experimental Results (v)

- **The maximal SET pulse width distribution for the overall CCN sequential element**
  - FFs and Block RAMs

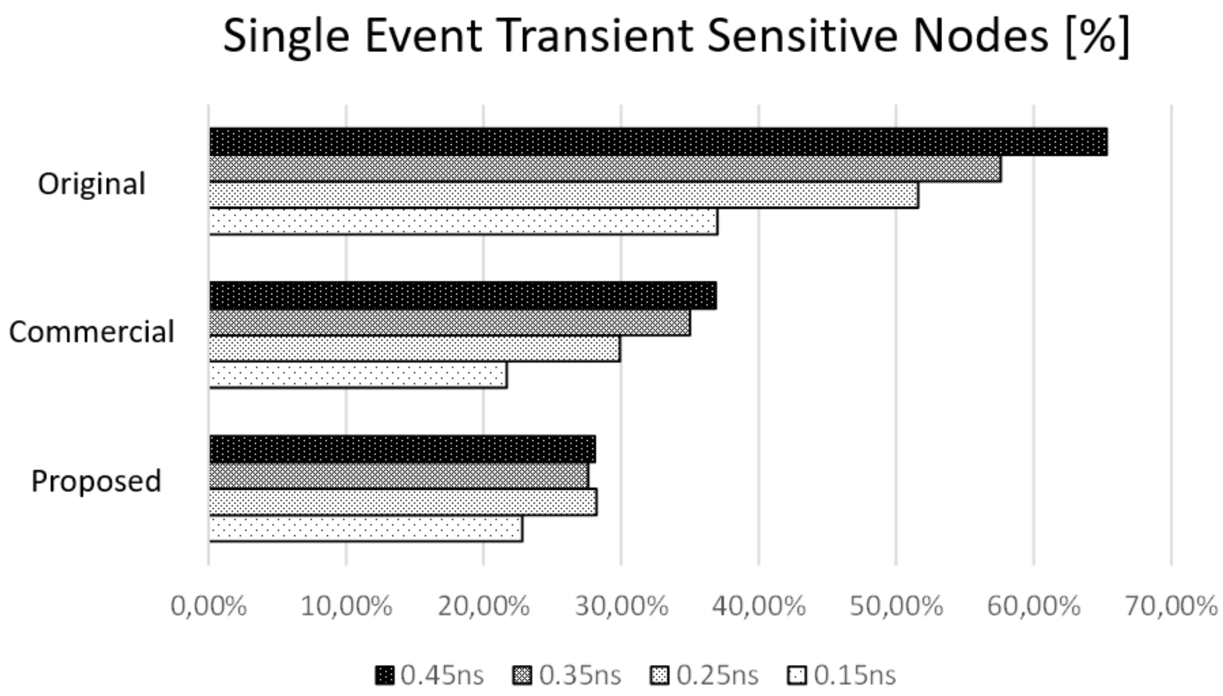


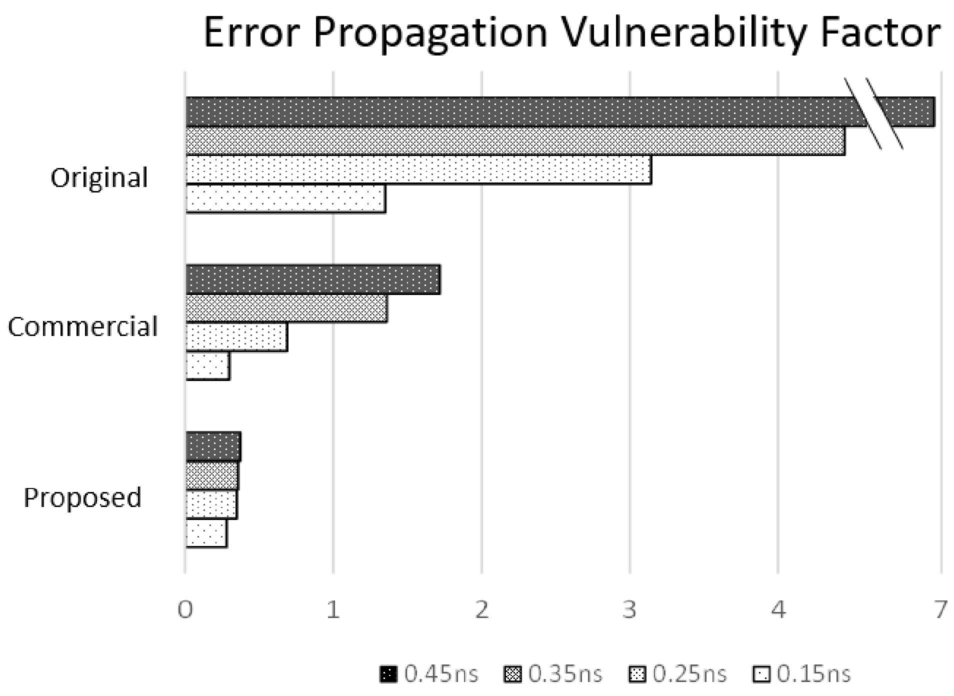
Experimental Results (vi)

- **Monte Carlo SET pulse width distribution obtained thanks to random fault injection on the CCN resources**



Sensitivity and Vulnerability Factors





Conclusions

- **Reliability evaluation of Neural Network must be performed considering structural and transient faults**
- **Faults intrinsic of other hardware can be analyzed via FPGA**
- **Sensitivity and Vulnerability Factors are crucial parameters that should be addressed for any reliability analysis of Neural Network**

Future works

- **Propose a comprehensive analysis involving other layers, network architectures and other accelerators such as TPU**
- **Compare the implementation tool also considering the synthesis and the mapping phases**
- **Compare results deriving by our approach with the results from network level injection and environmental tests**