Politecnico di Torino
Dipartimento di Automatica e Informatica

PhD in Computer and Control Engineering
XXXVI cycle

Supervisor
Matteo Sonza Reorda
Ernesto Sanchez

Reliability enhancement in GPU architectures

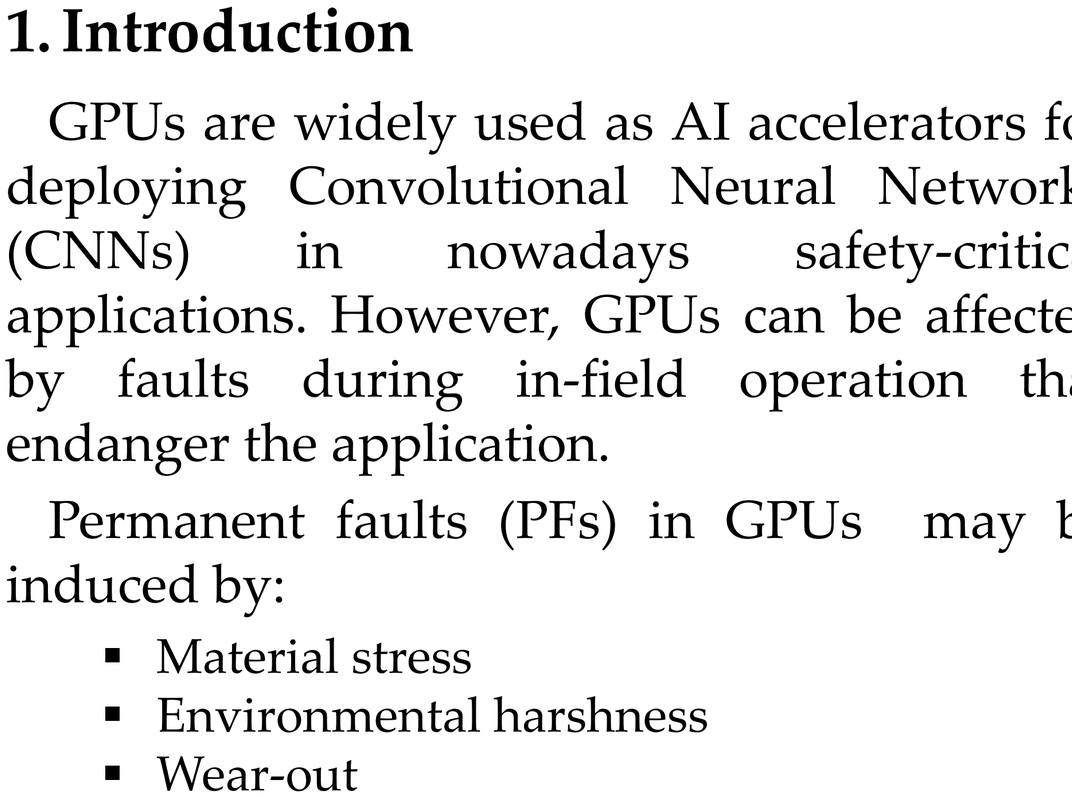
PhD Candidate: *Juan David Guerrero Balaguera*

1. Introduction

GPUs are widely used as AI accelerators for deploying Convolutional Neural Networks (CNNs) in nowadays safety-critical applications. However, GPUs can be affected by faults during in-field operation that endanger the application.

Permanent faults (PFs) in GPUs may be induced by:

- Material stress
- Environmental harshness
- Wear-out



The sensitivity estimation of CNNs to hardware faults in GPUs is mandatory to fulfill the requirements of the safety standards (e.g., ISO26262 for automotive domain).

2. Motivation

Reliability estimation of CNNs to PFs mainly resorts to Fault Simulation (F-Sim) campaign

❑ Application-level (usual approach)
- ✗ Neglects details of the underling Hardware
- ✗ Lacks accuracy evaluation of faults
- ✓ Fast simulation times

❑ Hardware-level (ideal scenario)
- ✓ Considers the hardware details of the GPU
- ✓ Accurate fault evaluation
- ✗ Prohibited simulation times

This research aims to propose effective F-Sim methods for impact evaluation of PFs in GPUs running DNNs workloads, looking for the trade-off between accuracy and simulation time.

3. Proposed Approach

Hardware injection through program transformation (HITPT)

- GPU Architectural-Level Fault Injection
- Fault propagation at the device speed
- Good trade-off between speed and accuracy

We developed a F-sim framework to emulate the PFs behavior in the register files (RFs) and functional units (FUs) of a GPU [1-3]

4. Results

- Faults located in the first ten register per thread of a GPU are the most sensitive for the reliability of a DNN
- Faults affecting the Integer cores mainly lead to crash or hang of the GPU
- Faults affecting the floating-point cores lead up to 40% of DNN accuracy degradation

5. Conclusions

This work describes a method able to provide for the first time an evaluation of the impact of PFs affecting a GPU while executing a CNN.

The results gathered via the proposed HITP evaluation demonstrate that more hardware vulnerabilities exist than predicted by the usual approaches based on injecting faults in the parameters of the CNN.

6. References

1. Juan David Guerrero, et.al, Evaluating the impact of Permanent Faults in a GPU running a Deep Neural Network, ITC-Asia,2022
2. Juan David Guerrero, et.al, Effective fault simulation of GPU's permanent faults for reliability estimation of CNNs, IOLTS, 2022
3. Esteban Rodriguez, et. al, A Multilevel Approach to Evaluate the Impact of GPU Permanent Faults on CNN's Reliability, International Test Conference, ITC, 2022.