Universidade: presente!

UFRGS PROPESQ

XXXI SIC

Salão UFRGS 2019 — Conhecimento · Formação · Inovação

21 . 25 . OUTUBRO . CAMPUS DO VALE

Selective Fault Tolerance for Register Files of Graphics Processing Units

Aluno: Ivan Peter Lamb

Orientador: José Azambuja

UFRGS

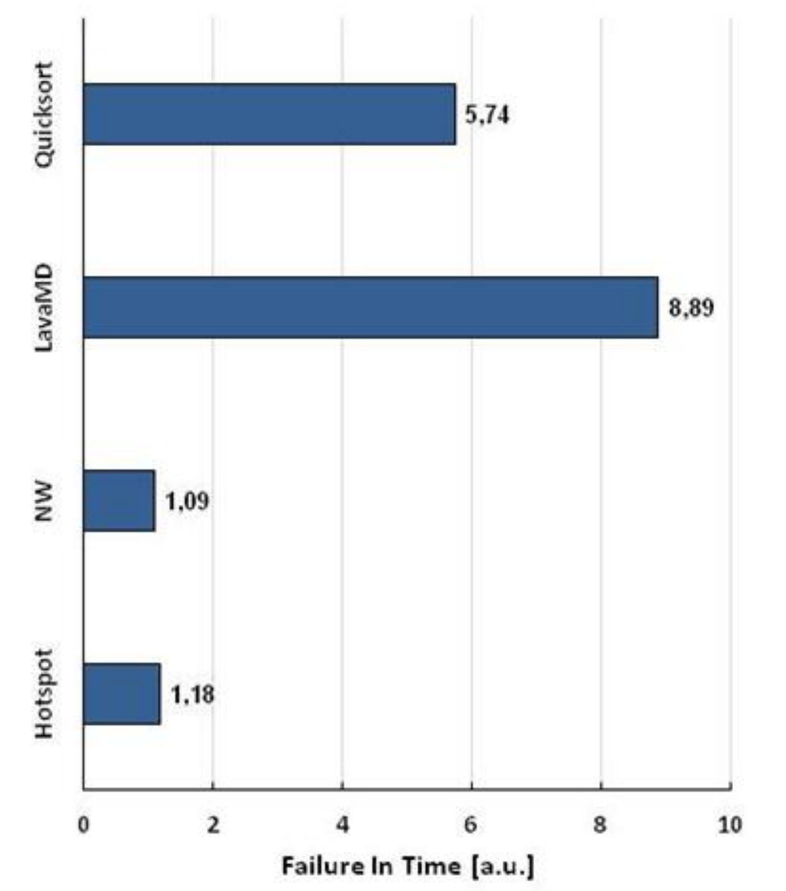
1. Introduction

- **Graphics Processing Units (GPUs)** reached **safety-critical applications**, such as automotive. In such applications, **fault tolerance techniques** are mandatory and have been applied to GPUs by means of **hardware** or **software** implementations
- Faults on electronic devices are mainly caused by energized particles which may cause **Single Event Upsets (SEUs)** in GPUs **registers** that provoke **Silent Data Corruption (SDC)** which leads the system to an **incorrect output**
- Considering that **a small margin of error can be considered safe in some applications,** this work proposes an **Approximate Computing (AC)** perspective **to relax register criticality** in order to improve **selective fault tolerance technique**

NVIDIA DRIVE™ PX is a powerful auto-pilot car computer designed to run the deep neural networks that will enable a car to see, think, and learn.

2. Reliability Evaluation

- **Device Under Test**
  - ✓ K20 and K40 NVIDIA Kepler GPU
  - ✓ Applications: Hotspot, NW, LavaMD, and Quicksort
- **Neutron Beam Experiment**
  - ✓ K40 NVIDIA Kepler GPU
  - ✓ Performed in Los Alamos Science Center (LANSCE)
  - ✓ Neutron flux between 1 and 25×10^5 $n/(cm^2/s)$
- **Register File Reliability Assessment**
  - ✓ SASSIFI: NVIDIA Fault Injection Tool
  - ✓ K20 NVIDIA Kepler GPU
  - ✓ 10.000 faults, 1 per execution
  - ✓ Target: application used registers
  - ✓ Random single bit-flip

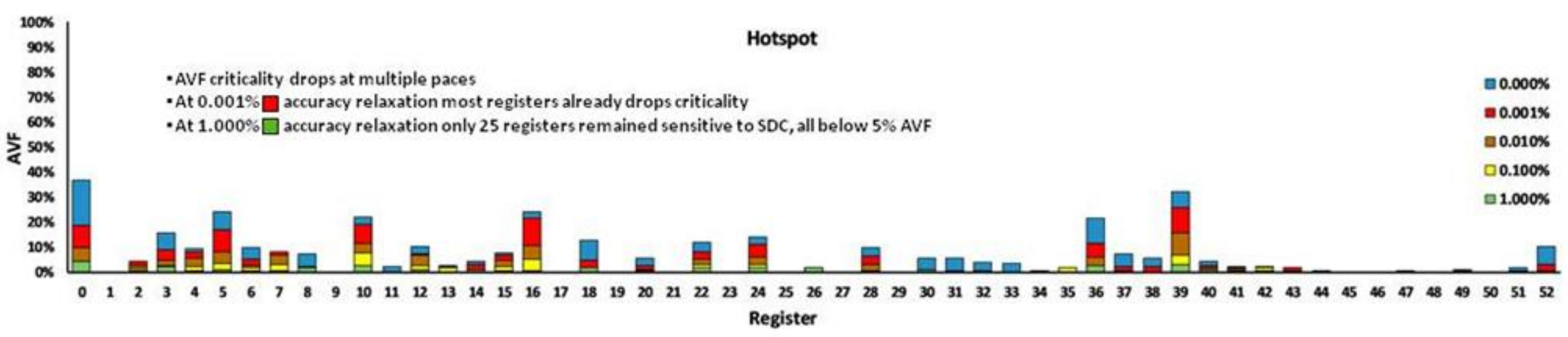


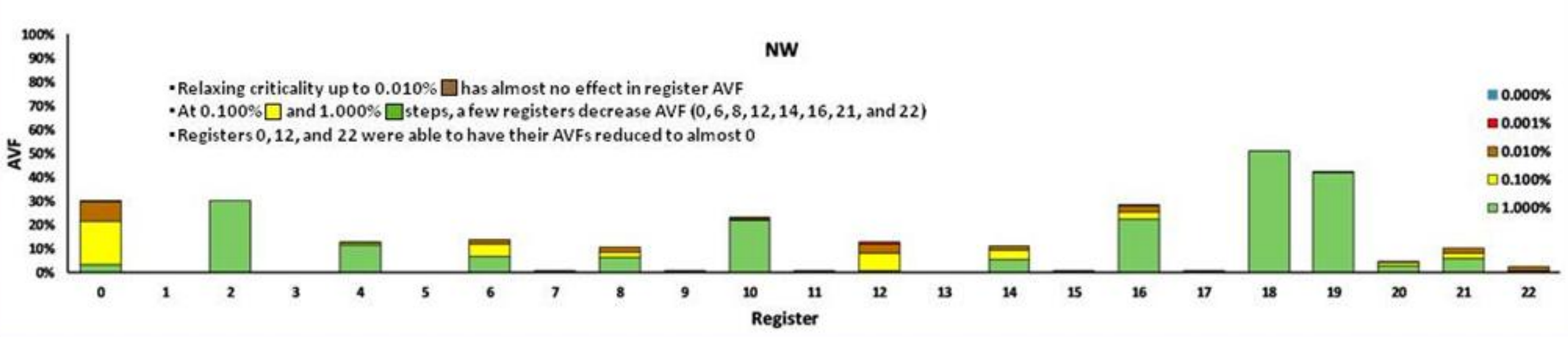
SDC FIT for all tested benchmarks normalized on a Poisson distribution

3. Relaxing Register File Methodology

- **Approximate computing** exploits the gap between the level of **accuracy required** by the application and the level of **accuracy provided** by the computing system
- By widening the required accuracy, we indirectly **relax register Architecture Vulnerability Factor (AVF).** By doing so, we aim **(1) to increase reliability** against SDC-induced-faults and **(2) to reduce the area overhead** in **selective fault tolerance techniques**
- As the required accuracy varies from application to application, **we chose 0% and a logarithmic scale varying from 0.001% to 1%** and evaluated the accuracy provided by the Kepler GPU system
- For Hotspot, NW and LavaMD, we relax accuracy by introducing a **percentage margin in which all individual results must be**. For Quicksort we relax accuracy by introducing a **percentage margin of total errors** in the output vector

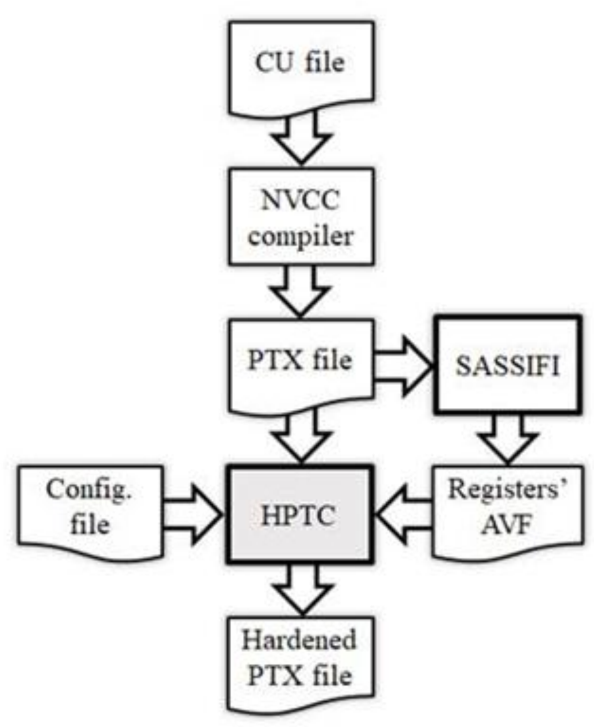
The graphics below show the individual register AVF with relaxed criticality





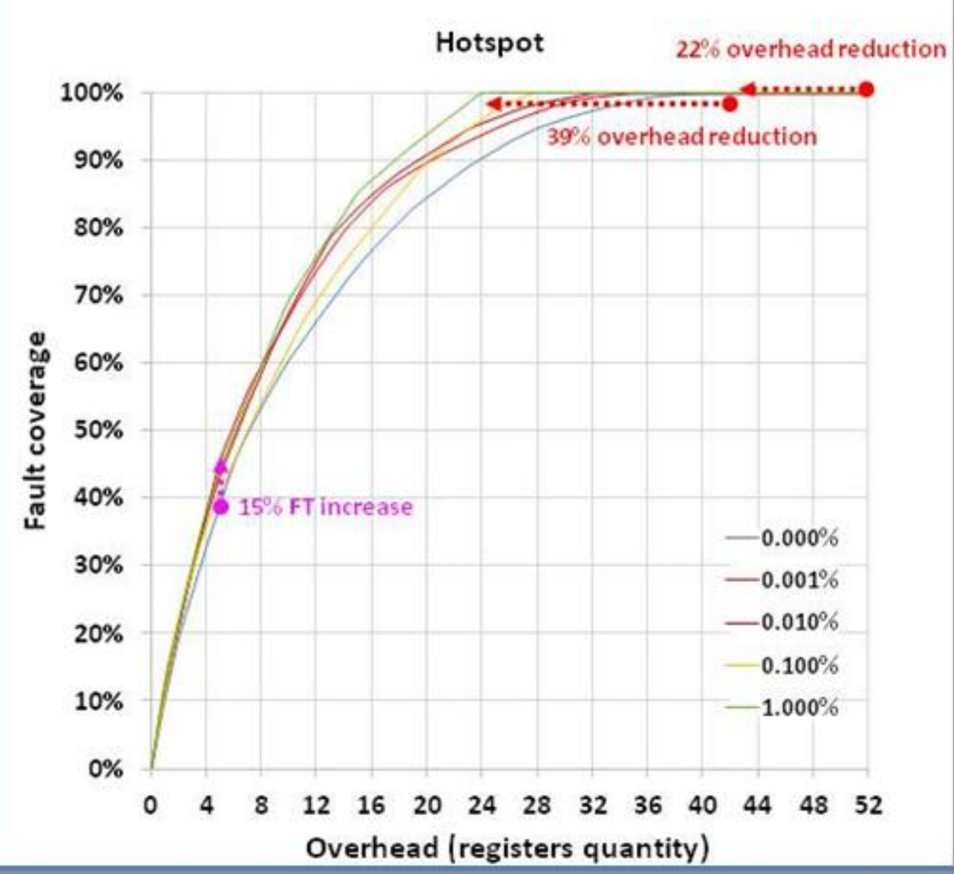
4. Selective Hardening Approach

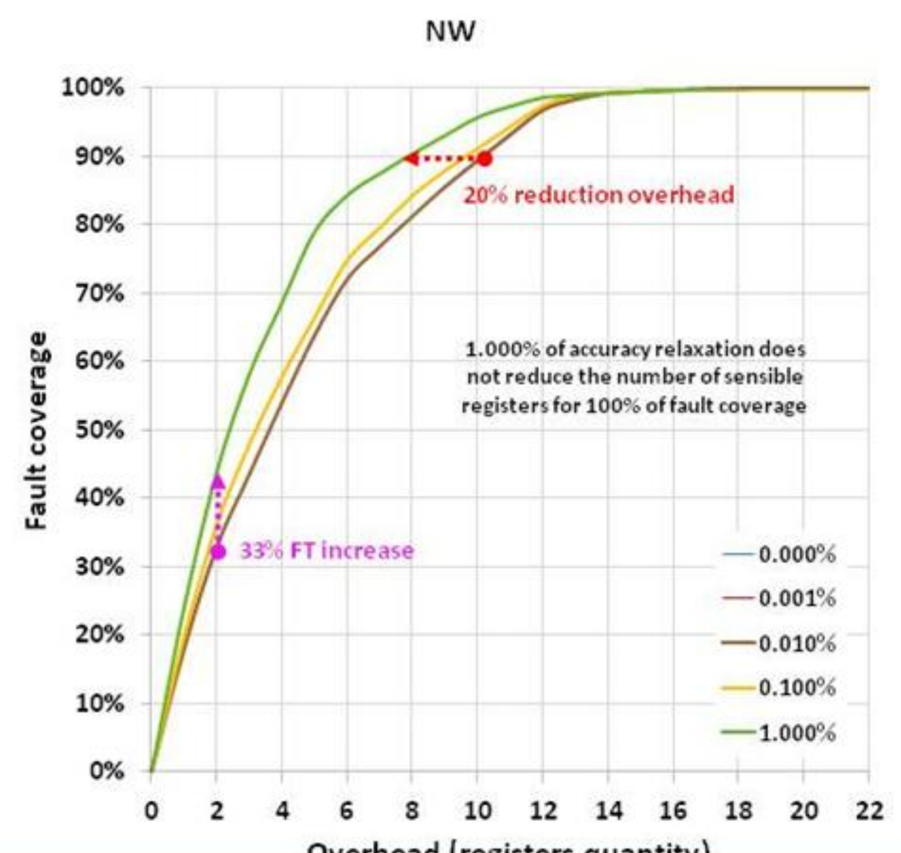
- **Selective hardening by hardware** replicates registers in a generic fashion, and **selective hardening by software** indirectly replicates registers by replicating instructions
- We **rank the most sensitive registers** considering AVF. The **registers' priority changes** according to the predefined accuracy relaxation
- To evaluate **software-implemented** selective fault tolerance, we must consider NVIDIA's **compilation Flow**. Previous works have validated that software-implemented techniques achieve the same fault coverage rates as hardware-implemented ones [1]



Software-implemented fault tolerance technique's flow [2]

The graphics below show fault coverage as a function of hardened registers





5. Conclusions and Future Work

- We proposed to decrease acceptance accuracy in order to improve selective fault tolerance techniques
- Results were able to **reduce overhead by an average of 42.4% while maintaining 100% fault coverage**, compared to selective hardening techniques
- **When lowering fault coverage constraints below 100%**, our approach presented higher gains, up to the point where, **at 10% hardened registers, we were able to increase fault coverage by an average of 77%**, compared to selective fault tolerance technique
- In the future, we intend to extend our approach to different GPUs and processor architectures

References

[1] M. Goncalves, F. Fernandes, I. Lamb, P. Rech, and J. Azambuja, *"Selective fault tolerance for register files of graphics processing units,"* IEEE Transactions on Nuclear Science, pp. 1–1, 2019.

[2] J. Azambuja, A. Lapolli, L. Rosa, and F. L. Kastensmidt, *"Detecting sees in microprocessors through a non-intrusive hybrid technique,"* IEEE Transactions on Nuclear Science, pp. 993–1000, 2011.