Politecnico di Torino

ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation
Doctoral Program in Computer and Control Engineering (36^{th}cycle)

Reliability Enhancement in GPU Architectures

By

Juan David Guerrero Balaguera

******

Supervisor(s):
Prof. Matteo Sonza Reorda, Supervisor
Prof. Ernesto Sanchez Sanchez, Co-Supervisor

Doctoral Examination Committee:
Prof. Luigi Carro, Referee, Universidade Federal do Rio Grande do Sul
Prof. Haralampos Stratigopoulos, Referee, Sorbonne Université, CNRS, LIP6
Prof. Antonio Miele, Politecnico di Milano
Prof. Leticia M. Bolzani Poehls, RWTH Aachen University
Prof. Massimo Violante, Politecnico di Torino

Politecnico di Torino
2024

Declaration

I hereby declare that, the contents and organization of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

Juan David Guerrero Balaguera
2024

* This dissertation is presented in partial fulfillment of the requirements for **Ph.D. degree** in the Graduate School of Politecnico di Torino (ScuDo).

Reliability Enhancement in GPU Architectures

Juan David Guerrero Balaguera

GPUs are important hardware accelerators for modern applications, particularly AI-based ones. They offer a high degree of parallelism, allowing multiple data to be processed simultaneously in a single chip. This is made possible by continuous technology scaling, resulting in higher transistor densities (e.g., 80 billion transistors for NVIDIA Hopper GPUs and 100 billion transistors for Intel GPUs). GPU devices offer great computational performance but can suffer from reliability issues associated with modern semiconductor technologies. As technology scales, several threats, such as accelerated wear-out, premature degradation, and high-temperature conditions, can increase the risk of hardware defects that lead to permanent faults. GPUs are now used in critical applications such as autonomous driving systems, aerospace, and avionics, among others. The reliability of GPUs is paramount, especially when used in such scenarios where they are required to operate correctly for longer lifetimes than typical consumer applications.

The reliability of GPU devices can be improved by implementing functional safety mechanisms (hardware or software) that can detect faults before they produce critical failures. These methods can ultimately reduce the probability of failure to acceptable levels. Unfortunately, hardware-based approaches require the addition of extra hardware structures to the device, which can increase costs, impact performance, or affect power consumption. Alternatively, the software-based self-testing (SBST) strategy is a flexible and noninvasive approach that offers in-field at-speed fault detection capabilities without hardware costs, leveraging the application's idle times to execute test procedures. Recently, several works have successfully demonstrated the feasibility of SBST for the development of Software Test Libraries (STLs), exploiting the inherent parallelism of GPUs for testing purposes and targeting functional units, memory modules, and control units.

Typically, the development of STLs for GPUs usually resorts to assembly languages, only. High-level programming languages (HLLs) for GPUs, such as CUDA C++ or OpenCL, simplify programming and are often the best, and sometimes the only, way to develop and encode applications. However, there are still several challenges and open questions when it comes to using HLLs for the development of STLs.

In this regard, this PhD thesis makes two main contributions. Firstly, it employs high-level (e.g., CUDA C++) or intermediate (e.g., CUDA PTX) programming languages to develop or map SBST strategies that are designed to test specific hardware components in GPUs. Secondly, it devises alternative strategies for compacting test programs, which help to reduce their memory footprint and speed up their test duration when used for in-field testing purposes.

On the other hand, techniques to identify permanent faults that can produce errors during the device's operative life are strongly required, since the typical simulation-based approaches result in prohibitive evaluation time. Additionally, these fault evaluations are crucial for two main reasons: first, they allow the identification of vulnerabilities in GPU's application regarding permanent faults, contributing to the development of effective software-based hardening strategies. Second, they can be used to assess the effectiveness of any software-based fault countermeasure against errors caused by permanent faults in GPUs.

Accordingly, this Ph.D. thesis proposes a method for evaluating the reliability of GPU applications in the presence of Silent Data Errors (SDEs) caused by permanent faults. The proposed method involves multi-level fault evaluations, which provide a better trade-off between accuracy (which is typically higher when we move closer to the hardware) and fast fault evaluations compared to other methods.

In conclusion, the thesis work proposes methods that aim to enhance the reliability of GPUs, taking a step forward from the current state-of-the-art. These methods include strategies for generating and compacting SBST for in-field GPU testing, as well as assessing the impact of permanent faults on GPU workloads. The effectiveness and limitations of these methods were evaluated through experiments on a set of representative benchmarks and compared with alternative solutions currently available.