Thesis Abstract: “New techniques for the Reliability Evaluation of AI-oriented Hardware Accelerators.”

- **PhD. Candidate:** Robert Alexander Limas Sierrra.
- **Supervisors:** Matteo Sonza Reorda and Josie E. Rodriguez C.
- **Cycle:** 38th.
- **Coordinator:** Fabrizio Lamberti.

Artificial Intelligence (AI) has become a cornerstone of modern computing, enabling applications in autonomous vehicles, robotics, healthcare, finance, and high-performance computing (HPC). These domains demand high computational throughput, energy efficiency, and low latency, which has driven the widespread adoption of specialized AI-oriented hardware accelerators. In practice, these platforms include devices such as graphics processing units (GPUs), deep learning accelerators (DLAs), and neural processing units (NPUs), often integrating dedicated tensor-acceleration engines to speed up matrix operations. They execute machine learning and deep learning workloads through massive parallelism, using architectures composed of many arithmetic units performing multiply–accumulate operations.

As these accelerators move beyond data centers and into safety-critical domains—such as autonomous driving, aerospace, and advanced robotics—their reliability becomes a fundamental design concern. AI-oriented accelerators are increasingly fabricated in deeply scaled semiconductor technologies that, while improving performance and power efficiency, also increase susceptibility to hardware faults. Radiation-induced upsets, process variations, and device aging can cause transient or permanent faults in logic and memory structures. Depending on where they occur and how they propagate, such faults may lead to silent data corruptions, where results are corrupted without detection, or detected unrecoverable errors, which may trigger exceptions, hangs, or application and system crashes. In safety-critical deployments and HPC environments, these outcomes can translate into mispredictions, accuracy degradation, or catastrophic failures.

The doctoral research presented in this thesis addresses these challenges by introducing new techniques for the reliability evaluation of AI-oriented hardware accelerators. The goal is twofold: first, to develop methodologies for accurate and scalable reliability assessment across multiple abstraction levels, and second, to design lightweight mitigation strategies, co-optimized across hardware and software, that enhance fault resilience under realistic constraints. Compared to the state of the art—often split between hardware-agnostic perturbation approaches, which are scalable but low fidelity, and low-level fault injection, which is accurate but expensive and difficult to generalize—this thesis contributes a unified cross-layer flow that links circuit-level fault mechanisms to accelerator-level propagation and application-level impact, while remaining tractable for large workloads and adaptable across numerical formats.

At the arithmetic level, the thesis performs fine-grain reliability characterization of floating-point and Posit arithmetic cores implementing addition, multiplication, and

multiply–accumulate operations. Through detailed fault-injection campaigns at gate and structural levels, the study identifies the substructures most responsible for error propagation and quantifies the magnitude and distribution of the resulting numerical corruptions. Rather than treating arithmetic units as black boxes or reporting only aggregate error rates, the proposed characterization connects observed error severity to specific internal blocks such as exponent or regime handling, normalization, and rounding. This enables feasible and format-aware mitigation decisions. The resulting vulnerability profiles explain why some floating-point designs exhibit rare but extreme outliers, whereas Posit designs may show higher activation rates with more bounded deviations.

At the accelerator level, the thesis investigates the reliability of tensor-core-style compute engines that accelerate general matrix multiplication operations, which are fundamental kernels in deep neural networks and convolutional neural networks. Faults are injected into internal dot-product units and local memories to study how dataflow organization, workload mapping, and numerical format affect fault propagation and error accumulation. Beyond reporting error rates, the thesis shows that tensor-core faults generate deterministic spatial corruption signatures tied to warp-to-dot-product-unit mapping and matrix–multiply–accumulate scheduling. This information is typically lost in application-level injection or overly abstract architectural models, yet it is essential to explain observed outcomes and to enable targeted runtime strategies that operate on tile fragments rather than duplicating entire kernels.

To reduce the computational cost of traditional fault-injection campaigns, the thesis proposes scalable evaluation frameworks combining analytical error models and a high-performance-computing-based execution environment. First, statistical fault models are developed to represent the impact of hardware faults as compact patterns of output corruptions in matrix multiplication results. Unlike fast simulators based on hardware-agnostic bit-flip assumptions, the proposed models are derived from low-level gate-level characterization and preserve both magnitude statistics and the spatial corruption patterns observed in tensor-core executions, enabling higher fidelity at a fraction of the simulation cost. Second, a distributed HPC-based environment is introduced to accelerate large fault-injection campaigns through containerized execution across multiple compute nodes, enabling analyses that would otherwise be impractical on a single workstation.

Beyond evaluation, the thesis explores practical mitigation strategies. On the hardware side, selective hardening is proposed as an efficient alternative to full redundancy. Instead of applying uniform replication across entire datapaths—often prohibitive for accelerator-grade arithmetic—the thesis leverages fine-grain vulnerability maps to protect only the blocks that dominate silent data corruptions and catastrophic numerical outliers. This targeted approach improves reliability under tight overhead constraints and provides a methodology to select where redundancy yields the highest reliability gain per area and power cost.

On the software side, a fault-tolerant matrix multiplication mechanism is introduced for tensor-core accelerators. Unlike classical kernel duplication and generic software redundancy, which often incur large slowdowns, the proposed mechanism exploits

tensor-core tiling and redundant execution at fragment granularity. It provides continuous detection with low overhead and activates fine-grain reconstruction only upon anomaly detection. This hardware-aware design yields predictable overhead while enabling correction of localized persistent faults without requiring hardware changes.

The key contributions of this thesis can be summarized as follows:

- Development of a unified cross-layer reliability evaluation methodology for AI-oriented accelerators that bridges circuit-level fault mechanisms, tensor-core architectural propagation, and application-level impact.
- Fine-grain fault analysis of FP and Posit arithmetic cores that maps error severity and activation to specific internal substructures, enabling feasible and format-aware hardening decisions beyond black-box reliability metrics.
- Architectural reliability assessment of tensor-core-style accelerators that identifies deterministic spatial fault signatures induced by scheduling and DPU mapping, capturing effects not represented by hardware-agnostic tensor perturbations.
- Hardware-aware analytical error models that preserve both magnitude statistics and spatial corruption footprints while reducing evaluation cost compared to full architectural fault injection.
- Mitigation strategies with bounded overhead: selective redundancy guided by vulnerability profiles, and tensor-core-aware GEMM detection/correction that avoids full kernel duplication.

Overall, the thesis advocates a cross-layer approach to reliability in AI-oriented hardware accelerators, integrating low-level fault modeling, architectural analysis, and hardware and software mitigation strategies. The key advance is the combined emphasis on accuracy, scalability, and deployability, providing practical tools and design guidelines for dependable AI computing in safety-critical and high-availability environments.