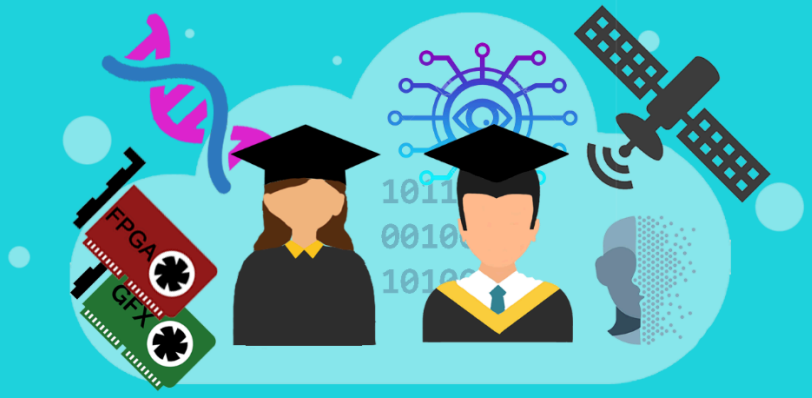


Diploma Thesis

Microprocessors and
Digital Systems
Laboratory



Architectural Vulnerability Analysis and Mitigation for DNN Accelerators

Reliability has emerged as a significant constraint for Large Language Models (LLMs) [1] and dynamic neural networks, especially as they migrate from controlled datacenters to edge and embedded paradigms. Error sources and aggressive efficiency techniques are capable of silently altering network results. The central challenge is to model how hardware- and architecture-level perturbations interact with the model's parameters and to establish accuracy-architecture tradeoffs. The main goal of the thesis is to **assess the impact of fault sources and aggressive optimization techniques on the accuracy of DNNs**. We will explore a variety of candidate architectures and fault sources [2]. The reliability analysis will consist of fine-grained fault injection at gate level using FPGAs.

Modern Candidate Accelerators:

- **NPU - AMD AI Engines:** AMD's XDNA NPU, integrated in Ryzen AI processors, is a 2D spatial array of VLIW/SIMD AI Engine tiles with local memories, connected through a programmable streaming interconnect [3].
- **Vectorized Processors:** Ara is an open-source RISC-V Vector processor coupled to the CVA6 core, where parallel lanes of register file slices and functional units process long vectors with a single instruction [4]. Its large vector register file and lane-based datapath expose distinct vulnerability profiles for DNN inference kernels.
- **Embedded GPUs:** Vortex is an open-source, RISC-V based GPGPU that supports OpenCL and CUDA through a minimal ISA extension [5]. Its configurable Verilog can run on FPGAs, enabling fault injection across GPU components.
- **CGRA:** A CGRA is a spatially programmable grid of lightweight processing elements and local memories that maps kernels as dataflow across the fabric, providing near-ASIC performance and energy efficiency with much greater flexibility than fixed hardware [6], [7].

Targeted Fault Models:

- **Radiation:** Particle strikes cause bit flips in flip-flops and SRAM arrays, leading to errors and Silent Data Corruptions.
- **Aging:** Degradation mechanisms such as BTI gradually increase transistor delays, causing timing violations and eventually permanent faults.
- **Thermal:** Hotspots from dense, sustained MAC activity increase delay, leakage and accelerate aging, making error rates workload- and mapping-dependent.
- **Undervolting:** Operating below nominal voltage yields large energy savings at the cost of timing and memory bit-cell failures, which error-tolerant layers can absorb [8].

The diploma thesis focuses on the following areas (but is not limited to):

- Studying the execution patterns of State-of-the-Art DNNs on the candidate accelerators
- Assessing the impact of hardware faults on model accuracy through FPGA/ASIC fault injection
- Vulnerability classification of the hardware structures [9]
- Exploration of error mitigation techniques (e.g., digital design, application mapping)

RELATED MATERIAL:

- [1] J. Guo, C. Chakrabarti, D. Fan, "SBFA: Single Sneaky Bit Flip Attack to Break Large Language Models," arXiv:2509.21843, 2025. <https://arxiv.org/pdf/2509.21843>
- [2] N. K. Salih, D. Satyanarayana, A. S. Alkalbani and R. Gopal, "A Survey on Software/Hardware Fault Injection Tools and Techniques," 2022 ISIEA, doi: 10.1109/ISIEA54517.2022.9873679
- [3] A. Rico et al., "AMD XDNA NPU in Ryzen AI Processors," IEEE Micro, vol. 44, no. 6, 2024.
- [4] M. Perotti, et al., "Ara2: Exploring Single- and Multi-Core Vector Processing With an Efficient RVV 1.0 Compliant Open-Source Processor," IEEE Transactions on Computers, 2024
- [5] B. Tine, et al., "Vortex: Extending the RISC-V ISA for GPGPU and 3D-Graphics," MICRO 2021
- [6] B. de Bruin, K. Vadivel, M. Wijtvliet, P. Jäskeläinen and H. Corporaal, "R-Blocks: an Energy-Efficient, Flexible, and Programmable CGRA," ACM TRETs, 2024. <https://doi.org/10.1145/3656642>
- [7] G. Alexandris, P. Chaidos, A. Maras, et al., "A Unified Framework for Mapping and Synthesis of Approximate R-Blocks CGRAs," <https://doi.org/10.48550/arXiv.2505.23553>
- [8] B. Salami, O. S. Unsal and A. Cristal Kestelman, "Comprehensive Evaluation of Supply Voltage Underscaling in FPGA on-Chip Memories," 51st IEEE/ACM (MICRO), 2018.
- [9] G. Papadimitriou and D. Gizopoulos, "AVGI: Microarchitecture-Driven, Fast and Accurate Vulnerability Assessment," IEEE HPCA, 2023.

PREREQUISITES:

- Familiarity with Machine Learning/Neural Network Model Architectures
- VLSI Digital Design (Verilog/SystemVerilog), Python, PyTorch/ONNX, Bash, C/C++, OpenCL

CONTACT INFORMATION:

- Panagiotis Chaidos, Ph.D Student, NTUA: (pchaidos@microlab.ntua.gr)
- Alexis Maras, Ph.D Student, NTUA: (amaras@microlab.ntua.gr)
- Georgios Alexandris, Ph.D Student, NTUA: (galexandris@microlab.ntua.gr)
- Panagiotis Eleftherakis, Ph.D Student, NTUA: (pelef@microlab.ntua.gr)
- Assistant Professor Sotirios Xydis, NTUA: (sxydis@microlab.ntua.gr)