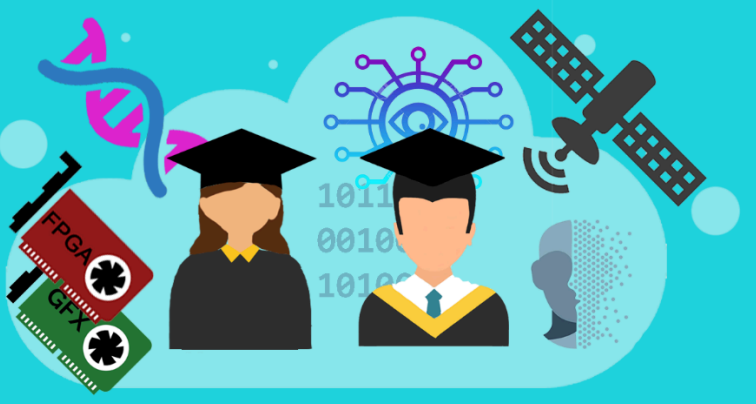
Diploma Thesis

Microproccessors and Digital Systems Laboratory



Hardware-Aware Reliability and Mitigation Assessment of LLMs and Dynamic NNs

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 from transient phenomena, wear-out mechanisms, and aggressive efficiency techniques are capable of deliberately altering network results. The central challenge is to model how hardware-level perturbations interact with the model's parameters and to establish accuracy–architecture trade-offs. The main goal of the thesis is to **assess the impact of fault sources and aggressive optimization techniques on the accuracy of LLMs and Dynamic Neural Networks.**

Faults can be categorized by their **permanence** into transient, leaving the hardware unchanged (radiation-induced upsets, voltage droops, thermal effects), and permanent, causing irreversible changes to the accelerator (device aging, stuck-at defects). They can also be categorized by **intent** into unwanted faults(manufacturing defects) and intentional approximations, made by the designer to explore trade-offs (quantization, approximate arithmetic). **Faults can thus come from several sources, manifest in different ways, and have different effects on a network, all of which need to be taken into account**.[3]

Modern Candidate Accelerators: Different underlying architectures yield different error manifestations, different dataflows and resulting network accuracy.

- **TPU:** A TPU is a specialized ASIC based on a systolic array architecture with tightly coupled on-chip memory, delivering high throughput and energy efficiency. Recent work shows that many parts of modern AI workloads can be mapped on a single systolic architecture.[2]
- **NPU:** With the term NPU (Neural Processing Unit) we can describe a variety of Domain Specific implementations, built for accelerating compute intensive parts of the input models. Usually it consists of custom implementations of GeMM or Matrix-Vector hardware operators. A SoTA and recent example of NPU frameworks is Google's Coral NPU[4]
- **Digital Compute-In-Memory:** Digital CIM executes multiply-accumulate operations inside or near memory arrays to minimize data movement, yielding large energy and latency reductions in memory-bound workloads, while retaining scalability.
- **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.

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

- Studying the execution patterns of State-of-the-Art LLMs and dynamic DNNs
- Modeling HW-Aware faults for modern DNN accelerators
- Assessing the impact of Hardware faults on model performance through FPGA/ASIC flow
- Thermal Aware modeling and mitigation

RELATED MATERIAL:

[1]Jingkai Guo, Chaitali Chakrabarti, Deliang “FanSBFA: Single Sneaky Bit Flip Attack to Break Large Language Models”, https://arxiv.org/pdf/2509.21843

[2] Jiawei Lin, Guokai Chen, Yuanlong Li, Thomas Bourgeat, “SystolicAttention: Fusing FlashAttention within a Single Systolic Array”, https://arxiv.org/pdf/2507.11331

[3]N. K. Salih, D. Satyanarayana, A. S. Alkalbani and R. Gopal, "A Survey on Software/Hardware Fault Injection Tools and Techniques," 2022 IEEE Symposium on Industrial Electronics & Applications (ISIEA), doi: 10.1109/ISIEA54517.2022.9873679

[4] https://developers.googleblog.com/en/introducing-coral-npu-a-full-stack-platform-for-edge-ai/

PREREQUISITES:

- Familiarity with Machine Learning/Neural Network Models Architectures
- Python, Pytorch/ONNX, Bash, Digital Design

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)
- Professor Dimitrios Soudris, NTUA: (dsoudris@microlab.ntua.gr)
- Assistant Professor Sotirios Xydis, NTUA: (sxydis@microlab.ntua.gr)