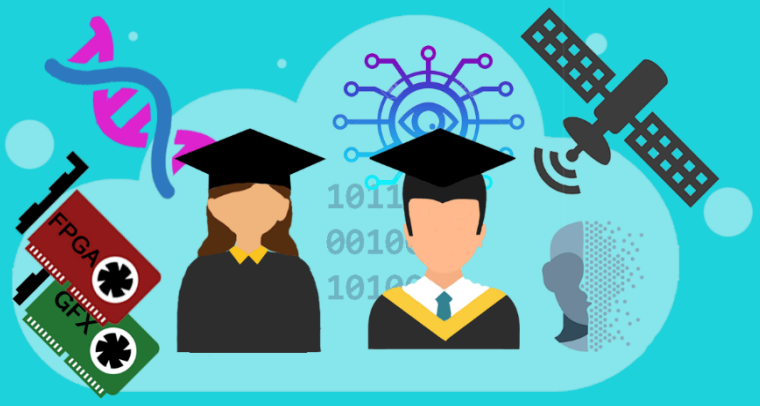
Diploma Thesis

Microproccessors and Digital Systems Laboratory



Exploring Reconfigurable Hardware Accelerators for Efficient Deep Learning at the Edge

The rapid and continuous generation of data by modern applications has driven a paradigm shift from centralized cloud computing toward *edge computing*, where data processing occurs closer to the data source. This transition aims to reduce latency, enhance privacy, and improve energy efficiency. However, deploying increasingly complex applications, particularly Deep Learning (DL) workloads such as Vision Transformers (ViTs) and Large Language Models (LLMs), on edge devices presents significant challenges. General-purpose hardware often fails to meet the stringent performance and energy-efficiency requirements of these models.

To overcome these limitations, **hardware accelerators** have become a key enabling technology, with solutions ranging from GPUs and TPUs to ASICs and FPGAs. Among these, Field-Programmable Gate Arrays (FPGAs) stand out due to their reconfigurability, parallel processing capabilities, and energy-efficient operation.

Recent advances have led to heterogeneous architectures that further enhance computational power while retaining the benefits of FPGAs. A notable example is the **Versal Adaptive Compute Acceleration Platform (ACAP)** by AMD/Xilinx, which integrates reconfigurable FPGA logic with high-performance AI Engines, which essentially are programmable ASIC-like processors.

Despite its promising architecture, maximizing the performance of modern deep learning workloads on Versal ACAP devices remains an open research challenge. Efficiently mapping the complex modern Deep Learning applications to the heterogeneous fabric requires algorithmic–hardware co-design approaches that balance performance and energy efficiency.

The goal of this thesis is to investigate and develop hardware accelerator architectures tailored for efficient deployment of modern deep learning models (e.g., LLMs) on Versal ACAP and FPGA platforms. We will explore algorithmic–hardware co-design strategies to optimize key computational kernels, aiming to surpass state-of-the-art implementations in terms of performance and energy efficiency, enabling powerful AI processing capabilities at the edge.

PREREQUISITES:

Knowledge of Digital Design, Experience with FPGAs, VHDL, C/C++, HLS, Python, Bash

RELATED MATERIAL:

[1] Zhuang, Jinming, et al. "CHARM: Composing Heterogeneous AcceleRators for Matrix Multiply on Versal ACAP Architecture." Proceedings of the 2023 ACM/SIGDA International Symposium on Field Programmable Gate Arrays. 2023.

[2] Wang, Chengyue, et al. "Reconfigurable Stream Network Architecture." Proceedings of the 52nd Annual International Symposium on Computer Architecture. 2025.

[3] Huang, Yingbing, et al. "New solutions on LLM acceleration, optimization, and application." Proceedings of the 61st ACM/IEEE Design Automation Conference. 2024.

[4] Wang, Hanrui, Zhekai Zhang, and Song Han. "Spatten: Efficient sparse attention architecture with cascade token and head pruning." 2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA). IEEE, 2021.

[5] Zeng, Shulin, et al. "Flightllm: Efficient large language model inference with a complete mapping flow on fpgas." Proceedings of the 2024 ACM/SIGDA International Symposium on Field Programmable Gate Arrays. 2024.

CONTACT INFORMATION:

Ilias Papalamprou, PhD Candidate [ipapalambrou@microlab.ntua.gr]

Dimosthenis Masouros Ph.D [dmasouros@microlab.ntua.gr]

Prof. Dimitrios Soudris [dsoudris@microlab.ntua.gr]