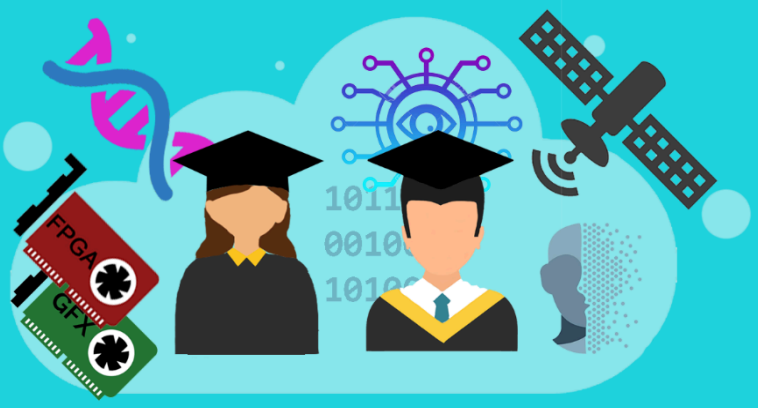
Diploma Thesis

Microproccessors and Digital Systems Laboratory



Exploring Reconfigurable Hardware Accelerators for Efficient Deep Learning at the Edge

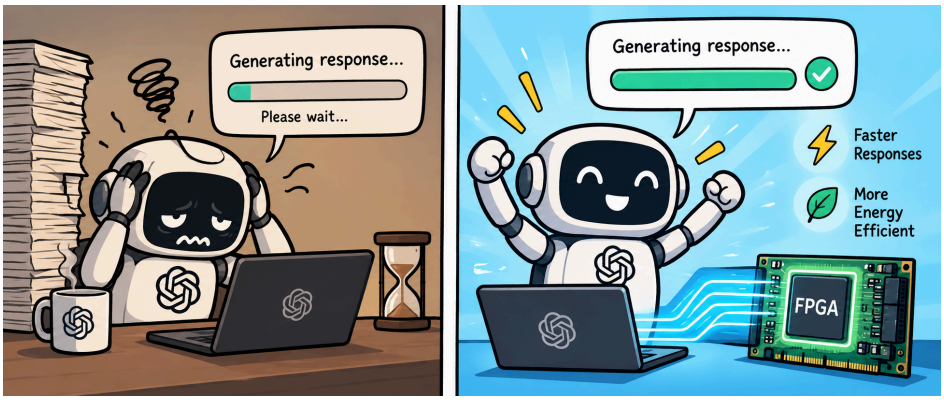
The rapid advancement of Artificial Intelligence (AI) is enabling a new generation of applications, ranging from Generative AI to Physical AI, where autonomous systems perceive, reason about, and interact with the physical world. These applications increasingly rely on complex Deep Learning (DL) models, including **Large Language Models (LLMs), Vision-Language Models (VLMs), and Vision Transformers (ViTs)**, introducing substantial computational demands.

The growing need for real-time processing, low latency, and energy efficiency has driven a paradigm shift from centralized cloud computing toward *edge computing*, where computation occurs closer to the data source. However, deploying complex DL workloads on resource-constrained edge devices remains challenging, as general-purpose hardware often struggles to meet their performance and energy-efficiency requirements.

To address these challenges, **hardware accelerators** have become a key enabling technology, ranging from GPUs and TPUs to ASICs and FPGAs. Among these, **Field-Programmable Gate Arrays (FPGAs)** offer reconfigurability, parallel processing capabilities, and energy-efficient operation. Recent advances have introduced heterogeneous architectures, such as AMD's **Versal Adaptive Compute Acceleration Platform (ACAP)**, which combines reconfigurable FPGA logic with specialized programmable AI Engines (AIEs).

Despite its promising architecture, efficiently mapping modern DL workloads onto Versal ACAP devices remains an open research challenge, requiring **algorithm–hardware co-design** to optimize computation, data movement, and energy efficiency.

The goal of this thesis is to investigate and develop reconfigurable 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.



Disclaimer: Generated using ChatGPT

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]