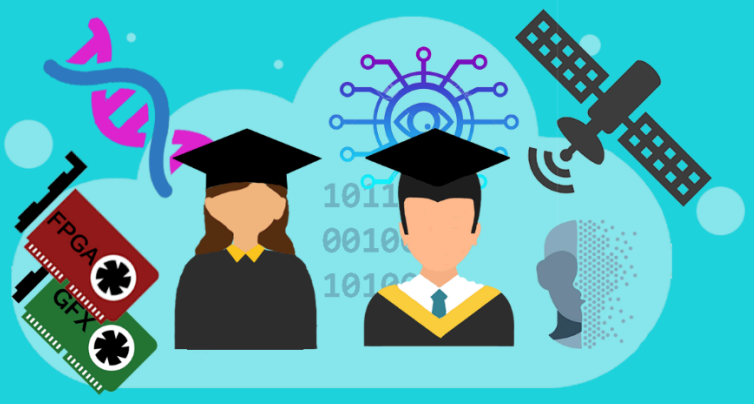
Diploma Thesis

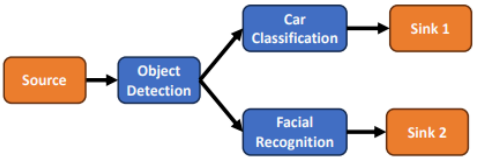
Microproccessors and Digital Systems Laboratory



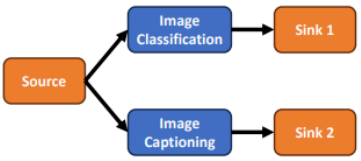
Heterogeneity-aware & SLO-driven AI Inference Serving

AI inference serving has become a cornerstone of modern intelligent systems, powering applications from real-time vision to large-scale recommendation engines. As models grow in complexity, traditional CPU-based infrastructures struggle to meet stringent performance demands, motivating the adoption of hardware accelerators such as GPUs, FPGAs, and TPUs. However, these devices introduce new challenges in scheduling and resource management due to their diverse capabilities and energy profiles. The resulting heterogeneity across cloud and edge environments makes efficient workload distribution and performance optimization increasingly complex.

In today's computing landscape, **energy efficiency** is as critical as raw performance. Data centers and edge deployments must balance operational costs and carbon footprint reduction while maintaining compliance with strict **service-level objectives (SLOs)** such as throughput and latency. Achieving this balance requires intelligent resource allocation strategies that can adapt to dynamic workloads without compromising user experience or violating SLO guarantees.



(a) Traffic analysis pipeline



(b) Social media pipeline

AI inference often operates through **multi-stage pipelines** that connect several dependent models, as seen in applications like traffic analysis or social media content processing [Figure]. Each stage, such as object detection, classification, or captioning, has distinct computational and latency characteristics and may benefit from different hardware types [1]. Managing these pipelines across heterogeneous accelerators introduces additional challenges in scheduling, data transfer, and SLO enforcement, but also provides opportunities for cross-stage optimization. Coordinated execution across devices can reduce end-to-end latency, improve energy efficiency, and enhance resource utilization.

In this thesis, we leverage an internal tool for efficiently converting and executing code on heterogeneous devices [2, 3], and develop a methodology for serving AI inference pipelines across diverse hardware platforms, in a staged-manner [4]. The proposed approach takes into account the unique performance characteristics of each model variant, such as latency, energy consumption, and initialization time, on different accelerators. The outcome of this work is a framework that aims to minimize energy consumption while maintaining strict service-level objectives (SLOs), effectively utilizing the heterogeneous hardware typically found in edge computing environments.

References:

[1] Ahmad, Sohaib, Hui Guan, and Ramesh K. Sitaraman. "Loki: A system for serving ml inference pipelines with hardware and accuracy scaling." Proceedings of the 33rd International Symposium on High-Performance Parallel and Distributed Computing. 2024.

[2] Leftheriotis, Aimilios, et al. "TF2AIF: Facilitating development and deployment of accelerated AI models on the cloud-edge continuum." 2024 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit). IEEE, 2024.

[3] https://github.com/aimilefth/CECAIServe

[4] Tzenetopoulos, Achilleas, et al. "Leveraging Core and Uncore Frequency Scaling for Power-Efficient Serverless Workflows." arXiv preprint arXiv:2407.18386 (2024).

Contact:

Achilleas Tzenetopoulos, Ph.D. candidate NTUA: atzenetopoulos@microlab.ntua.gr
Aimilios Leftheriotis, Ph.D. candidate UoP: aimilefth@microlab.ntua.gr
Dimitrios Soudris, Professor Microlab NTUA: dsoudris@microlab.ntua.gr