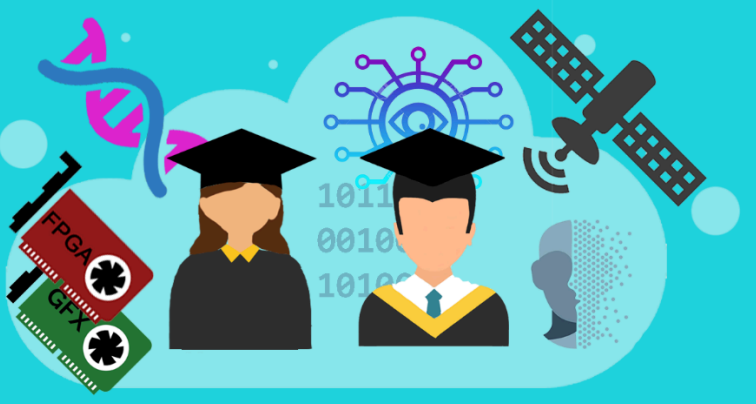
Diploma Thesis

Microproccessors and Digital Systems Laboratory



On-Premise LLM Serving on memory-constrained edge clusters

On-premise deployment is becoming an important setting for Large Language Model (LLM) serving [1], as organizations increasingly bring inference in-house for data privacy, cost control, and predictable availability. Unlike hyperscale data centers, these deployments often rely on **heterogeneous clusters** that combine different GPU generations, CPUs, and emerging accelerators such as NPUs, while operating under tight memory and resource constraints. Model weights, KV caches, and concurrent tenants compete for limited accelerator memory, often requiring computation and data to be partitioned or offloaded across heterogeneous devices [2]. At the same time, **shared workloads interfere through shared resources,** including accelerator compute and memory, CPU memory bandwidth, PCIe/interconnect bandwidth, caches, and power budgets.

The main goal is to develop **runtime**, **modeling**, and **compiler-level mechanisms** for **adaptive** and **efficient LLM inference** on **heterogeneous**, resource-constrained **systems**. The work will investigate how LLM workloads can be dynamically **partitioned** across available compute resources, how interference between co-located workloads can be **predicted**, and how heterogeneous systems can be **formally modeled** in terms of accelerator compute capability, memory capacity and bandwidth, and intra-node and inter-node communication characteristics. Such models can expose the relevant architectural and interconnect parameters as variables, enabling systematic **exploration of different hardware configurations** and execution strategies.

Beyond individual model invocations, the work will also consider **multi-stage LLM workflows and emerging agentic AI applications**, where a request may trigger multiple dependent model calls, tools, or inference stages with different performance characteristics. In these settings, the runtime can jointly adapt multiple execution knobs, including frequency scaling (DFS/DVFS), model and operator placement, quantization, batching, resource allocation, and CPU/accelerator offloading, to exploit slack across workflow stages, control interference, and balance latency, throughput, energy efficiency, and resource utilization.

The available diploma theses will focus on the following areas:

- Workload and hardware profiling of LLMs
- Development of analytical or learned performance models
- Interference prediction and formal optimization
- Compiler-Level conditional execution for control overhead minimization
- Dynamic expert placement in Mixture-of-Experts (MoE) models
- Integration on real-world LLMs (Llama, DeepSeek, Qwen)
- Integration on production grade serving systems (vLLM)
- Workflow-aware runtime optimization for compound and agentic AI

- Adaptive knobs tuning such as frequency scaling (DFS/DVFS), model placement, quantization, batching, and resource allocation.

RELATED MATERIAL:

[1] B. Li, Y. Jiang, V. Gadepally, and D. Tiwari, "LLM Inference Serving: Survey of Recent Advances and Opportunities," arXiv:2407.12391, 2024

[2] Xuanlin Jiang, Yang Zhou, Shiyi Cao, Ion Stoica, and Minlan Yu. NEO: Saving GPU Memory Crisis with CPU Offloading for Online LLM Inference. *MLSys, 2025*

[3] Shiyi Cao, et al., MoE-Lightning: High-Throughput MoE Inference on Memory-constrained GPUs. In ASPLOS 2025, https://doi.org/10.1145/3669940.3707267

PREREQUISITES:

- Familiarity with Machine Learning/Neural Network Model Architectures
- Linux, Python, C/C++

CONTACT INFORMATION:

- Panagiotis Chaidos, Ph.D. Student, NTUA: (pchaidos@microlab.ntua.gr)
- Georgios Alexandris, Ph.D. Student, NTUA: (galexandris@microlab.ntua.gr)
- Achilleas Tzenetopoulos, Ph.D., NTUA: (atzenetopoulos@microlab.ntua.gr)
- Assistant Professor Sotirios Xydis, NTUA: (sxydis@microlab.ntua.gr)