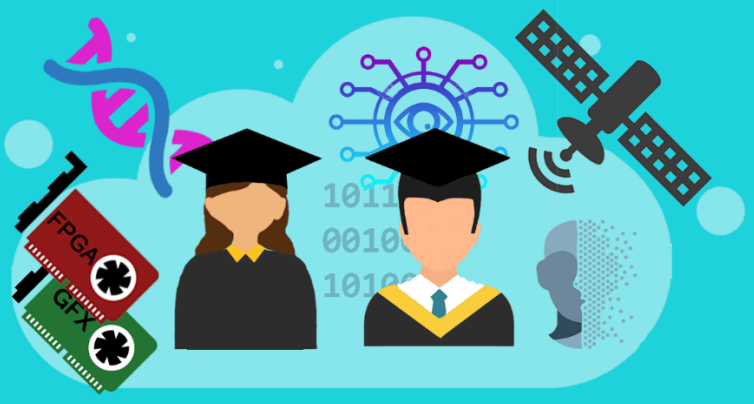
Diploma Thesis

Microproccessors and Digital Systems Laboratory



Dynamic Resource Managent of Multi-agent AI workflows

Agentic AI is a way of building systems where models act as goal-driven agents that can plan, use tools, and take multi-step actions. It is estimated that agentic AI will be incorporated into 33% of enterprise software by 2028 [1]. Multi-agent workflows are a concrete way to implement this idea by having multiple specialized agents collaborate in a pipeline or graph to solve complex tasks together: instead of one big model doing everything, we have several specialized agents (e.g., planner, retriever, code-writer, evaluator) that call each other in a dynamic graph.

At the same time, cloud platforms are moving towards a serverless model, where functions scale up and down automatically, pay-as-you-go, and can be placed on different machines or hardware accelerators [1] (CPUs, GPUs, possibly other accelerators). This thesis aims to bring these two worlds together: treat AI agents like serverless functions and orchestrate them with serverless-style techniques such as cold-start handling, autoscaling, and heterogeneous placement.

In such a system, different agents can have very different resource needs and performance behaviors. Some agents may require access to specialized hardware (e.g., GPUs or other accelerators), while others run comfortably on CPUs. Certain agents might benefit a lot from scaling out (more parallel instances increase throughput), or scaling up (more CPUs decrease latency) whereas others see diminishing returns from extra resources. In addition, some agents can process requests in batches very efficiently, trading a bit of waiting time for much better throughput and hardware utilization, while others are purely single-request and latency-focused. A key challenge in the thesis is to understand and model these differences, and then design orchestration policies that decide the resource allocation mix, i.e., hardware type, instance count, batching level, dynamically, during runtime.

The thesis will involve defining a few representative multi-agent workloads, and then **profiling and characterizing** them experimentally. Based on these observations and analysis results, the student will design a methodology for **dynamic resource management** of agents. The evaluation will focus on metrics such as **Service-Level-Objectives** satisfaction (e.g., end-to-end latency), throughput, and cost, from the perspective of a cloud provider that wants to run AI agents efficiently while still giving users fast responses.

References:

[1] Gartner Report: https://www.gartner.com/en/articles/intelligent-agent-in-ai

Contact:

Achilleas Tzenetopoulos, Ph.D. candidate NTUA: (atzenetopoulos@microlab.ntua.gr)
Dimosthenis Masouros, Post Doc NTUA
Sotirios Xydis, Ass. Prof., Microlab NTUA
Dimitrios Soudris, Professor Microlab NTUA