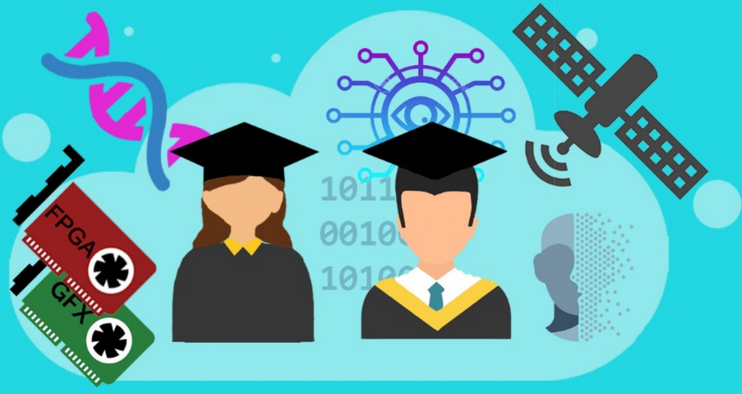
# Diploma Thesis

Microproccessors and
Digital Systems
Laboratory




## Improving inference engine GPU colocation using DNN based predictive monitoring

Nowadays, there is an ever-increasing number of artificial intelligence inference workloads pushed and executed on the cloud. To effectively serve and manage the computational demands, datacenter operators have provisioned their infrastructures with accelerators. Specifically for GPUs, support for efficient management lacks, as state-of-the-art schedulers and orchestrators (e.g Kubernetes), treat GPUs only as typical compute resources ignoring their unique characteristics and application properties. This phenomenon combined with the GPU over-provisioning problem leads to severe resource under-utilization. Even though prior work has addressed this problem by colocating applications into a single accelerator device, its resource agnostic nature does not manage to face the resource under-utilization and quality of service violations especially for latency critical applications.

To overcome the limitations of the resource agnostic state-of-the-art schedulers, the usage of real-time GPU metrics has been proposed. By processing the provided metrics in time-series format, schedulers are aware of the current resource state and thus they are able to efficiently colocate inference engines. It is evident that the processing of the acquired data, highly impacts the scheduling decisions. To get insight concerning the current as well as the future GPU state, Machine Learning (ML) and especially Deep Learning (DL) techniques have been employed. In this direction, Yeung et al. proposed a GPU utilization prediction engine for DL workloads that leads to an improvement of *61.5%* to GPU cluster utilization, showcasing the benefits of DL enhanced scheduling systems.

In this diploma thesis, we will design, implement and evaluate a model that is able to predict GPU metrics (utilization percentage, power consumption e.t.c.). As a first step, we will examine the runtime characteristics of a set of inference engines from state-of-the-art benchmark suites and gather the data for the training phase. Then, we will test various ML and DL models to find the most appropriate for each GPU metric. Finally, we will integrate our model to a Kubernetes scheduler and investigate different GPU colocation policies.

### PREREQUISITES:

Strong knowledge of **Bash**, **Python**

Desirable: Familiarity with **Machine Learning**, **Neural Networks**

## SKILLS YOU WILL LEARN:

In this diploma thesis, you will familiarize with **Docker** containers, **Kubernetes** and **Machine Learning** techniques.

## RELATED MATERIAL:

- Ferikoglou, Aggelos, et al. "Resource Aware GPU Scheduling in Kubernetes Infrastructure." 12th Workshop on Parallel Programming and Run-Time Management Techniques for Many-core Architectures 10th Workshop on Design Tools. 2021.
- Yeung, Gingfung, et al. "Towards GPU Utilization Prediction for Cloud Deep Learning." 12th USENIX Workshop on Hot Topics in Cloud Computing (HotCloud 20). 2020.

## CONTACT INFORMATION:


Aggelos Ferikoglou Ph.D, NTUA (aferikoglou@microlab.ntua.gr)

Dimosthenis Masouros Ph.D, NTUA (dmasouros@microlab.ntua.gr)

Assistant Professor Sotirios Xydis, HUA (sxydis@hua.gr)

Professor Dimitrios Soudris, NTUA (dsoudris@microlab.ntua.gr)