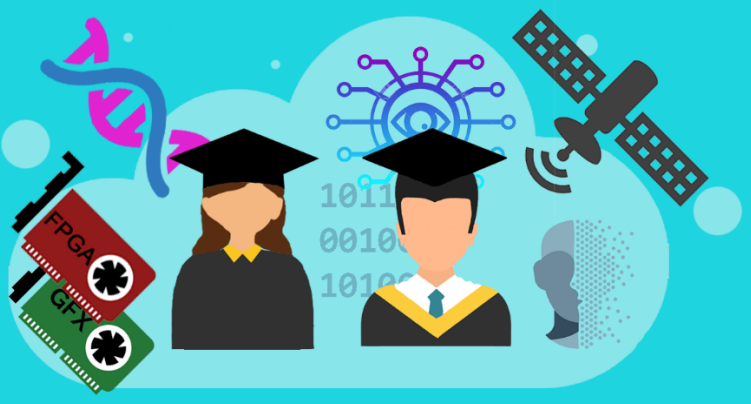
Diploma Thesis

Microproccessors and Digital Systems Laboratory



ML-based, interference-aware container orchestration on Kubernetes

Rather than hosting their applications on private infrastructures, organizations lately prefer to operate on service provider-managed resources on the Cloud. The established assumption of unlimited computing capacity offered by the Cloud Computing is both a fallacy and a pitfall for the users. Underneath the enhanced software stack provided, cloud resources comprise of fleets of distributed, heterogeneous physical machines.

Cloud service providers (CSPs), by exploiting virtualization and containerization technologies, employ resource multiplexing to enable economies of scale and to increase resource utilization in order to handle more and more client requests, decreasing their costs.

Nowadays, a great variety of applications gets scheduled and executed on public Cloud data-centers, from Machine Learning batch jobs and Deep Learning training and inference to web-servers and databases. Prior works [1],[2],[3] rely on either static or dynamic profiling of the target application to determine its sensitivity to interference. However, while the profiling window is kept short in time in order to keep the squandered time and resources low, it may not be representative enough for application's behaviour. Mars et. al explore the cumulative benefits of both interference and heterogeneity awareness [4]. Nonetheless, similarly to [5] and [1] authors focus on applications' pairwise interference when co-located on shared physical machines.

In clusters of 1000's of workloads, application-level monitoring, e.g., IPC, QPS, and exhaustive application co-location exploration are infeasible. In this work, instead of characterizing and evaluating the impact between different application combinations, we try to estimate the latency of any kind of deployed applications leveraging the system metrics. Could a map of applications' projection onto the low-level system metrics be adequate to predict latency and efficiently prioritize nodes for placement?

This diploma thesis will be a part that will hopefully conclude an ongoing work.The student will use various ML methods to predict execution time latency of various applications on heterogeneous machines in order to efficiently orchestrate them on a distributed Kubernetes cluster.

[1] Delimitrou, Christina, and Christos Kozyrakis. "Paragon: QoS-aware scheduling for heterogeneous datacenters." ACM SIGPLAN Notices 48.4 (2013): 77-88.

[2] Romero, Francisco, and Christina Delimitrou. "Mage: Online and interference-aware scheduling for multi-scale heterogeneous systems." Proceedings of the 27th International Conference on Parallel Architectures and Compilation Techniques. 2018.

[3] Yang, Hailong, et al. "Bubble-flux: Precise online qos management for increased utilization in warehouse scale computers." ACM SIGARCH Computer Architecture News 41.3 (2013): 607-618.

[4] Mars, Jason, and Lingjia Tang. "Whare-map: Heterogeneity in" homogeneous" warehouse-scale computers." Proceedings of the 40th Annual International Symposium on Computer Architecture. 2013.

[5] Mars, Jason, et al. "Bubble-up: Increasing utilization in modern warehouse scale computers via sensible co-locations." Proceedings of the 44th annual IEEE/ACM International Symposium on Microarchitecture. 2011.

Prerequisites:

- Linux, Bash/Shell scripting
- Experience on Python

Knowledge & Experience the student will acquire:

- A broader understanding of cloud computing architectures
- Apply Machine Learning methods to improve resource management decisions
- Work on technologies for automation and deployment
- Research, and become familiar with Kubernetes container orchestrator and its internals

Duration: At least 9 months

Contact:

Achilleas Tzenetopoulos Ph.D. student: (atzenetopoulos@microlab.ntua.gr)

Dimosthenis Masouros Ph.D. student: (dmasouros@microlab.ntua.gr)

Sotirios Xydis Ass. Prof.: (sxydis@microlab.ntua.gr)

Dimitrios Soudris Prof.: (dsoudris@microlab.ntua.gr)