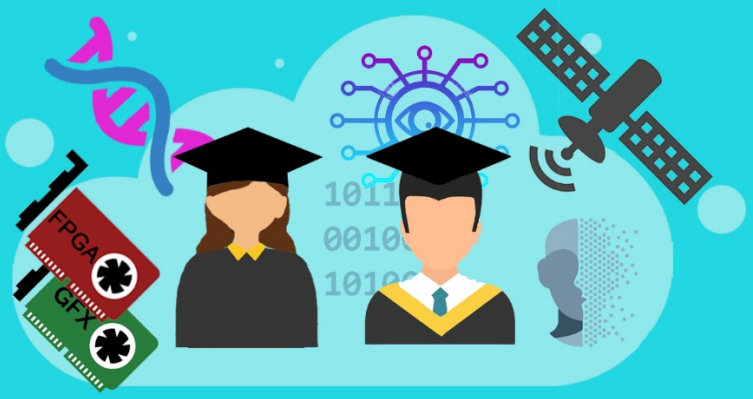
Diploma Thesis

Microproccessors and Digital Systems Laboratory



IMPROVING RESOURCE EFFICIENCY INSIDE DATA CENTER CLOUD ENVIRONMENTS

Nowadays, an increasing amount of computing is transferred to private and public data centers, such as Google Compute Engine, Windows Azure and Amazon's EC2. As a result, cloud providers have to deal with a huge amount of arriving workloads (up to millions) per second, which have to be scheduled and executed on server machines located inside the data center. These workloads can be divided in two main categories: Best Effort (BE) tasks, where the execution time is not a critical parameter of the application and Latency Critical (LC) tasks, where the execution time of the application must not exceed a certain deadline.

Several studies have established that the average server utilization in most data centers is low, ranging between 10% and 50%. For a hypothetical cluster of 10,000 servers, this idleness translates to a wasted capacity of 3,000 servers [1]. This underutilization can occur due to the unpredictability of the LC workloads. More specifically, to ensure that LC workloads meet their deadlines, cloud provides often dispose servers exclusively for them. However, LC workloads can present unpredictable spikes during the day, thus leaving those resources unexploited in periods of low traffic.

The goal of this diploma thesis is to explore the performance of various BE and LC workloads and dynamically improve the way that they are scheduled on the cloud servers, so that the resource efficiency inside the data center is optimized and the deadline of the LC workloads is not violated.

CONTACT INFORMATION: Dimosthenis Masouros demo.masouros@microlab.ntua.gr, Sotirios Xydis sxydis@microlab.ntua.gr, Prof. Dimitrios Soudris dsoudris@microlab.ntua.gr

[1] David Lo, Liqun Cheng, Rama Govindaraju, Parthasarathy Ranganathan, and Christos Kozyrakis. 2015. Heracles: improving resource efficiency at scale. SIGARCH Comput. Archit. News 43, 3 (June 2015), 450-462. DOI: https://doi.org/10.1145/2872887.2749475