

Master Thesis Proposal

Examiner: Prof. Vladimir Vlassov (vladv@kth.se)

Department of Computer Science, SCS division, KTH Kista

Academic supervisor: David Gureya (PhD student, daharewa@kth.se)

Efficient and QoS-Aware Co-location of Multiple Applications on Datacentre Nodes

In modern datacentres, workload consolidation is a widely used technique to improve resource utilization. This thesis focuses on a resource allocation problem of consolidating latency-critical applications with quality-of-service (QoS) requirements and throughput-oriented background (best-effort) applications that needs to achieve high performance. Current approaches to this problem consider the consolidation of applications in a single-socket system with a single slot for a multicore CPU. In contrast, multi-socket NUMA (Non-Uniform Memory Access) systems with multiple slots for CPUs constitute today an essential share of the nodes that comprise modern datacentres.

This thesis aims to explore machine learning-based multi-resource partitioning techniques that achieve two goals: (1) meeting the QoS requirements of all latency-critical applications, and (2) maximizing the performance of the best-effort applications.

Key Areas: online machine learning, NUMA systems, resource management, optimization