Ph.D. in Information Technology: Thesis Defense

July 20th, 2020

Online by Microsoft Teams – at 11.00

Anna PUPYKINA – XXXII Cycle

"Memory Management Techniques for Deeply Heterogeneous HPC Architectures"

Advisor: Prof. **Giovanni Agosta**

Abstract:

Resource allocation is a well-known problem, with a large number of research contributions towards efficient utilisation of the massive hardware parallelism using various exact and heuristic approaches. Nevertheless, memory management for heterogeneous systems is still a challenge, as these systems often feature separate physical spaces for the general-purpose part of the system and for the heterogeneous accelerators. Recent advances in memory allocation for homogeneous multicore architectures have focused on removing the need for application-specific allocators and improving scalability and allocation speed. Moreover, in many kinds of heterogeneous or accelerator-based systems, there is no dynamic memory allocation at all, with each accelerator endowed with its local memory. Thus, taking into account the difference between typical and deeply heterogeneous High Performance Computing (HPC) systems, memory management is becoming an essential part of resource management tools and requires novel algorithms to be developed.

This work explores the most recent state-of-art memory management techniques for HPC and Cloud Computing. A significant part of the memory optimisation research both in HPC and Clouds domain is focused on resource management techniques.
However, the degree of memory involvement in the provisioning and scheduling techniques is not very high, especially concerning the memory-processor interaction.

A problem to solve is the resource allocation optimisation in complex heterogeneous architectures taking into account memory utilisation.
Memory-centric Prediction-based Resource Allocation approach is addressed to solve this problem. The goal of the proposed approach is to manage the partitioning within a single heterogeneous node. Here, multiple accelerators coexist within a single node and can cooperate for a single application composed of multiple kernels, or they can be partitioned among different applications.
For choosing the optimal memory module in NUMA architecture regarding the multiple criteria, the multi-criteria memory analysis with fuzzy pairwise comparison is applied for overcoming the

subjectivity of criteria comparisons. Resources utilisation prediction is used to optimise the resource allocation in heterogeneous systems, taking into account the restrictions imposed by the runtime environment, such as the need for a fast short-term prediction with the minimal use of computational resources.

PhD Committee
Prof. **William Fornaciari**, DEIB
Prof. **Alessandro Cilardo**, Universita' di Napoli
Prof. **Dimitrios Soudris**, National Technical University of Athens