Ph.D. in Information Technology
Thesis Defense

September 18th, 2024
at 10:00 am
Aula Alpha – building 24

Bruno GUINDANI – XXXVI Cycle

Data-Driven Approaches for Managing Resources of Cloud, Edge, and High-Performance Computing Systems

Supervisor: Prof. **Danilo Ardagna**

Abstract:

The prevalence of complex software has surged in modern society, both in terms of strategic reliance and economic relevance. Simultaneously, the hardware required to support such software has grown in size, energy consumption, performance, and complexity. Notable examples of advanced architectures and paradigms for modern software are cloud computing, edge computing, and High-Performance Computing (HPC). A crucial aspect of executing complex software on machines with sophisticated hardware architectures involves selecting the appropriate configuration concerning software parameters and available hardware resources. Poor choices in these settings can result in noticeable performance reduction or substantial additional expenses. However, the software and hardware complexity poses challenges for fine-tuning these applications or predicting their performance based on input settings — tasks for which traditional white-box models are inadequate.

In this dissertation, we introduce several data-driven, black-box techniques for the performance modeling and the resource-constrained optimization of cloud computing, edge computing, and HPC systems. Black-box techniques are ideally suited to tackle these complex systems, as they do not require information about the internal workings of the system under study. The foundational elements of the proposed techniques are Bayesian Optimization (BO) and Machine Learning (ML). BO is one of the leading state-of-the-art techniques for black-box optimization of complex Information and Communication Technology (ICT) systems, and ML models can provide the required performance predictions with reliable accuracy. Both can infer the relationship between software/hardware input settings and output metrics of applications without needing any internal information.

In particular, this dissertation presents several tools to model application performance via ML, including (i) the open-source aMLLibrary for automatic parallel training and feature/hyperparameter selection, (ii) its integration in the OSCAR-P auto-profiling framework for Function-as-a-Service (FaaS)-based applications running in the edge computing continuum, and (iii) a framework for Graph Neural Network (GNN)-based thermal prediction in HPC centers. We also propose several algorithms integrating BO and ML models to optimize the execution of cloud and HPC applications while respecting resource or time constraints: they are the MALIBOO sequential algorithm and three parallel optimization extensions for HPC settings.

We validate our techniques in extensive experimental campaigns stretching several ICT domains, including but not limited to cloud-based big data applications, edge computing applications, neural networks, and molecular simulation software. Our trained ML models often achieve minimal error metrics, and our BO-based optimization techniques consistently outperform state-of-the-art solutions in finding optimal configurations and reducing the number of executions that do not respect constraints.

PhD Committee

Prof. **Gianluca Palermo, Politecnico di Milano**

Prof. **Ettore Lanzarone, Università degli Studi di Bergamo**

Prof. **Sotirios Xydis, National Technical University of Athens**