Ph.D. in Information Technology
Thesis Defense

October 1st, 2026
at 11:00 am
Room PT2 – building 20A

Roberto SALA – XXXVIII Cycle

Performance Modeling and Optimization of Edge-Cloud and HPC Systems

Supervisor: Prof. Danilo Ardagna

Abstract:

Modern computational infrastructures increasingly rely on a distributed computing continuum in which High-Performance Computing (HPC) systems, cloud platforms, specialized accelerators, and the widespread diffusion of computationally capable Internet of Things devices play complementary roles. This evolution supports complex and latency-sensitive workloads, including artificial intelligence applications, but it also introduces major challenges in resource management. System architects must balance energy efficiency, execution time, and operational costs while satisfying strict Quality of Service (QoS) constraints. Because evaluating configurations on real-world infrastructures is computationally expensive, manual tuning is infeasible, and traditional performance models often fail to capture the complex dynamics of distributed execution.

To address these limitations, this thesis develops complementary methods for profiling, modeling, and optimizing distributed workloads across the computing continuum. The contributions are structured around three main methodological areas. First, the dissertation introduces data-driven performance models for multi-component applications. It proposes an automated profiling framework that generates surrogate models for end-to-end execution times, a surrogate model for batched arrivals, and an energy-aware workload placement framework.

Second, the thesis focuses on the analytical characterization of caching systems operating under high request rates. It develops an exact Markovian model that formalizes the delayed hit phenomenon, bridging caching theory and queueing networks to capture the coupling between transient cache states and retrieval delays.

Finally, the thesis addresses the optimization of massive, discrete, and heavily constrained configuration spaces. It introduces Bayesian Optimization methods for discrete domains and parallel evaluation environments, including black-box constraints and batch selection under uncertainty. It also develops a fully asynchronous Artificial Bee Colony framework designed to scale across many concurrent workers. The thesis combines the two approaches by embedding the Bayesian Optimization routines into the asynchronous framework, yielding a scalable optimization architecture for evaluating and tuning complex HPC workloads under strict resource and QoS constraints.

PhD Committee

Prof. Monica Vitali, Politecnico di Milano

Prof. Michele Ciavotta, Università degli Studi Milano-Bicocca

Prof. Jonatha Anselmi, INRIA