Ph.D. in Information Technology
Thesis Defense

February 19th, 2024
at 17:00 pm
Sala Seminari Nicola Schiavoni

Federica FILIPPINI – XXXVI Cycle

Resource Allocation and Scheduling Problems in Computing Continua for Artificial Intelligence applications

Supervisor: Prof. **Danilo Ardagna**

Abstract:

Nowadays, Deep Learning (DL) and Artificial Intelligence (AI) are ubiquitous. They effectively apply to a wide range of products and solutions in largely heterogeneous fields, spanning from healthcare to, e.g., predictive maintenance and other industrial applications. At the same time, these years witness an accelerated migration towards distributed computing: Internet of Things devices as smart watches, smart city power grids, connected vehicles, smart homes generate huge volumes of data, whose processing requires mobility support, geo-distribution and, usually, effective strategies to reduce the operational latency. Privacy constraints, moreover, limit the possibility of analyzing data in the public Cloud, favouring the choice of on-premises resources or Edge devices with partial connectivity. Large Cloud datacenters, featuring powerful Virtual Machines often accelerated with GPUs, are exploited to process resource-demanding tasks, which benefit from the possibility of accessing ideally unlimited computational and storage resources according to pay-to-go pricing models. The effective management of such a complex environment, including heterogeneous Edge-to-Cloud resources interconnected in a so-called Computing Continuum, poses great challenges. The execution of AI inference workflows including diverse components needs to be optimized at design-time, selecting the most appropriate resources from the available computational layers, and then adapted at runtime, since fluctuating workload and environment conditions may determine sub-optimal scenarios where resources are saturated or under-utilized. Similarly, DL training applications executed in private or public Cloud clusters are to be effectively scheduled on suitable GPU-accelerated nodes to minimize the expected costs and meet the user-imposed due dates.
This dissertation addresses these three problems from different perspectives, proposing mathematical models and heuristic or Reinforcement Learning-based methods for the efficient and effective management of AI inference applications, both at design-time and runtime, and DL training jobs. Furthermore, it discusses techniques to develop analytical and Machine Learning-based performance models, crucial to accurately predict the applications response times on heterogeneous resources. The proposed tools are validated through experimental campaigns in both simulated and prototype environments, proving their effectiveness and applicability to practical scenarios.

PhD Committee

Prof. **Marco Gribaudo,** Politecnico di Milano

Prof.ssa **Claudia Canali,** Università di Modena e Reggio Emilia

Prof. **Diwakar Krishnamurthy,** University of Calgary