Research topic: Reinforcement Learning Across Computing Continua
Proponent: Danilo Ardagna, danilo.ardagna@polimi.it, Politecnico di Milano, Italy

Motivation

The worldwide public cloud market continues to grow and is expected to reach about 1,700 USD billions in 2029 [FortuneBusiness2023]. However, there is a general consensus that the most innovative applications will benefit from using computing resources at the periphery of the network where source data is produced. Many companies are evaluating the use of edge computing for data collection, processing and online analysis to reduce latency and data transfers. IoT devices will be more than 30 billions by 2025 and a growing number of use cases can benefit from applications spanning the cloud-edge continuum.

Given the inherently **distributed nature** of computing continua, coupled with an ever-increasing **complexity** in terms of **heterogeneity** and **dynamicity** of the **components** and the **limited capacity** of the edge layer, **guaranteeing the performance** of **mission-critical applications** is extremely **challenging**.

In this context, **Reinforcement Learning** (RL) techniques have recently attracted increasing attention given the possibility of **autonomously learning** the **optimal behavior** with little or even **no prior knowledge** of the **system** dynamics. By taking advantage of the possibility of distributed and/or hierarchical learning solutions, RL techniques have emerged as the most promising approach to cope with scenarios including dispersed, loosely coupled components which require fast, possibly local, decisions to provide consistent performance in face of uncertain and rapidly evolving environments.

Unfortunately, **current cloud-based/edge systems** implement **basic policies** for resource management, which trigger adaptation actions according to some threshold violations **without any performance guarantees** neither **providing costs** nor **energy consumption minimization**. One important **research direction** Europe should take is the **development** of **novel runtime management solutions based on RL,** enabling the execution of applications across various levels of the computing continuum, **trading-off application performance** (end-to-end latency and throughput) and **energy consumption** while **guaranteeing a reliable execution even under edge-cloud network partitioning.**

Current Status

Computing continuum is expected to solve the main issues of cloud computing in the support of the most innovative and modern applications by exploiting computing resources at the edge of the network. However, the deployment, execution and management of applications in the computing continuum is very challenging due to its inherently distributed nature and ever-increasing heterogeneity.

DEPLOYMENT AND ADAPTATION OF THE CONTINUUM INFRASTRUCTURE

The automatic deployment of applications in computing continua raises several challenges about monitoring and management of highly heterogeneous and distributed systems. Management and monitoring mechanisms should consider not only traditional resource utilization, but also network delays and communication overhead among system components. Moreover, orchestration strategies have to take into account the specific characteristics of the applications to be deployed or the deployment and the execution of applications in computing continuum scenarios, exploiting serverless computing and the Function as a Service (FaaS) paradigm. However, as surveyed in [Li2021], the existing challenges still wait for more advanced research and solutions to further explore the potentials of such computing model.

OFFLINE POLICY LEARNING

Due to the heterogeneity of the infrastructure and the high volatility of the workloads, modelling and optimising the performance of applications deployed in computing continua is a complex task: in literature, some approaches have been proposed on this topic, which tackle the problem from different perspectives. Some solutions rely on simulation-based analytical models, while other studies demonstrate the benefits of machine learning based black box approaches (e.g., not requiring knowledge about the system dynamics). The research community has also recently dedicated significant effort to component placement exploration in computing continua by relying on Mixed-Integer Linear Programming or proposing heuristic solutions based on, e.g., simulated annealing or tabu search.

ONLINE POLICY LEARNING

Distributed applications usually cope with varying working conditions regarding both the supporting infrastructure and the workloads. To keep consistent service levels, applications need adaptation abilities to dynamically change their configuration based on observed or expected conditions. The issue of managing distributed applications at runtime has been widely studied in the context of cloud computing. The same challenges along with additional ones (e.g., mobility, constrained energy and connectivity) must be faced in the computing continuum. As surveyed in [Liu2022], a variety of methodologies have been adopted to drive runtime adaptation, including heuristic methods, control theory and recently Machine Learning (ML). A key challenge towards effective online adaptation is the uncertainty that affects infrastructure conditions, performance models and workloads in practical scenarios [Kab2021].

ML techniques promise to automatically identify good control decisions [Sed2021,Fil2021]. Within the broad area of ML, RL well suits run-time application management and has been successfully applied to tasks adaptation and especially auto-scaling. RL has demonstrated the capability to solve problems of deployment, resource planning and load balancing in distributed computing environments. The main advantage consists in the fact that the agents learn to optimise the behavioural policy maximizing the long-term reward without having prior knowledge of the system but only by interacting with it. However, there are still several challenges in computing continua environments. In particular**, RL agents usually require a large number of interactions** with the environment to collect sufficient experience. This time-consuming process negatively affects the user experience in a complex environment [Gou2021]. For this reason, sim-to-real approaches are often used: agents are pre-trained in simulation environments and the resulting policies are then refined within the real deployment, reducing convergence times.

Research Challenges

EU research should investigate novel solutions based on RL enabling the execution of applications in computing continua with performance guarantees and energy consumption awareness.

In this context, nowadays we cannot rely on traditional applications deployment and runtime resource provisioning and adaptation solutions (often designed for centralized, mostly homogeneous and static environments) which are not adequate to cope with a diffused environment of dispersed, loosely coupled components which require fast, possibly local, decisions to provide consistent performance in face of an uncertain and rapidly evolving environment. While distributed solutions have been considered in the literature, they typically suffer from a slow-convergent learning process which negatively affects performance.

EU research should develop novel methods and approaches to implement smart and reactive deployment and runtime management of applications involving multiple components from the edge to the cloud. The frameworks should optimise the use of the underlying resources trading-off application performance (in terms of end-to-end latency or throughput) with energy consumption, while guaranteeing a reliable execution of the applications even under edge and cloud network partitioning.

Overall, we identify the following challenges:

C1 Resources Heterogeneity: The computing continuum involves heterogeneous environments ranging from large cloud data centres to resource-constrained edge nodes, which frequently become the system bottleneck given their limited capacity. This poses a challenge from the applications perspective, which might compete for the use of the underlying resources with different patterns and from their deployment and operation due to different hardware types (e.g., ARM in the edge, traditional x86 CPUs in the cloud), and different technology stacks (e.g., K3s, K8s).
C2 Highly Distributed Systems: The computing continuum scenario is characterised by a massively distributed nature , with many edge nodes and possibly multiple cloud providers, each interconnected with non-negligible network delays. The size of such infrastructure is in itself a challenge due to scalability/connectivity issues. Furthermore, the need to communicate application data and state information (e.g., system load) and to make decisions based on them may lead to inaccurate decisions due to stale information and may even lead to instability due to herding effect.
C3 Performance Guarantees in Highly Dynamic environments: applications spanning the various layers of the cloud-to-edge continuum are built to cope with strict latency constraints in executing certain functionalities (e.g., processing data from sensors in a self-driving car). Unfortunately, phenomena such as network unreliability (extremely frequent in the next generation 5G mmWave), device mobility, and limited computing capacity of nodes at the edge layer can have detrimental effects on latency. Being able to guarantee satisfactory QoS levels in these circumstances is, therefore, a considerable challenge.

This research will help to improve the operation of industrial applications by reducing costs and avoiding congestion and hazards. The development of open-source tools for edge-to-cloud resource management is advocated.

IMPACT
Although the phenomenon of edge computing is relatively recent, the IDC Europe Emerging Technologies Survey [IDC2022] states that a staggering **26% of the European organisations have already adopted an edge solution** of some kind. Moreover, the adoption trend suggests that **this number is expected to double over the next two years**. Another perspective that allows us to better understand the impact of edge computing, and its relationship with the cloud, on European organisations comes from the proportion of spending on edge computing out of total infrastructure expenditure, which reached 18% in 2021 (IDC's European Infrastructure Survey [IDC2022b]), with a strong upward trend compared to previous years driven by the increasing need for process automation and optimisation, improving customers/users experience or for automated threat-intelligence monitoring and prevention. **Cloud-edge continuum bridges the gap between emerging platforms** (like IoT, AI, streaming analytics, etc.) **and more traditional technology** (such as cloud), with the aim of **enabling new architectures** that exploit multiple connectivity solutions (e.g., 4G and 5G, but also LoRa or Bluetooth LE), and context-aware distributed real-time analytics, that minimise latency and **at the same time ensure privacy and trust**.
The continuity between cloud and edge also **enables other emerging scenarios**. For example, computation at the edge allows **robots** to process data locally, reducing time-to-action and ensuring that processes can remain active and secure even in the (temporarily) absence of the network. The critical role of low latency and high bandwidth in retrieving real-time data makes edge the perfect candidate for **AR/VR deployments**. **5G** heavily relies on distributed computing and storage capabilities, with edge being a critical component in both **public telecom networks and private network** deployments. Moreover, next generation mmWave 5G will provide **large bandwidth** but will **suffer performance variability.**
Also, **AI inferencing** is increasingly performed at the edge, reducing latency, therefore edge AI has tremendous capabilities to make predictions and recommendations or to process very complex and large data, enabling the possibility for edge devices to make real-time decisions. It is apparent how the **cloud-edge computing continuum** is an **incubator and accelerator for other forefront technologies**, from AI, to 5G, to robotics. Consequently, supporting the study in the field of cloud-edge computing fostering the development of resource management solutions able to provide QoS guarantees, means, in turn, supporting the process of technological modernization of organisations through the adoption of emerging technologies.

ECONOMIC IMPACTS
The edge market represents an exceptional economic opportunity for European enterprises. The **overall expenditure** of European enterprises in **edge-related solutions** is expected to ramp up from the €24 billion predicted for 2023 to €40 billion in 2025. Such growth is foreseen as a **steady trend** over the next few years, with a compound annual growth rate (CAGR) **between 2022-2025 of roughly 18%**. These numbers clearly show how this technology has already gained momentum, and that the hype generated will last for the next few years [IDC2022c].
Understandably, **the main driver of this growth lies in the opportunities generated by the realization of low latency use cases**, and deployments in locations that are either remote or without IT support staff made it possible by the effort of many new vendor propositions and a supply chain that is being built up.
Thus, the proposed research foster the development of solutions for an **extremely promising market**, which is growing fast, and necessitates advanced approaches optimising the runtime management of applications. In such a context, the availability of the right toolbox is pivotal to boost economic growth in a very dynamic and competitive environment.

TECHNOLOGICAL INNOVATIONS AND IMPACTS ON INDUSTRIAL APPLICATIONS
The data deluge, produced among others by mobile devices, IoT and cars and eventually routed to the cloud, poses serious **challenges to network infrastructures** that are required to provide a **high standard** of service in terms of **availability, efficiency and bandwidth**. In this context, **moving some computation toward the edge** of the network can **save core data centres from overcrowding** and the **network infrastructure from congestion**.
Organisations are, therefore, compelled to adopt **novel approaches, new operating models, and solutions for data governance and runtime management** of applications to ensure business continuity and to avoid possible service disruptions. As the progressive adoption of computing continua, the number of data processing services at the edge will increase; consequently, the **amount of data to be managed in the continuum will increase** exponentially. New data management approaches will need to be developed to ensure performance, reliability, and trustworthiness.
This will have a **deep impact on several sectors**, starting from **manufacturing**, where automation and intelligence at the edge can enhance monitor and maintenance systems by predicting bottlenecks and failures, or in **healthcare**, where, e.g., **edge AI can provide real-time patient data analysis** and enable instant response in life-critical systems.
Concrete solutions to cope with workload fluctuations, device mobility, edge-cloud network partitioning, limited battery capacity, implementing intelligent solutions to fulfill latency constraints, is one the main motivation fostering edge computing adoption.

References

[Fil2021] F. Filippini, D. Ardagna, M. Lattuada et al. ANDREAS: Artificial intelligence traiNing scheDuler foR accElerAted resource clusterS. FiCloud 2021: 388-393, 2021

[FortuneBusiness2023] https://www.fortunebusinessinsights.com/cloud-computing-market-102697

[Gou2021] M. Goudarzi, M. Palaniswami and R. Buyya, A Distributed Deep Reinforcement Learning Technique for Application Placement in Edge and Fog Computing Environments in IEEE Transactions on Mobile Computing, vol. , no. 01, pp. 1-1, 5555, 2021

[IDC2022] https://www.idc.com/eu/research/key-trends/emerging-technologies

[IDC2022b] https://www.idc.com/getdoc.jsp?containerId=EUR148267921

[IDC2022c] https://www.idc.com/getdoc.jsp?containerId=prEUR148783922

[IEA2023] https://www.iea.org/reports/data-centres-and-data-transmission-networks

[Liu2022] J. Liu et al. RL/DRL Meets Vehicular Task Offloading Using Edge and Vehicular Cloudlet: A Survey. IEEE Internet of Things Journal, vol. 9, no. 11, pp. 8315-8338, 2022

[Kab2021] H. M. D. Kabir, A. Khosravi, S. K. Mondal, M. Rahman, S. Nahavandi, R. Buyya. Uncertainty-aware Decisions in Cloud Computing: Foundations and Future Directions. ACM Comput. Surv. 54(4): 74:1-74:30, 2021

[Lin2021] C. Lin, H. Khazaei. Modeling and Optimization of Performance and Cost of Serverless Applications. IEEE Transactions on Parallel and Distributed Systems, 2021.

[Sed2021] H. Sedghani, F. Filippini, D. Ardagna. A Random Greedy based Design Time Tool for AI Applications Component Placement and Resource Selection in Computing Continua. EDGE'21, 32-40, 2021