Osmotic Computing Platform for Smart City Applications in the Cloud-Edge Continuum

Floriano Scioscia*, Giuseppe Loseto†, Davide Loconte*, Corrado Fasciano*, Filippo Gramegna*, Saverio Ieva*, Agnese Pinto*, Michele Ruta*
* Politecnico di Bari – Via E. Orabona 4, Bari (I-70125), Italy
{nome.cognome}@poliba.it
† Università LUM "Giuseppe Degennaro" – Casamassima BA (I-70010), Italy
loseto@lum.it

***Abstract*—The Cloud-Edge Intelligence (CEI) paradigm advocates decentralized data preprocessing, model training, and inference on devices across the edge of the network and in the cloud. This paper proposes a containerized Cloud-Edge microservice architecture that enables model training and prediction on both Edge and Cloud nodes, allowing for flexibility and dynamic adaptation to diverse requirements and resource availability. The framework incorporates an automatic task migration mechanism, leveraging opportunistic resource management and workload distribution between the Edge and the Cloud.**

***Index Terms*—Cloud-Edge Intelligence, Machine Learning, Microservice Architecture, Osmotic Computing, Cloud-Edge continuum**

I. Introduction

Smart city services and applications increasingly require technological platforms capable of integrating sensing, data storage, processing and analysis at large scales. Artificial Intelligence (AI) applications powered by wide Internet of Things (IoT) deployments allow training Machine Learning (ML) models on large data streams collections as well as providing inference (*i.e.*, prediction) capabilities with high accuracy. In early approaches, however, data transfer from IoT devices in the field to cloud computing infrastructures was needed for model training and inference. More recently, the adoption of the Edge Computing (EC) paradigm [1] has been exploiting the growing number of powerful processing devices available at the edge of local networks. This shift can enable improved efficiency and reduced latency by processing data closer to their sources. The core objective of EC is to relocate computing and communication resources from the cloud to the edge of networks in order to deliver services and perform quick (pre-)processing, protecting data privacy better and allowing faster responses for end users.

In this context, the *Cloud-Edge Intelligence* (CEI) research area aims to distribute ML model training and prediction tasks to edge or cloud nodes dynamically, according to application requirements, device state, and network conditions. Designing collaborative distributed platforms spanning the Cloud-Edge continuum requires dealing with the complexity of dynamic resource and service orchestration across edge and cloud layers. Additionally, run-time ML model training and prediction should be carried out at the edge when other application requirements become important as much as prediction accuracy, such as reducing response latency or maintaining data locality due to privacy concerns. As a consequence of these challenges, solutions with sufficient generality, maturity and efficiency are still missing.

This paper proposes a CEI platform based on the *Osmotic Computing* (OC) [2] paradigm. The proposal allows for the dynamic orchestration and provisioning of containerized microservices that implement the various logical components of a complete architecture in the Cloud-Edge continuum for the collection, preprocessing, storage, training, and analysis of data. Automatic deployment and migration are carried out in a flexible and dynamic way, taking into account the application workload needs, device status changes, and network topology modifications caused by *e.g.*, device mobility, energy supply interruptions, or network link congestion. The same ML tasks can be deployed either to edge nodes or to cloud infrastructure opportunistically [3], by taking advantage of resources currently available close to the job to be accomplished. The proposed osmotic framework is described in the next section, before conclusion.

II. Proposed Framework

Conventional architectures imposed limitations on the processing capabilities of edge nodes, restricting their role to basic data gathering from IoT sensing nodes. Edge node processing was primarily focused on tasks such as data stream encryption, transcoding, data stream combination, preprocessing, and summarization, meanwhile computationally demanding operations, such as ML model training and inference, were offloaded to the Cloud infrastructure.

This work proposes a microservice architecture illustrated in Figure 1. The proposal leverages the increasing capabilities of edge devices to perform ML tasks, encompassing both the utilization of pre-trained models and the execution of complete processing pipelines involving data preparation, feature extraction, training, and prediction. The architecture consists of distinct microservices deployed across heterogeneous devices, intelligently allocating different types of tasks across the cloud and edge layers based on their specific requirements and node capabilities. Microservices are packaged in containers and can be dynamically deployed to different devices. Orchestration and deployment strategy align with Osmotic Computing

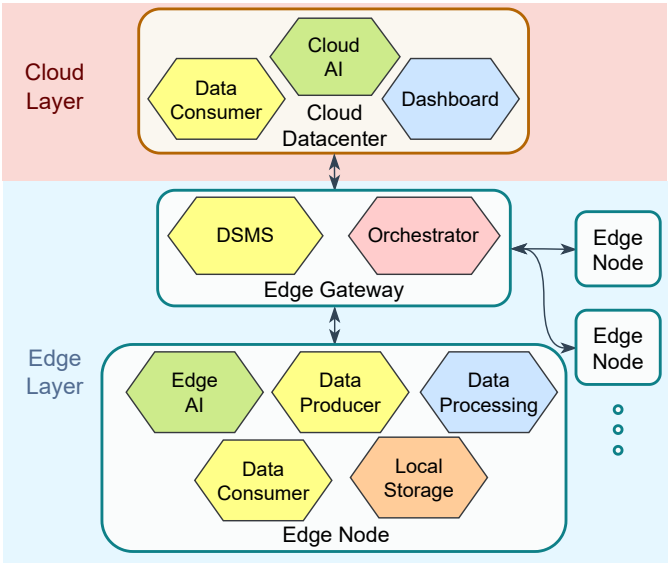

Fig. 1. Proposed architecture

principles [2], which considers the dynamic requirements of infrastructures and applications. Container provisioning adapts the virtual environment to the destination hardware, accommodating the high heterogeneity of physical resources. Dynamic administration of virtual network resources is also included for efficient service migration in the Cloud-Edge continuum, mitigating application failures and QoS degradation.

The framework architecture is composed of lightweight and loosely coupled microservices, serving as logical components with the following distinct responsibilities:

Data Stream Management System (DSMS): serves as the platform's Message Broker (MB) and manages asynchronous, event-driven communication.

Orchestrator: manages all the microservices and has the capability to migrate them in real-time between the edge and the cloud, based on resource availability and task requirements.

Data Producer and **Data Consumer:** they respectively send and receive preprocessed data through the MB.

Local Storage: provides temporary local storage for data collected from field sensors and IoT devices. It is strategically located close to data processing microservices to optimize latency and bandwidth while avoiding centralized data storage.

Data Processing: responsible for preprocessing data in preparation for ML model training tasks.

Edge Intelligence: it performs model training, validation and prediction for ML classification and regression tasks on edge devices, taking data from Local Storage.

Cloud Intelligence: this counterpart to Edge Intelligence microservice operates in the cloud, performs model training and prediction using data streams collected from various Data Producer instances.

Data Analytics: performs business intelligence analytics and visualization on the data collected in the cloud layer.

The proposed approach supports various composition patterns and fine-grained scalability to meet the specific requirements and constraints of applications. In particular, it facilitates continuous ML model improvement through a feedback loop, allowing less accurate models to be trained and used on edge devices while a more accurate model is trained in the cloud using larger amounts of data. The updated models can be sent to Edge Intelligence nodes as new data is gathered.

III. Conclusion and Perspectives

This paper has presented a novel platform for ML applications in the Cloud-Edge continuum, utilizing a microservices-based approach inspired by OC principles. Early experiments on a prototypical testbed [4] demonstrated the proposed framework can enable dynamic orchestration and deployment of microservices, granting flexibility, scalability, and opportunistic resource utilization.

The proposed platform is basically general-purpose and application-agnostic. Smart city applications include, but are not limited to: smart mobility and logistics, environment monitoring and waste management, smart buildings and infrastructures management, crowd management and urban safety. All these applications combine large-scale data processing and adoption of predictive models with requirements of data locality and low-latency responses.

Future work involves expanding the implementation of the prototype into a complete testbed comprising real-world sensor networks and IoT field devices. This testbed will serve as a platform to comprehensively assess the performance and scalability aspects of the proposed framework in realistic scenarios. Additionally, work will be devoted to expand the framework by incorporating semantic-enhanced ML approaches [5] and an orchestration service based on semantic matchmaking techniques [6], in order to perform a more flexible and explainable optimization of task allocation w.r.t. their specific requirements and context information.

Acknowledgments

This work has been partly supported by project BARIUM5G (Blockchain and ARtificial Intelligence for Ubiquitous coMputing via 5G), funded by the Italian Ministry of Economic Development.

References

[1] N. Abbas, Y. Zhang, A. Taherkordi, and T. Skeie, "Mobile edge computing: A survey," *IEEE Internet of Things Journal*, vol. 5, no. 1, pp. 450–465, 2017.

[2] M. Villari, M. Fazio, S. Dustdar, O. Rana, and R. Ranjan, "Osmotic computing: A new paradigm for edge/cloud integration," *IEEE Cloud Computing*, vol. 3, no. 6, pp. 76–83, 2016.

[3] W. Li, X. You, Y. Jiang, J. Yang, and L. Hu, "Opportunistic computing offloading in edge clouds," *Journal of Parallel and Distributed Computing*, vol. 123, pp. 69–76, 2019.

[4] G. Loseto, F. Scioscia, M. Ruta, F. Gramegna, S. Ieva, C. Fasciano, I. Bilenchi, and D. Loconte, "Osmotic Cloud-Edge Intelligence for IoT-Based Cyber-Physical Systems," *Sensors*, vol. 22, no. 6, p. 2166, 2022.

[5] M. Ruta, F. Scioscia, G. Loseto, A. Pinto, and E. Di Sciascio, "Machine learning in the Internet of Things: A semantic-enhanced approach," *Semantic Web*, vol. 10, no. 1, pp. 183–204, 2019.

[6] M. Ruta, F. Scioscia, I. Bilenchi, F. Gramegna, G. Loseto, S. Ieva, and A. Pinto, "A multiplatform reasoning engine for the Semantic Web of Everything," *Journal of Web Semantics*, vol. 73, p. 100709, 2022.