Management of Safety-Critical AI Services in the Compute Continuum

Lorenzo Colombi* , Mauro Tortonesi* *University of Ferrara, Ferrara, Italy
Email: {lorenzo.colombi, mauro.tortonesi}@unife.it

***Abstract*—The increasing complexity and criticality of AI-driven services across the compute continuum, spanning from edge devices to cloud datacenters, necessitates resilient, intelligent, and explainable management strategies. This work addresses the challenges of deploying and orchestrating safety-critical AI services in dynamic and resource-constrained environments, such as Industry 5.0 and Human Assistance and Disaster Recovery (HADR). We present a suite of complementary solutions, including a cloud-native MLOps platform tailored for SMEs, a semi-supervised federated learning framework (FedEdge-Learn), and novel semantic communication mechanisms that optimize data transmission using LLM-driven embeddings. Furthermore, we introduce an intent-based Zero-touch Service Management (ZSM) architecture, leveraging neuro-symbolic AI and collaborative intelligence to automate orchestration, model fine-tuning, and policy reasoning across federated Kubernetes clusters. These efforts pave the way for trustworthy, adaptive, and efficient AI service lifecycle management in environments characterized by disconnections, privacy constraints, and operational unpredictability. Future work focuses on extending the neuro-symbolic approach to support additional tasks, including dynamic node selection, optimized placement across the compute continuum with the goal of improving resilience and interpretability in distributed, resource-constrained environments like Industry 5.0 and HADR, while addressing challenges such as intermittent connectivity and evolving operational conditions.**

***Index Terms*—Zero-touch Service Management, MLOps, Industry 5.0,**

I. Introduction

The compute continuum concept envisions a unified, API-driven infrastructure that integrates compute, networking, and storage resources across all layers, from embedded and edge devices to cloud data centers [1]. Artificial Intelligence (AI) is increasingly intertwined with the compute continuum stack: advanced techniques such as deep reinforcement learning and continuum digital twin promise agile, large-scale resource optimisation across heterogeneous edge-to-cloud environments [2]. Machine Learning (ML) models, used to classify or predict resource usage patterns, also enable intelligent service management.

Moreover, thanks to the advancements in edge computing and related software, it is becoming possible to manage resources even in safety-critical environments, beyond cloud datacenters, such as Industry 5.0 and Humman Assistance and Disaster Recovery (HADR). These environments are those where system failures can lead to serious consequences, requiring high reliability and strict operational guarantees. In Industry 5.0, the widespread adoption of AI and ML is significantly reshaping manufacturing and production systems, operational efficiency, and optimized resource management [3]. [4]. On the other hand, the availability of edge-specific Kubernetes frameworks, such as KubeEdge, makes it possible to seamlessly manage services without taking care of the underlying hardware/software stack, even in HADR environments.

However, these environments present different and additional challenges, including ensuring consistent performance and robustness to changing conditions. Moreover, non-functional requirements like privacy, confidentiality, fairness, and explainability become especially crucial, as failures or biases in these areas can lead to severe consequences. Moreover, both in Industry 5.0 and HADR scenarios, nodes could possibly be disconnected from the network, and bandwidth could be limited [5]. Therefore, there is a need for stronger resilience to transient network partitions through dynamic task or even whole-cluster relocation. Lastly, these environments are typically made of many interconnected Kubernetes clusters, possibly run by different tenants, each one with specific policies in terms of user privilege and data privacy. Orchestration solutions for such federated setups remain immature, still struggling with automated placement, autoscaling, and "ping-pong" migrations [6].

Within this landscape, AI and ML technologies play a dual role. On one hand, they empower intelligent service management through diverse techniques, like for instance reinforcement learning and digital twins [2]. ML models can predict resource usage patterns to proactively manage service placement and reduce downtime [6]. Also, neuro-symbolic AI could be used to support intelligent, context-aware node selection, integrating machine learning, symbolic logic, and reinforcement learning to handle real-world complexity and continuously improve performance over time. On the other hand, the compute continuum supports AI growth by accommodating increasingly demanding workloads, including generative models [7].

While promising, these solutions remain in their early stages, and determining the optimal placement of services across heterogeneous, distributed, and potentially multi-cluster infrastructures continues to be a significant and unresolved challenge. For this reason, my research focused on addressing the challenges related to the AI service management and developing solutions, using many different but complementary

978-3-903176-75-1

approaches, specifically suited for safety-critical environments.

II. MLOps

Managing AI services in Industry 5.0 environments poses additional challenges, not typically found in standard ML deployments. These include the need to process real-time sensor data with strict latency requirements, integrate with legacy and resource-constrained devices [8].

To meet these requirements, we designed, in collaboration with Bonfiglioli, an industrial partner of the University of Ferrara, an MLOps platform to ease the adoption of MLOps techniques and management of ML-based services, including generative AI models, in particular for Small and Medium Enterprises (SMEs) [4]. The platform was then implemented using only open-source software in two different variations, following a building blocks approach and an all-in-one approach. Using the second implementation, the possibility of managing generative AI workload was also experimented [7].

Subsequently, to overcome the shortcomings of this initial design, starting from the absence of cloud-native technologies, limiting the adoption in a Compute Continuum scenario, we developed an improvement of the original platform using only cloud-native and container-based software, such as Kubeflow, in [9]. This platform was also designed to be deployed in a multi-Kubernetes cluster environment, which is typical of industrial environments.

Moreover, we tested our platform by creating an automated pipeline to automate the lifecycle management of the T2V-based Auto Encoder, which is another result of our industrial collaboration [10], [11]. This ML model has been developed to extract vector representation from multivariate time series and ease the anomaly detection in industrial machinery. To this end, we also explored the use of generative AI in Industry 5.0 and the related management challenges to the use of Generative AI. Specifically, in [12], a comparison between many state-of-the-art generative models is presented.

Lastly, neural networks often operate as opaque 'black boxes,' making it challenging to interpret or understand the reasoning behind their decisions, and AI explainability could be a requirement in a safety-critical environment. For this reason, we explored the use of Explainable Artificial Intelligence (XAI) in Industry 5.0, also in combination with Causal AI to determine the most significant feature. When applied in high-dimensional datasets, these techniques permit to increase the model performance, giving insight to the operators and improving the performance, since the limited number of features selected [13]

III. Federated Learning

An additional issue typical of HADR and Industry 5.0 settings is the difficulty of centralizing the collected data for privacy-related reasons or network issues. Moreover, the collected data are usually unlabelled [14]. At the same time, when AI models are deployed on the field for the first time, they lack enough data for effective training, and they have to wait for the collection of an appropriate amount of data. Therefore, the need for an effective distributed/federated training arises.

To enable model training in these distributed and constrained scenarios, we explored the use of Federated Learning (FL) to solve challenges related to the impossibility of collecting data in a centralized solution. In detail, we developed FedEdge-Learn, a semi-supervised FL framework, which is illustrated in Fig. 1, where clients train their models using only unsupervised techniques, while the central server shares a small amount of high-quality data to help the clients recognize the classes and decrease the time need to collect enough data to perform the first training. This solution improves model performance and reduces the time-to-market, leveraging edge training and preserving users' privacy [14].

Later, we also expanded our work in FLs and edge AI, experimenting with their intersection with Large Language Models (LLMs). Running LLMs at the edge offers several advantages, including enhanced privacy, reduced latency, and independence from constant cloud connectivity. These properties are especially important in federated settings, where data is inherently distributed and often sensitive.

In this context, we experimented with developing a simulated edge and distributed infrastructure for federated Parameter-Efficient Fine-Tuning (PEFT) using "small" LLMs. In the same work, we also compared a new token-based parameters FL aggregation, comparing it with the classical Federated Averaging [15]. This new aggregation method has been proposed to overcome the limitation of the classical Federated Averaging, which may not be well-suited for LLMs, as it fails to account for the differing informational value of training examples, since longer sequences could encapsulate more complex and informative structures, providing richer learning signals.

IV. Semantic communication

In demanding contexts such as HADR, the combination of intermittent connectivity, restricted bandwidth, and a highly dynamic environment necessitates the development of efficient communication optimization solutions.

In this context, it is possible to use modern ML embedding models and LLMs to extract semantic information from the collected data, even if they belong to complex data types like images and text. Subsequently, it is possible to use this semantic understanding, for example, by comparing semantic similarity, measuring the distance in vector space between two embeddings, to determine the relevance of the data and to choose which information has to be transmitted, reducing the network resource usage.

In collaboration with the Florida Institute of Human and Machine Cognition (IHMC), we worked on the communication optimization in HADR scenarios. We first developed a system that leverages an embedding model to filter out semantic redundancies in complex data types (e.g., image), reducing bandwidth usage [16]. Specifically, the designed system computes embeddings for all the collected images and computes the cosine similarity between the new embedding

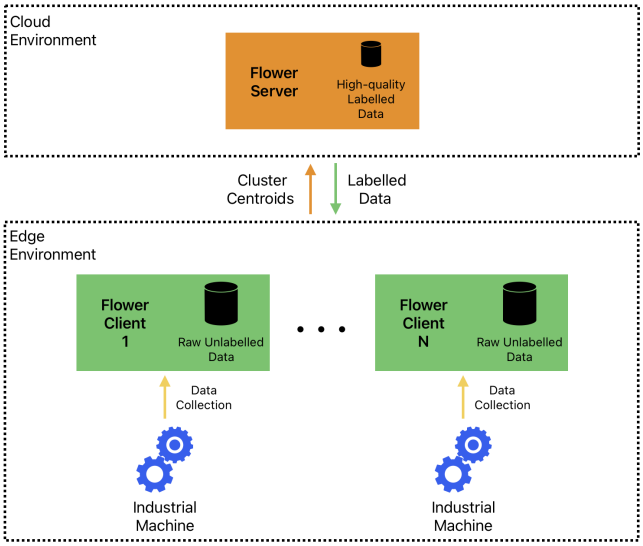

Fig. 1: Diagram of the Semi-Supervised Federated Learning Framework

and the others previously collected in a certain period. If the similarity exceeds a set threshold, the data is marked as redundant and not propagated, resulting in more efficient communication resource usage [16].

This work was subsequently extended toward proposing a semantic-aware publish/subscribe system. This system is designed to overcome the fixed metadata-based content-topic matching usually used in pub/sub systems. In [5], we propose the use of embedding models to automatically determine whether specific data is relevant to a topic, resulting in a more flexible and general topic-matching process. Moreover, since reaching this goal requires dealing with multiple modalities, specifically natural language and images, we compared two different implementations: one where a multi-model embedding model is used and a second where a two-step approach, which includes an image captioning model and a text embedding model, is used.

V. Neuro-Symbolic AI for Zero Touch Service Management

We are also actively working in collaboration with the University of Bologna on developing a Zero-touch network and Service Management (ZSM) solution tailored for Industry 5.0. ZSM offers a framework for fully automated management of networks and services, designed to operate independently of specific vendors. Its architecture supports flexible, vendor-neutral solutions, aligning with the industry's shift away from conventional, inflexible management approaches [17].

To address the challenges of orchestration in industrial environments, we are working on a novel framework that enhances Kubernetes (K8s) with Collaborative Intelligence (CI) principles. It introduces two main innovations: a CI-driven stateful service migration mechanism that optimizes application placement across the edge-to-cloud continuum with minimal downtime, and the Zoom-In functionality, which allows human operators to trigger dynamic model upgrades via a LLM-powered intent processor. Together, these features enable adaptive, efficient, and human-aware management of AI-driven services in complex, resource-constrained settings [18], [19].

In these works, intent-driven management plays a central role in automating service operations within the ZSM framework. By allowing high-level, outcome-focused expressions of management goals, it enables systems to autonomously determine and adjust optimal actions in real time [20]. We also introduced a LLM-based intent translation component.

We also continued the research jointly with the Budapest University of Technology and Economics, evaluating different state-of-the-art open source LLMs in intent translation. Moreover, we are also investigating the use of neuro-symbolic integration. In detail, we are working on an initial proposal of a neuro-symbolic system, illustrated in Fig. 2, where the natural language intent is translated into a machine-readable policy and then into a logic programming language, such as Scallop or Asnwer Set Programming (ASP), which is used to perform symbolic reasoning [21].

However, our solution is still in the early stages of development, and several real-world challenges remain unresolved, particularly the complex task of determining optimal service placement across the compute continuum, which is characterized by its heterogeneous and distributed nature. Optimal placement must consider a wide range of factors, including resource availability, latency sensitivity, energy consumption, data locality, and application-specific requirements. In real-world deployments, these variables are constantly changing, making static or rule-based approaches insufficient for sustained efficiency and adaptability.

VI. Future Directions

To address these challenges, the proposed neuro-symbolic algorithm could be further extended to incorporate support for intelligent and context-aware node selection. By combining symbolic reasoning with ML, this hybrid approach offers the potential to interpret high-level user intents, apply logical constraints, and still adapt to uncertain or dynamic system states.

More importantly, this reasoning process can be enhanced by integrating a solution-sorting mechanism capable of evaluating and prioritizing multiple candidate placements. Such a mechanism could incorporate AI-based techniques—including neural networks for pattern recognition or reinforcement learning to optimize decisions based on feedback over time.

This solution could be useful in a federated/distributed both to deploy FL service across a pool of nodes and in the client selection phase, where policies could be defined in a declarative ways.

Neuro-symbolic AI could be directly applied in HADR and Industry 5.0. In both cases, a neural network-based ML model could be used to extract features from data and represent them as logic facts or rules. Symbolic AI could be later used to

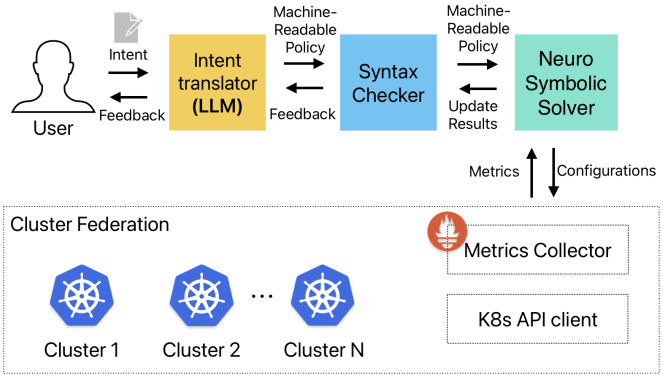

Fig. 2: Neuro-Symbolic Intent-based Service Management System Architecture

reason over these rules, resulting in a more interpretable and determinìstic process.

For example, in HADR, a convolutional neural network could be used to extract a graph representation of images, and using a logic programming language, it is possible to meet specific requirements. This could be useful in several ways, including visual question answering and enhanced topic matching.

Lastly, despite the research conducted on these topics, several areas remain open for exploration and would benefit from community feedback:

- **The Role of Neuro-Symbolic AI in Service Management:** Recent advances have shown promise in combining the robustness of symbolic reasoning with the adaptability of neural methods. However, their potential impact on service orchestration, fault tolerance, and decision-making in edge-cloud continua remains underexplored. What are the most promising pathways for integrating neuro-symbolic methods into service management architectures?
- **Managing Services Across Intermittently Connected Clusters:** Intermittent connectivity between clusters challenges assumptions of global consistency. What are the key challenges and research directions for enabling reliable service management under such conditions, and how can architectures be designed to remain resilient despite long-duration disconnections?

References

[1] M. Tortonesi, "The compute continuum: Trends and challenges," *Computer*, vol. 58, no. 3, pp. 105–108, 2025.

[2] D. Borsatti, W. Cerroni, L. Foschini, G. Y. Grabarnik, L. Manca, F. Poltronieri, D. Scotece, L. Shwartz, C. Stefanelli, M. Tortonesi *et al.*, "Kubetwin: A digital twin framework for kubernetes deployments at scale," *IEEE Transactions on Network and Service Management*, vol. 21, no. 4, pp. 3889–3903, 2024.

[3] G. Dimitrakopoulos, P. Varga, T. Gutt, G. Schneider, H. Ehm, A. Hoess, M. Tauber, K. Karathanasopoulou, A. Lackner, and J. Delsing, "Industry 5.0: Research areas and challenges with artificial intelligence and human acceptance," *IEEE Industrial Electronics Magazine*, 2024.

[4] L. Colombi, A. Gilli, S. Dahdal, I. Boleac, M. Tortonesi, C. Stefanelli, and M. Vignoli, "A machine learning operations platform for streamlined model serving in industry 5.0," in *NOMS 2024-2024 IEEE Network Operations and Management Symposium*. IEEE, 2024, pp. 1–6.

[5] R. Fronteddu, U. Ardinghi, L. Colombi, S. Dahdal, A. Morelli, M. Tortonesi, C. Stefanelli, and N. Suri, "Semantic information management systems," in *2025 International Conference on Military Communication and Information Systems (ICMCIS)*, 2025, pp. 1–9.

[6] P. Bellavista, S. Dahdal, L. Foschini, D. Tazzioli, M. Tortonesi, and R. Venanzi, "Kubernetes enhanced stateful service migration for ml-driven applications in industry 4.0 scenarios," in *2024 IEEE Annual Congress on Artificial Intelligence of Things (AIoT)*. IEEE, 2024, pp. 25–31.

[7] S. Dahdal, L. Colombi, M. Brina, A. Gilli, M. Tortonesi, M. Vignoli, and C. Stefanelli, "An mlops framework for gan-based fault detection in bonfiglioli's evo plant." *Infocommunications Journal*, vol. 16, no. 2, 2024.

[8] D. Tazzioli, R. Venanzi, and L. Foschini, "Stateful service migration support for kubernetes-based orchestration in industry 4.0," in *2024 IEEE Symposium on Computers and Communications (ISCC)*. IEEE, 2024, pp. 1–6.

[9] L. Colombi, I. Boleac, M. Brina, S. Dahdal, M. Tortonesi, M. Vignoli, and C. Stefanelli, "Multi-cluster mlops platform for industry 5.0," in *2025 IEEE Symposium on Computers and Communications (ISCC)*, 2025.

[10] L. Colombi, M. Vespa, N. Belletti, M. Brina, S. Dahdal, F. Tabanelli, F. Resca, E. Bellodi, M. Tortonesi, C. Stefanelli *et al.*, "Embedding models for multivariate time series anomaly detection in industry 5.0," *Data Science and Engineering*, pp. 1–17, 2025.

[11] L. Colombi, M. Vespa, N. Belletti, M. Brina, S. Dahdal, F. Tabanelli, E. Bellodi, M. Tortonesi, C. Stefanelli, and M. Vignoli, "Multivariate time series anomaly detection in industry 5.0," *arXiv preprint arXiv:2503.15946*, 2025.

[12] L. Colombi, M. Brina, M. Vespa, F. Tabanelli, S. Dahdal, E. Bellodi, R. Venanzi, M. Tortonesi, M. Vignoli, and C. Stefanelli, "Optimizing industry 5.0 machine learning-based applications via synthetic data generation," in *2024 IEEE 29th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*. IEEE, 2024, pp. 1–6.

[13] L. Colombi, N. Belletti, L. Ferrari, F. Tabanelli, S. Dahdal, M. Tortonesi, C. Stefanelli, and R. Venanzi, "Exploring explainable and causal ai for feature selection in industry 5.0," 2025, manuscript submitted for publication.

[14] L. Colombi, E. D. Caro, S. Dahdal, F. Poltronieri, F. Tabanelli, M. Tortonesi, C. Stefanelli, and M. Vignoli, "Fededge-learn: a semi-supervised federated learning framework for industry 5.0," in *2025 IEEE Symposium on Computers and Communications (ISCC)*, 2025.

[15] L. Colombi, M. Vespa, F. Resca, S. Cavicchi, E. di Caro, E. Bellodi, M. Tortonesi, and C. Stefanelli, "Investigating edge fine-tuning of large language models in a federated environment," 2025, accepted for publication.

[16] L. Colombi, S. Dahdal, E. Di Caro, R. Fronteddu, A. Gilli, A. Morelli, F. Poltronieri, M. Tortonesi, N. Suri, and C. Stefanelli, "Efficient data dissemination via semantic filtering at the tactical edge," in *MILCOM 2024 - 2024 IEEE Military Communications Conference (MILCOM)*, 2024, pp. 457–462.

[17] C. Benzaid and T. Taleb, "Ai-driven zero touch network and service management in 5g and beyond: Challenges and research directions," *Ieee Network*, vol. 34, no. 2, pp. 186–194, 2020.

[18] R. Venanzi, L. Colombi, D. Tazzioli, S. Dahdal, M. Tortonesi, and L. Foschini, "A collaborative intelligence-driven mlops framework for adaptive orchestration across the compute continuum," 2025, manuscript submitted for review.

[19] R. Venanzi, S. Dahdal, D. Tazzioli, L. Colombi, M. Tortonesi, and L. Foschini, "Intelligent zero-touch service migration for industry 4.0 cloud continuum deployments," 2025, manuscript submitted for review.

[20] K. Dzeparoska, J. Lin, A. Tizghadam, and A. Leon-Garcia, "Llm-based policy generation for intent-based management of applications," in *2023 19th International Conference on Network and Service Management (CNSM)*. IEEE, Oct. 2023, p. 1–7. [Online]. Available: http://dx.doi.org/10.23919/CNSM59352.2023.10327837

[21] L. Colombi, S. Cavicchi, F. Poltronieri, M. Tortonesi, C. Stefanelli, and P. Varga, "Investigating neuro-symbolic ai for intent-based service management," 2025, manuscript submitted for review.