UNICO I+D Project
6G-DATADRIVEN-05

# 6G-DATADRIVEN-05-E10

# Distributed orchestration for AI/ML (final release)


## Abstract

AI/ML driven orchestration of resources is a key enabler that promises to facilitate efficient operations of the connected industry. We provide a thorough review of current state of the art in distributed orchestration in this deliverable and propose a framework leveraging edge-AI to create smart agile orchestration agents. Our work in this deliverable also provides a basis for the edge-continuum for federated AI.



Financiado por la Unión Europea NextGenerationEU

GOBIERNO DE ESPAÑA MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES

Plan de Recuperación, Transformación y Resiliencia

# Document properties

| | |
|---|---|
| **Document number** | 6G-DATADRIVEN-05-E10 |
| **Document title** | Distributed orchestration for AI/ML (final release) |
| **Document responsible** | Carlos J. Bernardos (UC3M) |
| **Document editor** | Constantine Ayimba, Carlos J. Bernardos (UC3M) |
| **Editorial team** | Constantine Ayimba, Carlos J. Bernardos (UC3M) |
| **Target dissemination level** | Public |
| **Status of the document** | Final |
| **Version** | 1.0 |
| **Delivery date** | 31/12/2023 |
| **Actual delivery date** | 31/12/2023 |

# Production properties

| | |
|---|---|
| **Reviewers** | Antonio de la Oliva (UC3M) |

# Disclaimer

This document has been produced in the context of the 6G-DATADRIVEN Project. The research leading to these results has received funding from the Spanish Ministry of Economic Affairs and Digital Transformation and the European Union-NextGenerationEU through the UNICO 5G I+D programme.

# Contents



Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
UNICO I+D

# List of Figures



Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA
MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL
SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
UNICO I+D

# List of Acronyms

ADMM: Alternative Direction Method of Multipliers

AI/ML: Artificial Intelligence / Machine Learning

DNN: Deep Neural Network

DRL: Deep Reinforcement Learning

IoT: Internet of Things

NFV: Network Function Virtualization

QoS: Quality of Service

RNN: Recurrent Neural Network

SFC: Service Function Chain

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
UNICO I+D

# Resumen Ejecutivo

Este documento presenta la investigación más reciente sobre la orquestación distribuida para el proyecto 6G-DATADRIVEN-05. Además, también propone un marco actualizado que permite una orquestación robusta basada en inteligencia artificial y la implementación de algoritmos de inteligencia artificial/aprendizaje automático sobre el continuo Edge.

En resumen, las principales contribuciones de este entregable son las siguientes:

- un análisis detallado del estado del arte en orquestación distribuida;
- una taxonomía de la orquestación basada en inteligencia artificial/aprendizaje automático de última generación;
- un marco modular que aprovecha el continuo Edge distribuido para crear agentes de orquestación de inteligencia artificial escalables y modulares; y
- la definición de una entidad de análisis y entidades de procesamiento de datos que se pueden aprovechar para proporcionar la toma de decisiones de gestión basada en datos para la industria conectada.

Con especial atención al aprovechamiento de la inteligencia artificial en el contexto de la industria 4.0, la siguiente investigación se ha llevado a cabo en el marco de 6G-DATADRIVEN-05:

- una solución que permite usar inteligencia artificial para mitigar la interferencia inalámbrica en el control remoto de un brazo robótico (Groshev, Martín-Pérez, Guimarães, Oliva, & Bernardos, FoReCo: A Forecast-Based Recovery Mechanism for Real-Time Remote Control of Robotic Manipulators, 2022); y
- la formulación del problema de despliegue de servicios de robots para entornos de industria conectada (Khasa Gillani, 2022).

El resto del documento está redactado en inglés, de cara a maximizar el impacto del trabajo realizado en este proyecto.

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
UNICO I+D

# Executive Summary

This document provides the most up to date research on distributed orchestration for the 6G-DATADRIVEN-05 project. Further, it also proposes an updated framework enabling robust AI-based orchestration and deployment of AI/ML algorithms over the Edge continuum.

In brief, the key contributions in this deliverable are as follows:

- a detailed analysis of the state of the art in distributed orchestration;
- a taxonomy of state-of-the-art AI/ML based orchestration;
- a modular framework that exploits the distributed edge-continuum for creating scalable, modular AI orchestration agents; and
- the definition of an analytics entity and data processing entities that can be leveraged to provide data driven management decision making for the connected industry.

With particular regard to leveraging artificial intelligence in the context of industry 4.0, the following research has been carried out in the context of 6G-DATADRIVEN-05:

- a solution to mitigate the Wireless interference of remotely controlled robotic arms (Groshev, Martín-Pérez, Guimarães, Oliva, & Bernardos, FoReCo: A Forecast-Based Recovery Mechanism for Real-Time Remote Control of Robotic Manipulators, 2022); and
- the formulation of the problem related to the deployment of robotic services in connected industry (Khasa Gillani, 2022).

.

Financiado por la Unión Europea NextGenerationEU

GOBIERNO DE ESPAÑA MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES

Plan de Recuperación, Transformación y Resiliencia

UNICO I+D

# 1. Introduction

The explosion of network services with strict requirements of latency and/or throughput have spurred the innovation of flexible orchestration mechanisms that allocate resources in the most optimal fashion to meet these demands.

Many modern network services have adopted a cloud native architecture composed of interdependent components known as micro-services. The latter can be hosted on a myriad of distributed compute resources at different levels, from core to the edge of a network hierarchy. The decision on where each constituent component of a service function chain should be hosted is further complicated by the flux in resource capabilities and network conditions.

Legacy algorithms that aim to formally model orchestration as a suite of optimization problems have been found to be inadequate given the growing complexity of modern disaggregated networks and the cloud native application paradigms. These complexities have inspired research into smart orchestration algorithms that leverage AI/ML to optimise resource allocation.

In this document we provide a sweeping update on the state of the art in the smart orchestration algorithms presented in deliverable 6G-DATADRIVEN-05-E9. We survey bleeding-edge research into resource allocation mechanisms and classify them based on the principal methods they employ.

We further provide a revised proposal of a smart orchestration system that exploits synergies in cutting edge AI techniques in the state of the art. Our revised architecture considers the full plethora of the compute continuum that aligns well with the nature of modern disaggregated networks. Given the importance of reliable data in AI/ML training, our architecture includes data collection components that extract metrics from heterogenous network elements, performs requisite cross-referencing and normalisation before feeding them into the ML pipelines.

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA
MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL
SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
UNICO I+D

# 2. State of the Art in Distributed Orchestration and AI

The proper functioning of the plethora of modern network applications such as assisted driving powered by connected vehicles or multimedia streaming is largely determined by where, how and when they are allocated resources. Should the network be remiss in this basic function of allocating bandwidth, compute or priority on network flows, the consequences can be dire depending on the nature of the service.

In the realm of Distributed Orchestration, it is vital to efficiently distribute network resources among services for optimal functionality. This is achieved through the use of distributed algorithms, especially beneficial in large disaggregated networks. The approach involves dividing the orchestration problem, enabling each distributed agent to tackle a local issue. Through interactions, these agents exchange partial solutions, leading to the convergence of the system towards a global optimum.

In the subsequent paragraphs, we examine the state of the art in distributed orchestration covering the entire network spectrum from the cloud to the edge. Each problem addressed by each contribution is presented along with the proposed mechanism to solve it. We also provide insights into the gaps not handled by the proposals.

In order to address the challenge of joint optimization of service migration and edge deployment (Chen & al, 2022) propose an algorithm that breaks down the problem into a cascade of one-slot optimization problems. They solve the placement problem using a variant of Lyapunov optimization that leverages randomized rounding to handle the service migration problem. However, this approach does not consider load balancing which is crucial for service reliability.

(Bao, et al., 2023) propose pre-emptive orchestration based on traffic prediction as a mechanism to ensure service reliability. Their traffic prediction algorithm is based on a Gated Recurrent Unit model which relies on Recurrent Neural Networks (RNN). The orchestration mechanism they propose is based on load balancing theory which further guarantees reliability.

In (K. Zhu, 2023), the authors use weighted parameters to improve the convergence of a deep reinforcement learning based orchestrator. The approach proves quite efficient for allocating resources for Unmanned Aerial Vehicles and other energy-constrained network elements.

The authors of (Mohammed & Francesco, 2022) present a heuristic that leverages dominant resource fairness algorithm to allocate network resources to network slices. Their algorithm relies on two phases, the first at the central unit level and the other at the radio unit level. It achieves pareto optimality and works in polynomial time.

(Wen, 2020) leverage a genetic algorithm integrated with Hadoop to evaluate resource orchestration within a geographically distributed cloud environment. The emphasis is placed on addressing security needs by employing a metric that examines the misalignment between the security requirements of services and the security attributes of servers.

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
UNICO I+D

Deep Reinforcement Learning (DRL) is used by (Huang, et al., 2021) to create a scalable Service Function Chain (SFC) orchestrator. The technique also exploits aspects of federated learning to leverage local knowledge of available resources. The combination of these methods results in faster convergence in comparison with similar approaches. The approach however requires strict coordination among participating nodes.

(Chen, et al., 2022) investigate SFC orchestration problem in IoT. Their approach involves splitting the SFC orchestration into two phases: i) selecting the SFC to orchestrate; and ii) assigning resources to the SFC over the stratum. The second phase uses distributed Q-learning.

To facilitate the integration of trusted SFCs across the Edge and Cloud continuum, (Guo, Dai, Xu, Qiu, & Qi, 2020) introduce an architecture that amalgamates blockchain and DRL. In this approach, service providers authenticate services within a trusted chain for security assurance. Following authentication, the Service Function Chain (SFC) is orchestrated across a varied pool of Edge cloud facilities using a DRL-based solution. Each Edge facility is equipped with a DRL agent, allowing it to assess the feasibility of SFC deployment. If an agent determine infeasibility, the request is escalated to a cloud-level DRL agent for alternative solutions. This hierarchical approach is implemented across a distributed pool of Edge cloud facilities

The authors of (Chen & Xu, 2021) leverage multi-agent DRL for optimal resource allocation in a distributed setup. Each Edge server hosts a local DRL agent, learning the nuances of network and compute capabilities while managing local services. The distributed learning process converges towards a global optimal solution, accounting for local variations. To address the latter, they introduce N2O, a neural network orchestrator that communicates with a global actor, enhancing service deployment based on local insights. The study provides evidence of the stability and convergence of the proposed solution, demonstrating compliance with Lipschitz conditions.

(Dressler, et al., 2021) provide a vision of an extensively distributed pool of computing resources transcending traditional Edge boundaries and extending to mobile users who can potentially exchange their data processing and computing capabilities. Consequently, the authors introduce a comprehensive resource continuum that incorporates Cloud, Edge, and end-users' resources. The V-Edge architecture, applicable to any device, deploys micro-services as containers within the devices. In the event of network faults, it exhibits the capability to migrate or instruct microservice deployment to an end-user in close proximity to the consumers. Additionally, V-Edge anticipates the feasibility of training a distributed AutoML solution, adjusting its parameters for fully distributed training, updating, and orchestrating resources

The authors of (Malandrino F. , Chiasserini, Molner, & Oliva, 2022) investigate the efficiency of distributed machine learning across a network infrastructure, with regard to orchestration tasks. The study conceptualizes distributed machine learning as a collection of learning nodes, acting as servers tasked with minimizing an objective function. To enhance the process, these learning nodes rely on data supplied by information nodes in the network, functioning akin to servers equipped with

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
UNICO I+D

databases. The authors introduce the DoubleClimb algorithm, which determines the optimal number of training iterations to achieve the target performance while minimizing the exchanged data over the network. This approach enables the execution of distributed orchestration with minimal network overhead.

In a similar manner, (Malandrino, Chiasserini, & Giacomo, 2023) study the effect of distributed machine learning in a Cloud to Mobile continuum. In this approach they explore a split/distribution of a DNN into its different layers, and how to deploy each of them in different servers over the network. Their *RightTrain* algorithm selects the best nodes/servers to run each layer to achieve optimal training. The solution uses an expanded graph with all the possible deployments, and have optimality guarantees regarding the training performance of the AI/ML models.

(Miloud Bagaa, 2021) propose integrating AI in a cross-system network architecture that comprises the core network, servers, and end devices. The architecture adopts a system Network Intelligence Function (NIF) that collects from diverse sources in a distributed fashion to streamline network management tasks such as scaling the computational and networking resources allocated to services.

(Wang, Wei, Yu, & Han, 2022) introduce a joint optimization approach for communication, computing, and caching resources in multi-access edge network slicing. The aim of the study is to maximize utility for mobile virtual network operators while ensuring quality of service (QoS) in a two-level resource allocation problem. They propose a novel actor-critic mechanism called deep deterministic policy gradient that adapts intelligently to dynamic environments.

To ameliorate the problem of the large number of decision variables involved in NFV resource allocation in large networks, (Yu, Bu, Yang, Nguyen, & Han, 2020) propose a two-stage simplification of the problem: i) Bender decomposition that changes discrete variables into continuous ones; and ii) solving the associated dual problem. ADMM is then used in the second stage to solve the problem in a distributed manner, each node within the network substrate addresses a subproblem, subsequently updating the decision variables.

In a bid to address the problem of disclosure of sensitive network infrastructure data in distributed orchestration (Jason Hughes, 2022) propose slight randomizations in the data exchange. They exploit optimal transport theory to decide the resource allocation. They then employ ADMM to solve the associated optimal transport problem achieving a solution that iteratively converges to an optimal allocation.

The approaches considered above investigate the application of distributed orchestration techniques across the network substrate. The current state of the art treats the entire network hierarchy from cloud, edge to extreme edge (devices) as a continuum. There is therefore a growing trend to perceive the network as a comprehensive continuum that goes beyond legacy consideration of switches,

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
UNICO I+D

servers, and end devices. This inclination involves incorporating mobile phones and IoT devices into the orchestration processes for service-related tasks.

In Table 1 we provide a taxonomy of state-of-the-art distributed orchestration based on the technique(s) employed.

**TABLE 1: CLASSIFICATION OF DISTRIBUTED ORCHESTRATION TECHNIQUES IN THE STATE OF THE ART**

| Approach | Study | Technique Description |
|---|---|---|
| **Federated Learning** | (Malandrino F. , Chiasserini, Molner, & Oliva, 2022); (Malandrino, Chiasserini, & Giacomo, 2023) | All the learning nodes perform gradient descends based on local information, and their parameter updates are averaged and exchanged among all to find a global optimum, yet distributing the computation locally at each node. |
| **ADMM** | (Jason Hughes, 2022); (Yu, Bu, Yang, Nguyen, & Han, 2020) | Decomposition into sub-problems which can be iteratively solved using Lagrange multipliers in an alternating manner |
| **Architecture based** | (Dressler, et al., 2021); (Miloud Bagaa, 2021); | These studies propose cloud-native micro-service-based architecture with built-in orchestration mechanisms. |
| **Reinforcement learning** | (Chen, et al., 2022); (Guo, Dai, Xu, Qiu, & Qi, 2020); (Chen & Xu, 2021); (Wang, Wei, Yu, & Han, 2022) (K. Zhu, 2023) | These methods employ enhanced reward-based learning whereby an agent's action changes the state of the environment as a consequence of which it receives a reward signal which progressively directs it to the correct decisions. |
| **Genetic Algorithms** | (Wen, 2020); | Genetic algorithms test possible allocations of resources following a certain strategy/genome that is iteratively modified until optimality is attained. |
| **Heuristics** | (Mohammed & Francesco, 2022) (K. Zhu, 2023) | The work optimizes how to distribute the different layers of an AI/ML model to achieve convergence given the computational and network constraints of the underlying network stratum. |
| **Lyapunov optimization** | (Chen & al, 2022) | This technique is based on the theory of Lyapunov control that provides convergence and optimality guarantees on optimization metrics to be optimized e.g., the average latency of orchestrated services. |

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
UNICO I+D

# 3. Enhancements and lessons learned

With reference to the initial framework provided in deliverable 6G-DATADRIVEN-05-E9, the following were the lessons learned about the design implications for orchestration and AI for industry 4.0:

- *Scalability*: solutions should be scalable: to accommodate a growing number of devices, sensors, and AI components. This is crucial as the connected industry is dynamic and may expand over time.
- *Inter-operability:* this is key: to be compatible with heterogeneity of devices, protocols, and communication standards commonly used in industrial settings. Interoperability enables smooth integration with existing infrastructure and facilitates communication between diverse devices.
- *Robustness*: solutions need to be anti-fragile: Connected industry systems must be reliable to avoid disruptions in critical processes. Additionally, they need to incorporate resilience features to handle system failures, ensuring continuous operation even in the face of unexpected events.
- *Security*: data security should not be compromised, mechanisms must be incorporate to implement robust security measures to protect data, devices, and communication channels
- *Delay intolerance*: connected industry applications are latency sensitive given the speed of production lines and the cascading ramifications of delay.
- *Modularity*: solutions should be modular to adapt to evolving technologies and industry requirements without requiring an over-haul of the entire system.
- *Decentralised:* dis-aggregated schemes are best suited to meet the above challenges presented by connected industry. Principally, network and compute resources should facilitate the handling data closer to the source and performing local AI processing on devices (extreme-edge) or gateways (far-edge).

With the above considerations in mind our revised framework in this deliverable is leaner, more modular and simplified. In the following section, we delve into the design of this framework.

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
UNICO I+D

# 4. Distributed framework for AI/ML Orchestration at the edge continuum

We provide a federated AI based framework that enables a decentralised scalable approach to smart orchestration of resources at all layers of the edge continuum. We make enhancements that exploit local data sources to create models that can then be combined further up the edge continuum to produce more robust models. Moreover, the latter models can still exploit local runtime data to make context specific orchestration decisions at a granular level.

In this revised final framework, we simplify the interactions between the various levels of the edge continuum thereby facilitating their adoption by industry given that no drastic changes are required on their part. The de-centralized 5G network, with the Control/User Plane Split (CUPS) allow the management end-user data and enable the creation of new applications environments that can be deployed very close to the end-user. These features, coupled with network slicing, provide the basis for application migration to the edge and far edge (on-premise). The evolution towards fully autonomous self-managed networks is also facilitated by this design which can be easily extended to include coordinated closed-loops at various points of the hierarchy.

The framework outlined in this section aims to merge both domains to present and validate, within TRL7 environments, a reference architecture. This architecture tackles the challenge of integrating manufacturing industries through 5G technology, offering solutions for various aspects: handling massive Digital Twins (DTs), incorporating 5G networks into on-premises communications, offloading manufacturing applications to the cloud. Additionally, the framework conducts specific analyses on how these changes impact industry security standards. Overall, it focuses on integrating manufacturing data management into a standard IT ecosystem.

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
UNICO I+D

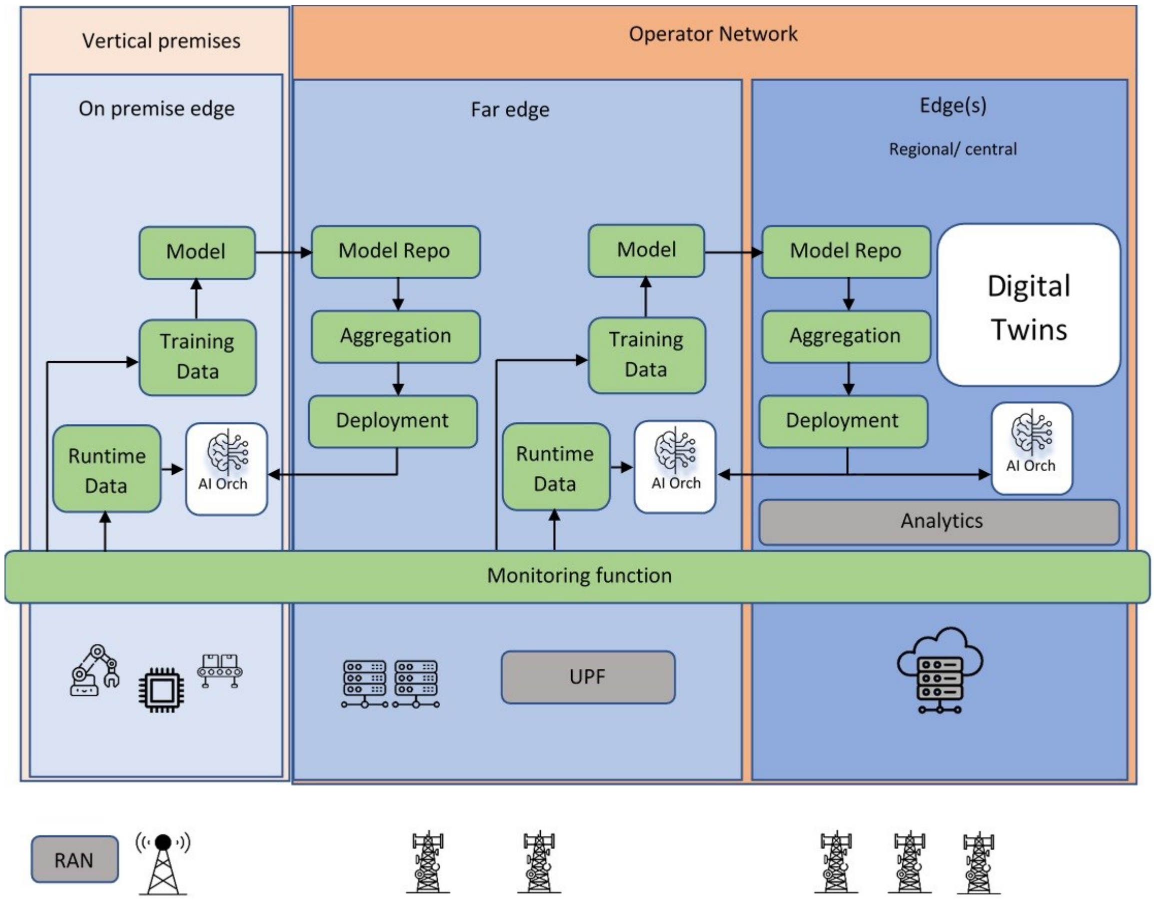


FIGURE 1: REVISED FRAMEWORK FOR DISTRIBUTED AI/ML ORCHESTRATION

Figure 1 depicts the revised framework. It realises our vision of a compact distributed environment, traversing the full range of the edge cloud continuum.

At a functional level, we consider the following:

- *Monitoring*: this takes input from various infrastructure elements such as AGV and Robotic Arms with Robot Operating System (ROS), IP Cameras, various sensors at the factory floor (on premise edge) network elements such as UPF and switch data as well as compute resource data from far edge and edge clouds.
- *Model training*: this follows a modular federated learning approach to exploit local data available at the vertical premises to train reinforcement learning agents without compromising sensitive commercial data.
- *Model aggregation* this module combines the various local models from the myriad extreme edge elements to create more robust models by exploiting learning from more distributed sources.
- *Analytics and predictive maintenance*: this component allows the vertical client to exploit data harnessed for learning to carry out complex analysis and management decisions based on forecasts. It also facilitates the creation of maintenance schedules for any equipment deemed to be operating close to collapse.

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
UNICO I+D

- *Digital Twins*: this component will provide a near real-time model of the current state of the IoT infrastructure replete with varying capabilities of all constituent nodes. This cyber-physical model will allow the operators to test configurations and check the ramifications before such changes are made on actual devices.

The framework provided in this document relies on the deployment of an on-premise 5G network integrated to the telco network. On-premise RAN resources will also provide inter-device communication capabilities to enhance device to device communication.

Given the heterogeneity of connected industry, the services deployed over the proposed framework will rely on open protocols to allow inter-operability between devices. This variegated nature of smart industry aligns well with the federated approach to machine learning that is the mainstay of our proposed framework.

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
UNICO I+D

# 5. Summary and Conclusions

In this deliverable, we provide bleeding edge, state of the art orchestration solutions ranging from legacy optimization approaches to federated learning exploiting robust techniques such as deep reinforcement learning. We provide a taxonomy that classifies these solutions based on the techniques employed by each.

We provide a scalable framework that envisions a continuum of Edge premises exploiting local data to create smart orchestration agents at the various levels of the continuum. The framework also allows the collected data to be further exploited in creating massive Digital Twinning solutions that are a key enabler of industry 4.0 services.

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
UNICO I+D

# 6. References

Applications, O. N. (2022). *Khasa Gillani, Jorge Martín-Pérez, Milan Groshev, Antonio de la Oliva, Carlos J. Bernardos, Robert Gazda.* arXiv.

Bao, B., Yang, H., Yao, Q., Guan, L., Zhang, J., & Cheriet, M. (2023). Resource Allocation With Edge-Cloud Collaborative Traffic Prediction in Integrated Radio and Optical Networks. *IEEE Access*, 7067-7077.

Chen, H., Wang, S., Li, G., Nie, L., Wang, X., & Ning, Z. (2022). Distributed Orchestration of Service Function Chains for Edge Intelligence in the Industrial Internet of Things. *IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS*.

Chen, L., & Xu, J. (2021). Seek Common While Shelving Differences: Orchestrating Deep Neural Networks for Edge Service Provisioning. *IEEE Journal on Selected Areas in Communications*.

Chen, X., & al, e. (2022). Dynamic Service Migration and Request Routing for Microservice in Multicell Mobile-Edge Computing. *IEEE Internet of Things Journal*, 13126-13143.

Dressler, F., Chiasserini, C. F., Fitzek, F. H., Karl, H., Cigno, R. L., Capone, A., . . . Rizzo, F. K. (2021). V-Edge: Virtual Edge Computing as an Enabler for Novel Microservices and Cooperative Computing. *IEEE Network*.

Groshev, M., Martín-Pérez, J., Guimarães, C., Oliva, A. d., & Bernardos, C. J. (2022). FoReCo: A Forecast-Based Recovery Mechanism for Real-Time Remote Control of Robotic Manipulators. *IEEE Transactions on Network and Service Management*.

Groshev, M., Martín-Pérez, J., Guimarães, C., Oliva, A. d., & Bernardos, C. J. (2022). FoReCo: A Forecast-Based Recovery Mechanism for Real-Time Remote Control of Robotic Manipulators. *IEEE Transactions on Network and Service Management*.

Guo, S., Dai, Y., Xu, S., Qiu, X., & Qi, F. (2020). Trusted Cloud-Edge Network Resource Management: DRL-Driven Service Function Chain Orchestration for IoT. *IEEE Internet of Things Journal*.

Huang, H., Zeng, C., Zhao, Y., Min, G., Zhu, Y., Miao, W., & Hu, J. (2021). Scalable Orchestration of Service Function Chains in NFV-Enabled Networks: A Federated Reinforcement Learning Approach. *IEEE Journal on Selected Areas in Communications*, 2558-2571.

Jason Hughes, J. C. (2022). Differentially Private ADMM-Based Distributed Discrete Optimal Transport for Resource Allocation. *IEEE GLOBECOMM 2022.* Rio de Janeiro.

K. Zhu, Z. Z. (2023). Auxiliary-Task-Based Energy-Efficient Resource Orchestration in Mobile Edge Computing. *IEEE Transactions on Green Communications and Networking*, 313-327.

Khasa Gillani, J. M.-P. (2022). *Orchestrating Networked Robotic Applications.* arXiv.

Malandrino, F., Chiasserini, C. F., & Giacomo, G. d. (2023). Efficient Distributed DNNs in the Mobile-Edge-Cloud Continuum. *IEEE/ACM Transactions on Networking*, 1702-1716.

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
UNICO I+D

Malandrino, F., Chiasserini, C. F., Molner, N., & Oliva, A. d. (2022). Network Support for High-Performance Distributed Machine Learning. *IEEE/ACM Transactions on Networking*.

Milan Groshev, J. M.-P. (2022). FoReCo: a forecast-based recovery mechanism for real-time remote control of robotic manipulators. *IEEE Transactions on Network and Service Management*, 12.

Milan Groshev, J. S.-P. (2022). Demo: FoReCo – a forecast-based recovery mechanism for real-time remote control of robotic manipulators. *SIGCOMM '22: Proceedings of the SIGCOMM '22 Poster and Demo Sessions* (pág. 2). Amsterdam: ACM.

Miloud Bagaa, T. T. (2021). Collaborative Cross System AI: Toward 5G System and Beyond. *IEEE Network*.

Mohammed, T., & Francesco, B. J. (2022). Efficient and Fair Multi-Resource Allocation in Dynamic Fog Radio Access Network Slicing. *IEEE Internet of Things Journal*, 24600-24614.

Wang, Z., Wei, Y., Yu, F. R., & Han, Z. (2022). Utility Optimization for Resource Allocation in Multi-Access Edge Network Slicing: A Twin-Actor Deep Deterministic Policy Gradient Approach. *IEEE Transactions on Wireless Communications*, 5842-5856.

Wen, Z. a. (2020). GA-Par: Dependable Microservice Orchestration Framework for Geo-Distributed Clouds. *IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS*.

Yu, Y., Bu, X., Yang, K., Nguyen, H. K., & Han, Z. (2020). Network Function Virtualization Resource Allocation Based on Joint Benders Decomposition and ADMM. *IEEE Transactions on Vehicular Technology*.

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA
MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL
SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
UNICO I+D