AI-Powered Orchestration-as-a-Service for 6G Networks: The 6G-CLOUD View

Jesus Perez-Valero, Gines Garcia-Aviles, Antonio Skarmeta, and Tao Chen

Abstract

As 6G networks grow more complex, managing resources and orchestrating services across diverse, dynamic, and energy-efficient environments becomes challenging. This paper presents a programmable Orchestration-as-a-Service (OaaS) Framework from the 6G-Cloud project that enables dynamic, scalable, and intelligent orchestration. The framework separates service orchestration from resource orchestration, supports network service–agnostic management, and integrates AI-driven optimization, digital twins (DT), and Cloud Continuum technologies. Finally, we demostrate the functionality and feasibility of the proposed architecture through an early-stage implementation, where an AI-powered use case for dynamic resource orchestration is validated, showing significant improvements in power efficiency and proactive resource scaling compared to baseline methods.

Introduction

The evolution of orchestration mechanisms, as defined by ETSI NFV and its advancements, is crucial for the next generation of 6G communication systems. Orchestration as a Service (OaaS) emerges as a key paradigm, enabling flexible and dynamic network service management by decoupling service orchestration from resource orchestration. This approach enhances the adaptability of 6G networks, allowing seamless integration with service-oriented architectures (SOA), dynamic resource allocation, and AI-powered operations—three foundational pillars of next-generation communication systems. In this work, a Network Service (NS) refers to an end-to-end service composed of multiple interconnected Network Functions (NFs) that together deliver a specific functionality (e.g., enhanced mobile broadband, network slicing). A Service Orchestrator (SO) manages the lifecycle of these NSs (handling their design, instantiation, scaling, and termination) based on high-level intents and SLAs. In contrast, a Resource Orchestrator (RO) focuses on allocating and managing the underlying physical and virtual resources (e.g., compute, storage, and networking) needed by NSs. These technologies are further complemented by the "Network of Networks" (NoN) and Cloud Continuum concepts, ensuring seamless integration of centralized, edge, and extreme-edge resources. However, a key pitfall of existing architectures is their inability to address the complexity of multi-domain scenarios, scalability, dynamic workload demands, and the need for energy-efficient operations.

Several projects have explored orchestration innovations for 6G. For instance, Hexa-X and Hexa-X-II introduced novel orchestration mechanisms [1], [2], [3], while ORIGAMI focused on service-based architectures [4]. 6G-BRICKS and EDGELESS investigated computational resource orchestration across the network [5], [6]. ADROIT6G integrates closed-loop functions into traditional management frameworks [7], and ETHER promotes an AI-driven MANO solution [8]. DETERMINISTIC6G develops programmable systems for deterministic service [9]. The 6G-Cloud project envisions a distributed, intelligent, and sustainable 6G architecture designed to overcome the limitations of earlier network designs. Its key architectural enablers include: an end-to-end service-based architecture unifying the RAN and Core networks; network services exposed via service-based interfaces; a Cloud Continuum integrating computing resources across all abstraction layers; Orchestration-as-a-Service (OaaS) that separates service and resource orchestration; and AI-as-a-Service (AIaaS) providing AI/ML-driven intelligence across all network domains.

As described in Figure 1 above, the 6G Cloud architecture will incorporate specific frameworks, namely the Management and Orchestration Framework (MOF), the Cloud Continuum Framework (CCF) and the AI/ML Framework (AIMLF) to address the challenges related to network flexibility, scalability and heterogeneity. Essential to this process is the inclusion of new interfaces that ensure effective security, integrity, and business relationships between stakeholders.

Based on these insights, we present a novel programmable MOF. Unlike traditional approaches, the MOF provides dynamic, scalable and customized orchestration capabilities, enabling seamless resource unification across centralized, edge, and extreme-edge environments. Its AI-driven functionalities ensure real-time optimization, proactive fault management, and energy-efficient operations. Additionally, the MOF

This work was supported by European Commission through the SNS JU Project 6G-CLOUD under Grant 101139073.

Jesus Perez-Valero (corresponding author), Gines Garcia-Aviles, and Antonio Skarmeta are with the Department of Information and Communications Engineering, University of Murcia, 30100 Murcia, Spain; Tao Chen is with the VTT Technical Research Centre of Finland, 02140 Espoo, Finland.

Digital Object Identifier: 10.1109/MCOMSTD.2026.3656636

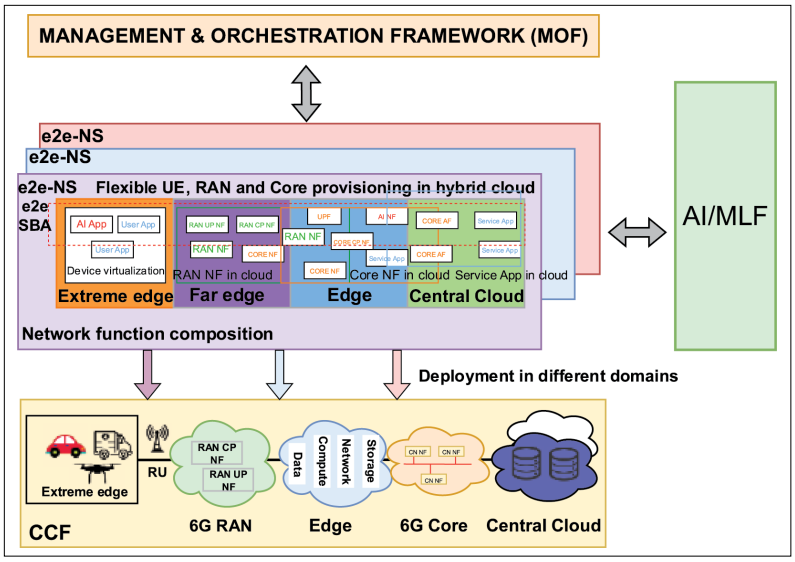

FIGURE 1. Overview of the AI-native 6G-CLOUD network architecture.

addresses the complexities of multi-domain and multi-stakeholder scenarios, offering enhanced flexibility, security, and reliability in the deployment and management of 6G network services.

The main contributions of this paper are as follows:

- **Programmable Orchestration as a Service Framework.** The proposed MOF introduces a highly modular architecture, which facilitates programmable orchestration through its Service-Oriented design and customizable control loops. This is achieved via dedicated SOs and ROs, which operate independently to handle specific NS. The separation of service and resource orchestration allows for dynamic, policy-driven resource allocation for specific NSs while maintaining low operational overhead, given that both orchestration entities will operate at different levels.
- **AI-Powered Service Orchestration as a Service.** The operations of the MOF are enhanced through AI-driven functions that can be either statically embedded or dynamically orchestrated on demand. These functions support critical tasks such as resource prediction, anomaly detection, and mobile compute availability forecasting, optimizing network efficiency and reliability. Within the MOF, AI-driven mechanisms assist the Operations Support System (OSS) and Business Support System (BSS) in network service (NS) reconfiguration, automated resource scaling, and orchestration of NS functions.

Limitations of ETSI NFV

Existing ETSI NFV-based orchestrators [5] employs a centralized and monolithic Service Orchestrator (SO) that handles all services, limiting customization and adaptability. In contrast, programmable OaaS introduces per-NS SOs, allowing fine-grained orchestration, tenant-specific configurations, and even third-party SO provisioning. Additionally, updating or replacing orchestration components in ETSI NFV frameworks often requires manual intervention or predefined templates, adapting to evolving network requirements cumbersome.

Another limitation of ETSI NFV-based orchestrators is the tight coupling of Service Orchestration (SO) and Resource Orchestration (RO) [6], leading to inefficiencies in multi-cloud, multi-domain environments. The SO focuses exclusively on end-to-end service lifecycle management, while RO dynamically optimizes infrastructure allocation across the Cloud Continuum, enhancing scalability and modularity. Moreover, traditional orchestrators rely on static, vendor-specific workflows and templates, limiting customization. OaaS introduces a programmable orchestration model, leveraging modular and API-driven SOs, enabling rapid adaptation to diverse service requirements.

While ETSI NFV orchestrators can incorporate AI/ML, their usage is constrained to specific use cases. OaaS inherently integrates AI/ML-driven orchestration, facilitating intelligent resource allocation, proactive fault detection, and network adaptation in real time. Additionally, ETSI NFV solutions provide only basic multi-tenancy support, leading to generalized orchestration that does not fully adapt to tenant-specific needs. They also lack comprehensive support for multi-stakeholder collaboration, such as interactions among cloud providers, operators, and third-party developers. OaaS enhances multi-tenancy by allowing per-tenant orchestration policies and enables multi-stakeholder governance, where each entity can enforce its own business logic, policies, and orchestration requirements.

Relation to SDOs and Gap Analysis

The proposed 6G-Cloud OaaS framework builds upon and extends concepts defined by major Standards Developing Organizations (SDOs), including ETSI, 3GPP, ITU-T, and IETF. Specifically, it aligns with the ETSI NFV-MANO architecture [11], [12] for virtualized network management and orchestration, the ETSI ZSM framework for zero-touch automation, and 3GPP SA5 specifications on service management and exposure. However, current standards exhibit several limitations in supporting AI-native, multi-domain, and cross-layer orchestration across the cloud-to-edge continuum. Table I summarizes how the proposed OaaS framework addresses the identified gaps.

Management and Orchestration Framework

The MOF is a central component of the 6G-Cloud architecture that manages distributed, heterogeneous network services efficiently by tightly integrating with the CCF. It separates service orchestration (handled by the MOF) from resource orchestration (managed by the CCF), enabling dynamic service lifecycle management and infrastructure operations. The MOF leverages programmable, AI-driven orchestrators from the AIMLF for real-time monitoring and optimization.

Architectural Building Blocks

The MOF provides full FCAPS (Fault, Configuration, Accounting, Performance, and Security) capabilities for virtualized network management, enabling real-time monitoring, fault detection, and performance and security optimization. Enhanced

SDO	Focus Area	Identified Gap	How 6G-Cloud OaaS Addresses It
ETSI NFV-MANO	Virtualized resource and service orchestration	Tight coupling between Service and Resource Orchestrators; limited AI integration; single-domain scope	Introduces decoupled SO/RO layers with AI-powered optimization and multi-domain orchestration
ETSI ZSM	End-to-end service automation (zero-touch)	Focused mainly on automation, lacks detailed AI lifecycle and DT integration	Embeds AIML Framework for closed-loop orchestration and proactive optimization
3GPP SA5	Network management and service exposure	Oriented toward 5G network slicing; limited abstraction for multi-cloud / edge orchestration	Extends service orchestration to 6G Cloud Continuum and enables NS-agnostic management
ITU-T FG-NET2030	Future network architecture vision	Conceptual guidance only; lacks specific orchestration mechanisms	Provides a concrete AI-native orchestration framework aligned with FG-NET2030 objectives
IETF/BBF	Service assurance and intent-based interfaces	Intent translation not coupled with real-time AI-driven orchestration	Integrates intent-based orchestration within MOF for SLA-driven service adaptation

TABLE 1. Mapping of the proposed OaaS framework to key SDO activities and addressed gaps.

by the AI/ML Framework, it supports autonomous fault prediction, anomaly detection, and adaptive performance tuning which is key for proactive resource scaling, NF placement, and orchestration. Dynamic, energy-aware resource orchestration is achieved through close interaction between the MOF and the CCF, leveraging real-time data across the cloud–edge continuum. Within the 6G-Cloud vision, the CCF manages resources through Resource Partitions (ResPs)—logical abstractions representing the infrastructure assigned to each NS. These ResPs are dynamically created according to service requirements (e.g., location, cost, energy efficiency, reliability, inter–data center delay), while the CCF ensures seamless interoperability and orchestration among multiple ResP instances.

Additionally, the MOF can use DT technology provided by the AIML Framework, which provides virtual simulations of network components and resources. Using DTs, the MOF can assess behaviour after executing, for example, NF placement or resource allocation, before implementing changes in the live environment, thus minimising disruptions and improving decision-making robustness. The DT simulations and AI/ML capabilities enhance the ability to predict and manage potential network bottlenecks and optimize resource usage.

The MOF includes a repository of NF and NS templates that enable streamlined, NS-agnostic deployment with minimal external interaction. It supports proactive resource scaling (adding, cloning, or removing NFs based on real-time AI/ML and DT analytics).

Global BSS (G-BSS)

The Global Business Support System (G-BSS) enables seamless integration with external entities like customers and partners, providing a user-friendly interface for service requests and management. It oversees order validation, SLA compliance, and triggers orchestration via the G-OSS to ensure timely service provisioning. Additionally, the G-BSS manages billing and charging, ensuring accurate billing based on resource and service usage.

Global OSS (G-OSS)

The Global Operations Support System (G-OSS) in the 6G-Cloud architecture manages operational functions such as real-time fault detection, configuration management, and performance monitoring to ensure high service availability. Working with the Main Service Orchestrator and service-specific orchestrators, it handles lifecycle management, load balancing, fault recovery, and SLA compliance. While management is tailored to each network service, the 6G-Cloud Service Orchestrator remains network-service agnostic.

Main Service Orchestrator (MSO)

The MSO is the central management entity in the 6G-Cloud architecture, responsible for high-level orchestration and lifecycle management of network services across the cloud continuum. It oversees the deployment of Network Services (NSs), Network Service Directors (NSDirs), and Service Orchestrators (SOs), ensuring proper initialization before delegating runtime operations to SOs and NSDirs. This separation of deployment and operation enhances modularity and efficiency. The MSO also coordinates with the Global OSS/BSS to process service requests, decompose them into orchestration domains, and manage the deployment of Virtual Network Functions (VNFs) and Service Function Chains (SFCs) across available resources [13].

Network Service Director (NSDir)

The NSDir manages the orchestration and lifecycle of individual network services within their own dedicated resource pools. Each NSDir includes a service-specific orchestrator and a management plane that handles performance monitoring, fault detection, and real-time optimization, ensuring reliability and isolation from other services. It

also integrates a dedicated OSS/BSS interface for operations such as billing, SLA management, and service monitoring. Instantiated per network service, the NSDir enables independent, customized, and resilient management of each service in the 6G-Cloud architecture.

Assets Repository (ARep)

The repository serves as a centralized storage for predefined Network Functions (NFs) and Network Service (NS) templates, enabling fast, consistent, and reusable service deployments. Each template includes service configurations, while the NSDir is instantiated at runtime to manage the service lifecycle. The MSO handles only deployment configurations, leaving runtime management to the NSDir. This design keeps the MSO lightweight and supports NS-agnostic, flexible, and efficient orchestration across different deployment contexts.

Communication Interfaces

Intra-System Communication Channels

The internal message bus of the MOF (depicted as the green internal bus in Figure 2) facilitates seamless communication between all components within the framework. This message bus supports the following processes:

- G-BSS to G-OSS communication to handle upcoming SLAs, ensuring that business and operational processes are aligned.
- G-OSS to MSO communication regarding specific SLA requirements, which are essential for resource planning, reservation, and service configurations.
- MSO to ARep interaction for the active management of available NFs, NSs, ASs, and templates, ensuring that the repository stays updated and accessible for orchestrators.
- Communication between all components and the IFG enables the sharing of state information with other frameworks within the infrastructure to maintain system-wide consistency and coordination.

The management and orchestration communication bus (depicted as the dark green line in Figure 2) connects the MSO with the management and orchestration planes of each NS Director (NSDir). It enables efficient coordination and resource optimization, supporting MOF–SO communication for system-wide orchestration decisions such as scaling, reconfiguration, and resource reallocation. This bus allows each NSDir to operate semi-autonomously while remaining integrated within the global orchestration framework.

The operations and business support communication buses (pink and purple lines in Figure 2) connect the G-OSS and G-BSS interfaces within each NS Director (NSDir). These buses enable independent management of operational and business tasks for each network service, separate from the global OSS/BSS. Through service-specific layers (NS-OSS/NS-BSS), the operations bus supports real-time monitoring, fault detection, and performance optimization, while the business bus manages billing, service usage, and SLA compliance, ensuring autonomous and responsive service management.

Inter-System Communication Channels

This communication bus is a critical element in integrating the MOF with the CCF, facilitating seamless data exchange, and allowing the MOF to retrieve essential information from the CCF while sending operational directives to coordinate processes across the continuum. The message bus not only supports real-time communication but also enables the MOF to optimise and manage resources based on the insights provided by the CCF. The expected interactions include:

- **Retrieval of Status Information From the CCF:** The MOF periodically requests status updates to obtain a real-time view of the available resources. This information is crucial for optimising the allocation of resources, and the MOF assures that resources are assigned in a way that maximises efficiency and meets the demands of the network.
- **Transmission of Placement Decisions to the CCF:** The MOF sends, for example, placement directives to the CCF, instructing it to deploy the necessary components

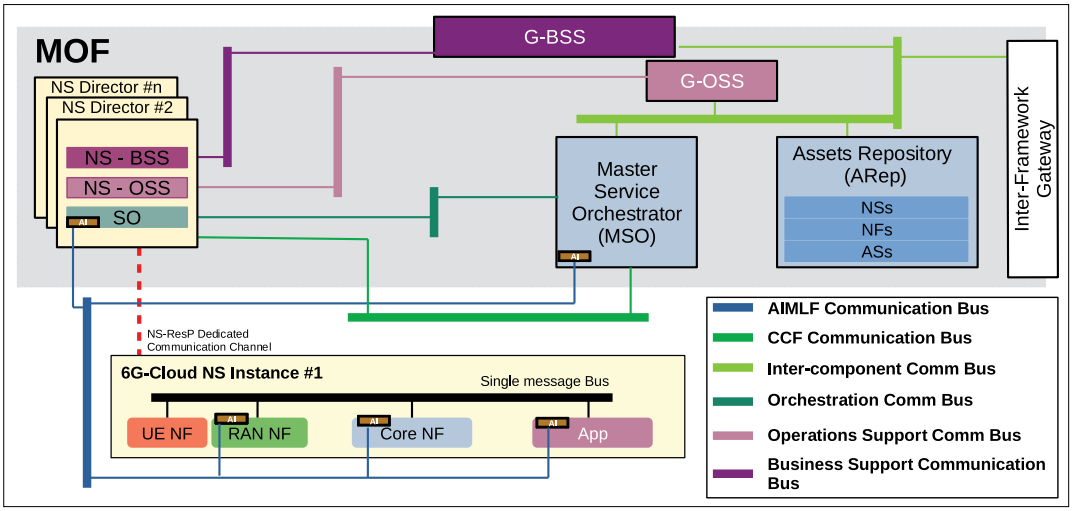

FIGURE 2. Architectural building blocks of the MOF Architecture.

for each NS. This includes assigning compute, storage, and network resources to the NS, ensuring they are optimally placed within the continuum to meet specified requirements.

- **Communication From the CCF to the MOF:** The CCF also provides the MOF with critical notifications and status updates. For example, the CCF might send information about the communication points where specific re-source partitions are ready to interact with the NS Director or other components. These communications ensure that the MOF knows the exact locations and statuses of resources, enabling it to manage deployments and orchestrate services without delay.
- **Business Interface Between Frameworks**: ensures that SLAs, billing, accounting and other economic aspects are communicated correctly. To achieve this, the MOF selects the parameters to be met for each NS independently, and the CCF provides a complete view of the business parameters through a dedicated API.

AI/ML Models Within the Framework

The AIMLF integrated within the MOF employs diverse learning paradigms to support various orchestration tasks. Supervised learning models handle resource forecasting, traffic prediction, and SLA violation detection, while unsupervised methods such as clustering and autoencoders address anomaly detection and performance analysis. Meanwhile, reinforcement learning (RL) and deep RL agents enable closed-loop, adaptive decision-making for dynamic resource orchestration. Model selection depends on the orchestration layer and data availability, with models continuously instantiated, retrained, and deployed to maintain optimal and adaptive performance across the 6G Cloud Continuum.

The inference latency of AI components is a critical factor that depends on model complexity and the required orchestration speed. Lightweight models enable millisecond-level inference for real-time tasks like anomaly detection, while more complex models used for resource planning or optimization may take seconds to minutes, fitting non–real-time operations. The AIMLF manages this latency–accuracy trade-off by selecting and deploying models that best align with each orchestration task's timing and precision needs.

Use-Case Scenario

We propose an end to end resource allocation scenario in the same fashion as in our previous work [14]. The scenario is illustrated in Figure 3, where distinct colors are used to indicate the functionalities assigned to each framework. The process begins with the monitoring of various metrics from resource partitions. These VNF compute metrics (e.g., CPU utilization) fluctuate over time due to the dynamic nature of the underlying cellular traffic. A self-healing loop leverages this data to first train machine learning (ML) models using historical datasets. During the inference phase, real-time samples are fed into the trained models to forecast future server load. Based on SLA or policy-driven thresholds configured via the G-OSS, prediction-based checks are then carried out to detect potential SLA violations or identify upcoming anomalies in network services (e.g., when an NS's total load approaches a predefined threshold). To proactively address such issues, the ROs execute corrective actions (such as dynamically allocating additional compute resources) to adapt to anticipated increases in demand.

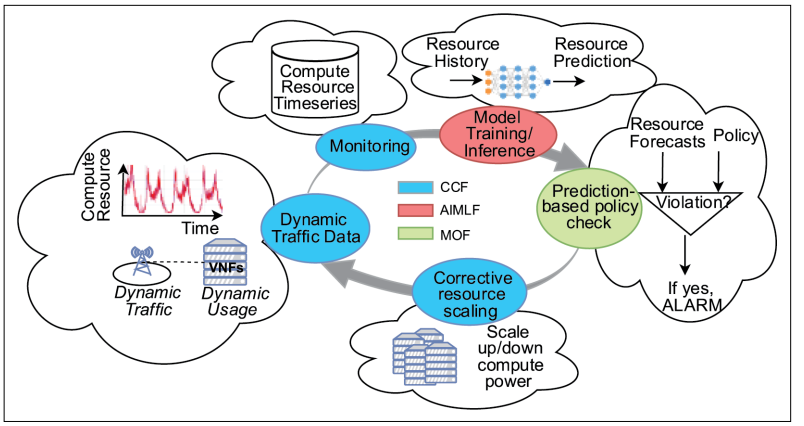

FIGURE 3. End-to-end resource allocation and self-healing workflow across the 6G-Cloud architecture where color coding highlights functional responsibilities within the orchestration loop (green: MOF, red: AIMLF, blue: CCF).

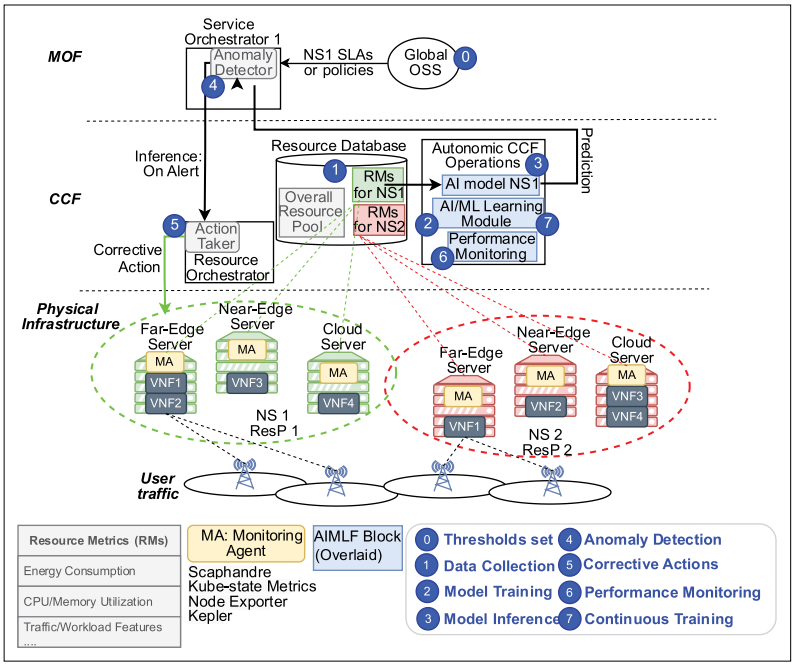

FIGURE 4. Step-by-step procedure for dynamic server activation, detailing AI-driven inference, anomaly detection, and corrective orchestration actions.

Resource Allocation Workflow

Figure 4 illustrates the system architecture, highlighting the key 6G-Cloud components involved in the proposed workflow. At the physical infrastructure layer, user devices connect through radio units to edge/cloud servers hosting the VNFs of various NSs. Resource partitions (shown in green for NS 1 and red for others) represent isolated allocations of computing resources. Monitoring Agents (MAs) are deployed on each server to collect time-series Resource Metrics (RMs). The CCF layer, assisted by ROs, manages the

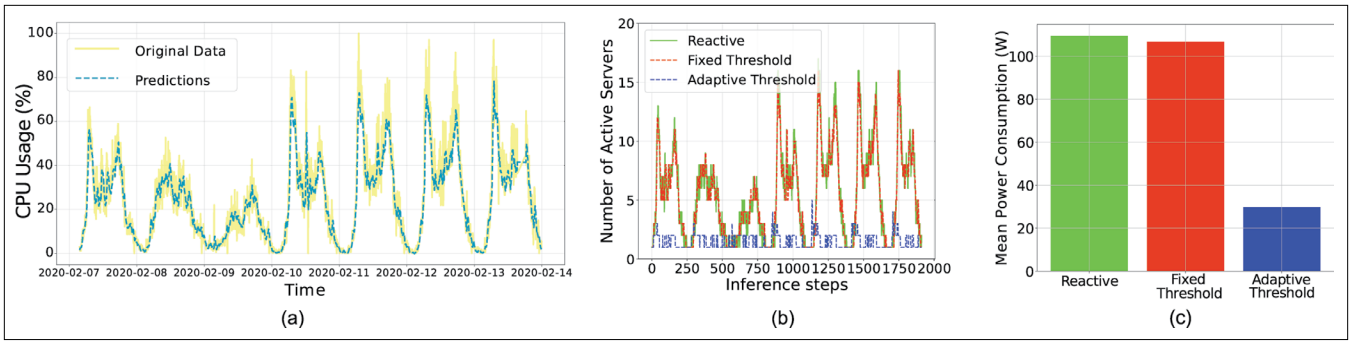

FIGURE 5. a) Actual versus predicted CPU usage over one week of data. b) Number of active servers obtained from three anomaly detection methods during 1000 inference steps. c) Corresponding power consumption for the three anomaly detection methods.

infrastructure, while the Autonomic CCF includes embedded AIMLF components (in blue). At the top, the MOF layer handles NS orchestration and SLA enforcement, including the detection of overloads and the triggering of scaling actions via the Ros. The proposed resource allocation workflow is unfolded as follows:

Step 1 The process begins by collecting the CPU utilization time series for NS 1's resource partition (ResP), which is then stored in the CCF's Resource Database along with contextual metadata, such as server type and location, to ensure data consistency.

Step 2 Historical CPU utilization data is used to train an ML model that predicts future CPU load for NS 1 over a specified lookback window. This offline training takes place within the AI/ML Learning Module, either in the CCF or on a central server, using data retrieved from the CCF Resource Database via the RO.

Step 3 The trained ML model performs inference on real-time data from the Resource Database, providing CPU utilization predictions to the MOF's Service Orchestrator (SO).

Step 4 The MOF's Anomaly Detector analyzes the predictions to detect potential CPU overloads, using thresholds defined by SLA or policy rules from the Global OSS.

Step 5 If the Service Orchestrator (SO) detects a CPU overload or SLA violation, the Action Taker triggers corrective actions such as dynamic scaling, activating additional servers or deactivating underutilized ones.

Step 6 The Performance Monitoring Client in the Autonomic CCF collects model inputs, predictions, and actual outputs, forwarding them to the PMon Server in the AIMLF. The server monitors prediction accuracy and triggers model retraining if performance degrades.

Step 7 The AI/ML Learning Module periodically retrieves updated Resource Metrics (RMs) from the Resource Database for retraining. This continuous process allows models to adapt to evolving workloads and conditions, with the latest versions stored in the AIMLF Models Database to maintain accuracy over time.

Preliminary Results

To provide initial insights[1], we evaluate the proposed dynamic server scaling use case against three different methods for anomaly detection:

Reactive (Without ML): A simple approach that adds a new server whenever CPU usage exceeds a fixed threshold (we use a green approach previously used in the literature [15]). It reacts only after the load increases and does not adapt to changing traffic conditions.

Proactive (Fixed Threshold): This method uses an LSTM modelto predict future CPU utilization. A new server is added if the predicted load for the next time step is expected to exceed the fixed threshold, enabling earlier responses to potential overloads.

Proactive (Adaptive Threshold): An advanced version that continuously updates its threshold based on recent predicted loads using a moving average. This allows the system to adapt to evolving workload patterns, activating or deactivating servers dynamically as predictions fluctuate.

The LSTM model was trained on real traffic traces from the Torino city network to forecast CPU utilization and detect potential resource-demand anomalies. The optimized model, featuring five hidden layers of 100 neurons each, was trained for 100 epochs on 16-sample sequences to predict CPU load five minutes ahead, using mean squared error (MSE) as the loss function. For power consumption modeling, each active server was assumed to have a fixed power cost, with parameters derived from real measurements in the TPC Power Consumption Database[2] using Intel NUC6i7KYK servers.

The effectiveness of the trained LSTM is illustrated in Figure 5(a), which compares the actual and predicted CPU usage over one week of data. The model accurately predicts traffic patterns, as evidenced by an MSE below 10^{-3}. Figure 5(b) shows that while the reactive and proactive-fixed threshold methods exhibit similar server usage due to accurate predictions and a shared static threshold, the proactive approach with an adaptive threshold significantly reduces the number of active servers by dynamically adjusting the threshold based on current traffic, which enables more flexible and stable server management. Finally, Figure 5(c) shows that an adaptive threshold achieves substantial power savings compared to the other methods by maintaining a lower number of active servers, demonstrating that dynamically adjusting the threshold is an effective strategy for significantly reducing power consumption.

Discussion

The proposed OaaS framework introduces important considerations in terms of security, privacy, and trustworthy AI-driven orchestration.

[1] Note that the current implementation represents an early-stage prototype of the proposed OaaS architecture, which is still under active development within the 6G-Cloud project.

[2] https://www.tpcdb.com/

Within this context, the AIMLF processes extensive volumes of telemetry and monitoring data that may include sensitive operational information. To safeguard data privacy, the framework employs federated learning and anonymized data sharing strategies, thus avoiding the central aggregation of raw data. Moreover, a component named Network Function Manager (NFM) acts as an intermediary that abstracts and sanitizes raw data before it reaches the orchestration layer, further reducing privacy risks. AI models integrated into the AIMLF are trained using privacy-preserving and domain-isolated datasets collected across the CCF. During inference, these trained models operate within secure execution environments embedded in orchestration components such as the MOF or the CCF, minimizing latency and exposure of sensitive information.

Each AI-driven orchestration decision is evaluated through a multi-layered safety mechanism before deployment. Actions are first validated in high-fidelity DT simulations, checked against predefined policy constraints, and then deployed incrementally using controlled rollout strategies. Any operation identified as unsafe or potentially disruptive is automatically blocked. This closed-loop, DT-validated approach, combined with multi-layer data abstraction, ensures that orchestration actions remain compliant with SLAs and do not compromise network stability or performance.

Furthermore, the distributed frameworks (MOF, CCF, and AIMLF) add layers of functional and data abstraction to ensure orchestration intelligence operates on contextualized, non-sensitive data. Their secure communication interfaces, reinforced with encryption, integrity checks, and trusted execution environments (TEEs), protect against tampering or malicious access.

Conclusion and Future Work

The MOF offers an innovative solution for service and resource orchestration in 6G networks, tackling challenges of scalability, flexibility, and efficiency. It enables dynamic, policy-driven, and low-overhead management. Through the integration of AI, digital twins, modular design, and programmable SOs, the MOF supports automation, proactive fault detection, and energy-efficient operations. Its multi-domain and multi-stakeholder compatibility ensures seamless interoperability, making the MOF a key enabler of intelligent and sustainable orchestration within the 6G-Cloud architecture. Future work within the 6G-Cloud project will focus on the development of a full-scale prototype and its extensive empirical validation against state-of-the-art benchmarks.

References

[1] J. Péerez-Valero et al., "AI-driven orchestration for 6G networking: The Hexa-X vision," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2022, pp. 1335–1340.

[2] M. A. Habibi et al., "The architectural design of service management and orchestration in 6G communication systems," in *Proc. IEEE INFOCOM Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Hoboken, NJ, USA, May 2023, pp. 1–2.

[3] S. Kerboeuf et al., "Design methodology for 6G end-to-end system: Hexa-X-II perspective," *IEEE Open J. Commun. Soc.*, vol. 5, pp. 3368–3394, 2024.

[4] L. E. Chatzieleftheriou et al., "Towards 6G: Architectural innovations and challenges in the ORIGAMI framework," in *Proc. Joint Eur. Conf. Netw. Commun. 6G Summit (EuCNC/6G Summit)*, Antwerp, Belgium, Jun. 2024, pp. 1139–1144.

[5] K. Ramantas et al., "6G-BRICKS: Building reusable testbed infrastructures for cloud-to-device breakthrough technologies," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Kuala Lumpur, Malaysia, Dec. 2023, pp. 751–756.

[6] C. Cicconetti et al., "EDGELESS: A software architecture for stateful FaaS at the edge," in *Proc. 33rd Int. Symp. High-Perform. Parallel Distrib. Comput.*, Jun. 2024, pp. 393–396.

[7] *SNS JU Project ADROIT6G, Deliverable 2.2 System Architecture-Initial*. Accessed: Jan. 24, 2026. [Online]. Available: https://adroit6g.eu/wpcontent/uploads/2024/09/ADROIT6G_D2.2_Architecture_v2.0.pdf

[8] NS JU Project ETHER, Deliverable 2.4. (Aug. 2024). *Final Report on ETHER Network Architecture, Interfaces, and Architecture Evaluation*. [Online]. Available: https://www.ether-project.eu/wp-content/uploads/sites/100/2024/10/ETHER Deliverable_D2.4_V1.0_final.pdf

[9] (Jun. 2024). *SNS JU Project DETERMINISTIC6G, Deliverable 1.1, Use Cases and Architecture Principles*. [Online]. Available: https://deterministic6g.eu/images/deliverables/DETERMINISTIC6G-D6.1-v0.3.pdf

[10] (Sep. 2024). *SNS JU Project 6G-Cloud, D2.1, Deliverable 2.1, Use Case Analysis, KPIs and Requirements to Service-Oriented Architecture Design*. [Online]. Available: https://www.6g-cloud.eu/index.php/services/service02/#

[11] *5G; Management and Orchestration; Architecture Framework*, Standard ETSI, (TS) 28.533, Jan. 2021. [Online]. Available: https://www.etsi.org/deliver/etsi ts/128500128599/128533/16.06.0060/ts_128533v160600p.pdf

[12] *Network Functions Virtualisation (NFV); Management and Orchestration; Report on Architectural Options*, Standard GS NFV-IFA 009, ETSI, Technical Specification (TS), Jul. 2016. [Online]. Available: https://www.etsi.org/deliver/etsi_gs/nfv-ifa/001099/009/01.01.0160/gs_nfv-ifa009v010101p.pdf

[13] S. F. Wassie, A. Di Maio, and T. Braun, "Deep reinforcement learning for context-aware online service function chain deployment and migration over 6G networks," in *Proc. 40th ACM/SIGAPP Symp. Appl. Comput.*, Mar. 2025, pp. 1361–1370.

[14] A. Giannopoulos et al., "AI-driven self-healing in cloud-native 6G networks through dynamic server scaling," in *Proc. IEEE 11th Int. Conf. Netw. Softwarization (NetSoft)*, Jun. 2025, pp. 43–48.

[15] J. Perez-Valero et al., "Performance trade-offs of auto scaling schemes for NFV with reliability requirements," *Comput. Commun.*, vol. 212, pp. 251–261, Dec. 2023.

Biographies

Jesus Perez-Valero (jesus.perezvalero@um.es) received the Ph.D. degree from the Universidad Carlos III de Madrid (UC3M) in 2024. He is currently a Post-Doctoral Researcher and a Lecturer with the Universidad de Murcia (UMU).

Gines Garcia-Aviles (gigarcia@um.es) received the Ph.D. degree in telematics engineering from the IMDEA Networks Institute, University Carlos III of Madrid. He is currently a Post-Doctoral Researcher and a Lecturer with the Universidad de Murcia (UMU).

Antonio Skarmeta (Senior Member, IEEE) (skarmeta@um.es) received the Ph.D. degree in computer science from the University of Murcia, Murcia, Spain. He has been a Full Professor and the Head of the Research Group ANTS, University of Murcia, since its creation in 1995. He has worked on and coordinated different European Union Research Projects.

Tao Chen (tao.chen@vtt.fi) received the Ph.D. degree in telecommunications engineering from the University of Trento, Italy, in 2007. Since then, he joined the VTT Technical Research Centre of Finland, and currently, he is a Senior Researcher in the connectivity research area. He is a Docent (Adjunct Professor) at the University of Jyväskylä, a Project Coordinator of the EU H2020 COHERENT Project, and a Board Member of the EU 5G PPP Steering Board. His current research interests include software-defined networking for 5G mobile networks, massive IoT in 5G, and dynamic spectrum access.