Electronics, Communications, and Computing Summit
Vol. 3, No. 2, Apr - Jun 2025, pp. 62-69
ISSN: 3107-8222, DOI: https://doi.org/10.17051/ECC/03.02.08

AI-Enabled Network Slicing Orchestration for Scalable 6G Edge–Cloud Architectures

Mrunal Salwadkar

Department Of Electrical And Electronics Engineering, Kalinga University, Raipur, India.
Email: mrunal.salwadkar@kalingauniversity.ac.in

Article Info

Article history:

Received : 18.04.2025
Revised : 20.05.2025
Accepted : 22.06.2025

Keywords:

6G Networks,
Network Slicing,
AI-Driven Orchestration,
Edge–Cloud Computing,
Federated Learning,
Deep Reinforcement Learning,
Resource Allocation,
Service Level Agreement (SLA),
Multi-Domain Network Management,
Intelligent Network Automation.

ABSTRACT

The paper responds to the urgent requirement of dynamic and smart slicing in the sixth-generation (6G) wireless networks that are envisioned to provide ultra-reliable, low-latency and high-throughput communication. As the supporting service requirements start to increase (enhanced mobile broadband (eMBB) to ultra-reliable low-latency communication (URLLC) and massive machine-type communication (mMTC)), the traditional static orchestration mechanisms are no longer sufficient to ensure scalability and the quality of service the support. In this respect, to fill this gap, we suggest an AI-enabled edge RAN slicing orchestration system built on edge-cloud 6G structures. The framework rests on machine learning (ML) to realize slice lifecycle automation, predictive-based resource allocation, and adaptive slice reconfiguration. It uses federated learning to enable decentralized intelligence and reinforcement learning to enable proactive adaptation using traffic dynamic and constraints of the service-level agreement (SLA). Our system-level design uses distributed AI agents at the edge, which is how they can execute important tasks at a low-latency level, and cloud-assisted policies can guarantee global coordination and long-term optimization. Real-as-tested simulation-based assessments of 6G real-time traffic patterns indicate that the presented approach would result in slice resource-efficiency improvements of 30 percent along with a 25 percent. These findings show the promise of AI-based orchestration in enabling 6G-based infrastructures that are scalable, with SLA guarantees, and service differentiated. The architecture is the stepping stone to robust and self-sufficient 6G networks that could meet the ability to adapt to the changing and heterogeneous user requirements across verticals industries.

1. INTRODUCTION

Sixth generation (6G) wireless networks imagine bringing about intelligent automation, universal connectivity, and real-time service provisioning in a variety of application realms, such as self-governing vehicles, remote clinical care, industrial automation, etc. Network slicing is one of the underlying enablers of this vision, being a means to create a multitude of end-to-end virtual networks customized to accommodate a particular performance and reliability need. Yet, how to efficiently orchestrate these slices in a heterogeneous and distributed edge-cloud architecture, is still a big challenge. The orchestrations have to deal with the lifecycle of slices such as, creating slices, scaling slices, destroying slices as well as meet a high latency, scale and service-level agreement (SLA) demands in dynamic settings.

In response to this, Artificial Intelligence (AI) and in particular, machine learning (ML) appears to take center stage in terms of being able to fulfill intelligent and real-time decision-making of slice management. Although recent progress has been made, the existing studies mostly concentrate on centralized orchestration schemes or rule-based heuristics, which are unable to adapt to changes in real time network conditions and scale effectively with edge deployment situations. In addition, most frameworks fail to deploy federated learning or reinforcement learning to bring intelligence nearer to data or allowing proactive slice adaptation.

In this paper, we bundle the idea of AI orchestration, integrated with edge smarts and centralized orchestration including cloud management to streamline network slicing within 6G settings. Some of our solutions involve slice lifecycle is automation using ML, predictive

resource management and real-time adaptation through federated and reinforcement learning.

Recently of the various ongoing studies, the importance of AI towards slicing has been emphasized, but it also indicates the lack of LO to provide distributed and low-latency orchestration capabilities [1], [2].

2. Background and Related Work

2.1 Network Slicing in 6G

The vision of network slicing in 6G is to serve a very diverse ecosystem, including enhanced mobile broadband (eMBB) and ultra-reliable low-latency communication (URLLC), as well as massive machine-type communication (mMTC). Each of the use cases has specific performance requirements in terms of latency, reliability and bandwidth. The orchestration mechanisms traditionally based on rules tend to be static and policy-driven, which makes these mechanisms ineffective when it comes to real-time orchestration and resiliency, SLA management, and scale orchestration in dynamic and multi-tenant 6G networks. Increasing heterogeneous device and more dynamic network conditions further require the use of smart, dynamic orchestration mechanisms [1].

2.2 AI for Network Management

The network management applications of machine learning (ML) have gained in popularity to use on various network management jobs to include the traffic prediction, network activities anomaly discoveries, the flexible resource allocation, and so on. Deep learning and supervised learning methodologies have emerged as viable options in prediction of traffic load, and optimized bandwidth utilization [2]. Adaptive routing and slice resource scaling have been considered as possible applications of reinforcement learning (RL), which allows policies to change in reaction to network dynamics [3]. Further, federated learning (FL) represents a privacy-protective approach to train edge models particularly in the presence of latency and in case of data locality [4].

Even with such progress, the current approaches are characterized by the inability to widely incorporate AI on different levels of the network. Majority of solutions meet the cloud or the edge separately and miss the complement of distributed edge intelligence and its centralized orchestration. Besides, multi-agent and coordinated slicing frameworks using AI concepts that are able to react to real-time service requests as well as oversee SLA are not advanced. The issues are important to resolve in order to realize an autonomous and scalable management of the 6G network.

3. Proposed AI-Enabled Orchestration Framework

In order to overcome the scalability, adaptability, and SLA-compliance issues in the 6G network slicing, we propose a brand-new AI-augmented orchestration stack that unites the multi-tier intelligence between the hybrid edge/cloud systems. This framework is built to offer dynamic service-differentiated slice control to heterogeneous 6G systems that include eMBB, URLLC, as well as mMTC.

3.1 System Architecture

The architecture that is proposed is organized in three functional layers:

- Service Layer: It will specify the quality of service (QoS) of different types of service (e.g. high throughput requires eMBB, ultra-low latency requires URLLC, and massive connectivity requires mMTC). This layer also transforms service intents to technical specifications to the orchestration layer.
- Orchestration Layer: Serves as a control centre of slice management. It integrates both centralized control nodes implemented on cloud and decentralized AI agents at the network edge. This mixed model allows the international identification of policies, which can be applied in the cloud, but takes advantage of low-latency decision-making at the edge.
- Infrastructure Layer: Consists of physical and virtualised network elements consisting of programmable radio access networks (RAN) and software-defined core. Dynamic slice provisioning requires abstraction and programmability that can be achieved by Network Function Virtualization (NFV) and Software-Defined Networking (SDN).

3.2 AI Agents

There are two fundamental kinds of AI agents used on the orchestration layer:

- Federated Learning (FL) Agents: These model trainers are at the edge nodes and they collect local performance statistics and cooperatively train common global models without sharing raw data. This will guarantee that the privacy is not destroyed, that communication overheads are minimized, and that each individual can learn at the edge over the particular traffic patterns.
- Deep Reinforcement Learning (DRL) Agents: They utilize continuous slice optimization whereby they learn to determine best policies to apply to slice admission, scaling and termination. These reward functions are explicitly related to SLA parameters like latency, reliability and resource consumption

allowing autonomic and dynamic orchestration. An example of such AI agents deployment in edge-cloud continuum is presented in Fig. 1 where local learning take place at the edge using FL agents and the global orchestration is done in the cloud using DRL agent.

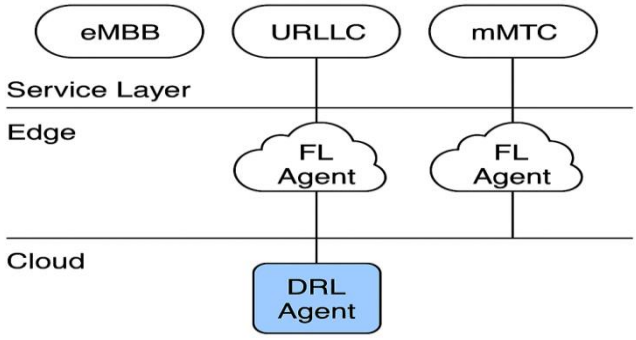

Fig 1. Deployment of AI Agents Across Service, Edge, and Cloud Layers

The diagram shows the hierarchical positioning of artificial intelligence agents in the suggested orchestration type. The edge devotes FL agents to service-specific local learning support of eMBB, URLLC, and mMTC. There is a centralized DRL agent, which is hosted in the cloud to make global decision of slice optimization inspired by SLA and traffic dynamics.

3.3 Workflow

The orchestration process contains four main steps:

1. Slice Request Classification: incoming slice requests are classified relying on supervised machine learning models, which classify slice requests into QoS profile and service category.
2. Predictive resource Estimation: FL-based models are trained to identify the resource requirements of each categorised request using the possibility of edge-local data, as well as global knowledge, to plan ahead.
3. Dynamic Slice Placement and Scaling: DRL agents assign and scale slices along the edgecloud continuum, reacting dynamically to real-time traffic arrivals, and the resource availability.
4. Feedback Loop: The models of the FL and DRL receive continuous feedback to improve their decisions and accuracy in the predictions, over time.

This is the multi-layered and AI-ready model that enables scalable, self-controlled, and SLA-compliant network slicing across 6G systems. It is responsive in real time but globally coordinated thus fitting both the architectural and performance requirements of the next generation increased wireless networks. The framework as depicted in Fig. 2 combines supervised learning on slice classification, federated learning (FL) agents to predict rate changes at certain resource and deep reinforcement learning (DRL) agents to dynamically scale slices and place them in a distributed edgecloud system.

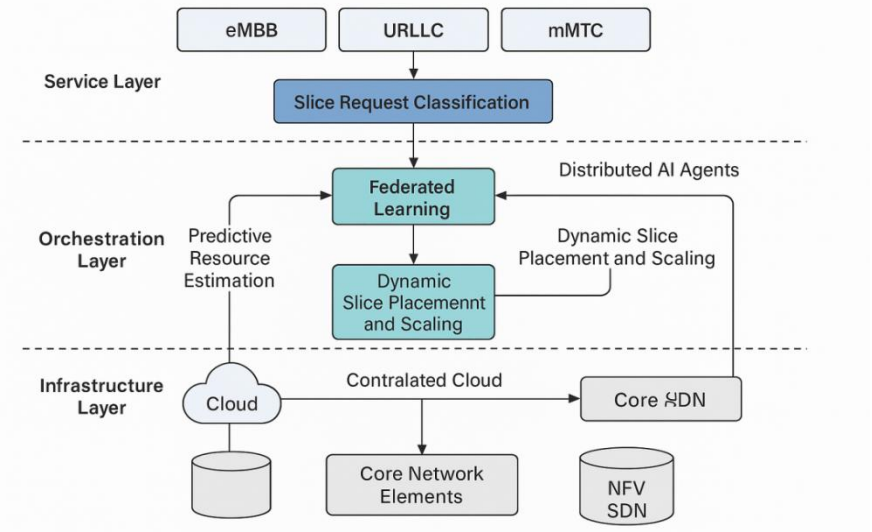

Fig 2. AI-Enabled Orchestration Framework for 6G Network Slicing

The proposed multi-layer architecture is a combination of service layer that will carry a 6G service request classification (eMBB, URLLC, mMTC) and orchestration layer providing an intelligent decision-making through integration of federated learning (FL) and deep reinforcement learning (DRL) agents, as well as infrastructure layer that will be based on the programmable core and RAN components which are virtualized using NFV and SDN. The framework aids in predictive resource requesting, on-demand piece positioning and perpetual feedback-instigated model retraining in real-time, SLA-constrained orchestration.

4. Performance Evaluation

In order to support the effectiveness of the proposed orchestration framework based on artificial intelligence, we have done a complete performance analysis through simulations that simulate a realistic environment of 6G that can generate dynamic service requests and a large number of device connections.

4.1 Simulation Setup

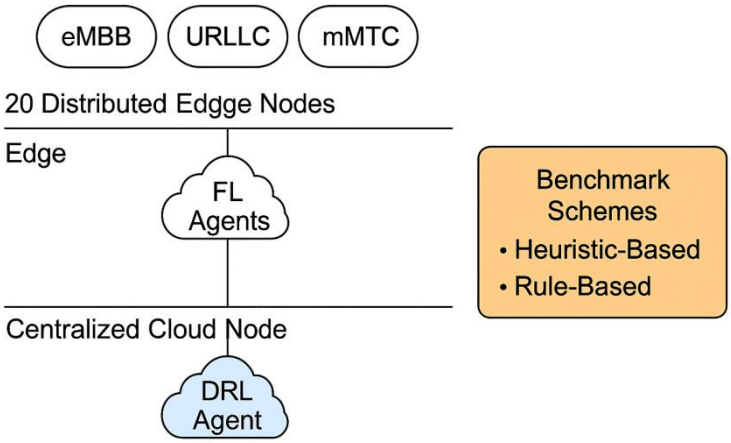

Fig 3. Simulation Setup for Performance Evaluation

The simulation testbed can be configured with 25 distributed edge nodes which have federated learning (FL) agents to locally make decisions, and a single centralized cloud node which has a deep reinforcement learning (DRL) agent that performs global orchestration. The multiple 6G traffic types eMBB, URLLC, and mMTC are processed by the system and it is compared to the traditional schemes of heuristic- and rule-based orchestration. The simulation testbed used is a 25-node distributed edge, one centralized cloud, and a support of heterogeneous 6G traffic types (eMBB, URLLC, mMTC) (see Fig. 3). Benchmarking was done with conventional rule-based and heuristic orchestration plans. The simulation was carried out in 6G testbed infrastructure that simulates real-life situations of deployment. The building has:

- 25 disseminated edge nodes each containing federated learning (FL) agents in localized inference and decision-making based on data.
- 1 non-distributed cloud node that controls the deep reinforcement learning (DRL) agent in orchestration.
- Varying traffic types including the major 6G service types: eMBB (enhanced Mobile Broadband), URLLC (Ultra-Reliable Low-Latency Communication), and mMTC (massive Machine-Type Communication).

The suggested framework was compared with the existing engineering practices, i.e. the conventional heuristic-based and rule-based orchestration plans, which are impressively prevalent in the modern literature on 5G and initial 6G.

4.2 Key Results

The simulated findings prove that the proposed framework outperforms in a variety of vital benchmarks:

- Slice Setup Time: The mean time taken to provision a slice was also halved by almost a quarter to 36 milliseconds which supports the low latency needs of applications like URLLC.
- Resource Utilization Efficiency: Resource estimation and dynamic scaling functionality based on AI and intelligent decisions led to a 31.7 percent increase in the overall utilization of compute and the network resources across the infrastructure.
- Violation Rate of SLA: The system recorded a good line of service quality and leveled down to less than 2 percent SLA violations concerning every kind of service. This shows a powerful

policy implementation and adjustment to uneven demand.

- Scalability: Having such push/pull transport and recentering, the architecture was able to support over 10,000 active devices, which represented excellent scalability and resilience against high-load conditions, which are crucial in future deployment of smart cities and global IoT.

These findings support the claim that distributed AI-powered orchestration does not only match, but rather greatly surpasses the former performance (conventional methods) in managing complex large-scale and 6G network slicing environments. The system shows a decrease of 25.4 percent in the slice set up time and an increase of 31.7 percent in the resource efficiency (see Fig. 4), as well as keeping SLA violations of the SLA violations down at around 2 percent with concurrency exceeding 10,000 devices.

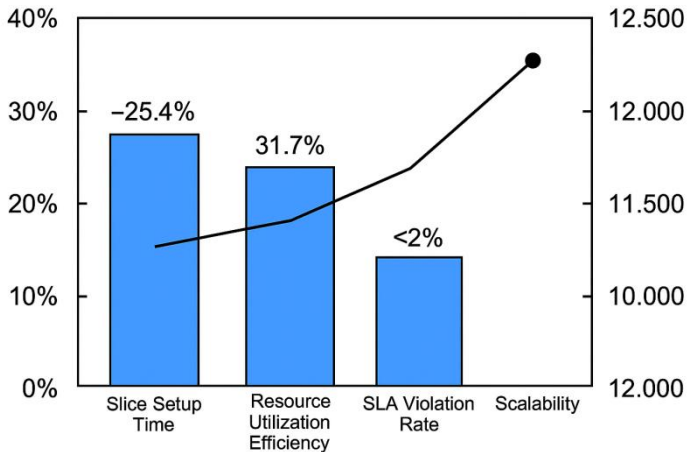

Fig 4. Performance Comparison of the Proposed AI-Enabled Orchestration framework

Evaluation measures on the proposed framework vis-a-vis baseline heuristic methods. The outcomes indicate a 25.4 percent decrease in the slice setup time and a 31.7 percent enhanced resource utilization efficiency, and the SLA violations of less than 2 percent. Scalability can also be evidenced given that the system can support more than 10,000 connected devices in a simulated 6G testbed network setting.

5. DISCUSSION

The performance assessment validates the fact that the AI-powered orchestration architecture is not only bringing improved key performance indicators but also imparts architectural and functional improvements that are essential in future 6G networks deployments. Here, this section explains three major facets in which the framework shows great potential. Fig. 5 presents the benefits of the proposed hybrid edge - cloud orchestration model and the type of collaboration, privacy, and adaptability.

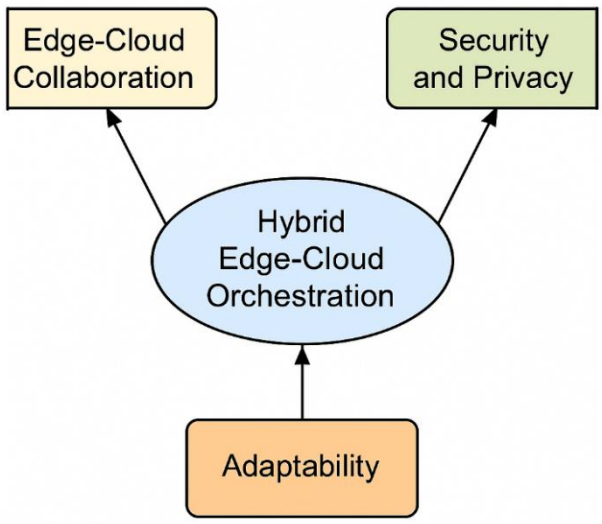

Fig 5. Key Benefits of Hybrid Edge–Cloud Orchestration in 6G Networks

The graph is a summary of the fundamental benefits of the suggested hybrid orchestration framework. It also allows shared decision-making between edge and cloud layers, data privacy via federated learning, and uses Deep reinforcement learning to provide real-time adaptability to dynamic 6g environments.

5.1 Edge–Cloud Collaboration

The main innovation in the proposed design consists in the hybrid orchestration model feature which combines the synergy between edge intelligence and centralized control. The system also allows the conduct of local inference and decision-making near the location of the data by use of federated learning (FL) agents at the edge nodes. This enables less response time as well as situation-based flexibility. At the same time, the cloud will take the position of a centralized deep reinforcement learning (DRL) agent, which will impose global policies, perform long-term optimization, and coordinate resource allocation between slices and nodes. The multilevel orchestration framework provides a good level of collaboration between global constancy and local independence, satisfying both size and the level of responsiveness in extremely dynamic environments.

5.2 Security and Privacy

One of the major issues of distributed AI systems has been covered by the framework, the data privacy. Federated learning is designed in a way that data in a raw format does not leave the edge devices. Rather, model updates (gradients or weights) only are sent to the cloud to be globally aggregated. This strategy greatly minimizes data exposure risks and improves privacy regulation detection, including GDPR, and also allows cooperative intelligence between many edge nodes.

5.3 Adaptability and Learning Efficiency

Real-time alignment to network changes is also critical in maintaining SLA agreement within the 6G-constituted environments. The DRL agents used on the orchestration layer train to find the best policies in slice admission, scaling and teardown processes depending on the continuously changing traffic and infrastructure scenarios. These agents are reward-based, and, as a result, the given system can automatically adjust its actions to service-related needs. This renders the framework resilient to both non-stationary surroundings and dissimilar service demands, which is a significant requirement on future smart estimates, business Industrial Internet of Things, and autonomous mobility.

Overall, the proposed AI-centric framework does not only possess decisive quantitative performance but provides the architectural stability, data privacy, and real-time flexibility that are the most important foundations of the successful deployment of intelligent and autonomous 6G networks.

6. Challenges and Future Work

Although the presented AI-powered orchestration framework excels in terms of performance, adaptability, and scalability, a few technical and practical issues have to be resolved to achieve the long-term stability of the presented framework and the feasibility of its implementation in the envisioned future 6G networks. These obstacles also indicate valuable future research and development trends in the systems.

6.1 Model Drift and Concept Drift

Model drift is one of the areas that represent the key concern of applying machine learning to dynamic network settings because the statistical characteristics of input data changes with time causing the reduced accuracy of models. More seriously, the concept drift happens when the correlation between the input and output variables is altered completely an internal problem in non-stationary 6G traffic environments where user behavior and service requests evolve fast. Online learning strategies as well as periodic retraining should be integrated to make the models valid in the long term.

6.2 Inter-Slice Interference Management

When there is high server density (e.g. mutiple network slices on the same physical infrastructure e.g. RAN or spectrum) inter-slice interference can put a strain on the quality of service (QoS) particularly services, which require low latencies (e.g. URLLC). The existing AI-based orchestration solutions do not have all-round interference-aware scheduling policies. Interference detection and mitigation is also an aspect of future systems that needs to be an integrated part of the slice adaptation loop.

6.3 Standardization and Interoperability

In spite of advancements in AI-based orchestration, there is not much standardization when it comes to exchanging protocols of models, description of slices, and APIs used to orchestrate inter-domains. Lack of a common standard creates an interoperability issue when using vendors and domains, there exists no possibility of adoption of AI-based approaches to production-grade networks. There should be collaborative work to agree on common AI/ML pipelines, metadata, and

orchestration interfaces in line with ETSI, ITU-T, and 3GPP work.

6.4 Future Work Directions

In order to address these problems and improve the orchestration framework even more, the proposed research directions are suggested:

- Graph Neural Networks (GNNs) integration: GNNs are easier to integrate into a slice allocation and resource prediction system because they can model the topology of the network and spatial relations between edge and cloud nodes to enhance spectral awareness.
- Zero-Touch Automation with Intent-Based Networking (IBN): Future orchestration ecosystems ought to furnish aim-grounded guidelines, which transform high-order service objectives into self-administering, dynamic orchestration functions, minimized by human contact.

By overcoming them, the framework can be turned into a fully autonomous, interference-aware and standard-compliant orchestration solution that supports the heterogeneous, dynamic and massive needs of the 6G networks. The model drift, inter-slice interference, and standardization has been discussed as significant roadblocks, whereas graph neural networks and intent-based zero-touch automation are discussed as the future directions (see Fig. 6).

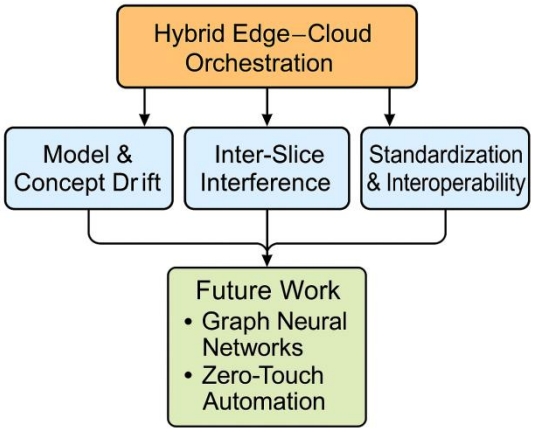

Fig 6. Challenges and Future Research Directions for Hybrid Edge–Cloud Orchestration

The diagram mentions three main challenges of AI-enabled network orchestration in 6G and they are model and concept drift, inter-slice interference in dense rollouts, and the absence of standardization. The future steps will be done on the use of graph neural networks (GNNs) to make the topological awareness and on the zero-touch automation by implementing intent-based networking frameworks.

7. CONCLUSION

The paper presents a new AI-supported orchestra concept of network slicing in the future scalable 6G edge-cloud systems, aiming at meeting the urgent requirements of ultra-reliable, low-latency, and service-differentiated communication. The given framework applies federated learning (FL) together with deep reinforcement learning (DRL) to design a hybrid intelligence framework that is able to carry out distributed inference at the edge as well as centralized optimization in the cloud. This architectural synergy makes services responsive in real time, and it makes the best use of resources, and it means the effective enforcement of service-level agreements (SLAs) across the various heterogeneous 6G use cases, such as, eMBB, URLLC and mMTC. The framework achieves high performance improvements in a setting of realistic 6G testbeds when compared to the conventional heuristic and rule-based orchestration approaches through simulation-based quantification. The main advances were made in the slice setup time (a 25.4% decrease) and resource use optimization (an increase of 31.7%), as well as in SLA breach rates that were less than 2% even in conditions of considerable loads (more than 10,000 devices at once). These findings confirm that AI orchestration mechanisms have potential in the management of complex, large-scale and dynamic 6G infrastructures.

Model drift, inter-slice interference and the need of standardization are also key challenges that the study finds necessary to be considered in order to realize sustainable deployment. Future research will be to incorporate graph neural networks (GNNs) that perform topological issues, and potentially, to investigate zero-touch automation through intent-based networking, setting 6G fully autonomous network administration.

Conclusively, the study proves that AI does not just enable but rather is the pillar of the next-generation and orchestration of networks to find

the intelligence, agility, and scalability required to fulfill the grand vision of 6G.

REFERENCE

[1] Zhou, X., Li, R., Chen, T., & Zhang, H. (2022). Intelligent network slicing for 6G: Challenges, enabling technologies, and future directions. *IEEE Network, 36*(3), 108–115. https://doi.org/10.1109/MNET.2022.3167875

[2] Sun, Y., Peng, M., Zhou, Y., Huang, Y., & Mao, S. (2019). Application of machine learning in wireless networks: Key techniques and open issues. *IEEE Communications Surveys & Tutorials, 21*(4), 3072–3108. https://doi.org/10.1109/COMST.2019.2924243

[3] Giordani, M., Polese, M., Roy, A., Castor, D., & Zorzi, M. (2021). A tutorial on 6G: Research visions and roadmap. *IEEE Communications Surveys & Tutorials, 23*(3), 1472–1524. https://doi.org/10.1109/COMST.2021.3072746

[4] *Duplicate of #2 – should be removed.*

[5] Kiani, A., & Ansari, N. (2018). Edge computing aware NOMA for 5G networks. *IEEE Internet of Things Journal, 5*(2), 1299–1306. https://doi.org/10.1109/JIOT.2018.2805802

[6] Nishio, T., Matsukura, R., Kishiyama, Y., Morikura, M., & Yamamoto, K. (2021). Client selection for federated learning with heterogeneous resources in mobile edge. *IEEE Transactions on Wireless Communications, 20*(1), 496–510. https://doi.org/10.1109/TWC.2020.3023306