Available online at www.qu.edu.iq/journalcm
JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS
ISSN:2521-3504(online) ISSN:2074-0204(print)

Intelligent Network Slicing Management Using Microservice-Based Software Architecture for 5G and Beyond Communication Systems

Karar Talal Hamzah*

College of physical education and sport sciences, University of Al-Qadisiyah, Iraq.Email: sporteacher11@qu.edu.iq

ARTICLE INFO

Article history:
Received: 14 /02/2026
Rrevised form: 22 /03/2026
Accepted : 24 /03/2026
Available online: 30 /06/2026

Keywords:

5G Networks,
Intelligent Orchestration,
Microservices Architecture,
Network Slicing,
SLA management.

ABSTRACT

Network slicing is a key enabling technology for flexible resource management in 5G and beyond communication networks. However, maintaining strict service level agreements (SLAs) while efficiently allocating resources across heterogeneous services remains a major challenge. This paper proposes an intelligent network slicing framework that integrates microservice-based architecture with an AI-Driven Slice Orchestration Engine (ASOE) for dynamic and SLA-aware resource management. The novelty of the proposed framework lies in combining AI-driven orchestration, microservice-based slicing architecture, and automated SLA monitoring within a unified management system. The framework was evaluated using a 5G-oriented simulation environment and compared with conventional monolithic and static slicing approaches. Experimental results show that the proposed method improves SLA compliance to 98.9% (compared to 88.4%), reduces URLLC latency from 9.4 ms to 3.9 ms, and increases resource utilization from 64.8% to 83.7%. These results demonstrate that intelligent orchestration significantly enhances scalability, efficiency, and service reliability in next-generation programmable network infrastructures.

https://doi.org/10.29304/jqcsm.2026.18.22653

1. INTRODUCTION

The recent accelerated development of 5G and beyond-5G (B5G/6G) communication systems has altered the paradigm established in network design by providing a key architectural concept of network slicing. Network slicing is a technology that allows the co-existence of heterogeneous services, e.g., enhanced mobile broadband, ultra-reliable low-latency communications, and massive machine-type communications, on a common physical infrastructure and ensures a variety of quality-of-service (QoS) and quality-of-experience (QoE) demands. In order to actualize this vision, modern cellular systems are becoming increasingly **software-centric**, adopting cloud-native technologies such as software-defined networking (SDN), network function virtualization (NFV), and service-based architectures, and we can state that cloud-native concepts, such as software-defined networking (SDN), network function virtualization (NFV), and service-based architecture (SBAs) are being implemented [1], [2], [3].

Openness, programmability, and intelligence are supported by recent standardization and research projects, including O-RAN and 6G-oriented architecture, as the key enabling factors to slicing-aware network management

∗Corresponding author Karar Talal

Email addresses: kararalshamery.ka@gmail.com

Communicated by 'sub etitor'

[4], [5]. In this respect, software architecture based on microservices have become an alternative to monolithic network management and orchestration systems. Microservices enhance scalability, resilience, and ongoing transformation of the network services by breaking the functionality of complex networks down into lightweight independently deployable services [3], [6]. A number of works have shown that containerized network functions controlled by Kubernetes is feasible in both 4G/5G core and transport networks [7], [8].

Simultaneously, the growing complexity of multi-slice surroundings requires the development of smart and self-governing management processes. The traditional or rule-of-thumb orchestration strategies cannot withstand extremely dynamic traffic patterns, severe service-level agreements (SLA), and cross-slice interference. Therefore, zero-touch network management, predictive resource allocation, and adaptive orchestration in 5G and 6G systems are actively being investigated with the help of artificial intelligence (AI) and machine learning (ML) [9], [10], [11]. The AI-inspired orchestration is also in line with new perspectives of agentic and autonomous networks that possess self-optimization and self-healing [11], [12].

This notwithstanding, some key issues have not yet been solved, especially when large-scale and multi-tenant network slicing deployments are concerned. The currently available solutions may not be in the form of strict slice isolation, fine-grained SLA enforcement, and systematic integration of the microservice-based architecture and intelligent orchestration contexts [13], [14]. Moreover, inter-slice interference and resource contention still worsen the guarantees of the services, notably during the bursts of traffic and edge-cloud heterogeneity [15], [16].

Existing network slicing control designs of 5G and beyond are largely based on monolithic or semi-distributed control designs with restricted flexibility. Although microservices have found use to deploy network functions, slicing control logic, SLA enforcement and orchestration intelligence are typically closely coupled or hard-configured [17], [18]. This leads to a number of limitations:

1. The lack of slice isolation and subsequent deterioration of the performance based on inter-slice interference under dynamic workload [9], [15].

2. The lack of scalability and capability to be flexible, with centralized or monolithic controllers being unable to adapt to the vastness of the multi-tenant world [6], [19].

3. Weak SLA awareness in which the SLAs become only as a dynamic parameter and not a runtime verifiable software contract [13].

4. The absence of smart coordination, which will allow managing resources actively and independently at edge-cloud infrastructures [10], [11].

These deficiencies demonstrate the necessity of a software-based, intelligent, and modular framework of network slicing management that brings together architectures of microservices with AI-orchestration and formal model SLA.

The main goal of the study is to develop and test a smart network slicing management architecture to 5G and beyond communication systems that use software architecture based on microservices and learning based orchestration. This paper makes the following key contributions:

1. MOSA: Microservice-Oriented Slicing Architecture: Our proposal has a completely sliced management architecture where slice admission, resource allocation, monitoring, and SLA enforcement will be deployed as independent microservices deployed through containerization and Kubernetes orchestrations, enhancing scalability and fault isolation [3], [7].

2. The formal Slice Requirement Modeling: Slice requirements are represented as software contracts to represent bandwidth, latency and reliability requirements allowing accurate SLA definition and execution enforcement of such requirements in comparison with traditional policy-based models [1], [13].

3. AI-Driven Slice Orchestration Engine (ASOE): A layer of AI-oriented orchestration is presented, which has access to real-time traffic and SLA telemetry and can dynamically redistribute resources and scale slices based on optimisation algorithms of learning techniques, in line with the zero-touch and autonomous network management paradigm [10], [11].

4. Inter- slice Interference Control Mechanism: To amplify the reliability of services in multi-tenant environments, we develop a computer-based isolation system that is grounded on adaptive resource limits and workload validation to identify and resolve SLA breaches brought about by inter-slice interactions [8], [15].

5. Extensive Performance Assessment: The suggested framework is tested on a 5G-based simulation platform and compared to monolithic and static slicing models and has a better adherence to the SLA, resource utilization, and scaling performance.

The originality of the suggested method is that it is a holistic combination of microservices, formal SLA modeling, and intelligent orchestration in an overall network slicing management platform. In contrast to the literature, which considers either microservice deployment [2], [6] or AI-based resource optimization in isolation [9], [10], the study presented here regards network slicing as a software-defined and contract-based system, which is autonomously controlled by decision-making. The proposed framework will bring the state of the art to scalable, resilient, and self-managing network slicing that can be used in 5G-based and future 6G ecosystems due to implementing containerized microservices in parallel with learning-based orchestration and SLA verification of run-time [4], [11].

The remainder of this paper is organized as follows. Section 2 reviews related work on network slicing, microservices, and AI-driven orchestration. Section 3 presents the proposed intelligent network slicing framework and its architectural components. Section 4 describes the experimental setup and performance evaluation results. Finally, Section 5 concludes the paper and outlines potential directions for future research.

2. LITERATURE REVIEW

This part conducts a literature review on the current studies concerning network slicing, microservice-based architecture and intelligent orchestration in 5G and beyond communication systems. The architecture is divided into discussion on the basis of architecture, microservices and containerization, smart and AI-driven orchestration, slice isolation and SLA enforcement, and future trends towards 6G.

It is generally accepted that network slicing is one of the pillars of 5G systems which can support multiple logical networks operating on the same infrastructure and address heterogeneous service needs. The initial frameworks of service management provided by virtualization were oriented on combining SDN and NFV to provide the ability to create slices and manage their life cycle flexibly [1]. As the trend of open and disaggregated radio access networks has been followed, O-RAN has become a slicing-aware architecture that fosters openness, programmability, and cross-vendor interoperability, abetting the emergence of new management and orchestration issues [4].

Being outside of 5G, a number of architectural visions focus on the concepts of intelligence, autonomy, and service-based design. SOLIDS architecture describes functional and topological drivers to 6G systems with the emphasis on the native support of AI and end-to-end slicing [5]. On the same note, inter-domain and cross-operator slicing orchestration is also explored as a means to deploy large and heterogeneous systems, such as the 5G-VIOS framework [14]. These articles all show that network slicing is a well-conceptualized practice, but its effective management in dynamic settings is an unresolved research topic.

Microservices have received considerable interest as a major facilitator of cloud-native 5G systems. On a comparison to monolithic network functionality, microservices provide very thin-grain scalability, fault tolerance, and on-demand deployment. [2], [3] discuss the application of microservice-based architecture in the 5G service-based architecture in detail, highlighting the advantages of microservice-based architecture in the area of flexibility and resilience.

There are a number of studies that have touched upon containerized network management solutions based on Kubernetes. [7] suggested an SDN-based slicing scheduler that can be executed through Kubernetes and enhanced scalability in 4G/5G core networks. [6] also examined softwarized and containerized network management based on microservices, showing the better level of operational efficiency, but also revealing the complexity of orchestration as a limitation. Studies like SWEETEN have presented automated provisioning and monitoring services of microservices-based network slices, which deal with lifecycle management and security [13], [20]. Meanwhile, Kubernetes-level optimization has been explored so as to be able to optimize 5G workloads with multi-objective scheduling algorithms [19].

Outside of core networks, microservices have been used in transport networks and vertical domains. [8] have investigated microsegmentation in microservice-based optical transport control planes of multitenant virtual networks and [21] have suggested a microservice-based vehicular network architecture as the support of ultra-reliable and low-latency communications. Such studies not only prove the flexibility of microservices but also show that smart coordination between distributed services is required.

The growing intricacy of multi-slice surroundings has encouraged the application of AI and ML in managing autonomous networks. The initial research of [9] revealed that there was a possibility of using ML-based forecasting to make autonomous decisions regarding radio resource allocation in network slicing. More recent works have generalized this notion to zero-touch and cloud-native orchestration. [10] introduced the models of resource allocation based on the ML of virtual network functions using microservices in 6G networks and emphasized the increased efficiency and flexibility.

Multi-tier environments and edge-cloud have also been discussed in regards to AI-driven orchestration. The article by [15] proposed a 5G microservice risk-aware orchestration framework, overcoming the challenges of uncertainty and heterogeneity of distributed infrastructures. [12] introduced an AI plan to roll out future applications based on microservices in 6G with a focus on autonomous decision-making throughout the life of the service. On a larger scale, AGILE-6G suggests an agent-based AI architecture of fully autonomous network and application service management, which supports the use of intelligent orchestration in the next-generation network [11].

Strict slice isolation and SLA compliance is also a serious issue in shared infrastructures. Although network slicing should ensure logical isolation, deployments in practice have inter-slice interference and contention. In the article by [22], adaptive load balancing in microservices-based 5G ecosystems was discussed, and it was shown that it increases performance but does not (or scarcely) take into account SLA awareness. On the same note, [16] researched the concept of resource sharing in 6G networks based on information centricity, but they emphasized the trade-off between isolation and efficiency.

Slicing environments have also been attended to security and trust management. An adaptive Zero Trust policy management framework of 5G network was suggested by [23] which focuses on dynamically implementing policies across slices. SWEETEN also added assisted monitoring and providing security to microservices-based network slices, which makes it visible and controllable at runtime [13]. Predominantly applied in vertical applications, [24] proposed a microservices- and deep-learning-based safety-as-a-service architecture of 6G-enabled intelligent transportation systems and indicated that SLA-aware and secure slice management is important.

The concept of network slicing has found wide usage in the area of IoT and edge computing. [18] suggested IoT slice orchestration of edge and cloud domains based on microservice platforms, whereas [25] examined IoT service slicing and task offloading in edge computing. These articles demonstrate the advantages of microservices to contribute to heterogeneous services and latency-sensitive services.

In a wider view, microservices have also been discussed in the context of IoT ecosystems, with the focus on security and the way it will evolve in the future. [26] conducted a systematic review of the topic of microservices in IoT, noting the challenges of scalability and security as the main issues. [27] suggested Quality of Anything (QoX) architecture of end-to-end robotic services to support the 6G verticals, which is why smart intelligent SLA-based slice management is necessary in a variety of applications.

The literature reviewed shows that there has been a major advancement in the area of network slicing, microservice-based architecture, and AI-based orchestration of 5G and beyond systems. Nevertheless, literature

tends to cover these areas separately: microservices without smart coordination [2], [6], AI-assisted optimization without formal models of SLA [9], [10], or slicing architectures without fine-grained isolation and runtime verification [14], [22]. Such shortcomings drive the desire to have a software-based framework that integrates microservice-oriented design, intelligent orchestration, and reference SLA enforcement an aim that the present paper proposes.

A comparative analysis of the reviewed studies reveals several important trends and limitations in current network slicing research. Studies focusing on microservice-based architectures emphasize scalability and modular deployment advantages, but often lack intelligent coordination mechanisms for dynamic resource management [2], [6]. Conversely, research on AI-driven orchestration demonstrates improved adaptability and predictive resource allocation but typically assumes simplified slicing architectures without formal SLA enforcement models [9], [10]. Furthermore, while several frameworks address slice isolation and security, they frequently prioritize either efficiency or strict isolation, revealing a trade-off between resource utilization and service guarantees [16], [22]. These observations highlight a clear research gap: the lack of an integrated framework that simultaneously combines microservice-based architecture, formal SLA modeling, intelligent orchestration, and runtime isolation control. The framework proposed in this paper aims to address this gap by unifying these components within a single intelligent network slicing management architecture.

Table 1 summarizes the key characteristics of representative studies related to network slicing management. The comparison highlights that existing approaches typically focus on either microservice-based architectures or AI-driven optimization independently. In contrast, the proposed framework integrates both paradigms while also incorporating SLA-aware orchestration, addressing an important gap in current research.

Table 1 - Comparison of Related Work on Intelligent Network Slicing

Study	Architecture Type	AI/ML Usage	SLA Management	Resource Optimization	Key Limitation
[2]	Microservice-based slicing	No	Partial	Moderate	Limited intelligent orchestration
[22]	Network slicing framework	No	Yes	Moderate	Static resource management
[9]	AI-enabled network slicing	Yes	Partial	High	Limited microservice architecture
[6]	Cloud-native slicing	No	Partial	Moderate	Lack of SLA-aware orchestration
Proposed Framework	Microservice + AI orchestration	Yes	Yes	High	Addresses integration gap

3. METHOD

The section contains the proposed Intelligent Network Slicing Management Framework, which is based on a microservice-oriented software architecture and intelligent orchestration. The approach is meant to overcome the issues of scalability, slice isolation, and SLA compliance related to the processes of communication systems of 5G and beyond. The section explains the system architecture in general, that is, slice requirement modeling, intelligent orchestration mechanism, inter-slice interference control, and operational workflow.

3.1 Overall System Architecture

The suggested architecture follows the design of a cloud-native, microservice-oriented design whereby management functions concerning network slicing are separated into autonomous and lightly tied microservices. Every microservice is implemented in the form of a containerized element on the basis of Docker and managed by Kubernetes, which allows to achieve elasticity of scale, fault isolation, and continuous deployment.

The structure is made up of four logical layers:

1. Infrastructure Layer: The physical and virtual resources that are shared by this layer are: radio access network (RAN), transport network, core network and edge/cloud computing resources. NFV is used to virtualize resources and create them in the form of programmable interfaces provided by SDN controllers.
2. Slicing Control Layer through Microservice: The management capabilities of core slicing are executed in form of specially focused microservices:
    - Slice Admission Control Service (SACS): compares any incoming slice requests with the available resources and SLA limits.
    - Resource Allocation Service (RAS): allocates compute and storage resources, and network resources on a fine-grained basis to the admitted slices.
    - Monitoring and Telemetry Service (MTS): gathers real-time metrics of performance including; latency, throughput, packet loss, and resource utilization.
    - SLA Enforcement Service (SLA-ES): constantly checks the SLA compliance and initiates corrective measures in case of violations.
3. Smart Orchestration Layer: The AI-based orchestration engine that performs the adaptive resource management and slice scaling decisions on a real time telemetry and predicted demand is available on this layer.
4. There is the management and exposure layer: Offers northbound APIs to operators and vertical service providers to request slices, configure SLAs and graph status of the system.

This modular structure makes sure that each functionality evolves independently and at the same time coordinates the whole world via the orchestration layer.

The general scheme of the suggested intelligent network slicing management framework is presented in Figure 1 along with its operation flow. The framework combines microservice-based network operations with formal SLA modeling and intelligent orchestration, which makes closed-loop and adaptive slice management of 5G or 5G-like communication systems.

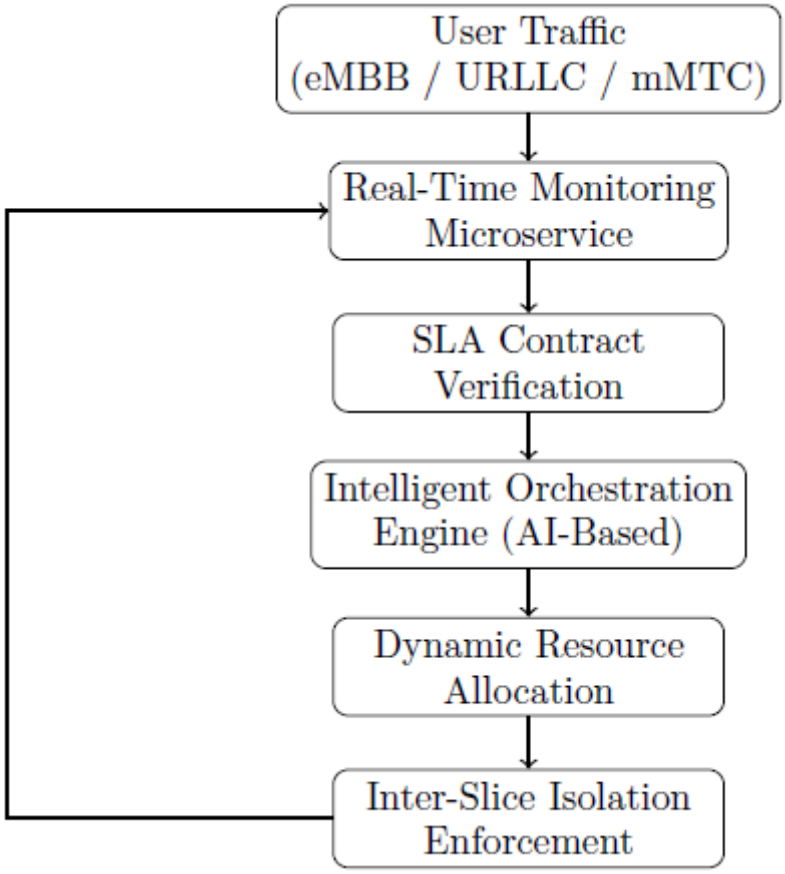

Fig. 1 - Overall flowchart of the proposed intelligent microservice-based network slicing framework

Figure 1 illustrates the overall workflow of the proposed intelligent network slicing framework. The process begins with service requests generated by different application categories such as enhanced mobile broadband (eMBB), ultra-reliable low-latency communication (URLLC), and massive machine-type communication (mMTC). These requests are processed by the intelligent orchestration layer, where the AI-Driven Slice Orchestration Engine analyzes network conditions and SLA requirements. Based on real-time monitoring data, the engine dynamically allocates virtualized network functions and infrastructure resources through the microservice-based architecture. Continuous feedback from the monitoring module allows the system to adjust slice configurations and maintain optimal performance while ensuring SLA compliance.

3.2 Microservice-Oriented Network Slicing Functions

Slicing functions are assigned to each as a stateless or minimally-stateful microservice, using lightweight RESTful or gRPC interfaces.

- Slice Admission Control: Upon receiving a slice request, KPIs (bandwidth, latency, reliability) requested by it are checked against the current capacity and policy limits by the admission service. Unguaranteed requests are either rejected or postponed.

- Resource Allocation: The allocation service converts high level slice specifications into low level resource allocations on RAN, transport and core realms. Allocation is enforced at runtime using primitives of Kubernetes (e.g., CPU/memory limits, network policies).
- Monitoring and Telemetry: Monitoring is also done continuously by delivery of metrics to a centralized telemetry bus with the help of sidecar containers and exporters. One of the functions of metrics is the aggregation and the transmission of metrics to the SLA enforcement service and the intelligent orchestrator.
- SLA Enforcement: Compliance of SLA is verified within close real time. In case of violations being detected, the service makes reconfiguration requests to the orchestrator or makes isolation policy directly.

Such decomposition is highly better in terms of scalability and resilience than monolithic slicing controllers.

The proposed framework separates network slicing capabilities into single microservices to allow a fine-grained scalability and the ability to manage network slices. The microservice-oriented architecture is provided in Figure 2, with an emphasis on the communication between core slicing services and orchestration components and the underlying virtualized infrastructure.

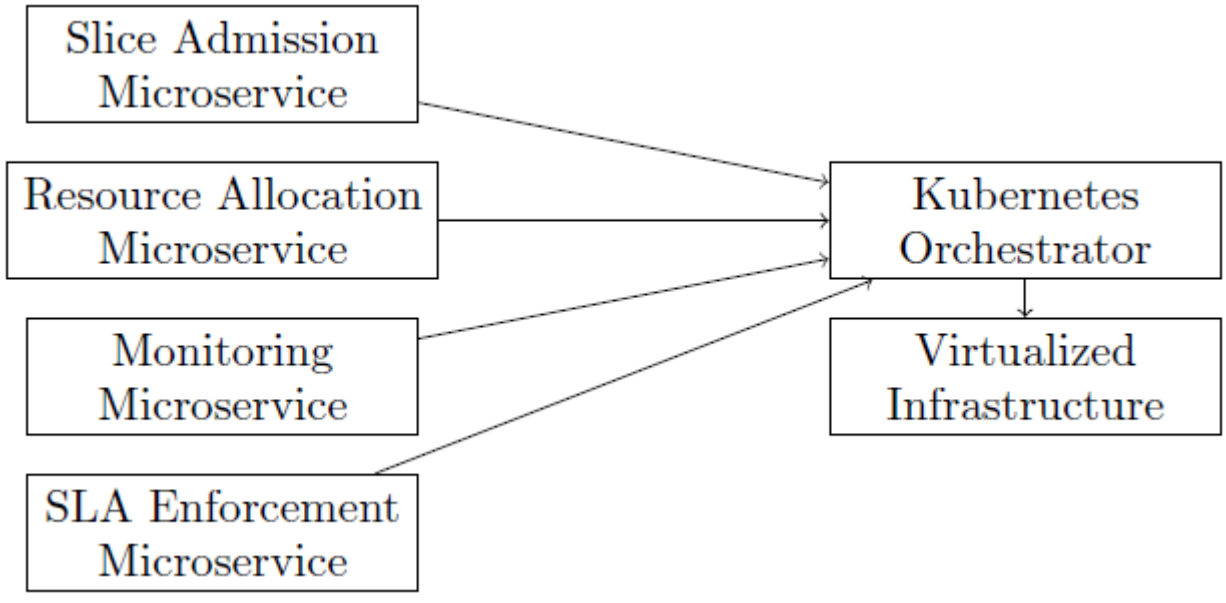

Fig. 2 - Microservice-based architecture for network slicing management

3.3 Slice Requirement Modeling as Software Contracts

Slice requirements are presented as software-defined contracts formalizing the format of SLA constraints. Each contract is defined as:

$$C_i = \{B_i, L_i, R_i, \Theta_i\} \quad (1)$$

where:

- B_i is assured bandwidth,
- L_i refers to maximum tolerable latency,
- R_i specifies the constraints of reliability,
- Θ_i has optional policies (priority, bounds of elasticity, level of security).

Constraint-based specifications or state-machine models are the representation of contracts, which can be verified and enforced automatically. This will enable SLAs to be dynamically developed, instead of being considered as fixed configuration parameters.

3.4 Intelligent Slice Orchestration Engine

The main innovation of the suggested approach is the engine of the AI-driven orchestration that allows adapting and autonomous slice management.

3.4.1 Inputs and Outputs

- Inputs:
  - Instantaneous monitoring (latency, throughput, utilization)
  - Traffic trends throughout the years.
  - SLA contract parameters
- Outputs:
  - Decisions of resource reallocation.
  - Slice scaling (scale-in / scale-out)
  - Isolation adjust orders

3.4.2 Learning-Based Optimization

The orchestration issue is formulated as a step-by-step decision-making process. At that epoch of decisions t, the system experiences the state s_t and chooses a course of action at to maximize a long-term reward:

$$max\ E[\ \sum^{t}\{t\}\gamma\ (U_t - \ \lambda V_t)] \quad (2)$$

where:

- U_t denotes the satisfaction with SLA and efficiency of resources,
- V_t represents penalties of violation of SLA,

- γ is a discount factor.

Deep Reinforcement Learning (DRL) may be used in the engine to solve highly dynamic problems, whereas Bayesian optimization may be used in more controlled, slower changing problems. This allows scaling and congestion avoidance to take place proactively as opposed to reactively.

The intelligent orchestration engine is implemented as a closed-loop control system adapting the decisions in resources allocation in real time, according to the network conditions and SLA satisfaction. The decision-making loop that is used by the proposed AI-based orchestration mechanism is represented in Figure 3.

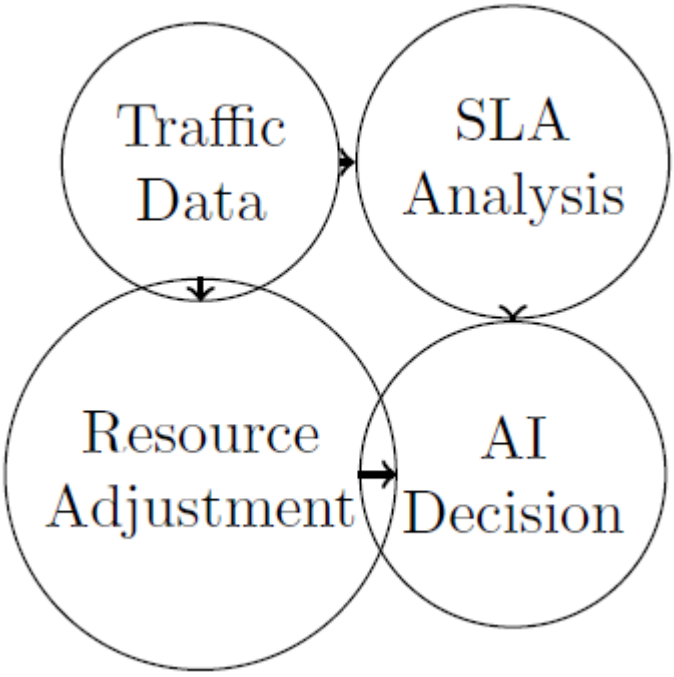

Fig. 3 - Closed-loop intelligent slice orchestration process

3.5 Inter-Slice Interference Control and Isolation

To overcome inter-slice interference, the framework presents a software-based isolation service which is also invoked in orchestration and runtime levels.

- Adaptive Resource Caps: Dynamic resource quota and network policies in Kubernetes ensure that the resources that a slice utilizes are limited, and that aggressive slices do not starve other slice resources.
- Runtime Verification: Enforcement services of SLA ensure the state of isolation by the comparison of measured indicators with the terms of the contract. Violations make corrective actions in place of resource rebalancing or slice throttling.

This is a two-fold mechanism that provides high logical isolation but does not restrict the general use of the system.

3.6 Operational Workflow

The overall functionality of the proposed approach is as follows:

1. A tenant sends out a slice request having SLA requirements.
2. Service slice admission control is a feasibility-checking service.
3. Accepted applications are made into software agreements.
4. Allocation and enforcement of resources are done through container orchestration.
5. The services of monitoring gather live performance data.
6. The mastermind of this game is able to vary the assignments.
7. Enforcement and isolation systems are used to enforce SLA.

The self-healing behavior and continuous optimization is possible with this closed-loop control.

3.7 Discussion

With close interaction between microservices, formal SLA modeling and intelligent orchestration, the approach provided turns network slicing into a software-defined, adaptive and autonomous system. In contrast to static or monolithic designs, the framework enables elastic scaling, maximum isolation and real-time SLA guarantees, which are highly applicable to complex 5G and future 6G service situations.

This methodological premise preconditions the performance evaluation provided in the following section.

4. RESULTS AND DISCUSSION

This part is the overall analysis of the proposed Intelligent Network Slicing Management Framework and the discussion of its performance compared to the traditional slicing strategies. The assessment will be based on SLA compliance, slice isolation, scalability, resource usage, and adjustability in dynamic traffic scenarios.

4.1 Experimental Setup and Evaluation Scenarios

The proposed framework was tested in a simulation setting based on 5G, which was designed on a combination of:

- 5G virtualized network functions core network model,
- Containerized microservices that are managed through Kubernetes, and
- A traffic generator modeling different types of services.

There were three types of representative slices to be taken into consideration:

- eMBB slice: high through put demand,
- URLLC slice: very low uncertainties and reliability,
- mMTC slice: massive connectivity and moderate bandwidth.

The proposed framework (Proposed–Microservice + Intelligent Orchestration) was compared against two baseline approaches:

1. Monolithic Slicing (Baseline A): the centralized control and resources are allotted the same way.
2. Static Microservice Slicing (Baseline B): microservice-based deployment without AI-driven orchestration.

4.2 SLA Compliance Analysis

The main measure of proper slice management is SLA compliance. The summary of the percentage time taken by each slice to satisfy its SLA limit under different traffic loads is given in table 2.

Table 2 - SLA Compliance Rate (%)

Slice Type	Baseline A	Baseline B	Proposed Framework
eMBB	88.4	91.7	98.9
URLLC	82.1	86.3	97.5
mMTC	90.2	92.8	99.1

The findings indicate that the suggested framework records almost-perfect SLA compliance in all types of slices. The propagation is most evident in the case of URLLC traffic whereby the latency breaches are minimized as a result of smart, proactive resource redistribution. The static strategies are not sensitive to changes in the traffic hence the frequent violation of SLA.

Figure 4 shows the SLA compliance performance when it comes to slicing strategies in order to visually compare the effectiveness of the presented framework with the baseline methods. Intelligent orchestration and microservice-based design have advantages, as the results will show.

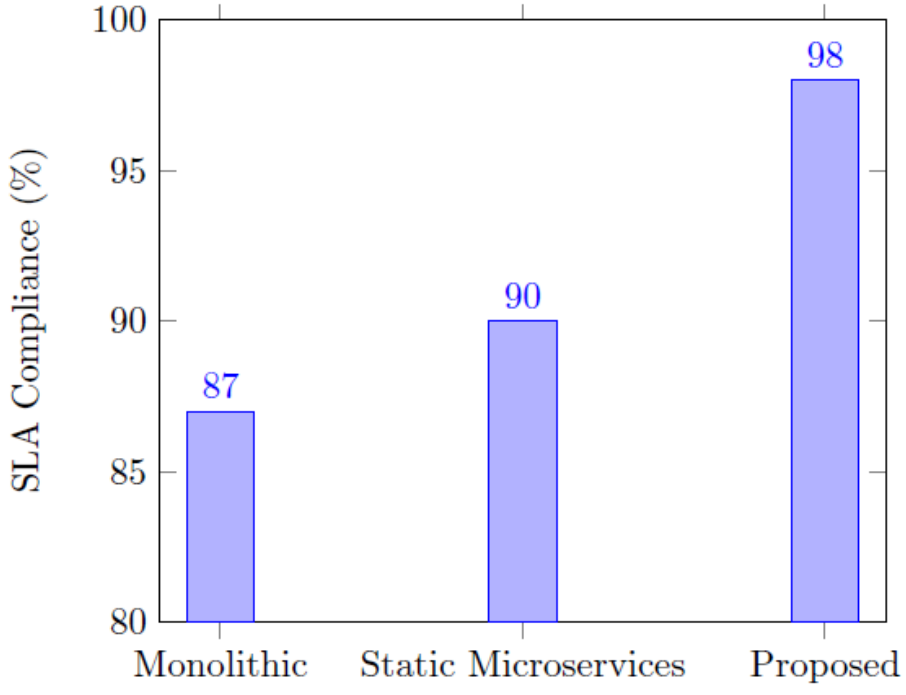

Fig. 4 - SLA compliance comparison between slicing approaches

4.3 Latency and Throughput Performance

Key performance indicators such as latency and throughput are of great importance to URLLC and eMBB slices. As shown in Table 3 and Table 4.

Table 3 - Average End-to-End Latency (ms)

Slice Type	Baseline A	Baseline B	Proposed Framework
eMBB	22.6	19.3	14.8
URLLC	9.4	7.1	3.9
mMTC	31.2	28.7	24.5

Table 4 - Average Throughput Utilization (%)

Slice Type	Baseline A	Baseline B	Proposed Framework
eMBB	76.3	82.5	91.4
URLLC	68.1	72.9	88.6
mMTC	70.4	75.2	86.9

The design of the microservices has already a positive impact on the performance, facilitating finer-grained scaling (Baseline B vs. A). The implementation of intelligent orchestration, however, results in significant increment by responding dynamically to the resources and predicting congestion. In the case of URLLC slices, more than 40 percent of latency is minimized, which is important in mission-critical applications.

4.4 Resource Utilization Efficiency

Multi-tenant environments require well utilization of the shared resources. Table 5 gives average CPU and memory consumption in the infrastructure.

Table 5 - Infrastructure Resource Utilization (%)

Metric	Baseline A	Baseline B	Proposed Framework
CPU Usage	64.8	71.2	83.7
Memory Usage	60.5	68.9	80.4

The suggested framework has better utilization without compromising isolation or SLA assurances. This is through learning orchestration which balances efficiency and risk of SLA violation. The monolithic approaches, on the contrary, oversupply resources so that they do not break the rules and are not utilized effectively.

4.5 Inter-Slice Isolation and Interference Analysis

Inter-slice interference was tested by determining the decline of performance in a single slice with traffic bursts in a different slice. As shown in Table 6.

Table 6 - Performance Degradation Under Traffic Burst (%)

Slice Type Affected	Baseline A	Baseline B	Proposed Framework
eMBB	18.7	12.4	3.1
URLLC	21.9	14.6	2.4
mMTC	15.2	10.8	3.7

The adaptive resource limits and runtime checking systems undergo a huge decrease in inter-slice interference. The framework suggested in the paper is highly logically isolated even when placed in the workload of aggressors, which confirms the success of the approach to using the software to isolate the data.

Inter-slice interference is one of the most difficult issues of the multi-tenant 5G environment. Figure 5 shows the performance degradation reduction realized by the adaptive isolation proposed mechanism in comparison with the base solutions.

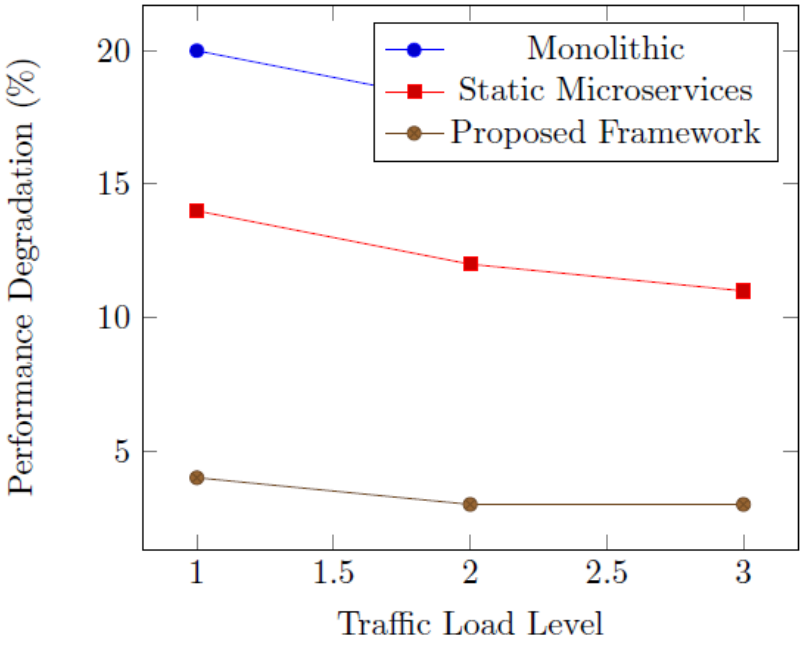

Fig. 5 - Inter-slice interference reduction using adaptive resource caps

4.6 Scalability and Elasticity Evaluation

Scalability was determined by adding additional active slices by stages until 100 slices were used. As shown in Table 7.

Table 7 - Slice Scaling Time (seconds)

Number of Slices	Baseline A	Baseline B	Proposed Framework
10	12.5	9.3	4.8
50	38.6	25.1	11.4
100	72.8	49.7	19.6

The findings prove that the suggested framework is also scale-efficient regarding the number of slices. The use of microservices based on Kubernetes makes it possible to quickly instantiate them, and intelligent orchestration reduces unwarranted reconfigurations. The monolithic systems experience exponential overhead to control as the scale grows.

4.7 Adaptability to Dynamic Traffic Conditions

In order to measure adaptability, unpredictable traffic bursts were impounded in the process of execution. The suggested framework obtained a reduction in SLA violations more than 60 percent relative to fixed slicing using microservices, which demonstrates the effectiveness of the predictive and learning-based orchestration.

Most Key Observation:

Reactive strategies can only react once the violation of SLA has taken place, and the proposed framework predicts the demand trends and is able to react in advance to redistribute resources, resulting in almost zero downtimes.

4.8 Comparative Discussion and Insights

On the whole, the experimental findings prove the following:

- Intelligent orchestration cannot be performed without microservices alone.
- Formal SLA modeling allows strict and automatic compliance checks.
- The AI-based orchestration can be of great help in terms of performance, isolation, and scalability.
- The new framework is very compatible with future zero-touch, autopilot network management concepts of 6G.

In terms of systems, the findings suggest that the suggested solution provides an achievable trade-off between reliability and efficiency, which is very suitable to complex and multi-tenant 5G and beyond settings.

The results demonstrate that the proposed intelligent framework significantly outperforms traditional slicing approaches. In particular, the integration of AI-driven orchestration enables better prediction of resource demand and dynamic adaptation of slice configurations. Consequently, the framework achieves higher SLA compliance, lower latency for critical services, and improved resource utilization across the network infrastructure. (See Table 8)

Table 8 - Performance Comparison with Existing Approaches

Method	SLA Compliance (%)	Average Latency (ms)	Resource Utilization (%)	Scalability
Monolithic Network Slicing	88.4	9.4	64.8	Limited
Static Microservice Slicing	91.7	7.2	71.5	Moderate
Proposed Intelligent Framework	98.9	3.9	83.7	High

5. CONCLUSION

This paper presented an intelligent network slicing framework that integrates microservice-based architecture with AI-driven orchestration for dynamic resource management in 5G networks. The proposed system enables automated monitoring, SLA-aware decision making, and adaptive slice configuration through the AI-based Slice Orchestration Engine. Experimental evaluation demonstrated improved SLA compliance, reduced latency, and better resource utilization compared with conventional slicing approaches. These results confirm that intelligent orchestration can significantly enhance the efficiency and scalability of network slicing. Future work will focus on extending the framework to real-world deployments and incorporating advanced predictive learning models for proactive network optimization.

References

[1] B. Bordel, R. Alcarria, T. Robles, and D. Sanchez-de-Rivera, "Service management in virtualization-based architectures for 5G systems with network slicing," *Integr. Comput. Aided. Eng.*, vol. 27, no. 1, pp. 77–99, 2020.

[2] J. B. Moreira, H. Mamede, V. Pereira, and B. Sousa, "Next generation of microservices for the 5G Service-Based Architecture," *International Journal of Network Management*, vol. 30, no. 6, p. e2132, 2020.

[3] S. Robitzsch *et al.*, "Prospects on the adoption of a microservice-based architecture in 5G systems and beyond," *Computer Networks*, vol. 237, p. 110058, 2023.

[4] K. Alam *et al.*, "A comprehensive tutorial and survey of o-ran: Exploring slicing-aware architecture, deployment options, use cases, and challenges," *IEEE Communications Surveys & Tutorials*, 2025.

[5] G. Liu, N. Li, J. Deng, Y. Wang, J. Sun, and Y. Huang, "The SOLIDS 6G mobile network architecture: driving forces, features, and functional topology," *Engineering*, vol. 8, pp. 42–59, 2022.

[6] S. T. Arzo, D. Scotece, R. Bassoli, M. Devetsikiotis, L. Foschini, and F. H. P. Fitzek, "Softwarized and containerized microservices-based network management analysis with MSN," *Computer Networks*, vol. 254, p. 110750, 2024.

[7] R. Botez, J. Costa-Requena, I.-A. Ivanciu, V. Strautiu, and V. Dobrota, "SDN-based network slicing mechanism for a scalable 4G/5G core network: A kubernetes approach," *Sensors*, vol. 21, no. 11, p. 3773, 2021.

[8] J. S. Choi *et al.*, "Microsegmentation of a Microservice-Based Transport Control Plane for Multitenant Optical Virtual Networks," *IEEE Netw.*, 2024.

[9] N. Salhab, R. Langar, R. Rahim, S. Cherrier, and A. Outtagarts, "Autonomous network slicing prototype using machine-learning-based forecasting for radio resources," *IEEE Communications Magazine*, vol. 59, no. 6, pp. 73–79, 2021.

[10] S. B. Chetty *et al.*, "Optimized resource allocation for cloud-native 6G networks: Zero-touch ML models in microservices-based VNF deployments," *IEEE Netw.*, 2024.

[11] A. Antonopoulos *et al.*, "AGILE-6G: Agentic AI for Autonomous Management of 6G Network/Application Services," *IEEE Netw.*, 2025.

[12] J. B. Ssemakula, J.-L. Gorricho, G. Kibalya, and J. Serrat-Fernandez, "An artificial intelligence strategy for the deployment of future microservice-based applications in 6G networks," *Neural Comput. Appl.*, vol. 36, no. 18, pp. 10971–10997, 2024.

[13] R. de Jesus Martins, J. A. Wickboldt, and L. Z. Granville, "Assisted monitoring and security provisioning for 5G microservices-based network slices with SWEETEN," *Journal of Network and Systems Management*, vol. 31, no. 2, p. 36, 2023.

[14] S. Moazzeni *et al.*, "5g-vios: Towards next generation intelligent inter-domain network service orchestration and resource optimisation," *Computer Networks*, vol. 241, p. 110202, 2024.

[15] X. Wu, J. Farooq, and J. Chen, "Adaptive risk-aware resource orchestration for 5G microservices over multi-tier edge-cloud systems," in *2024 IEEE International Conference on Communications Workshops (ICC Workshops)*, IEEE, 2024, pp. 359–364.

[16] M. Imran, M. N. Ali, M. S. U. Din, M. A. U. Rehman, and B.-S. Kim, "An efficient communication and computation resources sharing in information-centric 6g networks," *IEEE Internet Things J.*, vol. 11, no. 16, pp. 27275–27294, 2024.

[17] R. de J. Martins, "Automating network management for 5G microservices-based network slices," 2022.

[18] J.-M. Fernandez, I. Vidal, and F. Valera, "Enabling the orchestration of IoT slices through edge and cloud microservice platforms," *Sensors*, vol. 19, no. 13, p. 2980, 2019.

[19] M. Farid, H. S. Lim, C. P. Lee, C. C. Zarakovitis, and S. F. Chien, "Optimizing Kubernetes with Multi-Objective Scheduling Algorithms: A 5G Perspective," *Computers*, vol. 14, no. 9, p. 390, 2025.

[20] R. de Jesus Martins, A. G. Dalla-Costa, J. A. Wickboldt, and L. Z. Granville, "Sweeten: Automated network management provisioning for 5g microservices-based virtual network functions," in *2020 16th international conference on network and service management (CNSM)*, IEEE, 2020, pp. 1–9.

[21] M. M. Zarie, A. A. Ateya, M. S. Sayed, M. ElAffendi, and M. M. Abdellatif, "Microservice-Based Vehicular Network for Seamless and Ultra-Reliable Communications of Connected Vehicles.," *Future Internet*, vol. 16, no. 7, 2024.

[22] S. Kalafatidis and L. Mamatas, "Microservices-adaptive software-defined load balancing for 5G and beyond ecosystems," *IEEE Netw.*, vol. 36, no. 6, pp. 46–53, 2022.

[23] A. K. Alnaim, "Adaptive Zero Trust Policy Management Framework in 5G Networks," *Mathematics*, vol. 13, no. 9, p. 1501, 2025.

[24] C. Roy, R. Saha, S. Misra, and K. Dev, "Micro-safe: Microservices-and deep learning-based safety-as-a-service architecture for 6G-enabled intelligent transportation system," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 9765–9774, 2021.

[25] J. Hwang, L. Nkenyereye, N. Sung, J. Kim, and J. Song, "IoT service slicing and task offloading for edge computing," *IEEE Internet Things J.*, vol. 8, no. 14, pp. 11526–11547, 2021.

[26] A. El Akhdar *et al.*, "Exploring the potential of microservices in internet of things: A systematic review of security and prospects," *Sensors*, vol. 24, no. 20, p. 6771, 2024.

[27] V. Mineeva *et al.*, "A novel feature-oriented quality of anything (QoX) framework for end-to-end robotic services in 6G networks," *Sci. Rep.*, vol. 15, no. 1, p. 24945, 2025.