Electronics, Communications, and Computing Summit
Vol. 3, No. 2, Apr - Jun 2025, pp. 91-97
DOI: https://doi.org/10.17051/ECC/03.02.12

Latency-Aware and QoS-Optimized Service Function Chaining in Multi-Tier SDN-Enabled 5G Network Architectures

Nisha Milind Shrirao[1], Sumit Ramswami Punam[2]

[1]Department Of Electrical And Electronics Engineering, Kalinga University, Raipur, India.
Email: nisha.milind@kalingauniversity.ac.in
[2]Department Of Electrical And Electronics Engineering, Kalinga University, Raipur, India.
Email: sumit.kant.dash@kalingauniversity.ac.in

Article Info

Article history:

Received : 11.04.2025
Revised : 23.05.2025
Accepted : 27.06.2025

Keywords:

Service Function Chaining (SFC),
Software-Defined Networking (SDN),
5G Networks,
Latency-Aware Networking,
Quality of Service (QoS) Optimization,
Virtual Network Functions (VNFs),
Multi-Tier Network Architecture,
Reinforcement Learning,
Edge Computing,
Network Orchestration

ABSTRACT

The Service Function Chaining (SFC) is essential to service delivery of end-to-end Quality of Service (QoS) within the 5G network by dynamically orchestrating a group of Virtual Network Function (VNF). Nevertheless, the newfound 5G applications latency sensitivity in conjunction with the dynamicism of traffic loads presents very serious problems to effective service provisioning. The current paper describes a latency-aware and Quality of Service (QoS)-optimized SFC architecture that can be used in multi-tier Software-Defined Networking (SDN)-enabled 5G systems. The proposed architecture will rely on a hierarchical SDN control plane, which includes the edge, fog and core controllers, to facilitate a distributed management of network resources. In order to efficiently perform VNF placement and service path selection on-demand, we include a Deep Reinforcement Learning (DRL) agent which is trained with information about the state of the network in terms of the latencies, bandwidth and QoS attributes. The set of experiments on a multi-tier SDN setting proves that our mechanism has the ability to considerably lessen the end-to-end latency and QoS contravention and optimize the overall use of network resources better. The proposed model contains up to 43 percent reduction in latency and more than 97 percent QoS compliance in various traffic conditions when compared to conventional centralized and greedy SFC algorithms. These outcomes validate the viability and feasibility of the suggested solution and it is a very possible course of action to implement the given solution into the 5G and beyond-5G network infrastructures of the future.

1. INTRODUCTION

5G network advancement is inspired by ultra-low latency, greater reliability, and connections related to mammoth gadgets, which will empower sophisticated services, including self-driving cars, remote health care, and automation in industries. Service Function Chaining (SFC) is important to meet these objectives by dynamically coordinating a sequence of Virtual Network Functions (VNFs) i.e., firewalls, NATs, and deep packet inspection, which are application-specific. Nevertheless, the control of SFC over heterogeneous and distributed network layers with edges, fog, and core nodes that are formed in Software-Defined Networking (SDN)-enabled 5G networks is challenging since SFC demands have to face short latencies, unstable network conditions, and resource distribution across our infrastructure. Current SFC solutions are usually based on a centrally controlled mechanism or heuristic techniques that do not satisfy the latency and Quality of Service (QoS) requirements of today 5G applications. In addition, most of them are not adaptable and real time aware of tiered control architectures, which are critical towards scaling-up and responsiveness in mass deployment. Late research has also started investigating the machine learning-driven SFC solutions, but they either exclude multi-tiered SDN control architecture or common latency-QoS optimization strategy [1]. To overcome these shortcomings, this paper suggests a hierarchical SDN-based control plane and deep reinforcement learning-supported QoS-optimized latency-conscious SFC solution to real-time 5G networks that specifies intelligent VNF placement and path in the network environments.

2. RELATED WORK

Some methods have been suggested to streamline Service Function Chaining (SFC) in 5G network. Common centralized SFC mapping methods apply methods of computing the best approach of service and VNFs placement, which tends to depend on the global SDN controllers. As an example, Rehman et al. [1] proposed a QoS-aware SFC mechanism based on centralized controller to embody the VNF chaining decisions. Although they are effective when used in small scale networks, the techniques do not scale well with huge traffic loads common in large 5G implementations. Some other works deal with latency-insensitive optimization, in which the goal is to optimise resource efficiency or throughput without particular emphasis on latency. This is the case with the cost-aware placement algorithm proposed in Zhani et al. [2], which focuses on optimizing the overhead of its operations but cannot satisfy the ultra-reliable low-latency communication stringent latency and jitter demands of so-called ultra-reliable low-latency communication (URLLC) services.

In the recent past, adaptive and context-aware service chaining have been pursued in form of AI-driven SFC strategies. As an example, Liu et al. [3] utilized deep reinforcement learning (DRL) in order to enhance VNF placement decisions on-demand. The majority of those models though presume flat or centralized SDN control and fail to utilize the hierarchical control plane structures, which are vital in distributed 5G architectures to minimize signaling overheads and enhancing reaction time at the network edge.

To stand out among the other papers, the given paper introduced the concept of a latency-based QoS-optimized SFC model to be implemented in a combination of hierarchical SDN control and DRL-based orchestration. The framework overcomes existing constraints through facilitating distributed decision-making, latency-sensitivity in real-time, and scalability to a variety of 5G service requirements.

3. System Architecture

The system architecture proffered will be able to respond to latency-sensitive and QoS-critical applications in 5G settings using hierarchical Software-Defined Networking (SDN) control structure. Its architecture is partitioned into three functional planes, each performing unique orchestration, control, as well as resources such that when it comes to Service Function Chaining (SFC) operations it follows a pattern of the scalability, responsiveness, and efficiency.

3.1 Multi-Tier SDN Architecture

The system incorporates a three-level SDN control plane to cover the drawbacks of monolithic SDN controller in wide and, therefore, latency-sensitive network deployments called 5G:

- Edge Controllers: At access network level, these controllers control the real-time user traffic and local flow rules, and deploy latency-sensitive Virtual Network Functions (VNFs), e. g. firewall or intrusion detection system. This comes with ultra-low levels of latency response and quick adaptation to fluctuating service needs as they are close to the end-users.
- Regional (Fog) Controllers: These are in the middle tier, that aggregate decisions of various edge controllers. They streamline the use of resources in regional data centers or in fog nodes through coordination of mid-tier VNFs such as content filtering, video transcoding or caching. The local controllers inject geographically specific intelligence relieving the central controller and making it resilient during network partitioning.
- Central Controllers: We have central controllers and they exist in the heart of the network, these have a worldwide picture of the infrastructure and are involved in inter-regional coordination. They manage end-to-end SFC policy enforcement, global resource assigning and inter-domain orchestration. The layer also stores the historical and predictive analytics models to support the upper layer decision; migration of service and service scaling.

Distributed intelligence and minimal delay in signaling, required in deterministic latency and high service availability in 5G and beyond-5G networks, are beneficial in the hierarchical model.

3.2 Service Function Chain Composition

The SFCs are built based on requirements in specific applications of services offered. All of the service requests are presented in the form of directed acyclic graph (DAG), with the nodes representing a unit VNF and the edges representing the sequence of execution. These graphs are interpreted by the SFC Manager which operates in the SDN control plane and translates them into physical and virtual resources within the multi-tier network.

Mapping process consists of composing an appropriate set of VNF instances and the network paths underneath them so that end to end latency, bandwidth, jitter and other QoS requirements are met tightly. The SFC Manager also operates using both static policy (e.g. SLA constraints) and dynamic measurements (e.g. link congestion, CPU load, VNF execution time) to make decisions. In this paper, the role of a reinforcement learning based agent that works together with the SFC Manager is to facilitate adaptive, data-driven optimization of the VNF placement and hitchhiking, developing continuously in real time to adapt to network conditions. The proposed architecture would add considerable performance and scalability gains to SFC provisioning in complex 5G networks due to the efficient division of labor and hierarchical control structure and intelligent orchestration.

As shown in Figure 1, the hierarchical organization of this architecture and the relationship of various control levels as well as the relationship with the Service Function Chaining (SFC) Manager are well visualized.

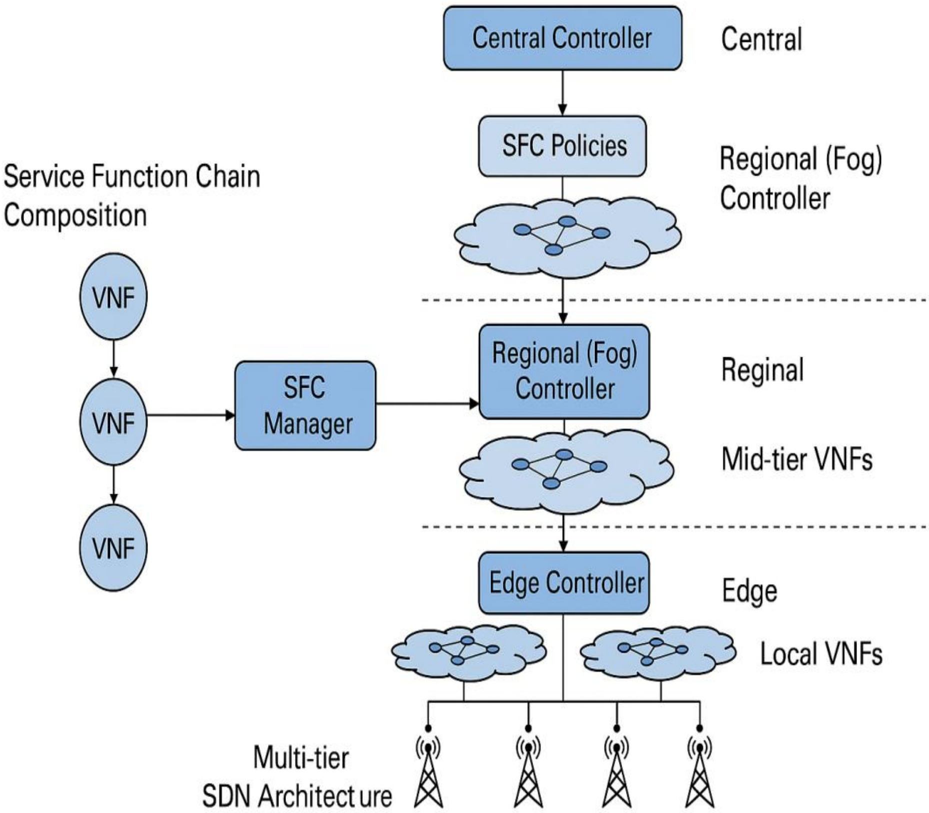
Central Controller
Central
SFC Policies
Regional (Fog) Controller
Service Function Chain Composition
VNF
VNF
VNF
SFC Manager
Regional (Fog) Controller
Reginal
Mid-tier VNFs
Edge Controller
Edge
Local VNFs
Multi-tier SDN Architect ure

Figure 1. Multi-Tier SDN Architecture for Latency-Aware Service Function Chaining in 5G Networks.

This schematic presents a three-tier set of control plans with central, regional (fog) as well as edge SDN controllers. It also provides how the SFC Manager will map VNF graphs to network functions given latency and QoS requirements.

4. Problem Formulation

The National Performance Optimization Initiative The main objective of the proposed latency-sensitive and Quality of Service (QoS) optimized Service Function Chaining (SFC) framework is that the user traffic in service chains must pass through a chain of Virtual Network Functions (VNFs) with the least amount of end-to-end latency to achieve the key Quality of Service (QoS) parameters such as packet loss, jitter and bandwidth. This part codifies the goal of optimization, as well as the constraints.

Objective Function

Let L_{ijk}denote the latency incurred when traffic flows from node i to node k via VNF j. The objective is to minimize the cumulative end-to-end latency across all selected service chains. The binary decision variables $x_{ij}$and y_{jk}represent the placement of VNF j at node i and the selection of a communication path between VNFs j and k, respectively.

$$\min_{x_{ij},\, y_{jk}} \sum_{i,j,k} L_{ijk} \,.\, x_{ij} \,.\, y_{jk} \text{ ----------- ---------(1)}$$

The aim of this is because the overall latency of all the paths of service functions is minimized.

QoS Constraints

The constraints on the optimization are as follows QoS:

- Packet Loss Constraint:

$$P_l \le P_l^{max} \text{ ----------------- ----- (2)}$$

The total packet loss P_l across the SFC must not exceed the maximum tolerable limit P_l^{max}, as defined by the service-level agreement (SLA).

- Jitter Constraint:

$J \leq J^{max}$ ---------------------------(3)

Jitter J, which denotes the variation in packet delay, must remain below the maximum allowable jitter J^{max} to ensure consistent streaming quality and real-time responsiveness.

- Bandwidth Constraint:

$B \geq B^{min}$ --------------------------------(4)

The available bandwidth B along the selected path must meet or exceed a minimum required threshold B^{min} to ensure throughput guarantees.

Decision Variables

- $X_{ij} \in \{0,1\}$:Indicates whether VNF j is placed at node i.
- $Y_{jk} \in \{0,1\}$:Indicates whether the traffic path between VNF j and VNF k is selected in the chain.

These binary parameters are simultaneously optimized in order to create a latency-friendly and QoS-controlled service chain.

Remarks

Such formulation facilitates the orchestration of VNFs using dynamic, scalable, and constraint aware of VNF functions. The real-time network state information is provided to the decision making process that is solved subsequently using a reinforcement learning based agent that learns on-its-own to find an optimal path minimizing latency (while respecting service constraints) in hierarchical SDN-controlled 5G network station environment.

5. Proposed Methodology

This paper presents a deep reinforcement learning (DRL)-based intelligent orchestration framework to deal with the complexity/latency sensitivity of Service Function Chaining (SFC) in the multi-tier SDN-enabled 5G networks. The methodology is based on two main components (i) Latency-Aware SFC Orchestrator, which interacts with SDN control plane, (ii) QoS-Aware Reinforcement Learning engine, which, in real time applies optimization of VNF placements and flow paths.

5.1 Latency-Aware SFC Orchestrator

The orchestrator would be the decision making module within which dynamically allocates the VNFs and decides the paths to follow when converting VNFs to the service. It is closely coupled to multi-tier SDN control plane (edge, fog, and core controllers), which allows it to track traffic flows, resource load, and the latency at the level of links across the network.

The orchestrator uses a Deep Q-Network (DQN) agent in order to implement intelligent, adaptive decisions. Such an agent trains one policy to correlate observed network states with the optimal action (i.e. VNF placements and routing decisions). In particular, the DQN:

- Determines the pattern of optimal VNF placement that examines the latency profiles, link utilization or compute capacity at candidate nodes.
- It is dynamically configured to choose end to end service paths that minimize delay and congestion across different traffic loads and quality of service requirements.

When DRL is used, the system evolves in time adjusting its choices based on trial-and-error exchange between the environment and the network.

5.2 QoS-Aware Reinforcement Learning

Reinforcement learning model is defined as Markov Decision Process (MDP), and the environment, i.e. network infrastructure, and the agent, i.e. orchestrator.

State Space

In the states s there is composite vector to represent:

- Current VNFs loads: CPU, memory and length of processing queues at every VNF node.
- Path delays(latencies): Measurements of the real-time delays between possible VNF pairs.
- QoS indicators: Packet loss rate, jitter and available bandwidth.

Action Space

Such actions can be node selection to deploy VNF as well as to form the next hop path of the service chain.

Reward Function

The reward R is defined to induce desired behavior of the agent, i.e., QoS-compliant and latency-efficient behavior:

$$R = -\alpha L - \beta P_l - \gamma J + \delta B - - - - - - - - - - - - - - - (5)$$

Where:

- L: End-to-end latency
- P_l : Packet loss
- J: Jitter
- B: Available bandwidth
- α, β, γ, δ: Weight coefficients calibrated according to the service type (e.g., URLLC requires low L and J, whereas eMBB emphasizes B)

The goal of the agent is to maximize the reward over time, implicitly, ensure bandwidth and jitter requirements without a time latency and packets loss.

This orchestration framework is learning-based and augers well with adaptive, scalable, and situation-aware SFC deployment, which is also well suited to the dynamic and hetereogeneous demands of future 5G and beyond networks.

6. Experimental Results

In accordance with the objectives of the proposed latency-aware and QoS-optimized SFC framework, extensive simulation-based performance analysis was carried out to access the effectiveness of the framework. In this section, the experimental scenario and the assessment criteria of the results are described along with comparative analysis of outcomes with the footprints of the baseline SFC strategies.

6.1 Experimental Setup

Mininet, an open-source network emulator was used to build the simulation environment jointly with the Ryu SDN controller as a way to apply and view the control logic. The virtualized network (which was mimicking it) had a 3 level topology SDN network, with 60 nodes (with an edge, fog, and core layer).

Three orchestration strategies of SFC were compared:

- DQN-SFC: A reinforcement learning agent based on Deep Q-Network is emulated, and is also integrated with multi-tier SDN control.
- Centralized SFC: A global network knowledge is used to compute the service chains by a monolithic SDN controller.
- Greedy SFC: Heuristically picks VNFs and paths making local decisions of shortests paths.

6.2 Evaluation Metrics

Evaluation of performance was done based on the following major indicators:

- End-to-End Latency (ms): Latency to packets which go through the entire VNF chain.
- QoS Violation Rate (%): A proportion of the service requests that do not comply with latency, jitter, or packet loss restrictions.
- CPU Resource Utilization: Percent- An average amount of computational utilization of all nodes that VNFs are hosted on.
- Computation Overhead (Qualitative): Orchestration complex and response time of the decision --making algorithms.

6.3 Results Summary and Analysis

Table 1. Performance Comparison of SFC Strategies

Method	**Avg. Latency (ms)**	**QoS Violation Rate (%)**	**CPU Utilization (%)**
Proposed (DQN-SFC)	13.4	2.1	68.3
Centralized SFC	23.7	8.4	54.1
Greedy SFC	18.2	6.9	61.0

According to Table 1, proposed DQN-SFC approach has the lowest latency and QoS violation rates, as well as is more resource efficient than baseline approaches. The obtained findings directly show that the initial DQN-SFC model is a better option in all of the assessed aspects:

- It results in about 43.5 percent and 26.3 percent improvement over the centralized solution and the greedy solution respectively confirming the latency-awareness of the learning agent.
- The rate of QoS violation was more than 75% which proves that reward approach is helpful in maintaining SLA compliance.
- The CPU usage is balanced, and has been optimized meaning the network supports VNFs better distributed in the various nodes of the network that ensures high scalability and there are no hotspots of being overloaded.

Even though the DQN-SFC model comes at the cost of moderate overheads in computations through these learning and inference cycles, it is rebalanced by huge gains in the efficiency of real time service delivery and utilization of resources. These results prove the viability and efficiency of the suggested solution to be implemented in the latency-sensitive 5g and beyond-5g networks.

Figure 2 compares the proposed DQN-SFC framework with the baseline approaches with respect to the three core metrics of the end-to-end latency, resource utilization, and QoS compliance, which shows that the proposed model outperforms the other methods substantially.

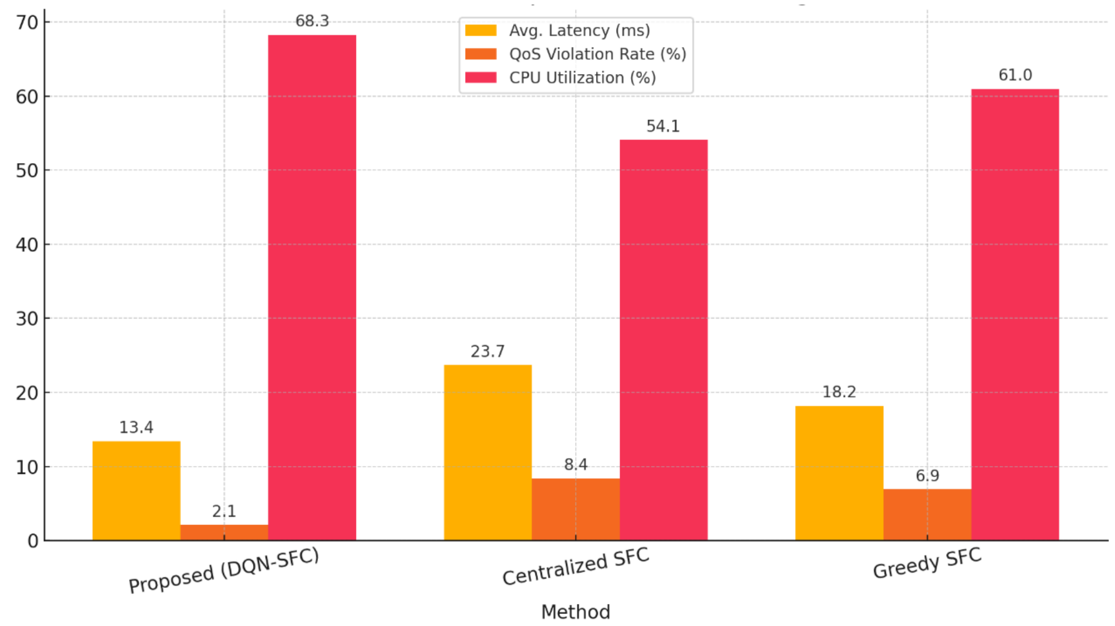

Figure 2. Comparative Performance of SFC Strategies.

This DQN-SFC model exhibits less latency, lesser QoS violations rates, and a better utilization of the CPU resources when compared to the centralized and greedy SFC approaches.

7. DISCUSSION

The experimental findings affirm that the multi-tier SDN architecture therein, coupled with a reinforcement learning -based SFC orchestration strategy can radically improve overall 5G network effectiveness in terms of latency, QoS compliance and resource efficiency. The hierarchical architecture with edge, fog, and core SDN controllers is effective to distribute the control logic and eliminate the centralized controller bottlenecks and allows making localized decisions to support latency-sensitive applications. This also automatically enhances network scalability and responsiveness particularly in dynamic and high density traffic conditions that are typical in 5G traffic conditions. Adaptive VNF placement and service path selection with the help of Deep Q-Network (DQN) has moderate computational overhead, but it has to be noted that most of the overhead is generated in the training session. This is however compensated by the fact that the model is capable of learning through network responses and adapting on the fly thus making better decisions and reliable service as time goes by. The model perfectly illustrates a positive trade-off between the complexity of learning and the efficiency of orchestration tasks and thus it can effectively be applied to the deployment of resource-limited settings where it is vital to implement intelligent adaptation.

Although the present contribution deals with fundamental issues regarding the reduction of SFC latency and QoS optimization, there is still a space to improve it. Future research may combine network sliceing techniques to enable service-level differentiation on verticals (automotive, healthcare, and IoT), and also instill Multi-Access Edge Computing (MEC) shackles in order to better the precision of VNF location in ultra-low latency instances. Besides, multi-agent reinforcement learning (MARL) framework scale-up of learning model can also be used to further decentralize the decisions and make them extra robust within heterogeneous copious networks.

8. CONCLUSION AND FUTURE WORK

In this paper, a brand new appropriate-latency and QoS-optimized Service Function Chaining (SFC) framework was proposed to the diverse-tier SDN-empowered 5G network. The presented solution will combine hierarchical SDN control structure and Deep Reinforcement Learning (DRL)-based orchestrator to rationally control placement of Virtual Network Functions (VNFs) and path selection. By the simulation-based assessment, the framework exhibited a substantial performance improvement in such areas as the smaller end-to-end latency, better QoS adherence, and reduced CPU resource consumption compared to the conventional centralized and heuristic-based SFC methods.

This work contains the most significant contributions as:

- A distributed and scalable control plane that employs edge, fog, and core SDN controllers in order to enhance response time in large 5G-scale deployments.
- The integration of a DQN-based learning model capable of adapting to real time

network and traffic conditions to provide dynamic intelligent orchestrations decisions.

- An end-to-end evaluation paradigm that ensures that the efficiency of the method holds true against some of the primary performance metrics such as latency, QoS violation rate, and resource usage.

In future, the works will be devoted to further expanding the framework to network slicing to deliver differentiated services as well as implementing Multi-Access Edge Computing (MEC) support, to further minimize latency to ultra-reliable low-latency communication (URLLC) applications. Furthermore, in decentralized 6G and beyond networks, multi-agent reinforcement learning (MARL) and federated orchestration have a potential to improve scalability, fault tolerance, and privacy aspects of the network itself.

REFERENCE

[1] Rehman, A., Malik, S. U. R., Jan, M. A., Khan, M. A., & Javaid, N. (2022). A QoS-aware service function chaining mechanism for SDN-enabled 5G networks. IEEE Access, 10, 48738–48752. https://doi.org/10.1109/ACCESS.2022.3169973

[2] Zhani, M. F., Zhang, Q., Simon, G., & Boutaba, R. (2021). VNF placement with replication for load balancing in NFV networks. IEEE Transactions on Network and Service Management, 18(3), 2823–2837. https://doi.org/10.1109/TNSM.2021.3089012

[3] Liu, J., Zhu, Z., Li, Y., & Miyazaki, T. (2022). Deep reinforcement learning-based online VNF placement and SFC embedding in SDN/NFV-enabled 5G networks. IEEE Transactions on Network and Service Management, 19(2), 1924–1939. https://doi.org/10.1109/TNSM.2022.3168985

[4] Gao, Y., Dong, L., Zhang, X., & Yu, H. (2023). Latency-aware VNF placement and SFC routing in 5G network slicing. IEEE Transactions on Mobile Computing, 22(1), 156–169. https://doi.org/10.1109/TMC.2022.3149157

[5] Liyanage, M., Said, S. B. H., Gurtov, A., & Yla-Jaaski, A. (2016). Software defined wireless networks: A survey. IEEE Communications Surveys & Tutorials, 18(4), 2713–2737. https://doi.org/10.1109/COMST.2016.2571118

[6] Koc, A. T., Wang, L., & Zhang, G. (2022). RL-SFC: Reinforcement learning-based SFC orchestration in SDN/NFV-enabled networks. IEEE Access, 10, 108734–108745. https://doi.org/10.1109/ACCESS.2022.3213426

[7] Gupta, A., Jha, R. K., & Dalal, U. D. (2023). Efficient VNF placement and SFC design in 5G edge networks using deep learning. IEEE Internet of Things Journal, 10(5), 4292–4304. https://doi.org/10.1109/JIOT.2022.3210365

[8] Ning, K., Li, Y., Zhao, Z., & Shen, X. (2021). Multi-agent deep reinforcement learning for VNF orchestration in large-scale SDN/NFV networks. IEEE Journal on Selected Areas in Communications, 39(8), 2335–2349. https://doi.org/10.1109/JSAC.2021.3075746