AI-Native Network Slicing for 6G Networks

Wen Wu, *Member, IEEE,* Conghao Zhou, *Student Member, IEEE,* Mushu Li, *Member, IEEE,*
Huaqing Wu, *Member, IEEE,* Haibo Zhou, *Senior Member, IEEE,* Ning Zhang, *Senior Member, IEEE,*
Xuemin (Sherman) Shen, *Fellow, IEEE*, and Weihua Zhuang, *Fellow, IEEE*

***Abstract*—With the global roll-out of the fifth generation (5G) networks, it is necessary to look beyond 5G and envision the 6G networks. The 6G networks are expected to have space-air-ground integrated networks, advanced network virtualization, and ubiquitous intelligence. This article presents an artificial intelligence (AI)-native network slicing architecture for 6G networks to enable the synergy of AI and network slicing, thereby facilitating intelligent network management and supporting emerging AI services. AI-based solutions are first discussed across network slicing lifecycle to intelligently manage network slices, i.e., *AI for slicing*. Then, network slicing solutions are studied to support emerging AI services by constructing AI instances and performing efficient resource management, i.e., *slicing for AI*. Finally, a case study is presented, followed by a discussion of open research issues that are essential for AI-native network slicing in 6G networks.**

***Index Terms*—6G, AI-native, network slicing, AI for slicing, slicing for AI, ubiquitous intelligence.**

I. Introduction

Compared with existing wireless networking including the fifth generation (5G), 6G is more than an improvement of key performance indicators (KPI) requirements, such as increased data rates, enhanced network capacity, and low latency. The 6G networks are envisioned to have the following unique features. First, space networks, e.g., low earth orbit (LEO) satellites, air networks, e.g., unmanned aerial vehicles (UAVs), and ground networks, e.g., cellular base stations (BSs), are integrated into a *space-air-ground integrated network (SAGIN)* to provide global coverage and on-demand services [1]. Second, resource virtualization using network slicing techniques and end user virtualization using digital twin techniques can facilitate *advanced network virtualization* to provide flexible network management [2], [3].[1] Third, intelligence penetrates every corner of networks, ranging from end users, the network edge, to the remote cloud, which results in *ubiquitous intelligence*. A number of network nodes are endowed with built-in artificial intelligence (AI) functionalities, thereby not only facilitating intelligent network management but also fostering AI services, e.g., deep neural network based applications. Hence, 6G networks are expected to create a new wireless networking ecosystem that brings societal and economic benefits.

The 6G networks will support a diverse set of services with different quality of service (QoS) requirements, such as multisensory extended reality and hologram video streaming. To support diversified services as established in 5G networks, network slicing is a potential approach to construct multiple logically-isolated virtual networks (i.e., slices) for different services on top of the common physical network [4]. The QoS requirements of different services can be guaranteed via cost-effective slice management strategies ranging from preparation, planning, and operation phases in the network slicing lifecycle.

Developing network slicing schemes faces many challenges in 6G networks due to their unique features. First, managing slices over space, air, and ground network segments in the SAGIN requires judicious coordination of heterogeneous network segments. Moreover, the 6G networks need to support a variety of new services while satisfying their different and stringent QoS requirements, which further complicates slice management. Hence, it is paramount to develop intelligent slice management solutions in 6G networks. Second, fuelled by powerful computing capability and advanced AI techniques, ubiquitous intelligence is fostering abundant AI services with new QoS requirements, such as data quality, inference accuracy, and training latency. Hence, it is necessary to construct customized network slices to support the emerging AI services in 6G networks.

In this article, we propose an *AI-native* network slicing architecture for 6G networks to facilitate intelligent network management while supporting emerging AI services. AI-native means that, as a built-in component in the network slicing architecture, AI exists not only in the software-defined networking (SDN) controller for managing network slices, but also in network slices as services for end users. Hence, the synergy of AI and network slicing in the proposed architecture is two-fold: On one hand, AI techniques can be applied to manage network slices, namely *AI for slicing*. The network slicing lifecycle including preparation, planning, and operation phases is introduced, along with specifying AI-based solutions for each phase. In addition, the detailed procedure of information exchange among end users, access points, and the SDN controller is presented; On the other hand, network slicing can be applied to construct customized network slices for various AI services, namely *slicing for AI*. Potential approaches such as AI instance construction and efficient resource management for AI services are introduced.

W. Wu is with the Frontier Research Center, Peng Cheng Laboratory, Shenzhen, Guangdong, China (email:w77wu@uwaterloo.ca).

C. Zhou, M. Li, H. Wu, X. Shen, and W. Zhuang are with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario, Canada (email:{c89zhou, m475li, h272wu, sshen, wzhuang}@uwaterloo.ca). *Corresponding author: Huaqing Wu*.

H. Zhou is with the School of Electronic Science and Engineering, Nanjing University, Nanjing, China (email: haibozhou@nju.edu.cn).

N. Zhang is with the Department of Electrical and Computer Engineering, University of Windsor, Windsor, Ontario, Canada (email: ning.zhang@uwindsor.ca).

[1]End user virtualization is to virtualize end users in network operation and management by characterizing users' behaviours and status (e.g., service demands and QoS satisfaction), which can be achieved using digital twin concepts.

The remainder of this article is organized as follows. In Section II, expected features of 6G networks are discussed, and then the AI-native network slicing architecture is proposed. The basic ideas of AI for slicing and slicing for AI are presented in Section III and Section IV, respectively. A case study is presented in Section V. In Section VI, the research directions are identified, followed by the conclusion in Section VII.

II. AI-Native Network Slicing for 6G Networks

A. Network Slicing

Network slicing is an emerging technology to support diversified applications in a cost-effective manner [4], [5]. The concept of network slicing can be traced back to the late 1980s [6]. Nowadays, network slicing is a key technology in 5G networks, supported by network function virtualization (NFV) and SDN techniques. Specifically, NFV enables virtualized resources and network functions for flexible resource management, while SDN facilitates centralized network management for network optimization. In 5G networks, network slicing has been defined in the 3rd generation partnership project (3GPP) Release 15 [7]. Moreover, in coming 6G networks, network slicing will continue evolving and play an increasingly important role.

The basic idea of network slicing is to create multiple logically-isolated network slices on top of the common physical infrastructure, which can achieve flexible and adaptive network management. Its benefits are three-fold: 1) Multi-tenancy - Multiple virtual networks can share the common physical infrastructure, thus reducing capital expenditures in the network deployment; 2) Service isolation - Multiple slices are constructed for different services via judicious resource management, such that service level agreements of different slices can be effectively guaranteed; 3) Flexibility - Network slicing can support flexible network management, as slices can be created, modified, or deleted on-demand.

B. Features of 6G Networks

From 5G to 6G, it is in general expected KPI requirements to be increased by at least an order of magnitude. According to a recent white paper [8], the KPI requirements of the 6G networks include 1 Tbps peak data rate, 20-100 Gbps user experienced data rate, 0.1 ms end-to-end latency, 10 million devices/km^2, and near 100% coverage. Such KPI requirements demand several candidate technologies, such as THz communications and AI [6]. The 3GPP working group will discuss 6G candidate techniques by the end of 2026, and the first 6G standard is expected to debut by 2030.

Distinguished from 5G networks, 6G networks have several features:

- SAGIN - While current ground networks provide good coverage in highly populated areas, 6G needs to provide universal coverage, including in rural areas, remote lands, and sparsely populated areas. To achieve this goal, 6G will exploit the altitude dimension. Space, air, and ground network segments are integrated into the SAGIN [1], [9], which can provide global coverage, facilitate on-demand services, and support high-rate low-delay services;
- Diversified services - Many services have stringent QoS requirements in different dimensions. Mobile virtual reality (VR) and hologram video streaming applications require a high data rate, e.g., the uplink data rate of mobile VR is up to 5 Gbps. Other applications may require ultra-high reliability, such as autonomous driving, industrial control systems, and robot/UAV swarm, e.g., the required reliability of autonomous driving is up to 99.999% [10];
- Ubiquitous intelligence - With caching capability, a large amount of data can be stored in the network. In addition, with the development of AI techniques, edge computing, and device computing, intelligence is pushed from the remote cloud to the network edge and end users. As such, AI will be integrated into 6G networks for intelligent network management by directly learning from extensive data in the network. Moreover, ubiquitous intelligence will foster a number of AI services in which AI is provided as services.

C. AI-Native Network Slicing Architecture

These features impose new challenges on developing network slicing schemes for 6G networks. Firstly, the SAGIN not only increases the number of integrated network segments but also introduces extra dynamics on network resource availability due to satellite mobility and UAV manoeuvrability. Network slicing schemes should accommodate for the large-scale SAGIN while taking dynamic resource availability into account. Moreover, supporting diversified services with stringent QoS requirements further complicates network slicing scheme design. Secondly, ubiquitous intelligence facilitates many emerging AI services which will be prevalent in 6G networks. Different from conventional services, facilitating AI services requires multiple steps, including collecting high-quality data samples, training satisfactory AI models, and performing low-latency model inference, which should meet diverse QoS requirements. How to satisfy such diverse QoS requirements for AI services remains a challenging issue.

To address the above challenges, an AI-native network slicing architecture for 6G networks is presented. As shown in Fig. 1, the architecture aims at integrating SAGIN and ubiquitous intelligence and supporting diverse services with stringent QoS requirements. Compared with network slicing for 5G networks, the proposed architecture has two new characteristics. Firstly, AI is integrated into SDN controllers to realize intelligent network slicing, such that a number of network slices with stringent QoS requirements can be managed efficiently and cost-effectively via AI techniques, which is referred to as *AI for slicing*. Secondly, emerging AI services are supported by network slicing. In addition to network slices for conventional services, new network slices are constructed for AI services on top of the common physical infrastructure, which is referred to as *slicing for AI*.

Two types of SDN controllers are deployed in the proposed architecture. One is the centralized SDN controller located at the cloud, which is to manage network slices. The other is the local SDN controller located at access points, which is

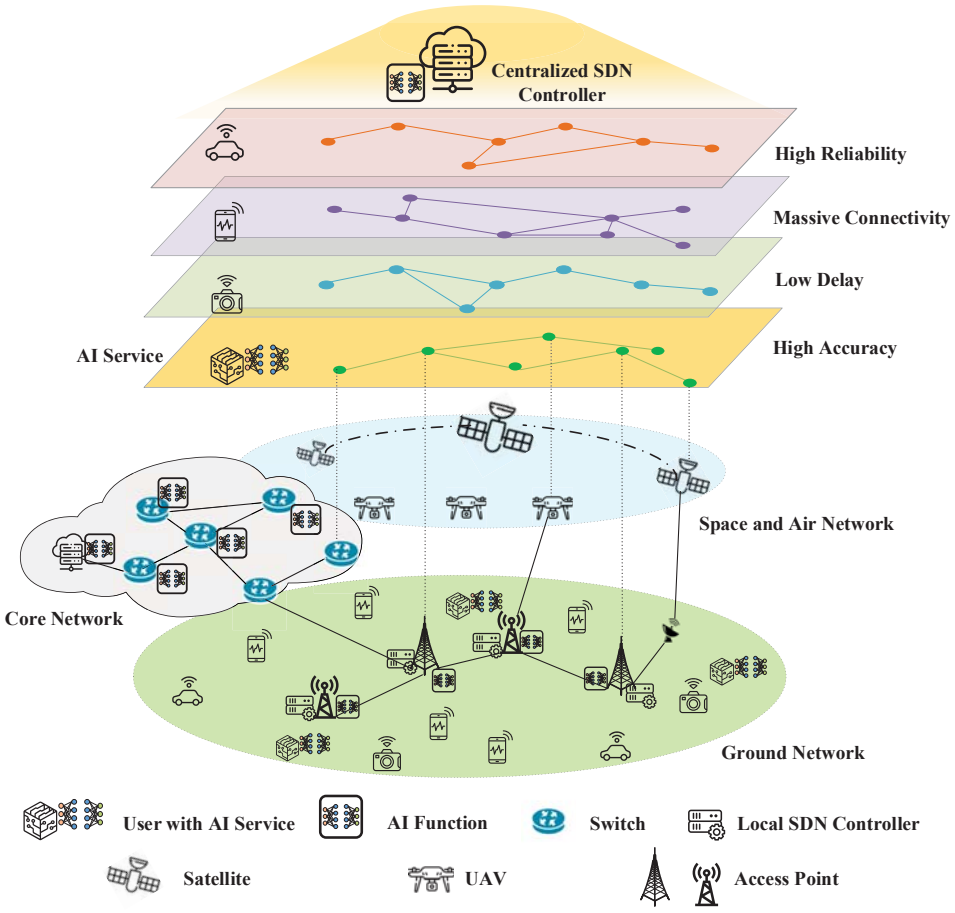


Fig. 1. An illustration of the AI-native network slicing architecture for 6G networks.

to schedule resources to end users within each network slice. The SDN controller in the following refers to the centralized SDN controller unless otherwise stated. In the following, we will illustrate the basic ideas of AI for slicing and slicing for AI in Section III and Section IV, respectively.

III. AI for Slicing

In this section, we introduce the network slicing lifecycle with three phases and then investigate potential AI solutions for each phase. Next, the corresponding procedure of information exchange in AI for slicing is discussed.

A. Network Slicing Lifecycle

The network slicing lifecycle consists of three phases: *preparation*, *planning*, and *operation*, as shown in Fig. 2. The centralized SDN controller is in charge of the preparation and planning phases, while the operation phase is coordinated by local SDN controllers.

1) Preparation Phase: This phase is to construct and configure network slices based on service requirements, data traffic, user information, and virtual network resource availability. To achieve the goal, the SDN controller conducts the following tasks:

- Service requirement extraction - This task is to classify services by extracting their QoS requirements, such as service delay, service priority, throughput, and reliability. The 3GPP has standardized specific service/slice type values for classified services, such as enhanced mobile broadband, ultra-reliable low-latency communications, and massive machine-type communications services [7];
- Network resource and function virtualization - Network resources, such as communication, computing, and caching resources, are pooled into virtualized resource blocks via advanced resource virtualization techniques. Similarly, network functions, such as firewall, network name translation, and domain name system, are separated from dedicated hardware network functions into virtualized network functions (VNFs). Through virtualization, the SDN controller can flexibly manage network resources and functions.

Once these tasks are completed, the SDN controller can construct network slices for each admitted slice request.

2) Planning Phase: This phase aims at reserving network resources to slices for service provisioning. The planning phase operates in a large timescale. Time is partitioned into multiple planning periods (windows) for each slice. The duration of each planning window depends on service demand and network dynamics, whose value ranges from several minutes to several hours. To achieve the goal, the following two steps are conducted in the planning phase:

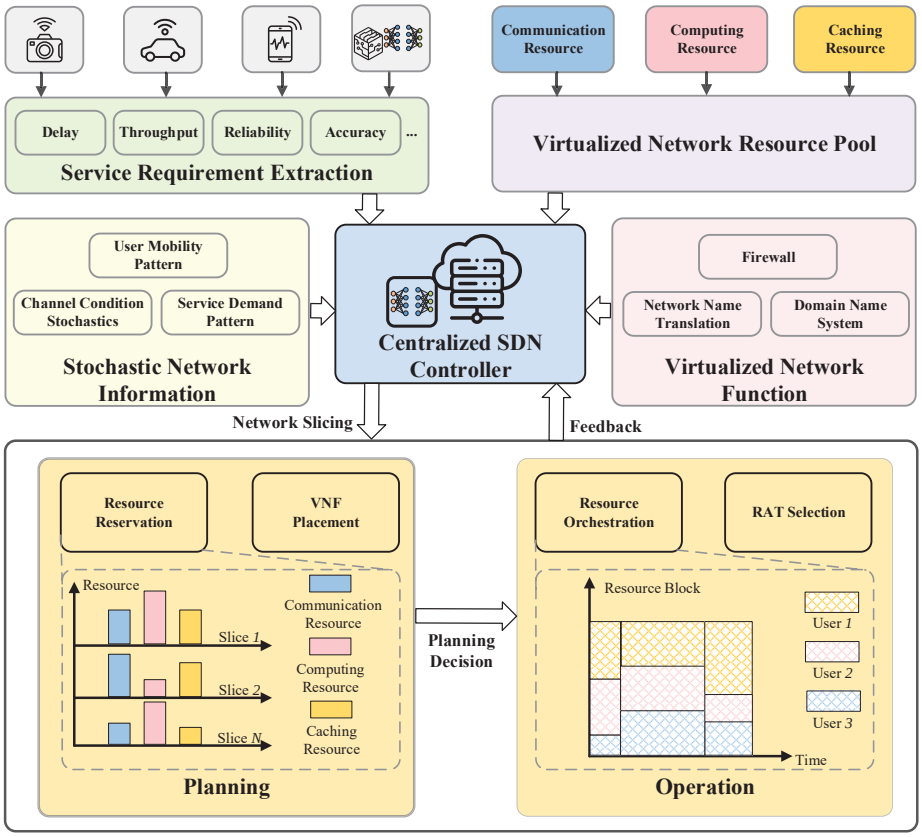


Fig. 2. An illustrative example for the network slicing lifecycle which includes preparation, planning, and operation phases.

- Service and network information collection - Benefiting from the global control functionality of the SDN controller, extensive network information can be collected from underlying physical networks, such as service demands, stochastic channel conditions, and user mobility patterns. The collected information is utilized for the following resource reservation decision making;
- Resource reservation - At the beginning of each planning window, the SDN controller adjusts the amount of reserved network resources for each slice based on the monitored slice performance. The reserved virtualized network resources of each slice are mapped to the physical network. At the end of each planning window, some system information is fed back to the SDN controller, such as resource utilization, system performance, and service level agreement satisfaction. Based on the feedback information, the SDN controller can adjust resource reservation decisions to accommodate dynamic network environments while guaranteeing QoS requirements.

3) Operation Phase: This phase is to schedule the service of a slice using the reserved resources for subscribed end users. The operation phase works in a much smaller timescale (e.g., 100 ms) than that in the planning phase. Specifically, under the coordination of the centralized SDN controller, local SDN controllers allocate network resources to end users in each slice according to their real-time data traffic. The operation decisions include selecting radio access technology (RAT), determining user association with specific radio access points, deciding proper protocol and associated parameters, and orchestrating resources among end users.

B. Roles of AI in Network Slicing

Although network slicing can facilitate service provisioning, managing a number of network slices incurs significant network management cost, especially in 6G networks. As shown in Fig. 3, AI-based network slicing is a potential solution in which AI plays different roles in different network slicing phases.

AI for preparation: In the preparation phase, AI needs to perform two tasks. 1) Service demand prediction - Based on historical data, service demand can be predicted via AI techniques, such as recurrent neural networks. Prior studies show that the service demand and resource usage of a slice can be accurately predicted [11]. The prediction results can be utilized for decision making in the planning phase. 2) Slice admission - The SDN controller admits slices to maximize network resource utilization considering resource availability and service demands. As the slice admission decision is binary, this problem is deemed as an integer optimization problem. In large-scale networks with complex resource availability distribution, conventional optimization solutions become complicated and intractable, while AI-based solutions are potential.

AI for planning: In the planning phase, AI can perform two tasks. 1) VNF placement - The SDN controller deploys VNFs

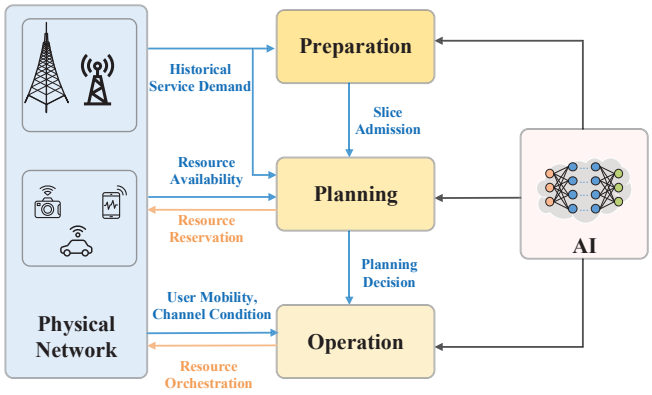


Fig. 3. The considered AI-based network slicing solution in which AI plays different roles in preparation, planning, and operation phases.

to support services in the network. The resources allocated for VNFs should be dynamically adjusted for time-varying service demands to guarantee service delay requirements. Deep learning methods can be applied to enhance resource utilization in dynamic network environments. 2) Resource reservation - The SDN controller reserves resources for different slices based on their service demands. Since data traffic loads are time-varying, the resource reservation should be adaptive to dynamic real-time demands, which can be addressed via reinforcement learning (RL) methods, such as deep deterministic policy gradient (DDPG).

AI for operation: Two exemplary operation tasks are as follows: 1) Resource orchestration - The reserved resources of a slice are allocated to end users. The decisions are determined based on real-time user mobility, service demands, etc. To efficiently utilize resources, RL methods can be applied for dynamic resource orchestration; 2) RAT selection - To maximize system utility, an optimal RAT is selected among multiple candidate RATs for each end user. Due to user mobility, user-perceived service performance of an RAT is stochastic. Such problem can be addressed by multi-armed bandit methods, e.g., contextual bandit.

C. Procedure of Information Exchange

The AI for slicing procedure involves the information exchange among end users, access points, and the SDN controller. The procedure is illustrated in Fig. 4 with steps as follows:

1) Access points collect user-level stochastic information, such as end users' service demand patterns, mobility patterns, and stochastic channel conditions;
2) Access points translate the user-level information into desired service-level information. For example, user density information can be obtained from processing user location information, and AI techniques can be used for such data abstraction, fusion, and analysis;
3) The processed service-level information is delivered to the SDN controller;
4) The SDN controller runs AI-based planning algorithms to make decisions based on the collected service-level information;
5) The determined planning decisions are sent back to all access points;
6) Access points enforce the received planning decisions, e.g., reserving network resources for corresponding slices;
7) End users in service report their real-time information to their associated access points, such as real-time service demands, channel conditions, and task data sizes;
8) Access points run the AI-based operation algorithm to allocate resources for end users based on real-time user-level information;
9) Service requests from end users are supported with the allocated network resources. For example, computation tasks can be offloaded to access points using communication resource and then processed using computing resources. For each operation slot within a planning window, Steps 7-9 are repeated;
10) Access points monitor slice performance in the network given the enforced planning decisions by measuring end users' satisfaction rates across all operation slots within a planning window;
11) Access points report network performance to the SDN controller;
12) The SDN controller makes the planning decision for next planning window and adjusts the planning policy based on the feedback information.

Note that in the preceding procedure, Steps 1-6 and Steps 10-12 are in the network planning phase, and Steps 7-9 are in the operation phase.

IV. Slicing for AI

The slicing for AI is to utilize network slicing to support AI services while satisfying QoS requirements. Potential solutions include constructing and selecting AI instances and efficient resource management in the AI service lifecycle.

A. AI Instance

There are diversified implementation options for supporting AI services. An AI service can be implemented via different kinds of algorithms, training manners, and network resource allocation. For example, objective detection services can be implemented via ResNet32, Inception-v3, AlexNet, or VGG16 algorithms. Hence, the primary issue of supporting an AI service is to determine an appropriate implementation option in the network.

We introduce the concept of *AI instance* to address the issue, as shown in Fig. 5. An AI instance of an AI service represents an implementation option for an AI service. The basic idea is to construct multiple candidate implementation options and then select an appropriate one based on network environments. The procedure of the conceptual AI instance management framework consists of two steps. 1) AI instance construction - The network operator constructs multiple candidate AI instances for each AI service based on available network resources and service requirements. An AI instance may include (i) the AI algorithm which specifies the implementation

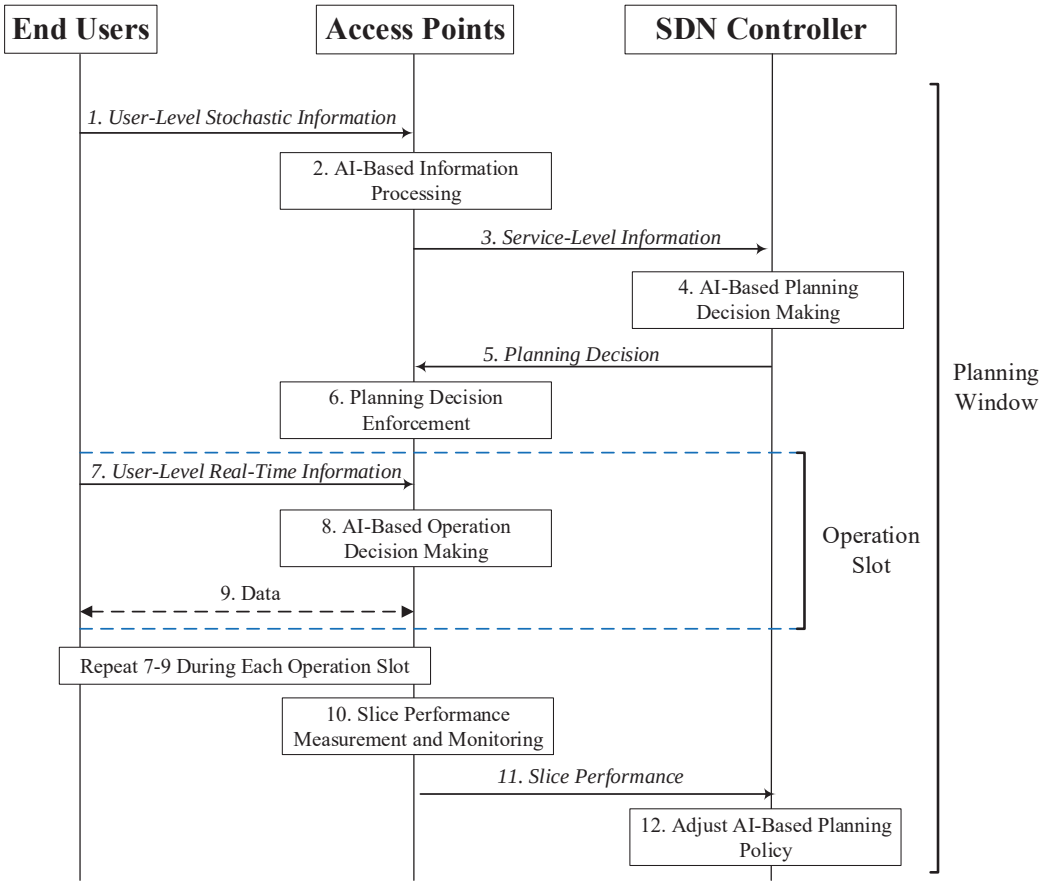


Fig. 4. Procedure of information exchange in AI for slicing.

algorithm and the corresponding neural network architecture, (ii) the training manner of the AI algorithm, e.g., centralized or distributed training, and (iii) the amount of the required network resources. 2) AI instance selection - In this step, the AI service provider selects an appropriate AI instance among candidate AI instances based on user service preference (e.g., privacy preservation preference). If an AI instance is selected, the AI service will be executed using the AI algorithm and the corresponding required amount of network resources by the AI instance. In summary, the idea of AI instance provides flexibility for AI service management.

B. Resource Management in AI Service Lifecycle

Running AI services includes three stages: data collection, model training, and model inference, i.e., AI service lifecycle [12], [13]. Specifically, *data collection* is to collect data via communication links, and the collected data can be stored in network edge servers. Based on the collected data, an AI model can be trained in the *model training* stage. The model training can be implemented in either a centralized or a distributed manner. For example, multiple devices can work collaboratively to train a global model via federated learning. Next, well-trained AI models are deployed to execute specific computation tasks, which is referred to as *model inference*. The model inference can be performed in multiple manners. For example, device-edge collaborative inference approaches can allocate and process computation tasks at different network nodes to achieve a low inference latency.

The performance of an AI service depends on all the three stages in the AI service lifecycle. For example, model inference accuracy depends on multiple factors, such as the quality of the collected data, the number of training iterations, and the approach of model inference. Meanwhile, all these three stages consume multi-dimensional network resources. As a result, to optimize the performance of AI services, network resources should be jointly allocated for these three stages. The reserved network resources in AI slices should be further allocated to these three stages to satisfy their corresponding QoS requirements.

V. Case Study

In this section, a case study is provided on AI-assisted resource reservation, aiming at reducing long-term overall system cost.

A. Considered Scenario

We consider an air-ground integrated network for providing autonomous driving services to vehicles traversing a highway segment. For the considered highway segment with a length

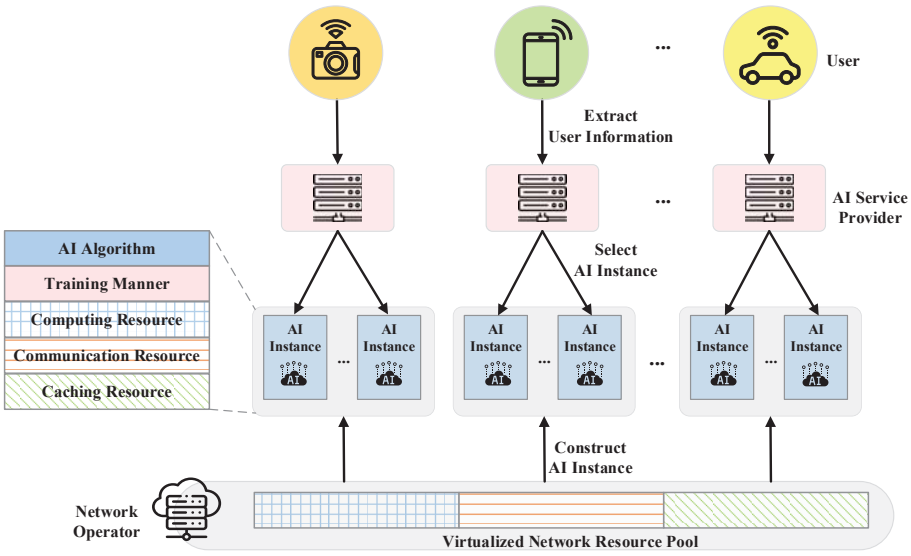


Fig. 5. The conceptual AI instance management framework for AI services.

of 2 km, two BSs are uniformly deployed along the highway with a separation distance of 1 km, and one UAV is deployed in the centre hovering at a height of 100 meters. When an autonomous vehicle is driving on the highway, extensive computation-intensive tasks are required to be processed. For prompt task processing, all access points are equipped with edge computing servers, and vehicles can associate to the nearest access point and upload their computation tasks. We consider a delay-sensitive autonomous driving service, e.g., object detection, whose delay requirement is 100 *ms* for road safety [14]. The data size and computation intensity of a task are set to 0.6 Mbit and 6×10^8 cycles, respectively. The service delay is characterized by the queuing theory since task arrivals are assumed to follow a Poisson process with rate $\lambda = 1$ packet/sec. To guarantee the service delay requirement, a network slice is constructed, in which spectrum and computing resources are reserved.

Resource reservation decisions are made to minimize the overall system cost considering vehicle traffic dynamics. The overall system cost is defined as $C = \sum_{t=1}^{T}(\omega_r C_r^t + \omega_s C_s^t + \omega_d C_d^t)$, which is a weighted summation of three cost components across all T planning windows. 1) Resource reservation cost C_r^t accounts for the amount of the reserved spectrum and computing resources at BSs at planning window t. The spectrum resource is allocated in a unit of subcarrier of 5 MHz, and the computing resource is allocated in a unit of virtual machine (VM) instance with a processing rate of 10×10^9 cycles/sec; 2) Slice reconfiguration cost C_s^t accounts for the difference between two consecutive resource reservation decisions [15]; 3) Delay requirement violation penalty C_d^t refers to the penalty once service delay exceeds the delay requirement. These weight parameters are set to $\omega_r = 1$, $\omega_s = 20$, and $\omega_d = 200$, respectively. The planning window size is set to one hour.

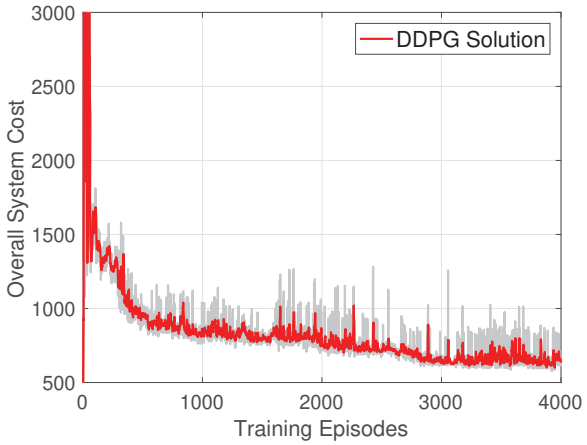
We propose a DDPG-based solution to minimize the overall system cost [15]. In this solution, both actor and critic networks are fully-connected neural networks with four layers, the numbers of neurons in two hidden layers are 128 and 64, respectively, and their learning rates are set to 2×10^{-4} and 2×10^{-3}, respectively. For performance comparison, we adopt an optimization-based solution, named *myopic resource reservation*, in which network resources are reserved to minimize the resource reservation cost at each planning window while satisfying the delay requirement.

B. Simulation Results

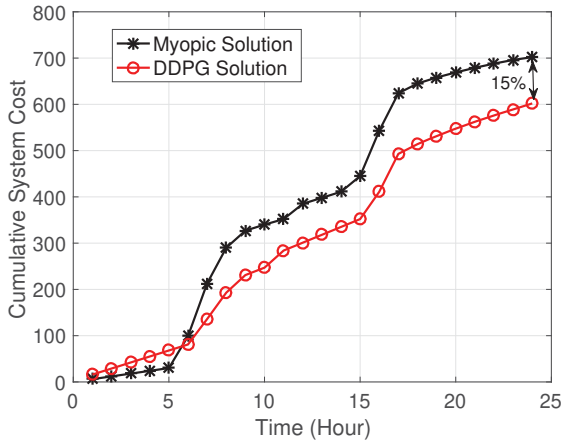
We evaluate the performance of the proposed DDPG-based solution based on real-world highway vehicle traffic flow trace collected by Alberta Transportation.[2] As shown in Fig. 6(a), we first present the convergence performance of the proposed DDPG-based solution. A five-point moving average is applied to process raw simulation points to highlight the convergence trend (i.e., red curve). It can be seen that the DDPG-based resource reservation solution has converged after 4,000 training episodes.

Next, as shown in Fig. 6(b), the cumulative system cost within one day is presented. It can be observed that the DDPG-based solution can reduce the cumulative overall system cost within one day by around 15% as compared to the myopic solution. The reason is that the proposed DDPG-based solution is able to minimize the long-term overall system cost, while the myopic solution minimizes the short-term system cost, which incurs prohibitive slice reconfiguration cost due to frequent adjustment of network resource reservation in highly dynamic vehicular networks. The simulation results show that the proposed AI-based resource reservation solution can achieve a low system cost.

[2]Alberta Transportation: http://www.transportation.alberta.ca/mapping/.



(a) Convergence performance



(b) Cumulative system cost within one day

Fig. 6. Performance evaluation of the proposed DDPG-based resource reservation solution.

VI. Open Research Issues

In the following, we discuss some open research issues pertaining to AI-native network slicing.

A. Joint Design of Network Planning and Operation

Planning and operation are performed and coupled in two different timescales. Specifically, the planning phase is performed at a large timescale (e.g., minute level) to reserve resources for different slices based on service demands, while the operation phase is performed at a small timescale (e.g., sub-second level) to allocate the reserved resources to on-demand users within each slice. Achieving the optimal network slicing performance requires a joint optimization design of planning and operation.

B. Data Management Framework

The cornerstone of AI-native network slicing is abundant data that can be used for AI model training. In 6G networks, data is widely distributed in the network. Due to limited communication resources, the cost of collecting a large amount of data cannot be neglected. In addition, the collected data is required to be processed to mine valuable information for network management. For example, abundant historical behaviour data from individual users can be analyzed to predict spatio-temporal service demand distributions. Hence, establishing a data management framework to collect and analyze data is necessary.

C. Prediction-Empowered Network Slicing

With the development of advanced AI technologies, the data traffic in the network can be predicted. How to effectively leverage the power of prediction for network slicing is an interesting topic. Since the prediction is imperfect, the prediction error may degrade the performance of network slicing. How to evaluate the impact of prediction errors on system performance and to develop corresponding solutions are important research issues.

VII. Conclusion

In this article, we have proposed the AI-native network slicing architecture to facilitate intelligent network management and support AI services in 6G networks. The architecture aims at enabling the synergy of AI and network slicing. The AI for slicing is to help reduce network management complexity, while adapting to dynamic network environments by exploiting the capability of AI in network slicing. The slicing for AI is to construct customized network slices to better accommodate various emerging AI services. To accelerate the pace of AI-native network slicing architecture development, extensive research efforts are required, such as in the identified research directions.

Acknowledgements

This work was financially supported by Natural Sciences and Engineering Research Council (NSERC) of Canada. The authors would like to thank Jie Gao, Qihao Li, and Kaige Qu for many valuable discussions and suggestions throughout the work.

References

[1] N. Zhang, S. Zhang, P. Yang, O. Alhussein, W. Zhuang, and X. Shen, "Software defined space-air-ground integrated vehicular networks: Challenges and solutions," *IEEE Commun. Mag.*, vol. 55, no. 7, pp. 101–109, Jul. 2017.

[2] X. Shen, J. Gao, W. Wu, M. Li, C. Zhou, and W. Zhuang, "Holistic network virtualization and pervasive network intelligence for 6G," *submitted to IEEE Commun. Surveys Tuts.*, 2021.

[3] R. Minerva, G. M. Lee, and N. Crespi, "Digital twin in the IoT context: A survey on technical features, scenarios, and architectural models," *Proc. IEEE*, vol. 108, no. 10, pp. 1785–1824, Oct. 2020.

[4] X. Shen, J. Gao, W. Wu, K. Lyu, M. Li, W. Zhuang, X. Li, and J. Rao, "AI-assisted network-slicing based next-generation wireless networks," *IEEE Open J. Veh. Technol.*, vol. 1, no. 1, pp. 45–66, Jan. 2020.

[5] W. Zhuang, Q. Ye, F. Lyu, N. Cheng, and J. Ren, "SDN/NFV-empowered future IoV with enhanced communication, computing, and caching," *Proc. IEEE*, vol. 108, no. 2, pp. 274–291, Feb. 2020.

[6] X. You *et al.*, "Towards 6G wireless communication networks: Vision, enabling technologies, and new paradigm shifts," *Sci. China Inf. Sci.*, vol. 64, no. 1, pp. 1–74, Jan. 2021.

[7] A. Kaloxylos, "A survey and an analysis of network slicing in 5G networks," *IEEE Communications Standards Magazine*, vol. 2, no. 1, pp. 60–65, Mar. 2018.

[8] N. Rajatheva *et al.*, "White paper on broadband connectivity in 6G," *arXiv:2004.14247*, 2020, [Online]. Available: https://arxiv.org/abs/2004.14247.
[9] C. Zhou, W. Wu, H. He, P. Yang, F. Lyu, N. Cheng, and X. Shen, "Deep reinforcement learning for delay-oriented IoT task scheduling in space-air-ground integrated network," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 911–925, Feb. 2021.
[10] C. Campolo, A. Molinaro, A. Iera, and F. Menichella, "5G network slicing for vehicle-to-everything services," *IEEE Wireless Commun.*, vol. 24, no. 6, pp. 38–45, Dec. 2017.
[11] C. Gutterman, E. Grinshpun, S. Sharma, and G. Zussman, "RAN resource usage prediction for a 5G slice broker," in *Proc. ACM MobiHoc*, Catania, Italy, 2019.
[12] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.
[13] M. Li, J. Gao, C. Zhou, X. Shen, and W. Zhuang, "Slicing-based AI service provisioning on network edge," *IEEE Veh. Technol. Mag.*, 2021, to appear.
[14] S.-C. Lin, Y. Zhang, C.-H. Hsu, M. Skach, M. E. Haque, L. Tang, and J. Mars, "The architectural implications of autonomous driving: Constraints and acceleration," in *Proc. ASPLOS*, 2018, pp. 751–766.
[15] W. Wu, N. Chen, C. Zhou, M. Li, X. Shen, W. Zhuang, and X. Li, "Dynamic RAN slicing for service-oriented vehicular networks via constrained learning," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 7, pp. 2076–2089, 2021.

BIOGRAPHIES

Wen Wu (S'13-M'20) earned the Ph.D. degree in Electrical and Computer Engineering from University of Waterloo, Waterloo, ON, Canada, in 2019. He received the B.E. degree in Information Engineering from South China University of Technology, Guangzhou, China, and the M.E. degree in Electrical Engineering from University of Science and Technology of China, Hefei, China, in 2012 and 2015, respectively. He worked as a Post-doctoral fellow with the Department of Electrical and Computer Engineering, University of Waterloo. He is currently an Associate Researcher at the Peng Cheng Laboratory, Shenzhen, China. His research interests include 6G network, pervasive network intelligence, and network virtualization.

Conghao Zhou (S'19) received the B.E. degree from Northeastern University, Shenyang, China, in 2017 and received the M.S. degree from University of Illinois at Chicago, Chicago, IL, USA, in 2018. He is currently working toward the Ph.D. degree with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada. His research interests include space-air-ground integrated networks and machine learning in wireless networks.

Mushu Li (S'17-M'21) is a postdoctoral fellow with the Department of Electrical and Computer Engineering, University of Waterloo, Ontario, N2L 3G1, Canada, where she also received her Ph.D. degree in electrical engineering in 2021. She received her M.A.Sc. degree from Ryerson University in 2017. She was the recipient of Natural Science and Engineering Research Council of Canada Graduate Scholarship in 2018, and Ontario Graduate Scholarship in 2015 and 2016, respectively. Her research interests include machine learning in wireless networks and network slicing.

Huaqing Wu (S'15-M'21) received the Ph.D. degree from University of Waterloo, Ontario, Canada, in 2021. She received the B.E. and M.E. degrees from Beijing University of Posts and Telecommunications, Beijing, China, in 2014 and 2017, respectively. She received the prestigious Natural Sciences and Engineering Research Council of Canada (NSERC) Postdoctoral Fellowship Award in 2021. She is currently a postdoctoral research fellow at the McMaster University, Ontario, Canada. Her current research interests include vehicular networks with emphasis on edge caching, wireless resource management, space-air-ground integrated networks, and application of artificial intelligence (AI) for wireless networks.

Haibo Zhou (M'14-SM'18) received the Ph.D. degree in information and communication engineering from Shanghai Jiao Tong University, Shanghai, China, in 2014. He is currently an Associate Professor with the School of Electronic Science and Engineering, Nanjing University, Nanjing, China. He was named the 2020 highly cited researcher in cross-field. He was a recipient of the 2019 IEEE ComSoc Asia-Pacific Outstanding Young Researcher Award. He is currently an Associate Editor of the IEEE IoTJ, IEEE Network, IEEE WCL. His research interests include resource management in VANET, 5G/B5G wireless networks and SAGINs.

Ning Zhang (M'15-SM'18) is currently an Associate Professor in the Department of Electrical and Computer Engineering at University of Windsor, Canada. He received the Ph.D degree in Electrical and Computer Engineering from University of Waterloo, Canada, in 2015. His research interests include connected vehicles, mobile edge computing, wireless networking, and machine learning. He is a Highly Cited Researcher. He serves as an Associate Editor of IEEE Internet of Things Journal, IEEE Transactions on Cognitive Communications and Networking, and IEEE Systems Journal; and a Guest Editor of several international journals, such as IEEE Wireless Communications and IEEE Transactions on Industrial Informatics.

Xuemin (Sherman) Shen (M'97-SM'02-F'09) is currently a University Professor with the Department of Electrical and Computer Engineering, University of Waterloo, Canada. His research focuses on network resource management, wireless network security, social networks, 5G and beyond, and vehicular ad hoc networks. He is a Canadian Academy of Engineering Fellow, a Royal Society of Canada Fellow, and a Chinese Academy of Engineering Foreign Fellow. He received the R.A. Fessenden Award in 2019 from IEEE, Canada; the James Evans Avant Garde Award in 2018 from the IEEE Vehicular Technology Society; and the Joseph LoCicero Award in 2015 and Education Award in 2017 from the IEEE Communications Society.

Weihua Zhuang (M'93-SM'01-F'08) has been with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada, since 1993, where she is currently a professor and a Tier I Canada Research Chair in Wireless Communication Networks. Dr. Zhuang is a Fellow of the Royal Society of Canada, the Canadian Academy of Engineering, and the Engineering Institute of Canada. She is also an elected member of the Board of Governors and VP-Publications of the IEEE Vehicular Technology Society.