AI-Native Network Slicing for 6G Networks

Wen Wu, *Member, IEEE,* Conghao Zhou, *Student Member, IEEE,* Mushu Li, *Student Member, IEEE,* Huaqing Wu, *Student 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 sixth generation (6G) networks. The 6G networks are expected to have space-air-ground integrated networking, advanced network virtualization, and ubiquitous intelligence. This article proposes an artificial intelligence (AI)-native network slicing architecture for 6G networks to facilitate intelligent network management and support emerging AI services. AI is built in the proposed network slicing architecture to enable the synergy of AI and network slicing. AI solutions are investigated for the entire lifecycle of network slicing to facilitate intelligent network management, i.e., *AI for slicing*. Furthermore, network slicing approaches are discussed to support emerging AI services by constructing slice 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.

Index Terms

6G networks, AI-native, network slicing, AI for slicing, slicing for AI, ubiquitous intelligence, space-air-ground integrated network.

I. Introduction

Compared with existing wireless networking including the fifth generation (5G), the sixth generation (6G) is more than an improvement of key performance indicators (KPI) requirements,

W. Wu, C. Zhou, M. Li, H. Wu, W. Zhuang, and X. Shen are with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (email:{w77wu, c89zhou, m475li, h272wu, sshen, wzhuang}@uwaterloo.ca);

H. Zhou is with the School of Electronic Science and Engineering, Nanjing University, Nanjing 210023, China (email: haibozhou@nju.edu.cn);

N. Zhang is with the Department of Electrical and Computer Engineering, University of Windsor, Windsor, ON N9B 3P4, Canada (email: ning.zhang@uwindsor.ca).

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 with low earth orbit (LEO) satellites, aerial networks with unmanned aerial vehicles (UAVs), and terrestrial networks with cellular base stations (BSs) will be integrated into a *space-air-ground integrated network (SAGIN)* to provide global coverage and on-demand services [1]. Second, network resource virtualization and end user virtualization techniques will form *advanced network virtualization* to provide flexible network management. Third, intelligence penetrates every corner of 6G networks, ranging from end users, the network edge to the cloud, which results in *ubiquitous intelligence*. The network nodes will be 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 (DNN) based applications. Hence, 6G networks are expected to create a new wireless networking ecosystem that will bring societal and economic benefits.

The 6G networks are expected to support a diverse set of services with different quality of service (QoS) requirements, such as multisensory extended reality, autonomous driving, and hologram video streaming. To support diversified services as established in 5G networks, network slicing is a potential approach to construct multiple slices for different services on top of the common physical network infrastructure [2]. The QoS requirements can be satisfied via cost-effective slice management strategies ranging from preparation, planning, and scheduling phases in the network slicing lifecycle.

Advanced network slicing faces many challenges in 6G networks due to their unique features. First, managing slices over satellite, airborne and terrestrial networks in the SAGIN requires judicious coordination of heterogeneous network segments. 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 emerging AI services with new QoS requirements, such as model accuracy and learning speed. 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 that manages network slices, but also in network nodes in the slices that provide services. Hence, the synergy of AI and network slicing in the proposed architecture is two-fold: On one hand, AI methods are applied to manage network slices, namely *AI for slicing*. Particularly, the network slicing lifecycle including preparation, planning, and scheduling phases is introduced, along with specifying AI methods for each phase. The detailed procedure of information exchange among users, access points, and the SDN controller is presented; On the other hand, network slicing is applied to construct customized network slices for various AI services, namely *slicing for AI*. Potential approaches such as slice instance construction and resource management for AI services are introduced.

The remainder of this article is organized as follows. In Section II, some 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 on AI for slice planning is presented in Section V. In Section VI, the research directions are identified, followed by the conclusion of this work in Section VII.

II. AI-Native Network Slicing for 6G Networks

A. Network Slicing

Network slicing is an emerging technology to manage complex and large-scale networks and support diversified applications in a cost-effective manner [2], [3]. The concept of network slicing can be traced back to the late 1980s [4]. Network slicing has become advanced technologies in 5G networks, supported by SDN and network function virtualization (NFV) techniques. Specifically, NFV enables virtualized resources and network functions on top of the physical network for flexible resource management, while SDN facilitates network resource orchestration in a centralized manner, thereby optimizing network performance. In 5G networks, network slicing has been defined in the 3rd generation partnership project (3GPP) Release 15 [5]. Moreover, in future 6G networks, network slicing is expected to continue evolving and play an increasingly important role.

The basic idea of network slicing is to create multiple virtual networks (i.e., slices) on top of a common physical network infrastructure, in order to achieve flexible and adaptive network management. The main benefits of network slicing are three-fold: (1) Multi-tenancy - multiple virtual network operators share the common physical network infrastructure, which can reduce

capital expenditures in the network deployment; (2) Service isolation - various logically isolated slices are constructed for different services via resource orchestration, such that the service level agreements of different slices can be guaranteed; (3) Flexibility - network slicing supports flexible network management, as slices can be created, modified, or deleted on-demand, resulting in cost effectiveness.

B. Features of 6G Networks

From 5G to 6G, it is in general expected that KPI requirements will be increased by at least an order of magnitude. According to a recent white paper [6], 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 requirements demand several candidate technologies, such as THz communications, intelligent reflecting surface, and AI [4], [7], [8]. By the end of 2026, the 3GPP working group will discuss various 6G candidate techniques. The first 6G standard is expected to emerge by 2030.

Distinguished from 5G networks, there are several features of 6G networks as follows:

- SAGIN - While the current cellular 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 would exploit the altitude dimension. Space networks (e.g., LEO, medium earth orbit, and geosynchronous earth orbit satellites), aerial networks (e.g., UAVs and balloons), and ground networks (e.g., cellular and WiFi networks) are integrated into the SAGIN [1], [9], to provide global coverage, facilitate on-demand services, and support high-rate low-delay services;
- Abundant services with stringent QoS requirements - Many new applications will have stringent QoS requirements in different dimensions. Mobile augmented reality, 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 [10]. Some other applications require extremely high reliability, such as autonomous driving, industrial control systems, and robot/UAV swarm, e.g., the required reliability of tele-operated driving systems is up to 99.999% [11];
- Ubiquitous intelligence - With caching capability, a large amount of data can be stored in the network. In addition, with the development of advanced AI techniques, edge computing

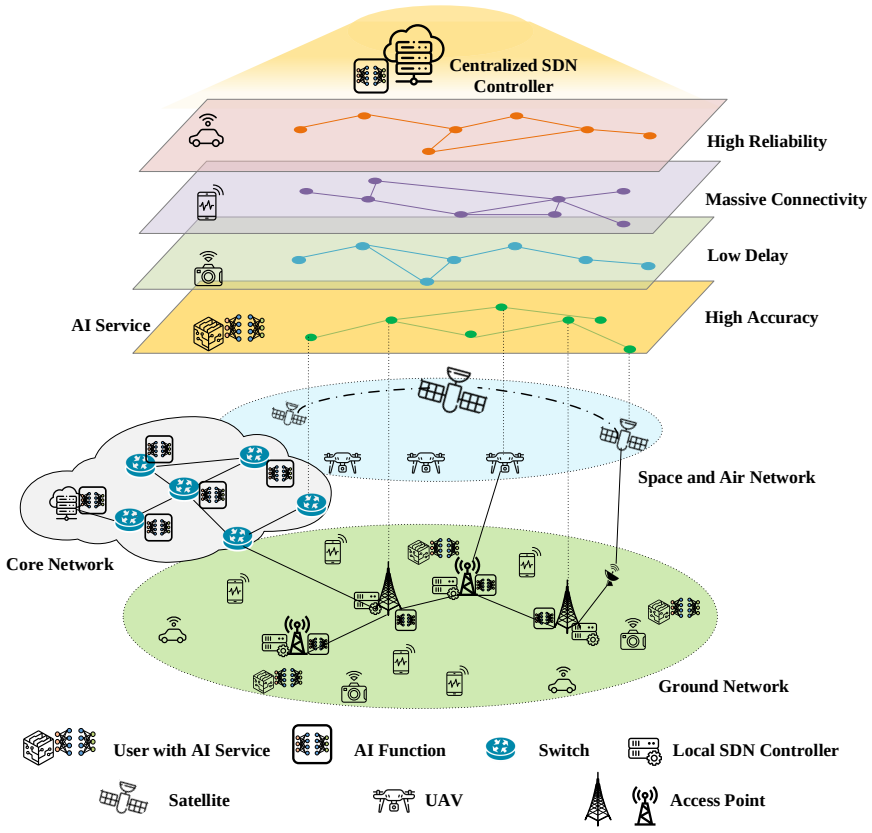


Fig. 1. An illustration of the AI-native network slicing architecture for 6G networks.

and on-device computing capability, intelligence is pushed from the cloud to the network edge and mobile devices. As such, AI should be integrated as a built-in component in 6G for intelligent network management by directly learning from extensive data. In addition, ubiquitous intelligence fosters centralized and distributed learning applications.

C. AI-Native Network Slicing Architecture

The potential features impose new challenges on applying network slicing strategies for 6G networks. Firstly, the SAGIN increases not only the number of network components but also the slice scope due to global network coverage, requiring efficient management of a large number of network slices. Secondly, the stringent QoS requirements of emerging services call for effective network slice management while adapting to dynamic network environments. Thirdly, ubiquitous intelligence will support many cutting-edge AI services which are expected to be dominant in future 6G networks, requiring constructing customized network slices for emerging AI services.

To address these challenges, an AI-native network slicing architecture for 6G networks is introduced. 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 the SDN controller at the cloud and local controllers at the access points to realize intelligent network slicing. As such, a large number of network slices with stringent QoS requirements can be managed efficiently and cost-effectively via AI methods, which is referred to as *AI for slicing*. Secondly, in addition to network slices for conventional services, new network slices are constructed to support emerging AI services on top of the physical SAGIN, which is referred to as *slicing for AI*. Note that the SDN controller in the proposed architecture consists of multiple controllers organized in tiers for implementation. Specifically, the centralized SDN controller located at the cloud is to manage all the network slices, while other local SDN controllers located at the access points are to schedule resources for the covered end users.

In the following, we will illustrate the basic ideas of AI for slicing in Section III and slicing for AI in Section IV, respectively.

III. AI FOR SLICING

In this section, we first introduce the three phases in network slicing lifecycle, and then investigate potential AI solutions in these phases. Finally, the corresponding procedure of information exchange is discussed.

A. Network Slicing Lifecycle

The network slicing lifecycle consists of three phases: *preparation*, *planning*, and *scheduling*. The operation in each phase is as follows.

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

- Extract service requirements - The operation 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 [5];

- Virtualize network resources and functions - 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 (VNF). Through virtualization, the SDN controller can flexibly manage the resources and network functions.

Through these operations, the SDN controller constructs a network slice for each admitted slice request.

2) Planning Phase: This phase is to reserve network resources for each slice before service provision. The planning process operates in a large timescale. Time is partitioned to planning period (window) for each slice. The duration of each planning window varies from slice to slice, depending on dynamics of each slice. It can range from 5 minutes to 2 hours, for example [12]. To achieve the goal, the following two steps should be conducted:

- Collection of service and network stochastic information - Benefiting from the global control functionality of the SDN controller, the extensive network information can be collected from underlying physical networks, such as service demand patterns, stochastic channel conditions, and user mobility patterns. The collected information can be utilized for the following planning decision making;
- Resource reservation - The SDN controller monitors the performance of existing network slices and adjusts the reserved network resources for each slice in the next planning window. The centralized SDN controller maps the allocated resources of each slice to the physical network infrastructure. Then, at the end of each planning window, some information from the system are feedback 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 the network slicing decisions to accommodate dynamic environments while guaranteeing QoS performance.

3) Scheduling Phase: This phase is to schedule the service of a slice using the reserved resources for subscribed end users. The scheduling process operates in a much smaller timescale

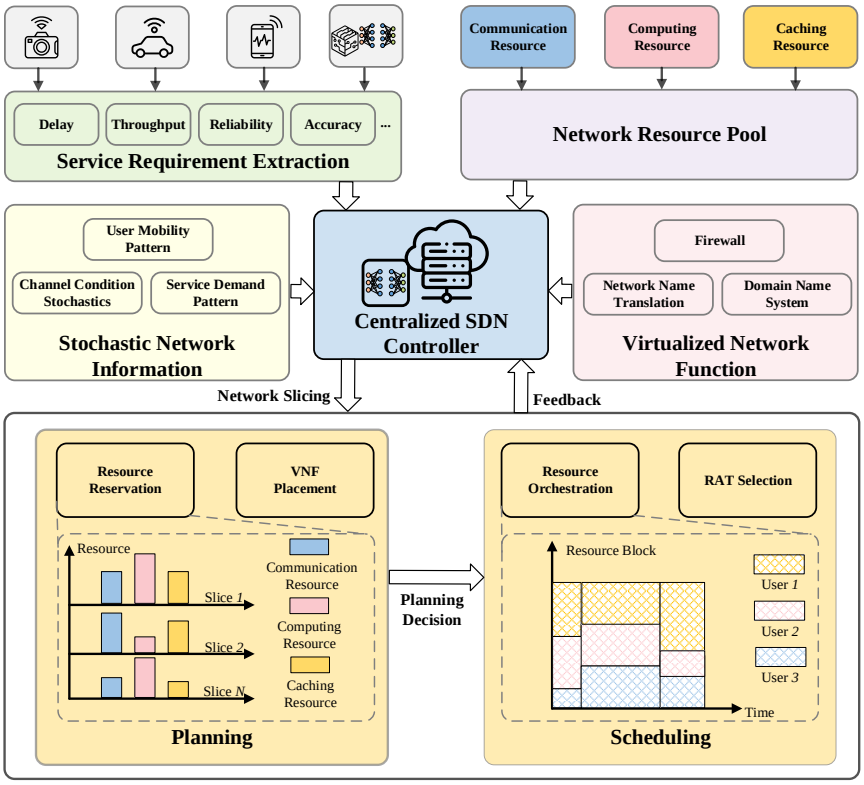


Fig. 2. An illustrative example for the network slicing lifecycle.

(e.g., 100 ms) than the planning window. Specifically, under the coordination from the centralized controller, local SDN controllers allocate resources for real-time services of each slice to end users according to users' real-time data traffic. The scheduling decisions include selecting radio access technology (RAT), determining user association with specific radio access points (e.g., macro BSs or small BSs), deciding proper protocol and associated parameters, and orchestrating resources for end users (e.g., BS transmission power, bandwidth, computing cycles, and cache space).

B. Roles of AI in Network Slicing

Although the network slicing facilitates service provisioning, managing multiple network slices in 6G networks can have significant network operation cost. AI-based network slicing is a potential solution to address the issue. As shown in Fig. 3, AI plays different roles in different phases of network slicing, as detailed in the following.

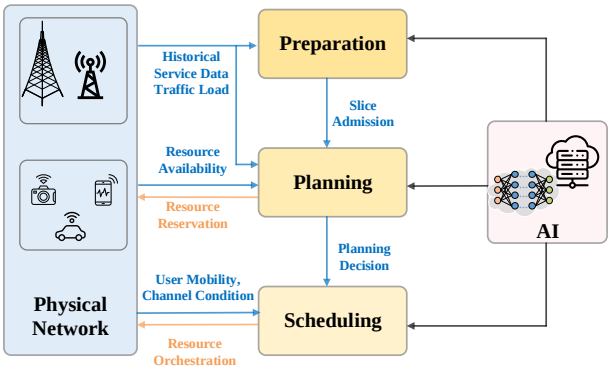


Fig. 3. An illustration of AI for slicing.

AI for Preparation: In the preparation phase, AI needs to perform two important tasks: 1) Service demand prediction - based on the historical data, service demand can be predicted via AI methods, such as recurrent neural networks (RNNs). Prior studies show that such service demand and the resource usage of a slice can often be predicted with a high accuracy [13]. The prediction results can be utilized for decision making in the preparation phase; 2) Slice admission - the SDN controller admits slices to maximize system resource utilization, based on the network resource availability and service demand of these slices. As the decision variable of slice admission is binary, this problem is deemed as an integer optimization problem. In a large-scale network with complex resource availability distribution, traditional optimization methods become complicated, such that deep learning is an alternative approach to address the problem.

AI for Planning: In the planning phase, AI needs to perform two tasks: 1) VNF placement - the SDN controller deploys VNFs to support services in the network. The resources allocated for VNFs should be dynamically adjusted for changing service demands to guarantee satisfactory service delay performance. A reinforcement learning method can be applied to interact with the dynamic environments, thereby enhancing resource utilization; 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 also be addressed via reinforcement learning methods, such as Q learning and other cutting-edge versions (e.g., deep deterministic policy gradient (DDPG), and deep Q learning algorithms).

AI for Scheduling: Two exemplary AI tasks in the scheduling phase are: 1) Resource or-

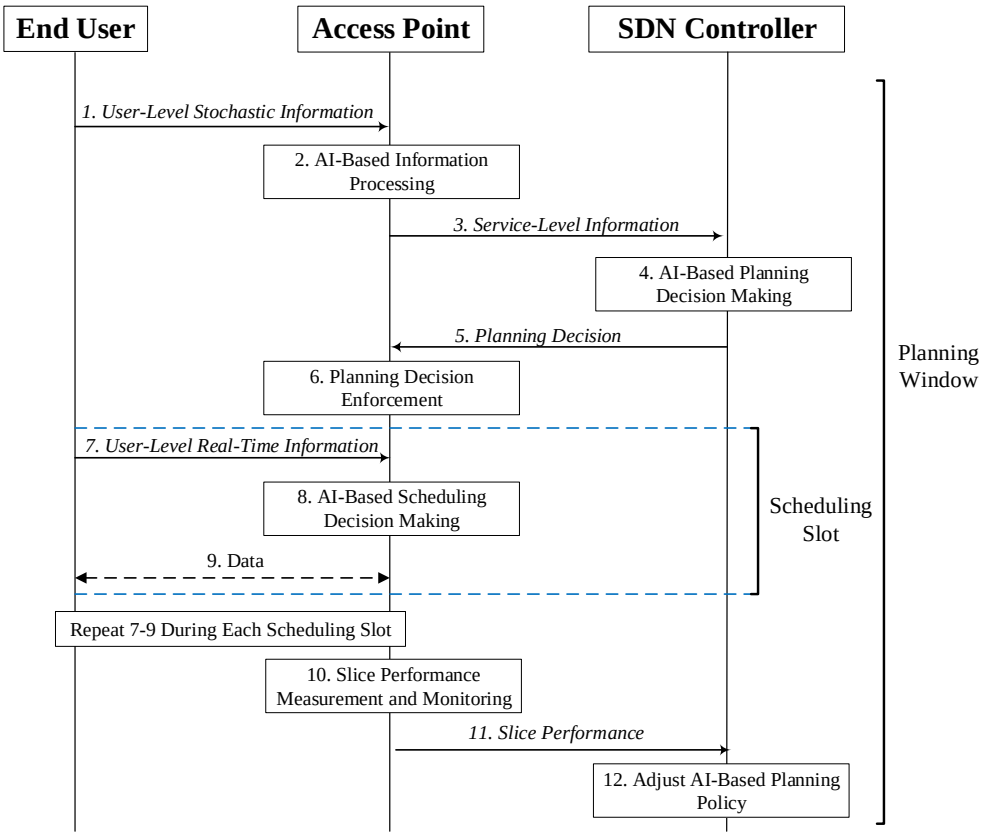


Fig. 4. Procedure of information exchange in the AI-native network slicing architecture.

chestration - the reserved resources of a slice are allocated to end users, given the planning decision results. The resource orchestration decision is determined based on dynamic real-time user mobility, channel conditions, data loads and so on. To efficiently utilize resources, reinforcement learning solutions can be developed for dynamic resource orchestration; 2) RAT selection - To maximize system utility, an optimal RAT is selected among multiple candidate RATs for an end user. Due to dynamic wireless networking environment and user mobility, the user-perceived service performance of an RAT is stochastic. As such, the RAT selection problem can be addressed by multi-armed bandit algorithms, e.g., contextual bandit.

C. Procedure of Information Exchange

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

1) The access point collects the user-level stochastic information, such as the user service demand patterns, mobility patterns, and stochastic channel conditions;
2) The access point translates the user-level information into desired service-level information. For example, user density information of a region can be obtained from processing user location information. In this step, AI techniques can be used for data abstraction, aggregation and analysis. For example, RNN can be applied for service demand prediction;
3) The processed service-level information is delivered to the AI-empowered SDN controller;
4) Based on the collected service-level information, the AI-empowered SDN controller runs planning algorithms to make decisions;
5) The planning decisions are sent back to associated access points;
6) Each access point enforces the determined planning decisions, i.e., reserving network resources for corresponding constructed slices;
7) 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) Based on real-time user-level information, the access point runs the AI-based scheduling algorithm to allocate resources to each user;
9) The service request from each user is supported via the allocated resources. For a computation task, it is uploaded by using communication resource and then processed by using computing servers. For each scheduling slot, Steps 7-9 are repeated;
10) The access point monitors the slice performance in the network given the enforced planning slicing decisions by measuring the users' satisfaction rates across all the scheduling slots;
11) The access point reports the network performance to the AI-empowered SDN controller;
12) Based on the feedback information, the AI-empowered SDN controller adjusts the planning policy.

In the preceding procedure, Steps 1-12 are in the planning phase and Steps 7-9 are in the scheduling phase.

IV. Slicing for AI

Slicing for AI is to utilize network slicing to support AI services while satisfying QoS requirements, which is implemented via constructing slice instances and resource management.

A. *Slice Instance Construction for AI Services*

Supporting AI services has diversified implementation approaches. The implementation of an AI service can be realized via different kinds of algorithms, training manners, and allocation of heterogeneous network resources (including computing, communication and caching). For example, objective detection services can be implemented via ResNet32, Inception-v3, AlexNet, or VGG16 algorithms. Hence, the primary issue of constructing slices for AI services is to determine an appropriate implementation approach.

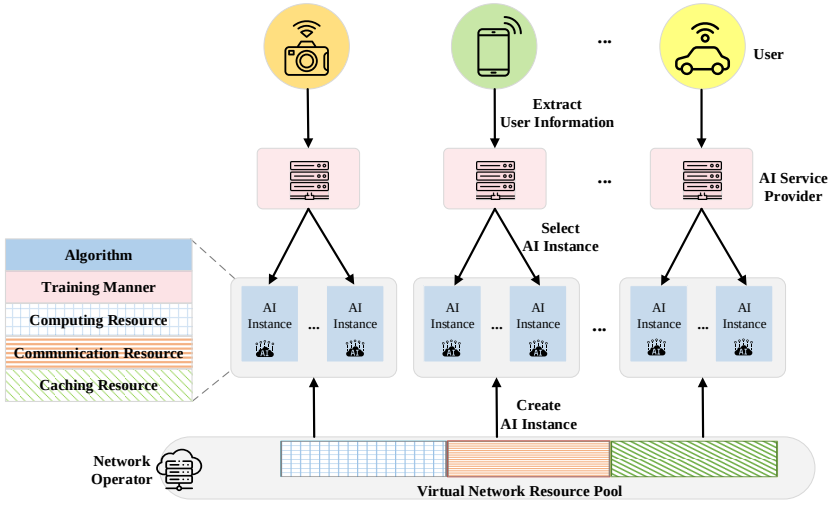


Fig. 5. The AI instance framework for flexible AI service management.

We introduce the concept of *AI instance* to address the issue, as shown in Fig. 5. An AI instance represents an implementation approach of the AI service. The basic idea is to select one AI instance from multiple candidate implementation alternatives. The operation of the AI instance framework consists of two steps: 1) AI instance construction - the network operator creates multiple candidate AI instances for each AI service based on available virtualized network resources and AI service requirements. Specifically, an AI instance is composed of multiple attributes, including the algorithm type which specifies the implementation algorithm and the corresponding neural network architecture, the training process which specifies the training manner of the specified implementation algorithm and the amount of the required network resources. Even if the amount of the overall required network resources is the same, different implementation algorithms may consume network resources at different physical locations; 2)

AI instance selection - this step is to associate the AI service provider with an appropriate AI instance. For this purpose, the AI service provider observes the status and collects requirements of its current subscribed users, and then selects an AI instance among candidate AI instances provided by the network operator. In summary, the proposed concept of AI instance has the advantage of flexible AI service management. The network operator can dynamically manage AI services via creating, modifying and deleting candidate AI instances on demand.

B. Resource Management in AI Service Lifecycle

The execution of AI services has multiple stages. The AI service lifecycle includes three stages: data collection, model training, and model inference [14]. Firstly, the data collection stage is to collect desired data in 6G networks. The collected data can be stored and then abstracted or refined for the following model training stage. Secondly, the model training stage is to train an AI model based on collected data. The model training can be cloud-based centralized learning or distributed learning. For example, multiple devices can work collaboratively to train a global model in a federated learning paradigm. Thirdly, the model inference stage is to complete specific inference tasks, given the pre-trained AI models. The inference can be performed in a collaborative manner, such as using a device-edge-cloud orchestrated model inference approach in which tasks are transferred and processed at different network nodes. All the three stages consume multi-dimensional network resources. Moreover, the performance of AI services depends on all the three stages in the AI service lifecycle. The model accuracy depends on the quality of the collected data, the number of training iterations, and the model inference approaches. As a result, to optimize the performance of AI services, the resources should be judiciously allocated for the three stages.

V. Case Study

In this section, a case study is provided on AI for planning, aiming at reducing system cost.

A. Networking and Service Scenario

We consider a highway scenario, in which five BSs are uniformly deployed along with a highway segment of 5 km. The vehicles upload computation-intensive tasks to the roadside BSs for processing. In the physical vehicular network, two network slices are constructed for two

types of services: 1) delay-sensitive autonomous driving service with a maximum tolerable delay requirement of 100 ms; 2) delay-tolerant high-definition map creation service without a specific delay requirement. In the planning phase, communication and computing resources are reserved for the two slices at each planning window of one-hour duration. We use real-world highway vehicle traffic flow trace collected by Alberta Transportation.[1]

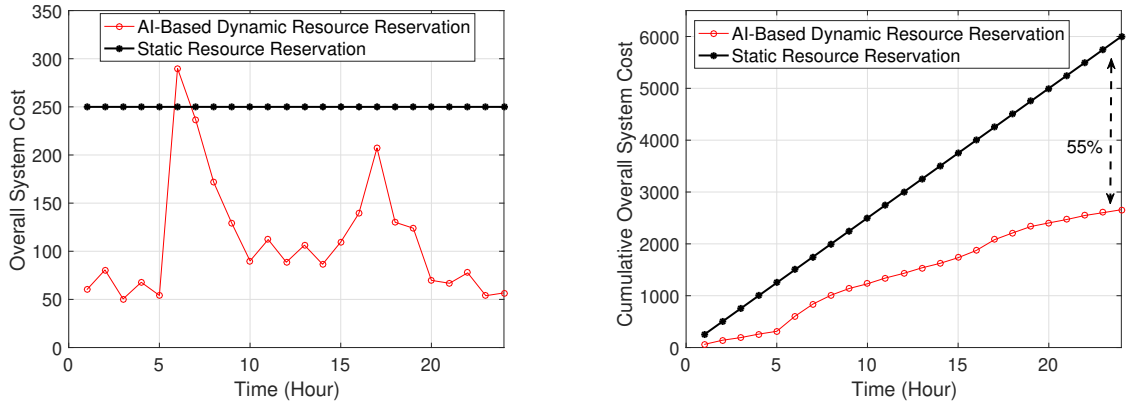
To handle vehicle traffic dynamics, we adopt an AI-based solution, i.e., a DDPG policy, for proactively making dynamic resource reservation decisions. Both actor and critic networks in the DDPG policy are fully-connected neural networks with four layers, and the number of the neurons in two hidden layers are 128 and 64, respectively. The learning rates of the actor and critic networks are set to 10^{-4} and 10^{-3}, respectively. We compare the proposed policy with a non-AI static resource reservation policy which reserves the same amount of resources all the time. Specifically, to satisfy the service delay requirement, the static policy makes resource reservation decisions according to the maximum vehicle traffic flow within a day.

The performance in the planning phase is measured by the overall system cost which is a weighted summation of four cost components: 1) operation cost that accounts for communication and computing resource reservation; 2) slice reconfiguration cost that accounts for adjusting resource reservation across two successive planning windows; 3) delay constraint violation penalty, which occurs once the service delay exceeds the maximum tolerable delay constraint; and 4) system revenue obtained from a low service delay [15]. The corresponding weights of four cost components are set to 1, 5, 200, and 25, respectively.

B. Simulation Results

As shown in Fig. 6(a), we first compare the overall system cost performance between the AI-based resource reservation policy and the static policy at different hours within a day. It can be seen that the AI-based policy outperforms the static policy with a lower system cost at most of the hours. The reason is that the proposed AI-based solution can dynamically adjust resource reservation with respect to vehicle traffic flow dynamics, such that the operation cost can be reduced. However, at peak vehicle traffic hour, the AI-based policy incurs a larger system cost than the static policy. This is because the AI-based policy needs to increase the amount of the

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



(a) System cost at different hours in one day

(b) Cumulative system cost within one day

Fig. 6. Performance comparison between the AI-based dynamic resource reservation policy and the non-AI static resource reservation policy.

reserved resources to adapt to vehicle traffic flow at the peak hour, which results in a significant slice reconfiguration cost. However, the static policy does not have any slice reconfiguration cost.

We show the cumulative overall system cost within one day in Fig. 6(b). It can be observed that the AI-based policy can significantly reduce the cumulative overall system cost within a day by around 55% for the given vehicle traffic dynamics. This is because the proposed AI-based policy can adapt to vehicle traffic flow dynamics via interacting with the unknown network environments, thereby obtaining the optimal planning policy to minimize the long-term system cost.

VI. Open Research Issues

Although AI-native network slicing is potentially a disruptive technology for 6G networks, the related research is still at an early stage. In the following, we discuss some open research issues pertaining to AI-native network slicing.

A. Joint Design of Planning and Scheduling

Planning and scheduling are operated and coupled in two different timescales. Specifically, the planning phase operates at a large timescale (e.g., minute level) to reserve resources for different slices based on service demands, while the scheduling phase operates 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 scheduling. Moreover, when both the planning and scheduling decision making are empowered by AI algorithms, the joint design faces potential divergence concerns of cascading AI algorithms for decision making.

B. Data Management Framework

The cornerstone of AI-native network slicing is abundant data that can be used for AI model training. In future 6G networks, data is widely distributed in the entire network. Due to limited communication resources, collecting a large amount of data incurs a significant cost. 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 for AI-native network slicing architecture.

C. Prediction-Empowered Network Slicing

With the development of advanced machine learning technologies, especially DNN, the data traffic in the network can be predicted. How to effectively leverage the power of prediction for network slicing is an interesting topic. In addition, 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

The authors would like to thank Jie Gao (Marquette University) 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, 2017.

[2] 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, 2020.

[3] 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, 2020.

[4] 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, 2021.

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

[6] N. Rajatheva *et al.*, "White paper on broadband connectivity in 6G," *arXiv:2004.14247*, 2020, [Online]. Available: https://arxiv.org/abs/2004.14247.

[7] M. Giordani, M. Polese, M. Mezzavilla, S. Rangan, and M. Zorzi, "Toward 6G networks: Use cases and technologies," *IEEE Commun. Mag.*, vol. 58, no. 3, pp. 55–61, 2020.

[8] K. Letaief, W. Chen, Y. Shi, J. Zhang, and Y.-J. A. Zhang, "The roadmap to 6G: AI empowered wireless networks," *IEEE Commun. Mag.*, vol. 57, no. 8, pp. 84–90, 2019.

[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, 2021.

[10] GSMA, "Cloud AR/VR whitepaper," 2019.

[11] C. Campolo, A. Molinaro, A. Iera, and F. Menichella, "5G network slicing for vehicle-to-everything services," *IEEE Wirel. Commun.*, vol. 24, no. 6, pp. 38–45, 2017.

[12] X. Foukas, M. K. Marina, and K. Kontovasilis, "Orion: RAN slicing for a flexible and cost-effective multi-service mobile network architecture," in *Proc. ACM MobiCom*, Snowbird, Utah, USA, 2017.

[13] 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.

[14] M. Li, J. Gao, C. Zhou, X. Shen, and W. Zhuang, "Slicing-based AI service provisioning on network edge," *IEEE Veh. Technol. Mag.*, 2021, submitted.

[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.*, 2020, DOI: 10.1109/JSAC.2020.3041405.