Intelligent and Converged Networks
2020, 1(3): 281–294

ISSN 2708-6240
DOI: 10.23919/ICN.2020.0019

An intelligent self-sustained RAN slicing framework for diverse service provisioning in 5G-beyond and 6G networks

Jie Mei, Xianbin Wang*, and Kan Zheng

Abstract: Network slicing is a key technology to support the concurrent provisioning of heterogeneous Quality of Service (QoS) in the 5th Generation (5G)-beyond and the 6th Generation (6G) networks. However, effective slicing of Radio Access Network (RAN) is very challenging due to the diverse QoS requirements and dynamic conditions in the 6G networks. In this paper, we propose a self-sustained RAN slicing framework, which integrates the self-management of network resources with multiple granularities, the self-optimization of slicing control performance, and self-learning together to achieve an adaptive control strategy under unforeseen network conditions. The proposed RAN slicing framework is hierarchically structured, which decomposes the RAN slicing control into three levels, i.e., network-level slicing, next generation NodeB (gNodeB)-level slicing, and packet scheduling level slicing. At the network level, network resources are assigned to each gNodeB at a large timescale with coarse resource granularity. At the gNodeB-level, each gNodeB adjusts the configuration of each slice in the cell at the large timescale. At the packet scheduling level, each gNodeB allocates radio resource allocation among users in each network slice at a small timescale. Furthermore, we utilize the transfer learning approach to enable the transition from a model-based control to an autonomic and self-learning RAN slicing control. With the proposed RAN slicing framework, the QoS performance of emerging services is expected to be dramatically enhanced.

Key words: Radio Access Network (RAN); network slicing; network resource management; intelligent network

1 Introduction

With the rise of the Internet of Everything (IoE), a plethora of vertical services and applications, ranging from high-precision manufacturing, intelligent transportation systems, and smart home to virtual reality, have been introduced to improve our lives, society, business, and industry operations. These vertical services with far-reaching impacts are characterized by an extremely diverse set of Quality of Service (QoS) requirements[1]. However, supporting emerging vertical applications well exceeds the capabilities of the existing 5th Generation (5G) networks. Specifically, the concurrent provisioning of heterogeneous vertical services calls for an extremely flexible, adaptive, and intelligent network architecture, which directly contradicts today's "one-size-fits-all" network design paradigm. Therefore, in the design of 5G-beyond and the 6th Generation (6G) networks, new network design and operation approaches have attracted research attention, including network slicing and Artificial Intelligence (AI) assisted communication and networking technologies.

The fundamental concept of network slicing is to divide one physical network into multiple virtual logical networks, referred to as network slices, coexisting over

• Jie Mei and Xianbin Wang are with Electrical and Computer Engineering, Western University, London, ON N6A 5B9, Canada. E-mail: jmei28@uwo.ca; xianbin.wang@uwo.ca.
• Kan Zheng is with the Intelligent Computing and Communication (IC2) Lab, Wireless Signal Processing and Network (WSPN) Lab, Key Laboratory of Universal Wireless Communication, Ministry of Education, Beijing University of Posts and Telecommunications (BUPT), Beijing 100876, China. E-mail: zkan@bupt.edu.cn.
∗ To whom correspondence should be addressed.
Manuscript received: 2020-07-30; revised: 2020-10-31; accepted: 2020-11-28

a common shared physical network substrate, with the purpose of achieving different QoS provisions in each virtual network. Network slicing can be considered as a convergence of Software Defined Networking (SDN) with Network Functions Virtualization (NFV). By creating the control plane, SDN can provide a global view of network infrastructure and the capability of programmable network control. On the other hand, through NFV, network functions and resources are not restricted to dedicated network infrastructures. By combining SDN and NFV, the dimensionality of each network slice during network slicing can be customized to best fulfill the specific QoS requirements[2]. Unfortunately, effective slicing of Radio Access Network (RAN) is still very challenging due to the network dynamics, performance isolation of network slices as well as diverse QoS requirements of different services. Furthermore, the radio resource management also poses technical challenges on the RAN slicing, considering the scarcity of radio resources and increased inter-cell/inter-tier interference caused by spatial multiplexing of the spectrum. As a result, effective and dynamic RAN slicing will bring an unprecedented level of network complexity, which also makes the traditional mode based approach intractable and ineffective[3,4].

Therefore, we need a paradigm shift in 5G-beyond and 6G networks from conventional network slicing schemes, in which the network merely performs network operation for specific scenarios and always requires manual intervention to resolve unforeseen situations, into a completely new intelligent network slicing framework that can autonomously self-sustain the high QoS performance of diverse services under the highly dynamic and complex network conditions, which was elaborated in our previous work[5]. Following this paradigm, in this paper, we propose a hierarchical multi-layer RAN slicing framework to enable the network operation with self-sustained capability. Based on our understanding, the self-sustained network slicing is based on the following three key pillars:

- **Self-management of network resources with multiple granularity:** Resource granularity directly affects the flexibility to modify the network resource management according to wide varieties of network conditions and QoS requirements. By adopting multiple time and resource granularities, self-management of network resources can be decomposed into multi-levels with the reduced management complexity, which includes network-level resource planning, next generation NodeB (gNodeB)-level slice configuration adaption, and packet scheduling level slicing. Self-management with differentiated control granularity can enable an extremely flexible and adaptive network architecture.
- **Self-optimization of Key Performance Indicators (KPIs) of the network:** The aim of the self-optimization is to enable the automatic optimization of KPIs (e.g., spectrum efficiency, energy efficiency, and QoS metrics), that is, the configuration of network, the parameters of RAN functions, and the network resource management scheme are autonomously and continuously adapted to the highly dynamic network environments. Self-optimization for network slicing is an important area for further improvement due to the fact that the current optimization of network slicing mainly focuses on static network scenario.
- **Self-learning to adapt control strategy rapidly under unforeseen circumstances:** In the area of network slicing, most AI-enabled works still need to learn from scratch when facing with a new scenario, which is inefficient due to the large amount of training data required, especially in scenarios with high mobility and fast varying network conditions. With self-learning capability, we can gain "expert knowledge" by exploiting datasets in previous scenarios. By leveraging the "expert knowledge", the network controller can learn new scenarios with limited training data and modify RAN slicing control strategy quickly.

The remainder of the article is organized as follows: Section 2 provides a comprehensive overview on the network slicing technology. Challenges faced by RAN slicing are discussed in Section 3. A self-sustained RAN slicing framework is then presented in Section

4, followed by a detailed explanation of the proposed framework in Section 5. Finally, conclusions are drawn in Section 6.

2 Network slicing: Prior arts

According to the description of 3GPP TR 28.801[6], a network slice instance includes a set of network functions and the supporting network resources, which are arranged and configured to form a complete logical network that meets specific network characteristics required by a service instance[7]. Each network slice is an independent part of the RAN, transport, and Core Network (CN). In 5G and 5G-beyond networks, network slices can be deployed with high flexibility and elasticity thanks to NFV. Through NFV, different mobile network components, spanning the CN and RAN, will be virtualized as Virtualized Network Function (VNF)[8,9], which can run on top of a virtualization system hosted in clouds[10,11]. Current research in this area can be separated into two parts: The first one deals with slicing CN, and the latter focuses on slicing RAN.

2.1 Slicing CN

In the case of 4th Generation (4G) of mobile phone mobile communication technology standards system, while network functions of CN are implemented in dedicated hardware as 3GPP Release 8 Specifications, Taleb et al.[11] proposed the concept of a mobile carrier cloud that places network functions as virtualized machines hosted in clouds. In the 5G era, 3GPP defines a service-based network architecture in which mobile back-haul/core services are provided by VNF, including main components of the 5G Core[11]. CN slices are to be realized through the deployment of a combination of VNF, e.g., network registration and mobility management (i.e., Access and Mobility management Function, AMF), user plane forwarding and QoS handling (i.e., User Plane Functions, UPF), 5G connectivity service handling (i.e., Session Management Function, SMF), and so on.

Due to the inevitable limitation of network resources, current researches are mainly focusing on the deployment of CN slices and management of VNFs. Due to the interdependence of VNFs in each network slice, there are placement constraints for placing the VNFs in the serving clouds. At the same time, the VNFs of each network slice have heterogeneous resource requirements, i.e., Central Processing Unit (CPU), caching, communication, and storage resource[11,12].

2.2 Slicing RAN

With the concept of Cloud-RAN, RAN components are dividing RAN into Base Band Unit (BBU) and Remote Radio Head (RRH), whereby BBU runs as software and RRH will be kept deployed in the field[12]. In 5G networks, by transforming the functionality of BBU into a set of VNFs, RAN is evolving towards more flexible deployments by splitting the BBU into two entities, namely Central Unit (CU) and Distributed Unit (DU). In this context, CU hosts time-tolerant RAN functions, while DU hosts time-sensitive RAN functions, such as MAC and physical layer functions[13].

While network slicing is well studied in the CN, there are remaining challenges in the RAN slicing that need to be addressed. The stochastic nature of wireless networks, multi-dimensional QoS requirements of services, highly dynamic service traffic, and inevitable limitation of resources available in networks contribute to the challenge.

Considering the impact of RAN slicing on the radio interface protocol architecture in 5G networks, an RAN slicing orchestration framework is proposed to create pools of resources that are shared and allocated among RAN slices[14,15]. From the viewpoint of radio resource allocation, considering both the cellular network topology and the multi-cell interference, several automated admission control and network wide resource allocation schemes for RAN slicing were proposed for providing throughput guarantees in each slice[16–18]. Furthermore, in the heterogeneous RAN, a dynamic RAN resource management framework was proposed, which jointly considered the priority of RAN slices, base-band resources, front-haul and backhaul capacity, QoS, and interference[19].

Since the 5G networks are characterized by a high level of complexity, which makes traditional mathematical approaches untenable, recently, there have

been a few works towards the intelligent RAN slicing. A service demand-aware resource allocation scheme was proposed based on Deep Reinforcement Learning (DRL) to realize a dynamic and efficient spectrum allocation per network slice[20]. A two-timescale radio resource scheduling strategy was proposed in Ref. [21], where the Deep Learning (DL) algorithm was utilized to predict traffic flow on a long-term timescale, and the reinforcement learning method was used to perform radio resource management.

However, based on our knowledge and literature review, the above works are still not entirely suitable for future wireless networks, due to the following limitation: Most proposed network slicing schemes are limited by conventional model based approaches, which only perform well for ideal and specific scenarios. However, it makes the existing network slicing schemes lack robustness in the real deployment scenarios.

On the other hand, conventional model based approaches are derived from communication theory. However, the mathematical model is regulated by an inherent trade-off between accuracy and tractability. In network slicing, the network dynamics and time variation pattern are extremely difficult to be modeled in both an accurate and a tractable way. Therefore, it is necessary to enable the automated network slicing process without the requirements of precise modeling of the network dynamics.

3 Challenge faced by RAN slicing

The objective of RAN slicing is to satisfy specific QoS requirements of different services in a flexible and efficient manner. However, challenges arise in the design of an RAN slicing framework for the multi-cell scenario, which can be elaborated as follows:

• **Heterogeneous QoS requirements of diverse services:** According to 3GPP standards, the emerging services can be classified into three types: ultra-Reliable and Low-Latency Communication (uRLLC), enhanced Mobile BroadBand (eMBB), and massive Machine-Type Communication (mMTC). Provisioning of these services requires diverse data rate, reliability, and latency from networks[1]. Specifically, mMTC focuses on providing ubiquitous connectivity to a large number of low-complexity and low-power devices, while uRLLC is about providing ultra-low latency and high-reliability wireless connections. For instance, autonomous driving, one typical uRLLC service, requires a latency below a few milliseconds and reliability of close to 100%. In comparison, eMBB service mainly focuses on a high data rate and thus can tolerate longer latency and lower reliability. Therefore, it is hard to convert the diverse QoS requirements into the network resource demands, that is, how to quantify the multi-dimensional network resource demands of each service at every Transmission Time Interval (TTI). Because of these characteristics, simply increase of the data throughput for each service cannot efficiently fulfill QoS requirements. As a result, the RAN slicing framework need to intelligently and accurately estimate the network resource demands of slice according to the QoS requirements of service.

• **Spatial-temporal dynamics of network:** Due to the mobility of users and inherent service patterns, the status of the network is fluctuating from both spatial and temporal dimensions. Firstly, the distribution of user location and its mobility pattern can be varied in different regions according to lifestyles and work habits of people[22], which introduces spatially inhomogeneity of service traffic among different cells. On the other hand, the temporal dynamics of cellular networks contain both long-term fluctuations (i.e., the dynamics of service traffic) and short-term fluctuations (i.e., the dynamics caused by wireless channel). Specifically, it is shown that for a specific service, its service traffic patterns have strong daily patterns, which will result in long-term fluctuation of networks (in the level of hours)[6, 23]. Thus, the service traffic pattern over a region will significantly vary from different time of a day. To follow the spatial-temporal dynamics of the network, we need to design the RAN slicing control with multiple time, spatial, and network resource granularities to improve the effectiveness of network slicing.

• **Interference in wireless network environments:** When in a single-cell scenario, performance isolation of network slices can be strictly guaranteed by assigning orthogonal radio resources to each slice without considering in-band interferences. However, when in a multi-cell scenario, because of the scarcity of

spectrum, frequency reuse takes place among different cells to exploit the multiplexing gain. Then radio resources in each cell will suffer from the dynamic and unpredictable interferences from neighboring cells, making radio resources become inhomogeneous in communication capacity. Thus, it is more challenging to achieve performance isolation among slices in multi-cell scenarios. To this end, efficient utilization and multiplexing of radio resources are critical to ensure RAN slicing performance.

• **Signaling overhead cost by RAN slicing control:** Due to the rapid change and high complexity of the network, if the network controller directly optimizes the QoS performance of each Vehicle-to-Everything (V2X) service request, the real-time service is hard to guarantee due to the costed latency related to processing complexity. At the same time, the network controller will also need a large amount of mobile data to gain satisfying performance, which will incur the overwhelming signaling overhead. Therefore, it is necessary to balance the trade-off between the cost of signaling overheard and the global optimality of RAN slicing control. To address this concern, we need to perform RAN slicing control with multiple time granularities to reduce the amount of signaling overhead, which is an essential issue in designing the RAN slicing framework.

• **The high complexity of RAN slicing control:** In the network slicing, the precise modeling of a network environment is key for effective timely decision-making control and improving KPIs of the network. However, the traditional mathematical model is regulated by an inherent trade-off between accuracy and tractability. Specifically, the network dynamics are difficult to be mathematically modeled in both an accurate and tractable way[24]. Thus, it is highly desirable to utilize AI technologies to improve the effectiveness of network slicing without the requirements of a precise mathematical model. However, it is ineffective to utilize conventional AI algorithms in a "plug and play" way. Therefore, we should customize AI algorithms according to the characteristics of network slicing.

4 Design of RAN slicing framework with self-sustained capability

To tackle these challenges, we discuss how to design an intelligent network slicing architecture with high flexibility and self-sustained capability, which aims at optimizing and maintaining the QoS performance of services. As shown in Fig. 1, we consider a multi-cell wireless network. In this scenario, serval network slices are established, which share the same communication and computing resources in the network. Assume that the multi-cell RAN is split into three typical network slices, that is, eMBB slice, uRLLC slice, and mMTC slice.

4.1 Preliminary: Network architecture

In order to perform centralized control of the multi-cell RAN slicing, as shown in Fig. 1, a cloud-based network architecture with SDN and NFV technology is presented by integrating various network infrastructures in the multi-cell RAN[5,25]. The proposed architecture has two advantages. By using the NFV, different kinds

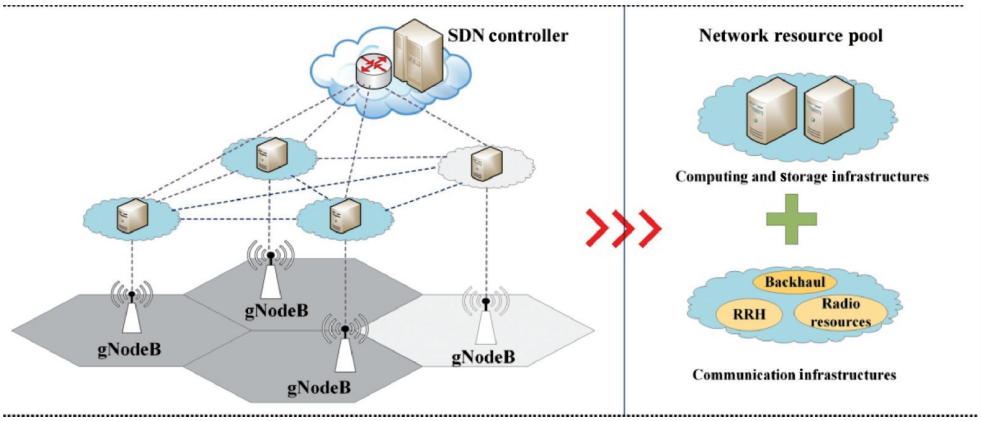

Fig. 1 An illustration of network architecture in multi-cell RAN scenario.

of resources are virtualized from dedicated network infrastructures. Then a network slice can be defined as a proper collection of resources, thereby further increasing flexibility. On the other hand, because of the centralized architecture and virtualization of network resources, AI-enabled technologies can be conveniently implemented. Three network components exist in the considered cloud-based framework as follows:

(1) Network resource pool: The network resources are mapped to the multi-dimension resource pool. According to its capability and ability, in the network resource pool, resources are further divided into three dimensions: communication, computing, and caching. Because computation infrastructures are heterogeneous, computation resources are virtualized uniformly without concern about configuration details. Moreover, radio resources are abstracted by a multidimensional grid of space, time, frequency, and transmit power. Furthermore, the resource pool can provide multiple resource granularities to support the operation of RAN slices.

(2) gNodeB: A gNodeB, defined by 3GPP, refers to the next-generation base station in the future networks, which accommodates the service requests inside its coverage area (i.e., one cell). As mentioned before, due to the limited radio resources, it is challenging to sufficiently provide orthogonal radio resources to each gNodeB in the multi-cell scenario. As a result, some gNodeBs will share the same radio resources, but this comes at cost of inter-cell interference. To address this issue, we can assign the same set of radio resources to multiple cells separated by an enough distance, which can reduce interference to an acceptable level.

(3) SDN controller: The SDN controller, hosting service entities in close physical proximity for users, can ensure the low end-to-end response time of service provisioning. Furthermore, the set of gNodeBs in the multi-cell scenario is directly connected and managed by the SDN controller. All gNodeBs share the same network resource pool and the SDN controller has centralized control through allocating and managing multi-dimension resource to maintain the operation of network slices.

4.2 Proposed RAN slicing framework

As shown in Fig. 2 and Table 1, an intelligent RAN slicing framework with multi-level and multi-control granularities is proposed, where PRB represents physical resource block. This framework consists of three levels:

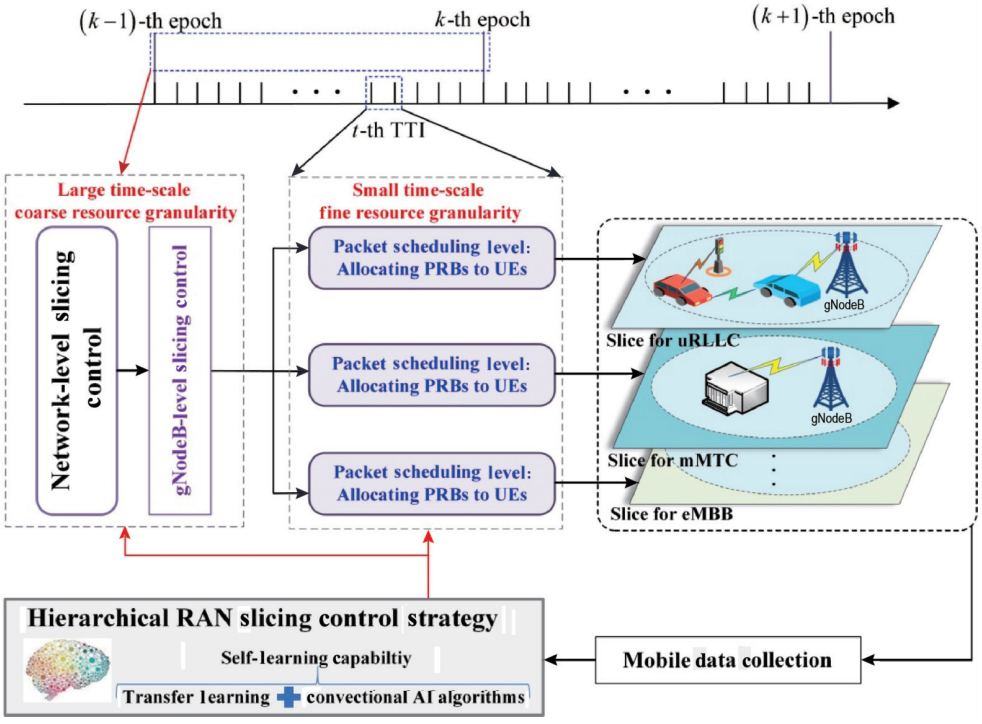

Fig. 2 A conceptual diagram of the proposed RAN slicing framework.

Table 1 Control granularity of the proposed RAN slicing framework.

Control level	Function	Resource granularity	Time granularity	Spatial granularity	Complexity
Network-level	Assign network resources to each gNodeB	Computation resources for RAN function operation; communication resources	Large timescale	Whole area	High
gNodeB-level	Adjust slice configuration by setting Guaranteed Bit Rate (GBR) and Maximum Bit Rate (MBR)	–	Large timescale	Single cell	Medium
Packet scheduling level	Allocate radio resources to active users	PRB allocation for packet transmissions	Small timescale	Single active user	Low

network-level, gNodeB-level, and packet scheduling level. At the network level, network resources (i.e., computation, caching, and communication resources) are assigned to each gNodeB at a large timescale. At the gNodeB-level control, each gNodeB adjusts the configuration of each slice in the cell at the large timescale. At the packet scheduling level, each gNodeB allocates radio resource allocation among users (UEs) in each network slice at a small timescale.

As we can see, the proposed RAN slicing framework with differentiated time-resource control granularities can flexibly deal with the time-varying traffic loads among gNodeBs. It is worthy to note that, under the proposed RAN slicing framework, RAN slicing control is determined by the control strategy, denoted as

$$\pi = \{\pi_N, \pi_g, \pi_P\},$$

where π_N denotes the network-level control policy, π_g stands for the gNodeB-level control policy, and π_P represents the packet scheduling level control policy. Furthermore, the control strategy π is represented by an Artificial Neural Network (ANN), which is generated by AI algorithms.

4.2.1 RAN slicing at network level

The network-level control policy π_N can set computational and communication capacity to support the operation of gNodeBs according to the dynamic of service traffics. The space domain granularity of network-level control is the whole geographical scope of the multi-cell RAN scenario. By analyzing network-wide observations of networks (e.g., the spatial distribution of the service traffic volume, user distribution, and so on), the SDN controller can achieve a global view of the whole network. With this knowledge, the SDN controller can predict large-scale user distribution as well as the hotspot area in the network. Then the SDN controller assigns the network resources to each gNodeB to ensure its operation and avoid the traffic overload.

Furthermore, the network-wide observation of network normally has obvious change on a long-term timescale (in the level of seconds)[25]. SDN controller should adapt the configuration of network slices on a long-term timescale, which can also lower signaling overhead and computation burden[26]. The network-level control decides the network resource assignment among gNodeBs, which includes the following:

• **Communication capacity at each gNodeB.** Firstly, the SDN controller assigns the radio resources (i.e., PRB) to gNodeBs by considering the expected spatial distribution of the traffic. Here, PRBs will be partially reused among gNodeBs to exploit multiplexing gain. Specifically, a set of orthogonal PRBs are assigned to gNodeBs with close distance, and PRB reuse will take place among gNodeBs with far distance, which can mitigate inter-cell interference[26]. Secondly, the bandwidth of front/ backhaul links is also determined by the network-level control policy π_N[27].

• **Computation capacity at each gNodeB.** In 5G-beyond and 6G wireless networks, the BBU of each gNodeB is virtualized and centralized as a BBU pool, where each virtual BBU associates with one gNodeB. In the proposed framework, the computation capacity of the virtual BBU is decided by the SDN controller according to service traffic volume in the coverage of gNodeB. Herein, the computation capacity of the virtual BBU is quantified in the form of the maximum number of CPU cycles per second. Specifically, the computation capacity

of virtual BBU determines the maximum baseband processing abilities of gNodeB, such as modulation, coding, and radio resource management. Furthermore, by abstracting these baseband processing abilities, we can map the computation capacity of a virtual BBU to its supported maximum total throughput of gNodeB[28].

4.2.2 RAN slicing at gNodeB level

Given the assigned network resources determined by the network-level control policy π_N at the large timescale according to the gNodeB-level control policy π_g, each gNodeB will tune the configuration of slices in its serving cell based on the dynamics of service traffic in the cell, in order to improve the QoS performance of services. It is noteworthy that gNodeB-level control is not directly involved in the real-time radio resource scheduling. Essentially, the gNodeB-level control has two main functions:

• **Ensuring QoS requirements of slices:** It transforms the dynamics of service traffic as well as the QoS requirements of services into the data rate constraint of users in each slice. Here, the GBR of users in each slice is used to ensure the QoS performance.

• **Ensuring performance isolation among slices:** It guarantees that the traffic overload in one slice does not negatively affect the QoS which is experienced by UEs in other slices. Here, each slice is imposed with the MBR of UEs to ensure performance isolation.

The gNodeB-level control is very important in the proposed RAN slicing framework. When current configurations of the network slice cannot fulfill the desired QoS requirements of service, the gNodeB-level control policy will tune the slice configuration to improve the QoS performance of services.

It should be noted that gNodeB-level control is a "soft" slicing control. Existing RAN slicing control schemes are normally "hard" slicing control. Hard slicing control assigns a fixed number of network resources to the agent, which can ensure QoS isolation but at the cost of flexibility and multiplexing gain. In contrast, soft slicing control at the gNodeB-level enables dynamic sharing of network resources among gNodeBs, which realizes the flexible radio resource management with high efficiency.

4.2.3 RAN slicing at packet scheduling level

Based on the configuration determined by the gNodeB-level, according to the packet scheduling level control policy π_P, each gNodeB executes the PRB allocation to users based on the instantaneous service request. The scheduling process operates at a time resolution given by the so-called TTI, which is the smallest time granularity in the system and currently is 1 ms in Long Term Evolution (LTE). Apart from allocating PRBs to users, packet scheduling level control policy π_P also responses for selecting the suitable physical layer parameters used in the transmissions (e.g., scheduling priority, modulation, and coding scheme[14]) according to the QoS requirements of services[29]. Then based on the QoS requirements of three considered services (i.e., mMTC, uRLLC, and eMBB), we propose the following sets of network slices:

• **Slice for the uRLLC service:** This slice should guarantee ultra-high reliable and ultra-low latency requirements of the uRLLC service. The SDN controller should utilize a semi-persistent scheduling function to avoid latency induced by signaling exchange. Furthermore, to support low latency transmission, one option is to reduce the number of symbols in each TTI, that is, mini-slot (e.g., 0.125 ms) is proposed as short-TTI for meeting the latency requirement of uRLLC.

• **Slice for the eMBB service:** eMBB users usually generate service requests to access the Internet with high average data rate requirements. To improve the overall Quality-of-Experience (QoE) and reduce backhaul bandwidth use in eMBB transmissions, the SDN controller should cache popular contents in the storage infrastructure in each gNodeB.

• **Slice for the mMTC service:** This slice supports the exchange of large amounts of messages between Internet-of-Thing (IoT) devices. The SDN controller should choose suitable access mechanisms and resource scheduling methods, which can avoid channel congestion and ensure high communication reliability.

4.3 Learning to control: RAN slicing powered by self-learning

Among the main promises of AI algorithms for RAN slicing control is its capability to enable the satisfying

QoS performance of network slices starting from a blank slate[29]. However, when deployment of new network slices or the network condition is changing, an important limitation of the paradigm of existing AI-based works should be generally carried out from scratch for each new case to obtain a new RAN slicing control strategy π. The motivation of self-learning enabled RAN slicing control lies in the consideration that the "learning-from-scratch" method is unaffordable for a multi-cell RAN scenario, due to a large amount of required training data, latency requirement as well as the related intelligent processing complexity[30].

To this end, the model should have self-learning ability, which can enable fast knowledge transferring from pre-trained models for different scenarios. To achieve this goal, we can utilize transfer learning to realize the self-learning for the RAN slicing control. In the broadest sense, transfer learning studies how to transfer the prior knowledge that is used in a given context into a different but related context, to execute a new task.

4.3.1 Paradigm of transfer learning

Technically, Transfer Learning (TL) is a research problem in machine learning that focuses on storing prior gained knowledge while solving one problem and applying it to a different but related problem. Formally, the framework of TL is defined as follows:

A domain, D, is defined as

$$D = \{X, \Pr(x \in X)\},$$

which is a two-element tuple consisting of feature space X and marginal probability $\Pr(x \in X)$, where x is a sample data point.

On the other hand, since transfer learning can be applied to a variety of learning problems, including classification, prediction, and reinforcement learning, we introduce a generic notion of a learning task below. The learning task Γ is defined as

$$\Gamma : x \in D \rightarrow y \in Y,$$

which maps sample data point x to output y. It is assumed that the learning task Γ is realized by an ANN.

Given a source domain D_{S}, a source task Γ_{S}, a target domain D_{T}, and a target task Γ_{T}, TL aims to help improve the learning of the target task Γ_{T} in the target domain D_{T} using the knowledge in D_{S} and Γ_{S}.

TL aims at solving the dilemma that an ANN cannot be trained well with the limited dataset as well as the stringent latency requirement of intelligent processing, by using the conventional learning model trained by large-scale labeled datasets, namely the source domain datasets, which are similar to the given datasets for the task, namely the target domain datasets.

In basic, the methods of TL can be divided into four categories: sample-based TL, mapping-based TL, network-based TL, and adversarial-based TL[31]. In this paper, we mainly focus on the network-based TL methods, which implement the transfer of knowledge by first training an ANN to execute the source task Γ_{S} in the source domain D_{S} and then refining the parameters of the obtained ANN to execute the target task Γ_{T} in the target domain D_{T}.

The common approach of network-based TL is to perform two-stage training. At first, the ANN is pretrained to execute the source task, yielding a tentative parameter set of the ANN. Next, in the second stage, the ANN is re-trained in the target domain. This approach is suitable for the situations in which a lot of datasets are available in the source domain, whereas only a few datasets are available in the target domain.

In the wireless communication networks, a large amount of training sets are hard to acquire, at the cost of consuming a large amount of signaling overheads for exchanging the sensing information and inevitably increasing the intelligent processing latency. Meanwhile, these signaling overheads will also occupy too many radio resources, leading to a low utilization efficiency of the radio resource. Without the TL technique, the learning procedure will be generally carried out from scratch for each new task, which means a large amount of signaling overheads are needed for each new task. Luckily, we can utilize the TL technique to significantly alleviate this issue. Although the source task needs a large amount of training sets, the target task only needs a small amount of training data, which greatly reduce the amount of signalling overheads for training.

4.3.2 Embedding transfer learning in RAN slicing

Wireless networks usually exhibit changing patterns over time. The changing wireless network environments will

require self-learning ability to continuously mine new features from wireless network environments without forgetting old but essential patterns[32].

Based on the paradigm of TL, as shown in Fig. 3, the self-learning for the RAN slicing control can be seen as a transfer learning problem, which aims at extracting expert knowledge from the previous scenario and using it to adapt the RAN slicing control strategy π.

Initialization: In the initial operation of the RAN slicing framework, the controller has no historical dataset for the training of the ANN model. To address this issue, we perform the RAN slicing operation by using the conventional mathematical based network resource management schemes with the sub-optimal but stable performance. Meanwhile, the controller collects data generated in operation. When the size of the dataset is large enough, we can train the RAN slicing control strategy π in an online way. Then we can get the parameters of the model for transfer learning in the future.

• **Trigger condition:** As mentioned in Section 3, the service traffic in the cellular networks has strong daily patterns. This attribute will result in significant changes in service traffic volume every few hours. When the service traffic volume is greatly changed (referred to as the new scenario), it is necessary for us to re-train the RAN slicing control strategy π.

• **Knowledge transfer:** Firstly, we have an RAN slicing control strategy π^{old} for the previous scenario. Specifically, the expert knowledge obtained in the previous scenario is converted to the parameters of π^{old}. Then we can use the parameters of π^{old} to initialize the parameters of the RAN slicing control strategy π^{new} for the new scenario.

• **Fast retraining:** After parameter initialization, we can train the RAN slicing control strategy π^{new}. Because π^{new} has been properly initialized, π^{new} can be fine-tuned quickly with a small training set.

As we can see, the self-learning enabled by transfer learning is very general and can be applied to varieties of AI algorithms. Therefore, self-learning can be successfully used to facilitate the implementation of an intelligent RAN slicing framework, especially by reducing the amount of required training data and validation purposes. It has essential practical value, because in the context of multi-cell RAN, the acquisition of a large amount of mobile data may not be practical.

5 Case study: Applying proposed RAN slicing framework for V2X service provision

5.1 Vehicular network scenario

To demonstrate the application of the proposed RAN

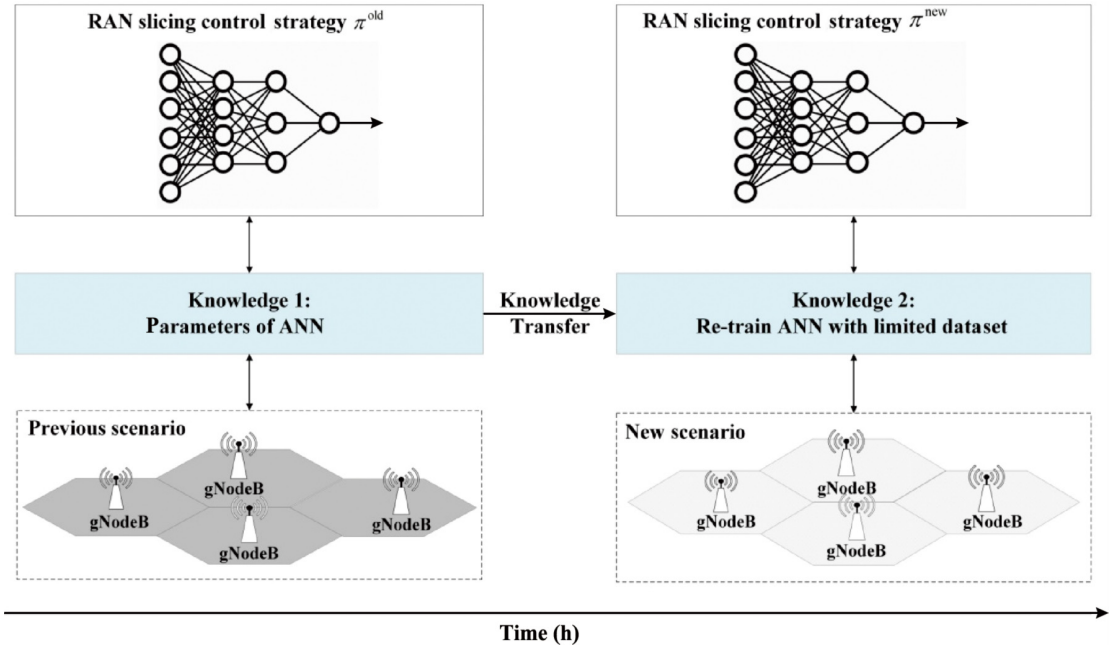

Fig. 3 Principle of embedding transfer learning in the RAN slicing control.

slicing, we consider a vehicular network in an urban area based on the Cellular Vehicle-to-Everything (C-V2X) standard. There are two kinds of nodes in this network: (1) vehicles requiring content services, and (2) a number of gNodeBs delivering content information. Two types of network slices will be established by the network operator to support time-critical driving safety-related information service (i.e., driving safety-related slice) and the bandwidth-consuming infotainment services (i.e., infotainment-related slice).

5.2 Operation workflow of proposed RAN slicing framework

The network-level and gNodeB-level control strategies are first implemented on the SDN controller. The packet scheduling level control strategy is then deployed on each gNodeB, respectively. Furthermore, the network-level and gNodeB-level control strategies are running in coarse time granularity (in the level of seconds). The packet scheduling level control strategy is running in fine time granularity (in the level of a millisecond). Specifically, the proposed RAN slicing control strategy is trained in the beginning with conventional DRL in an online/on-policy way. Figure 4 shows the interaction between the three-level RAN slicing controls in the proposed framework.

Firstly, based on the spatial distribution of vehicle and the corresponding service traffic, the SDN controller performs the network-level control by assigning the network resources (i.e., communication and computation resources) to each gNodeB. For instance, for the gNodeB covering the area with high vehicle density, the SDN controller will assign more network resources, and vice versa.

After the network-level control is completed, each gNodeB monitors the vehicle density and service traffic inside its coverage area. For the fluctuations of service traffic statistics (e.g., the average number of service traffic per slot), gNodeB reports the status of service traffic to the SDN controller and then performs an adaption of slice configuration (i.e., adjusting the GBR and MBR of user in the slice). For instance, when the traffic burst of the infotainment-related slice happens, the gNodeB-level controller will reduce the MBR of the infotainment-related slice to avoid traffic congestion, which may induce a negative impact on the QoS performance of the safety-related slice.

Then for the packet scheduling level control, each gNodeB makes the scheduling decision per time slot and allocates PRBs to active vehicles. Furthermore, the achievable data rate of the active vehicle in each network slice is restricted by the GBR and MBR, which can both ensure the QoS performance of each slice and avoid traffic overload in vehicular networks.

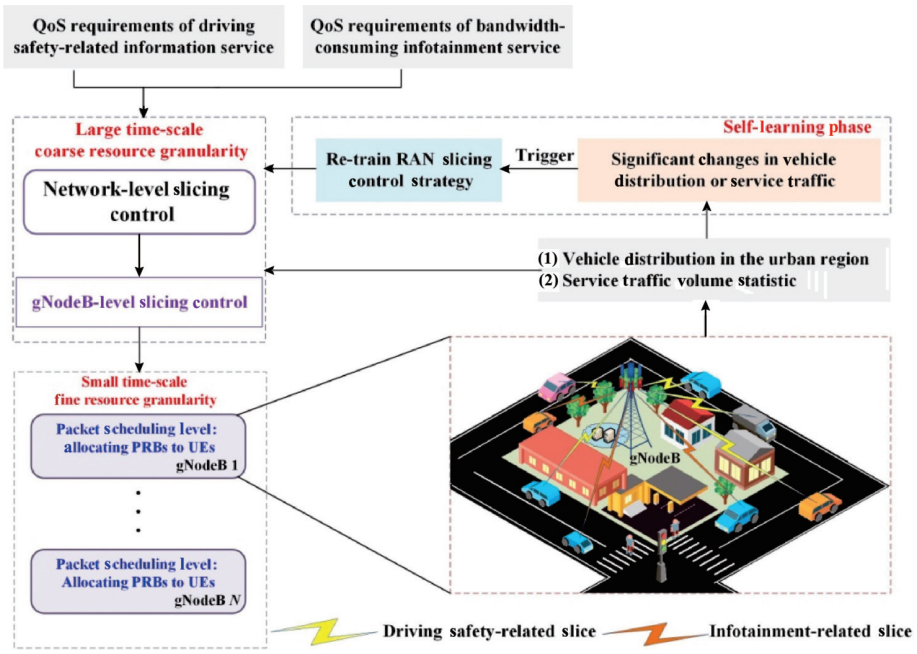

Fig. 4 An illustration of the proposed RAN slicing framework for vehicular service provision.

As mentioned, the service traffic of vehicular networks has daily patterns, which results in significant changes in service traffic volume every few hours. Therefore, when the vehicle traffic distribution in the urban scenario is significantly changed, the control strategy, denoted as π^{old}, cannot perform well for this new network condition. In this case, the self-learning phase is triggered, and the proposed RAN slicing framework needs to learn the new RAN slicing control strategy, denoted as π^{new}. For this purpose, the parameter set of the control strategy π^{old} is first used as the initial parameter set of the control strategy π^{new}. By this way, the prior knowledge of slicing control, interpreted as the parameter set of π^{old}, is transferred to the control strategy π^{new}. Then with prior knowledge, the control strategy π^{new} can be retrained by conventional DRL approaches with a small dataset.

As we can see, the proposed RAN slicing framework has the following main advantages:

• With multiple time and resource granularities, the RAN slicing framework can flexibly adapt to the changing network condition of vehicular networks.

• With the self-learning capability, the proposed RAN slicing framework can quickly learn and adapt its control strategy when the service traffic statistic in the urban scenario has significant variations.

6 Conclusion

Network slicing is a promising direction in meeting the diverse QoS requirements of emerging applications. In this paper, we have presented the design of a self-sustained RAN slicing framework for concurrent service provisioning. We have proposed a new intelligent RAN slicing framework with self-sustained capability. Although we have presented a preliminary study on the key challenges and solutions, much more studies are needed in order to bring them into practice. For instance, current related works mainly focus on the designing framework of network slicing and optimization of slicing control but neglect to integrate network slicing with new physical layer technologies, such as Orthogonal Time Frequency Space modulation (OTFS), Non-Orthogonal Multiple Access (NOMA), massive Multiple-Input Multiple-Output (MIMO), meta-surfaces, and deployment of Unmanned Aerial Vehicles (UAVs). With these new physical layer technologies, the transmission latency of service requests can be further decreased due to the improved spectrum efficiency. Nevertheless, the self-sustained RAN slicing framework can certainly play an important role in the 5G-beyond and 6G networks.

References

[1] C. Yang, W. M. Shen, and X. B. Wang, The internet of things in manufacturing: Key issues and potential applications, *IEEE Syst. Man Cybern. Mag.*, vol. 4, no. 1, pp. 6–15, 2018.

[2] E. C. Strinati, S. Barbarossa, J. L. Gonzalez-Jimenez, D. Ktenas, N. Cassiau, L. Maret, and C. Dehos, 6G: The next frontier: From holographic messaging to artificial intelligence using subterahertz and visible light communication, *IEEE Vehicular Technol. Mag.*, vol. 14, no. 3, pp. 42–50, 2019.

[3] Z. G. Zhang, Y. Xiao, Z. Ma, M. Xiao, Z. G. Ding, X. F. Lei, G. K. Karagiannidis, and P. Z. Fan, 6G wireless networks: Vision, requirements, architecture, and key technologies, *IEEE Vehicular Technol. Mag.*, vol. 14, no. 3, pp. 28–41, 2019.

[4] R. H. Wen, G. Feng, J. H. Tang, T. Q. S. Quek, G. Wang, W. Tan, and S. Qin, On robustness of network slicing for next-generation mobile networks, *IEEE Trans. Commun.*, vol. 67, no. 1, pp. 430–444, 2019.

[5] J. Mei, X. B. Wang, and K. Zheng, Intelligent network slicing for V2X services toward 5G, *IEEE Netw.*, vol. 33, no. 6, pp. 196–204, 2019.

[6] *Study on Management and Orchestration of Network Slicing for Next Generation Network*, 3GPP, TR 28.801 V15.1.0, 2018.

[7] D. Naboulsi, M. Fiore, S. Ribot, and R. Stanica, Large-scale mobile traffic analysis: A survey, *IEEE Commun. Surv. Tutorials*, vol. 18, no. 1, pp. 124–161, 2016.

[8] Network Functions Virtualisation (NFV), *Architectural Framework*. Sophia Antipolis, France: ETSI, 2013.

[9] R. Mijumbi, J. Serrat, J. L. Gorricho, N. Bouten, F. De Turck, and R. Boutaba, Network function virtualization: State-of-the-art and research challenges, *IEEE Commun. Surv. Tutorials*, vol. 18, no. 1, pp. 236–262, 2016.

[10] *System Architecture for the 5G System*, *Stage 2* (*Release 15*), 3GPP, TS 23.501, 2018.

[11] T. Taleb, M. Corici, C. Parada, A. Jamakovic, S. Ruffino, G. Karagiannis, and T. Magedanz, EASE: EPC as a service to ease mobile core network deployment over cloud, *IEEE Netw*, vol. 29, no. 2, pp. 78–88, 2015.

[12] M. Bagaa, T. Taleb, A. Laghrissi, A. Ksentini, and H. Flinck, Coalitional game for the creation of efficient virtual core network slices in 5G mobile systems, *IEEE J. Sel. Areas Commun.*, vol. 36, no. 3, pp. 469–484, 2018.

[13] H. Halabian, Distributed resource allocation optimization in 5g virtualized networks, *IEEE J. Sel. Areas Commun.*, vol. 37, no. 3, pp. 627–642, 2019.

[14] R. Ferrus, O. Sallent, J. Perez-Romero, and R. Agusti, On 5G radio access network slicing: Radio interface protocol features and configuration, *IEEE Commun. Mag.*, vol. 56, no. 5, pp. 184–192, 2018.

[15] I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini, and H. Flinck, Network slicing and softwarization: A survey on principles, enabling technologies, and solutions, *IEEE Commun. Surv. Tutorials*, vol. 20, no. 3, pp. 2429–2453, 2018.

[16] V. Sciancalepore, M. Di Renzo, and X. Costa-Perez, STORNS: Stochastic radio access network slicing, in *Proc. IEEE Int. Conf. on Communications*, Shanghai, China, 2019, pp. 1–7.

[17] B. Khodapanah, A. Awada, I. Viering, A. N. Barreto, M. Simsek, and G. Fettweis, Slice management in radio access network via iterative adaptation, in *Proc. IEEE Int. Conf. on Communications*, Shanghai, China, 2019, pp. 1–7.

[18] P. Caballero, A. Banchs, G. de Veciana, X. Costa-Pérez, and A. Azcorra, Network slicing for guaranteed rate services: Admission control and resource allocation games, *IEEE Trans. Wirel. Commun.*, vol. 17, no. 10, pp. 6419–6432, 2018.

[19] Y. L. Lee, J. Loo, T. C. Chuah, and L. C. Wang, Dynamic network slicing for multitenant heterogeneous cloud radio access networks, *IEEE Trans. Wirel. Commun.*, vol. 17, no. 4, pp. 2146–2161, 2018.

[20] Y. X. Hua, R. P. Li, Z. F. Zhao, X. F. Chen, and H. G. Zhang, GAN-powered deep distributional reinforcement learning for resource management in network slicing, *IEEE J. Sel. Areas Commun.*, vol. 38, no. 2, pp. 334–349, 2020.

[21] M. Yan, G. Feng, J. H. Zhou, Y. Sun, and Y. C. Liang, Intelligent resource scheduling for 5G radio access network slicing, *IEEE Trans. Vehicular Technol.*, vol. 68, no. 8, pp. 7691–7703, 2019.

[22] Y. P. Xiao, J. W. Lai, and Y. B. Liu, A user participation behavior prediction model of social hotspots based on influence and Markov random field, *China Commun.*, vol. 14, no. 5, pp. 145–159, 2017.

[23] A. Zappone, M. Di Renzo, and M. Debbah, Wireless networks design in the era of deep learning: Model-based, AI-based, or both? *IEEE Trans. Commun.*, vol. 67, no. 10, pp. 7331–7376, 2019.

[24] Y. N. Liu, X. B. Wang, G. Boudreau, A. B. Sediq, and H. Abou-Zeid, Deep learning based hotspot prediction and beam management for adaptive virtual small cell in 5G networks, *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 4, no. 1, pp. 83–94, 2020.

[25] J. L. Li, W. S. Shi, P. Yang, Q. Ye, X. S. Shen, X. Li, and J. Rao, A hierarchical soft RAN slicing framework for differentiated service provisioning, *IEEE Wirel. Commun.*, doi: 10.1109/MWC.001.2000010.

[26] M. Zambianco and G. Verticale, Interference minimization in 5G physical-layer network slicing, *IEEE Trans. Commun.*, vol. 68, no. 7, pp. 4554–4564, 2020.

[27] R. Ferrus, O. Sallent, J. Perez-Romero, and R. Agusti, On 5G radio access network slicing: Radio interface protocol features and configuration, *IEEE Commun. Mag.*, vol. 56, no. 5, pp. 184–192, 2018.

[28] Y. L. Lee, J. Loo, T. C. Chuah, and L. C. Wang, Dynamic network slicing for multitenant heterogeneous cloud radio access networks, *IEEE Trans. Wirel. Commun.*, vol. 17, no. 4, pp. 2146–2161, 2018.

[29] M. Yan, G. Feng, J. H. Zhou, Y. Sun, and Y. C. Liang, Intelligent resource scheduling for 5G radio access network slicing, *IEEE Trans. Vehicular Technol.*, vol. 68, no. 8, pp. 7691–7703, 2019.

[30] C. Y. Zhang, P. Patras, and H. Haddadi, Deep learning in mobile and wireless networking: A survey, *IEEE Commun. Surv. Tutorials*, vol. 21, no. 3, pp. 2224–2287, 2019.

[31] M. Z. Chen, U. Challita, W. Saad, C. C. Yin, and M. Debbah, Artificial neural networks-based machine learning for wireless networks: A tutorial, *IEEE Commun. Surv. Tutorials*, vol. 21, no. 4, pp. 3039–3071, 2019.

[32] A. Zappone, M. Di Renzo, M. Debbah, T. T. Lam, and X. W. Qian, Model-aided wireless artificial intelligence: Embedding expert knowledge in deep neural networks for wireless system optimization, *IEEE Vehicular Technol. Mag.*, vol. 14, no. 3, pp. 60–69, 2019.

Xianbin Wang received the PhD degree in electrical and computer engineering from National University of Singapore in 2001. He is a professor and Tier 1 Canada Research Chair at Western University, Canada. Prior to joining Western University, he was a research scientist/senior research scientist at Communications Research Centre Canada (CRC) between July 2002 and December 2007. From January 2001 to July 2002, he was a system designer at STMicroelectronics. His current research interests include 5G and beyond, Internet-of-Things, communications security, machine learning, and intelligent communications. He has over 400 peer-reviewed journal and conference papers, in addition to 30 granted and pending patents and several standard contributions. He is a fellow of Canadian Academy of Engineering, a fellow of Engineering Institute of Canada, a fellow of IEEE, and an IEEE distinguished lecturer. He has received many awards and

recognitions, including Canada Research Chair, CRC President's Excellence Award, Canadian Federal Government Public Service Award, Ontario Early Researcher Award, and six IEEE Best Paper Awards. He currently serves as an editor/associate editor for *IEEE Transactions on Communications*, *IEEE Transactions on Broadcasting*, and *IEEE Transactions on Vehicular Technology*. He was also an associate editor for *IEEE Transactions on Wireless Communications* between 2007 and 2011 and *IEEE Wireless Communications Letters* between 2011 and 2016. He was involved in many IEEE conferences including GLOBECOM, ICC, VTC, PIMRC, WCNC, and CWIT, in different roles such as symposium chair, tutorial instructor, track chair, session chair, and TPC co-chair. He is currently serving as the vice chair of IEEE London Section and the chair of ComSoc Signal Processing and Computing for Communications Technical Committee.

Jie Mei received the BS degree from Nanjing University of Posts and Telecommunications (NJUPT), China in 2013. He received the PhD degree from Beijing University of Posts and Telecommunications (BUPT) in 2019. Since August 2019, he has been a postdoctoral associate at Electrical and Computer Engineering, Western University, Canada. His research interests include intelligent communications and V2X communication.

Kan Zheng received the BS, MS, and PhD degrees from Beijing University of Posts and Telecommunications, China in 1996, 2000, and 2005, respectively. He is currently a full professor at Beijing University of Posts and Telecommunications, China. He has rich experiences on the research and standardization of new emerging technologies. He is the author of more than 200 journal articles and conference papers in the field of vehicular networks, Internet-of-Things, machine learning, and so on. He holds editorial board positions for several journals and has organized several special issues. He has also served in the Organizing/TPC Committees for more than ten conferences, such as IEEE PIMRC, IEEE SmartGrid, and so on.