International Journal of Future Generation Communication and Networking
Vol. 13, No. 3, (2020), pp. 684-708

A Network Segmentation in Next Generation Wireless Systems with the Help of Ai Techniques

Jasti Swarupa, SK Pujitha, Bejjanki Pradeep Kumar
[1]*Assistant Professor, Deptof IT, Vignan's Lara Institute of Technology& Science, A.P., India*
[2]*Assistant Professor, Deptof CSE, Vignan's Lara Institute of Technology& Science, A.P., India*
[3]*Assistant Professor, Deptof IT, Vignan's Lara Institute of Technology& Science, A.P., India*
Email: swarupa.jasti.77@gmail.com, pujithapu533@gmail.com, bpkn@rocketmail.com,

Abstract
The integration of communications with specific scales, various radio get right of entry to technologies, and diverse network assets renders next-era wireless networks (NGWNs) tremendously heterogeneous and dynamic. Emerging use instances and applications, together with system to machine communications, self sufficient driving, and factory automation, have stringent necessities in terms of reliability, latency, throughput, and so on. Such necessities pose new demanding situations to structure design, community control, and useful resource orchestration in NGWNs. Starting from illustrating these challenges, this paper presents at imparting an excellent know-how of the overall structure of NGWNs and 3 particular research problems under this structure.
First, we introduce a network-slicing based totally architecture and give an explanation for why and in which artificial intelligence (AI) need to be integrated into this structure. Second, the motivation, studies challenges, present works, and potential future instructions related to applying AI-based totally processes in three studies troubles are defined in detail, i.e., bendy radio access community slicing, computerized radio access generation selection, and cell facet caching and content shipping. In summary, this paper highlights the benefits and potentials of AI-primarily based tactics in the research of NGWNs.
Keywords : Next-generation wireless networks, heterogeneous networks, network slicing, machine learning, radio network slicing, radio access technology selection.

1.INTRODUCTION

A. Next-Generation Wireless Networks: Visions & Challenges

The evolution of mobile communications from the first to the fifth generation (5G) has revolutionized many aspects of human society in the past four decades. Expediting this evolution, the next-generation wireless networks (NGWNs) are envisioned to be the cornerstone for a vast number of novel applications, ranging from remote surgery to smart cities. Following the classification of services into enhanced mobile broadband (eMBB), massive machine-type communications (mMTC), and ultra-reliable and low-latency communications (URLLC) [1], the NGWNs will support even more diversified services with various throughput, latency, and reliability requirements [2]. Meanwhile, thanks to improved reliability and connection density, the NGWNs are expected to attract enterprise users, in addition to conventional mobile communication users, by supporting use cases such as autonomous driving and factory automation [3], [4].

The above evolution has been shaping wireless networks towards becoming increasingly heterogeneous and dynamic [5]. For instance, NGWNs will incorporate various components such as device-to-device (D2D), vehicle-to-everything (V2X), and mobile edge computing (MEC), with different radio access technologies including cellular, Wi-Fi, and dedicated short-range communications (DSRC), as well as different access points such as cellular base stations (BSs), road-side units (RSUs), and unmanned aerial vehicles (UAVs). Each component in the integrated heterogeneous communication networks can have a unique focus and a corresponding set of performance metrics. For example, V2X communications must handle highly dynamic

ISSN: 2233-7857 IJFGCN

communication channels and rapidly changing network topology, while D2D communications require decentralized channel access control and communication resource allocation with high energy efficiency. As the heterogeneous and dynamic characteristics are inevitable results of supporting ever-growing demands for increasingly-diverse communication services, they impose significant challenges in the architecture design, network deployment, and network management in NGWNs.

Designing the architecture for NGWNs that can handle diversified services and maximize infrastructure and resource utilization efficiency is the first major challenge. Achieving the goals of increasing network capacity and accommodating highly diverse services with stringent quality of service (QoS) requirements necessitates innovations in network architecture. Network densification via deploying ultra-dense small cells can improve network capacity [6]. However, it does not provide a solution to scalable management of heterogeneous networks, but creates additional challenges such as extra infrastructure deployment cost, low cell utilization efficiency, and inter-cell interference. The integration of terrestrial and space networks has been proposed for providing seamless communication coverage [7]. Such integration, however, poses a further challenge in network management considering the dynamic trajectory of UAVs, the orbits of satellites, and the resulting impact on the service range and communication channels. A cloud/Fog-radio access network (RAN) based architecture, which incorporates the paradigm of cloud and fog computing into wireless networks, has also been proposed [8]. However, such an architecture focuses on improving energy efficiency, reducing cost, and alleviating data traffic on the fronthaul rather than satisfying diversified service requirements in complex heterogeneous networks.

The second challenge is how to achieve scalable and intelligent network management that can adapt to dynamic network environments. Network environments can change rapidly due to user mobility, time-varying channel conditions, dynamically changing traffic load distribution, and temporal variations of content popularity. Up to the current generation of wireless communication services, the problem of handling a dynamic environment has been studied mostly on a small scale, i.e., from the perspective of individual or several mobile users or base stations. One example is opportunistic spectrum access that targets individual secondary mobile users for them to access channels in a dynamic network environment, [1]. Another example is the dynamic deployment of virtual machines in cloud-fog computing systems based on computing task arrival patterns [1]. Nevertheless, managing NGWNs requires the development of scalable and adaptive models and approaches that suit large-scale problems and heterogeneous network architectures, which should include both centralized and decentralized network control components.

Last but not least, effective real-time network resource orchestration in the presence of multi-dimensional resources, many service types, and unknown traffic models is another challenge. The NGWNs will integrate functionalities of networking, caching, computing, sensing, and control [2]. Correspondingly, the resources in NGWNs will extend beyond the conventional communication resources (i.e., bandwidth, time, and/or transmit power), and include computing and caching resources. As a result, adaptive and flexible network resource orchestration becomes crucial, considering the surging growth in data traffic and increasingly diversified and stringent QoS requirements. Conventional centralized resource allocation can become inadequate in certain parts of NGWNs. For example, resource allocation in microcells and D2D communications may need to be decided locally in order to reduce signaling overhead and response time [3]. In addition, conventional approaches that rely on instantaneous network information, such as channel state information, and focus on optimizing an instantaneous performance metric, such as instantaneous data rate, can become inapplicable when such information is unknown. In NGWNs, exploiting the spatial-temporal traffic patterns while achieving service differentiation and maintaining massive connectivity will be a major challenge in network resource orchestration.

ISSN: 2233-7857 IJFGCN

B. Network-Slicing Based Architecture

Network slicing is an important network architecture innovation in 5G that is also expected to be inherited in the next generation [4]–[5][6]. Network slicing enables the coexistence of multiple isolated and independent virtual (logical) networks, i.e., slices, on the same physical network infrastructure. The advantages of network slicing are multifold. First, through the multiplexing of the virtual networks, network slicing supports multi-tenancy, i.e., multiple virtual network operators (VNOs) sharing the same physical network infrastructure [7]. This reduces capital expense in network deployment and operation. Second, network slicing provides the potential to create customized slices for different service types with various QoS requirements, which can achieve service differentiation and guarantee service level agreement (SLA) for each service type. Third, as slices can be created on-demand and modified or annulled as needed, network slicing increases the flexibility and adaptability in network management [1].

The enabling techniques for implementing network slicing are software-defined networking (SDN) and network function virtualization (NFV). SDN leverages the cloud computing paradigm in network management, such that the network has a centralized controller to dynamically steer and manage traffic flow and orchestrate network resource allocation for performance optimization [1]. An SDN controller provides the abstract set of resources and control logic for establishing slices, and a slice can be viewed as an SDN client context [2]. Therefore, SDN facilitates the pre-defining of slice blueprints as well as the on-demand creation of slice instances based on the corresponding service characteristics and requirements. NFV implements network functions, e.g., firewall, load balancing, address translation, etc., as software instances, known as virtual network functions (VNFs), running on virtual machines on top of general servers (referred to as NFV nodes) without requiring specialized hardware [2]. Thus, a network service in NFV can be considered as a component of a network slice, while a network slice contains one or more VNFs [3]. NFV complements SDN in implementing network slicing since SDN establishes control plane functions that enable slicing while NFV provisions services and manages the life cycle of network slices and orchestrates slice resources through realizing VNFs [4].

Despite its popularity in both academia and industry, the slicing of RANs faces several challenges. For example, determining the optimal slicing granularity, i.e., whether or not there should be a slice for each type of service, each set of QoS requirements, each VNO, or some combination of the aforementioned, is an open problem [5]. In addition, effective admission control that strikes a balance among infrastructure utilization, service provisioning in each slice, and the revenue of network operator calls for further investigation [6]. Last, the monitoring of slice SLA and the slice adaption based on traffic dynamics can be challenging, considering that the resource allocation among slices aims at slice isolation. The aforementioned challenges, generally involving making optimal decisions in a dynamic environment with unknown information, may not be solved following conventional model-based methods. Therefore, although network slicing will continue to be an important part of the NGWNs, additional innovations in the network architecture are necessary for addressing the above challenges.

C. Integrating Artificial Intelligence

The past decade has witnessed remarkable advances in the research and applications of artificial intelligence (AI). Research in machine learning (ML), one of the most powerful AI tools, has been progressing rapidly to embrace a wide range of applications including voice recognition, image processing, and self-driving vehicles. The rapid advances in ML, boosted by the progress in hardware technology specialized to support AI, paves the path for applying AI in NGWNs [7]. A major advantage of ML is its ability to handle complicated problems, which renders ML a powerful tool that suits the dynamic, heterogeneous, and decentralized features of NGWNs. Applying ML can potentially yield benefits such as improved performance and faster convergence in network management automation and performance optimization in large-scale systems.

ISSN: 2233-7857 IJFGCN

ML methods include supervised learning, unsupervised learning, and reinforcement learning (RL), each of which suits a different group of research problems in wireless communications. Supervised learning relies on labeled data to learn the mapping from the input to the output, and can be used to analyze network data, learn network characteristics, and estimate network parameters [2]. Applications of supervised learning in communications and networks include traffic classification [2], smart offloading [3], sub-6 GHz to millimeter wave (mmWave) frequency handover [31], and mmWave beam alignment [3,2]. Unsupervised learning identifies patterns and attributes hidden in data for inference and prediction without using labeled data. Potential applications in communication networks include spectrum sensing [3] and traffic volume prediction [3,4]. RL iteratively learns the optimal decisions, based on the feedback of network state information, to maximize a cumulative reward in the long term. RL methods are particularly suitable for decision making in a dynamic environment. The applications of RL include protocol design [3,5] and user scheduling with resource allocation [3,6].

Due to their potential applications and benefits, applying ML methods is gaining momentum in the research and development (R&D) of communication networks to enhance system performance, flexibility, and scalability. For network data analysis, ML can handle the heterogeneity and spatial-temporal diversity in the data for network design and management [3,7]. For user mobility management, ML is a tool for analyzing the mobility pattern of mobile users for location-based services [3]. For network resource management, ML-based methods can be applied to model and study the joint allocation of communication, caching, and computing resources [3] or the joint problem of content caching and delivery [4].

As mentioned earlier, the heterogeneous and dynamic characteristics of NGWNs demand powerful tools to automate and optimize network slicing. From existing studies in literature, it can be seen that ML is potentially a promising tool for this purpose. Applying ML in network slicing can provide the innovations required to address the aforementioned challenges in the network architecture and resource orchestration and, thereby, help fulfill the great prospect of NGWNs.

The rest of this paper is organized as follows. Section II provides a description of the overall architecture. Sections III to V discuss three research problems in a network-slicing based architecture, as well as related research efforts and, in particular, AI-based approaches. Section III focuses on RAN slicing. In Section IV, we present radio access technology (RAT) selection and user association automation. Section V investigates content caching and content delivery. Section VI concludes this study. Table 1 lists the acronyms used in this paper.

ISSN: 2233-7857 IJFGCN

TABLE 1 List of Acronyms

3GPP	3rd Generation Partnership Project	5G	Fifth Generation
ABC	Always-best-connected	AI	Artificial Intelligence
AP	Access Point	AR	Augmented Reality
BBU	Baseband Unit	BS	Base Station
CapEx	Capital Expenditure	CNN	Convolution Neural Network
CRAN	Cloud Radio Access Network	D2D	Device-to-device
DNN	Deep Neural Network	DCF	Distributed Coordination Function
DL	Downlink	DSRC	Dedicated Short-range Communications
eMBB	Enhanced Mobile Broadband	HARQ	Hybrid Automatic Repeat-request
HNN	Hopfield Neural Network	IoT	Internet of Things
LRU	Least Recent Used	LSTM	Long Short-term Memory
LTE	Long Term Evolution	MADM	Multiple Attribute Decision Making
MDP	Markov Decision Process	MEC	Mobile Edge Computing
ML	Machine Learning	mMTC	Massive Machine-type Communications
NFV	Network Function Virtualization	NGWN	Next-generation wireless network
OpEx	Operational Expenditure	POMDP	Partially Observable Markov Decision Process
QoS	Quality of Service	RAN	Radio Access Network
RAT	Radio Access Technology	RL	Reinforcement Learning
RNN	Recurrent Neural Network	RSU	Road-side Unit
SDN	Software-defined Networking	SLA	Service Level Agreement
UAV	Unmanned Aerial Vehicle	UE	User Equipment
UL	Uplink	URLLC	Ultra-reliable and Low-latency Communications
V2X	Vehicle-to-everything	VNF	Virtual Network Function
VNO	Virtual Network Operator	VR	Virtual Reality

II. Network Architecture

This section presents the overall network architecture based on network slicing and discusses where and how AI can be applied. Due to the challenges mentioned in the introduction, the NGWN architecture is expected to have the following properties [4]:

- Flexible and scalable, to support a wide range of service types and QoS requirements, and to support scalable slice management after the deployment of slices;
- Automated and adaptive, to support automated RAN and cloud network resource allocation and adaptation based on data traffic and network performance, and to support automated slice creation, slice performance monitoring, and slice adaption;
- Open and modularized, to support customized slices defined or operated by VNO, and to open certain network management functions to third parties.

A network-slicing based AI-assisted network architecture satisfying the above properties is illustrated in Fig. 1. This architecture employs two-tier controllers, with a logical centralized SDN controller placed at a central cloud, and local SDN controllers at individual RANs. Each local controller is connected to the infrastructure in its corresponding RAN and responsible for collecting the network information and making local decisions in network operations. VNFs are deployed at servers connected to radio heads, APs, storage facilities, local data servers, etc. In the context of RAN, VNFs consist of baseband unit (BBU) functions, e.g., compression and encryption procedures and hybrid automatic repeat-request (HARQ) [4,2], [3]. Accordingly, network slicing translates to the placement of VNFs into various slices (subject to physical infrastructure constraints and QoS requirements), the establishment of the logical topology of the VNFs in each slice, and the mapping from the VNFs to the underlying physical infrastructure. In this architecture, computing becomes especially important due to the virtualization of network functions since the placement of VNFs in the slices is essentially the allocation of required computing resources.

ISSN: 2233-7857 IJFGCN

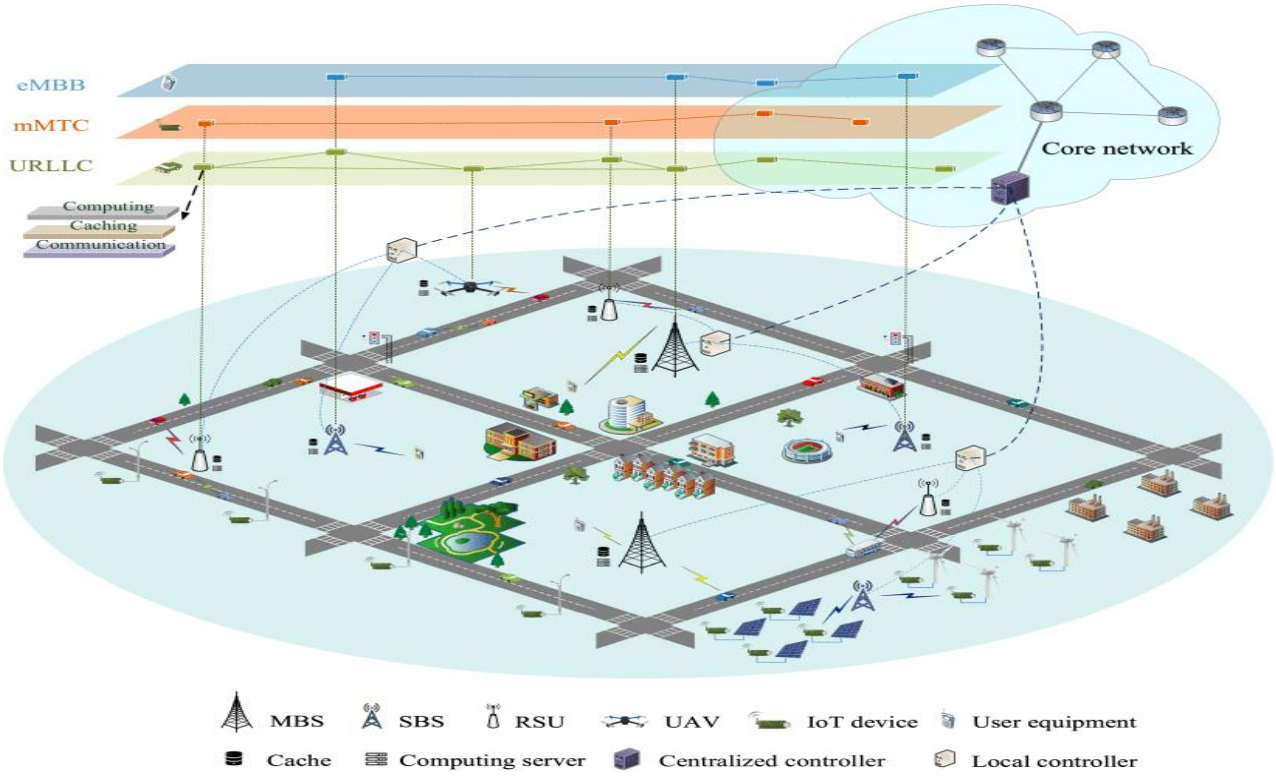

Figure 1. An illustration of network-slicing based NGWN architecture,with three example network slices.

The key functional components and their relations corresponding to the architecture in Fig. 1 are shown in Fig. 2. While end-to-end (E2E) connections span both wireless segment(s) and the core network, this illustration focuses on the wireless domain. The centralized SDN controller is responsible for slice blueprint definition and end-to-end slicing based on the information collected from local controllers. Local controllers are responsible for assisting the centralized controller in the slicing of their corresponding RANs. After a slice is deployed, the corresponding local controller is responsible for orchestrating slice resources among end users as well as monitoring slice status for resource utilization and QoS satisfaction. In addition, local controllers can be involved in slice adaption, while the centralized SDN controller may or may not be involved depending on the service type and the use case. The network status and operation data, aggregated from all slices, are collected by local controllers and either processed locally or forwarded to the centralized SDN controller for analysis. The analysis results will be used to update slice deployment and slice adaption. The relation between centralized and local controllers introduces an important question: whether and when should the centralized controller be involved in specific network management and resource allocation tasks under this network architecture? Evidently, involving the centralized SDN controller in such tasks can take advantage of the global network information for making optimal network management and resource allocation decisions. However, decision making via the centralized SDN controller can incur significant signaling overhead. Therefore, a balance between the tasks for the centralized and local controllers, which depends on the type of tasks and the type of a slice, should be investigated.

ISSN: 2233-7857 IJFGCN

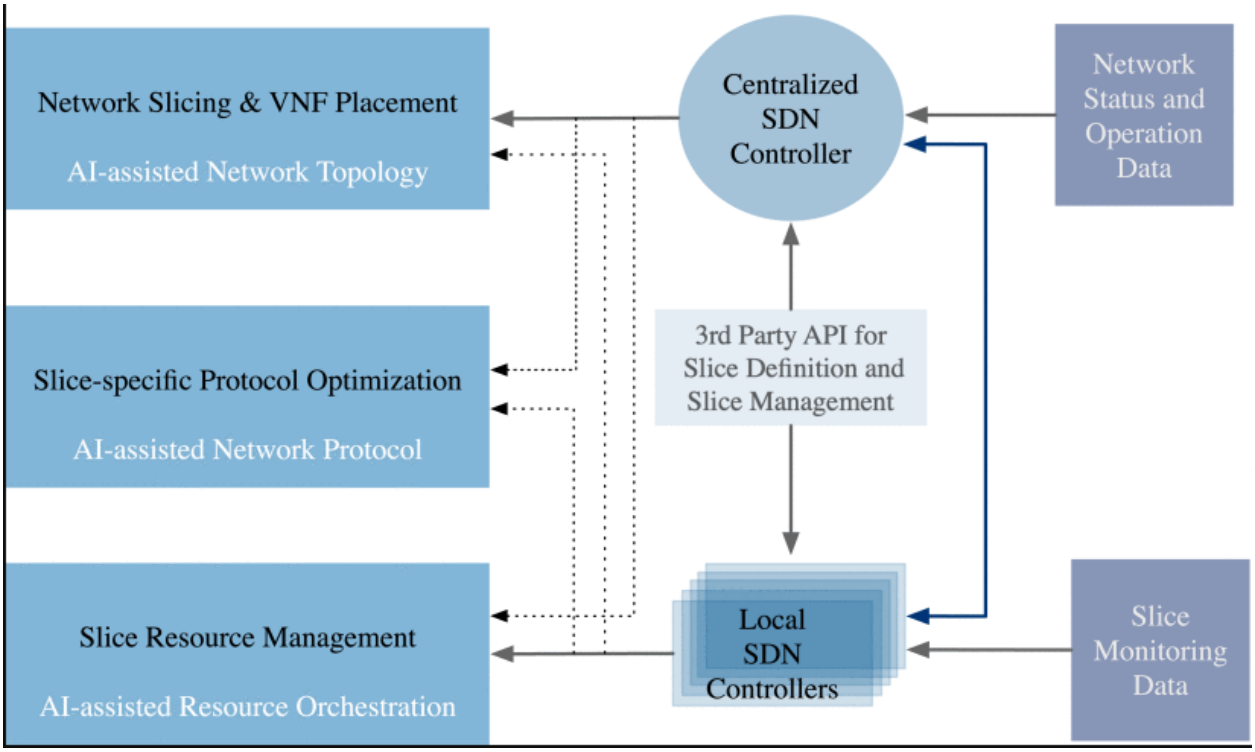

Figure 2. The functional architecture of a network-slicing based AI-assisted next-generation RAN.

The three blocks in Fig. 2, i.e., network topology, network protocol, and network resource orchestration, form a closed loop, which reflects the interplay between the two levels of network management: network planning and network resource scheduling [5,2]. Network planning, including the initial resource reservation for all slices, corresponds to the block of network topology in Fig. 2. Meanwhile, network resource scheduling consists of the network protocol and network resource orchestration blocks in Fig. 2, where the resource orchestration applies within each slice. As shown in Fig. 3, network planning admits slice requests, reserves resources for the admitted slices, and determines the placement of required VNFs in each slice. Based on the result of network planning, network resource scheduling further allocates resources in a slice to individual network users dynamically. The resulting SLA violation and resource utilization in the admitted slices are monitored, based on which the network planning may be adjusted in the future.

ISSN: 2233-7857 IJFGCN

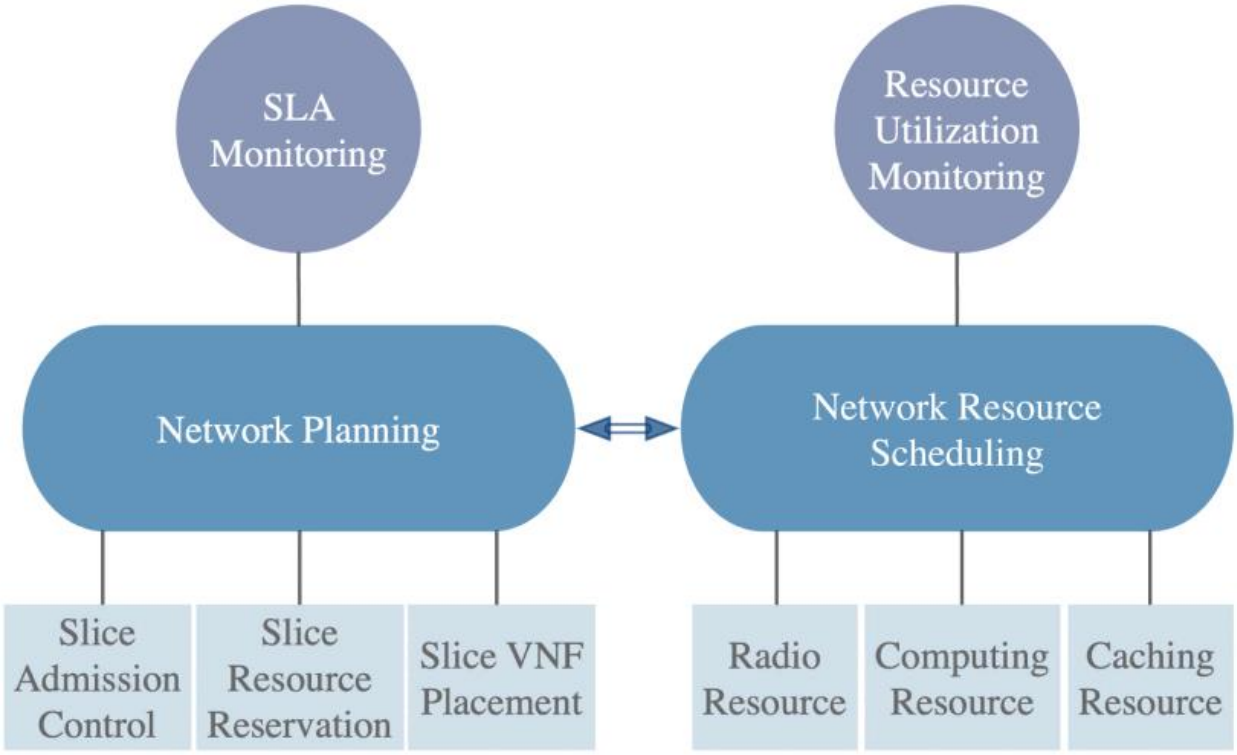

Figure 3. The interplay between network planning and network resource scheduling in the slicing based architecture.

The role of AI in the network architecture includes exploiting the slice SLA monitoring data and the slice resource utilization data to facilitate slice deployment, slice adaption, and slice update. Due to the heterogeneous and dynamic characteristics of NGWNs, conventional approaches based completely on statistical models are likely to become intractable or too slow, if not both. The results are suboptimal if the statistical models are inaccurate. In addition, the required statistical models are unavailable for many new use cases and emerging applications. By contrast, AI-based approaches can potentially make use of the aforementioned monitoring data for both slice and network performance optimization. Through the application of AI for data analysis and decision making in both the centralized and local SDN controllers, the slicing based architecture in Fig. 1 can be empowered by AI.

In the following sections, we investigate the RAN slicing framework, RAT selection automation, and content caching and delivery, respectively, under this AI-assisted slicing based network architecture.

III.RAN Slicing Framework

RAN slicing is deemed as the most promising technology in 5G networks and beyond, by providing a flexible and scalable network architecture to support a variety of services attached to manifold QoS requirements. By slicing the shared physical wireless networks into multiple isolated logical networks, RAN slicing can dynamically and elastically allocate network resources to provide tailored services for isolated logical networks. Building on the shared physical network infrastructure, RAN slicing is a cost-effective solution for network management. A study reported that RAN slicing can reduce capital expenditure (CapEx) and operational expenditure (OpEx) by up to 60 billion USD worldwide within the next five years [48]. These benefits motivate the study on RAN slicing for NGWNs. Extensive industry efforts have been devoted to RAN slicing framework ratification. For example, network slicing has been introduced as one of the key features of international mobile telecommunication (IMT)-2020 network [3]. The 3rd generation partnership project (3GPP) has conducted extensive studies on the slicing based architecture for 5G networks [5]. Multiple proof-of-concept systems on RAN slicing have been developed and evaluated based on real-world network

traffic data. In this section, we first present the research challenges of RAN slicing. Then, existing works on RAN slicing are reviewed, which are summarized in Table 2. Finally, potential benefits and challenges of emerging AI-based RAN slicing are discussed.

TABLE 2 Summary of Literature on RAN Slicing

Topic	Related Work	Objective	Approach
Resource Allocation	[21]	Maximize the proportional sum rate of all the users via spectrum slicing in two-tier cellular networks	Optimization
	[44]	Maximize the network utility via spectrum slicing and transmit power allocation in vehicular networks	Optimization
	[26]	Maximize network revenue for URLLC and eMBB services in cellular networks	Optimization
	[45]	Maximize the network revenue by intelligently admitting network slice requests	RL
	[46]	Maximize the network revenue via dynamically slicing time-varying spectrum resource in indoor neural-host small cells	DDPG
	[47]	Minimize service latency in a sliced RAN by computing resource allocation and task transmission scheduling	Deep learning
	[48]	Maximize the long-term utility of the service provider via channel allocation in a sliced RAN	Multi-agent stochastic learning
	[49]	Maximize the utility of individual service provider by joint slicing computing and communication resources	Multi-agent double deep Q learning
Traffic Prediction	[50]	Predict the average service traffic load in a LTE testbed	LSTM
	[51]	Predict the maximum service-specific traffic load for each slice based on real-world 5G network data	Deep learning
	[45]	Predict the service-specific traffic load based on customized user mobility pattern	Unsupervised learning

A. Research Challenges in RAN Slicing

The RAN slicing in NGWNs can be divided into two steps: 1) Slice creation – Various over-the-top services with different QoS requirements request for creating slices to guarantee service isolation. After receiving a slice creation request, the network controller decides to accept or reject the request based on the availability of network resources. Once a slice is admitted, a new slice will be created based on the slice templates and network function instances; and 2) Resource orchestration – Network resources are allocated to admitted slices in order to meet their SLAs. Since emerging real-time mobile services, such as virtual reality (VR), augmented reality (AR), and autonomous driving, may consume multiple-dimensional resources (communication, computing and caching), a slice would be allocated with multiple virtualized network resources. These virtualized network resources would be mapped to the physical network infrastructure via a resource mapping algorithm. An illustrative example of the RAN slicing is shown in Fig. 4. The NGWNs become more complicated due to diverse network resources, heterogeneous network topology (e.g., cellular BSs, drone BS, WiFi APs), and differentiated QoS requirements (e.g., URLLC, eMBB, mMTC[1]). These characteristics create challenges for RAN slicing to support diverse services.

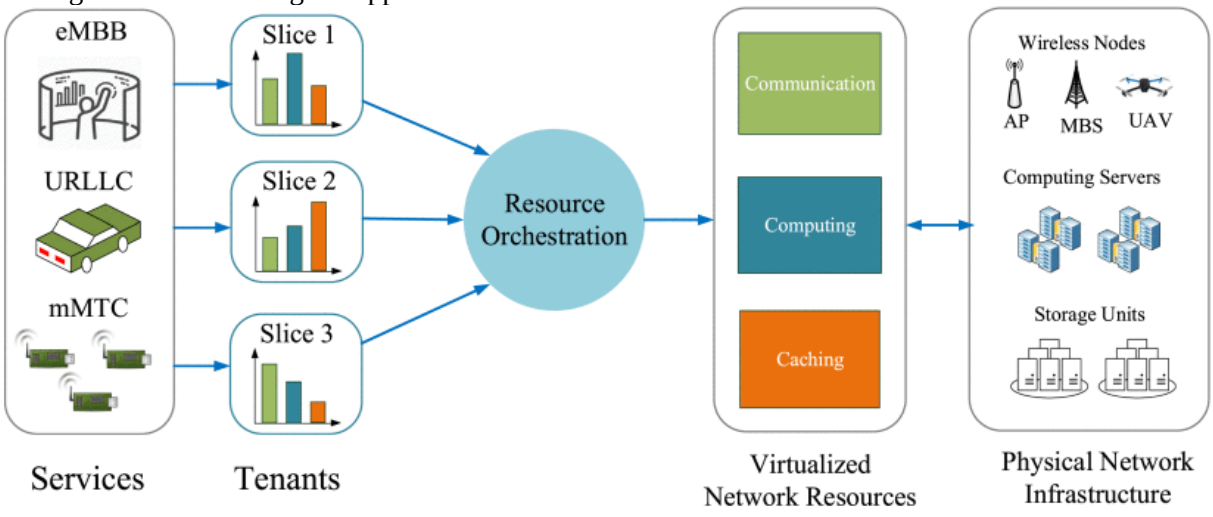

Figure 4. The RAN slicing in NGWNs.

The goal of RAN slicing is to provide customized services for mobile users with differentiated QoS requirements in heterogeneous networks. Hence, the key issue of RAN slicing is how to efficiently allocate network resources while meeting the user QoS requirements. As shown in Fig. 5, multiple-dimensional network resources of the shared network infrastructure are allocated to each slice in a slicing window.[2] Based on *a priori* service-specific traffic statistics, the communication resources can be sliced in terms of radio spectrum bandwidth, the computing resources can be sliced in terms of CPU computing power, and the caching resources can be sliced in terms of storage unit for each slice. Hence, RAN slicing should jointly allocate multiple network resources (e.g., communication, computing, and caching) to optimize the network utility, while satisfying the differentiated QoS requirements of customized services. Due to the heterogeneous network infrastructures and differentiated QoS requirements, RAN slicing faces the following unique challenges:

- Resource interplay – Since a service may consume multiple network resources, there exists an inherent tradeoff among the network resources. For example, in computing offloading services, the service latency consists of two elements: task transmission latency and task processing latency. If a user associates with a remote MEC server having abundant computing resources for task processing, a high task transmission latency will incur. On the other hand, if a user associates with a nearby MEC server having insufficient computing resources, it takes a longer time for task processing. In such a manner, the allocation of computing and communication resources is coupled with each other in the exemplary computing offloading services. Similarly, the allocation of multiple network resources is intertwined, which complicates the RAN slicing. A joint multiple network resource allocation scheme should be judiciously designed to maximize network welfare;

- Strict QoS requirements – Compared with traditional 4G networks, 5G networks and beyond have stricter QoS requirements, including a higher throughput and a lower latency. Especially, the typical URLLC service in 5G requires ultra-high reliability (e.g., 99.999%), which is much stricter than that of other services. In addition, the payload of data packets in URLLC services is usually small, such as 32 bytes [2]. The transmission performance of short-length packets cannot be characterized by the traditional Shannon theory which is suitable for long-length packet transmission due to a large transmission overhead. Instead, the finite block length channel coding theory should be applied to characterize the achievable rate for short-length packets [3]. Traditional QoS provisioning is unsuitable for short-length packet URLLC services with ultra-high reliability. Thus, an accurate QoS provisioning for URLLC services is desired in the RAN slicing framework;

- User mobility – Due to the high network density, users may frequently move out the coverage of its associated network infrastructure, which results in a dynamic network topology. For example, high-mobility vehicle users can trigger handover frequently. The dynamic network topology changes the service traffic distribution, rendering previously optimal slice allocation suboptimal over time, degrading network performance, and may even violate users' QoS requirements. When the network performance degrades to a threshold, adjusting existing slices or creating new slices will be triggered, which incurs slice reconfiguration overhead. Thus, dynamic yet efficient RAN slicing to accommodate user mobility remains a challenging issue.

ISSN: 2233-7857 IJFGCN

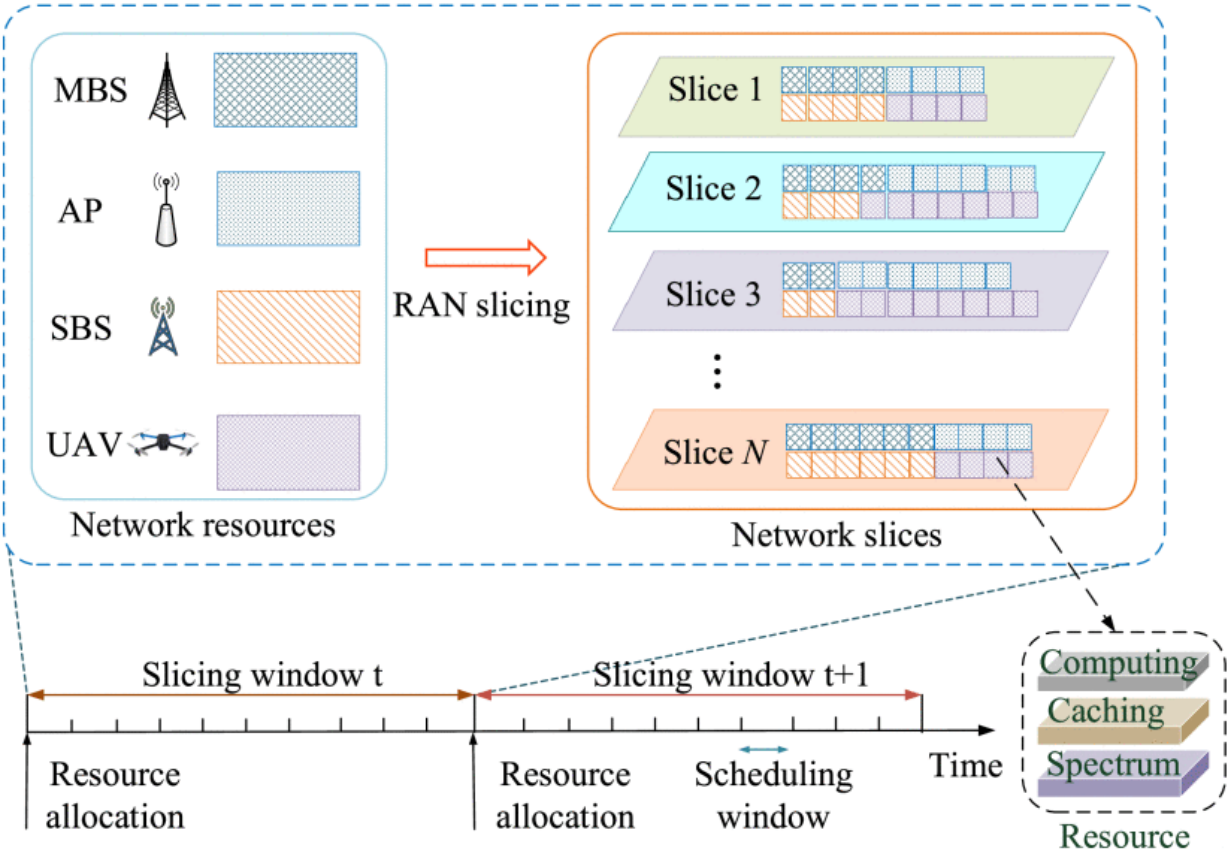

Figure 5. The network resources of the shared network infrastructure are allocated to each slice via RAN slicing in each slicing window.

B. Existing Approaches

Extensive research efforts have been devoted to RAN slicing in different contexts due to its advantages in reducing network operation cost and improving resource utilization. Based on the known service-specific traffic statistics, a communication resource slicing strategy is proposed to support both machine-type users and mobile users, by allocating the spectrum in heterogeneous networks [2], in which bandwidth resource and user association are jointly allocated to maximize network utility. The results show that the proposed slicing strategy can effectively boost the network utility compared with benchmark schemes. A communication resource slicing strategy is developed to provide customized services in the context of vehicular networks [44]. These works mainly formulate a RAN slicing problem as an optimization problem with the objective of maximizing network utility, while satisfying the QoS constraints of admitted slices. By resorting to optimization theory, these complicated optimization problems can be solved by classic iterative optimization algorithms. Another line of work focuses on the communication resource slicing from the perspective of a network operator with an objective of maximizing the operator's revenue in different scenarios, such as in cellular networks [6] and indoor neutral-host small cell networks [4,6]. Sub-optimal algorithms are applied to solve these complicated slicing problems. The existing works address the communication resource slicing from different perspectives. With emerging services and new use cases, further investigation is needed for RAN slicing that incorporates multiple-dimensional network resources. However, a multiple-resource slicing problem is much more complex than an individual resource slicing problem, taking account of network dynamics in traffic load and user mobility. In addition, existing works mainly deal with services attached to relatively loose QoS requirements, and hence developing RAN slicing to support strict URLLC services requires further investigation. In summary, model-based optimization methods are widely applied to solve RAN slicing problems, which can effectively manage resources in a small-scale network under simplified network statistical models.

ISSN: 2233-7857 IJFGCN

Existing model-based optimization methods suffer from two limitations: i) the prerequisite of *a priori* accurate traffic model – Service demand statistic models are usually assumed to be known in advance and accurate in most of the existing works, such as a known Poisson process to model service traffic of mobile users, which do not hold in practical time-variant wireless networks, especially in highly-mobility scenarios; and ii) high computational complexity – With the dense deployment of wireless networks, efficient RAN slicing for a large-scale network (e.g., tens to hundreds of BSs and APs) is required. Applying existing iterative model-based optimization methods can be unsuitable since the computational complexity greatly increases with the network scale, such that slicing algorithms will take a long time to converge. These limitations undermine the practicality of the existing model-based optimization methods. Hence, an efficient RAN slicing strategy for large-scale networks without accurate *a priori* traffic model, is of paramount importance.

C. AI-Based RAN Slicing

With the development of advanced AI techniques, model-free AI-based methods become promising techniques to provide potential benefits to address the difficulties with unknown traffic models and high computational complexity. In the following, the potential benefits and challenges of AI-based RAN slicing are discussed in detail.

AI-based methods can provide two potential benefits for RAN slicing. On one hand, we can use AI-based methods to provide accurate service-specific traffic prediction. Only with such accurately predicted service-specific traffic, RAN slicing can effectively facilitate network resource allocation to accommodate service demands in the near future. Recent studies show that AI-based methods, such as deep neural network (DNN) and long short-term memory (LSTM), are capable of accurately forecasting service-specific traffic load. For example, a DNN is used to predict aggregated data traffic in cellular networks based on historical service requests in. For fine-grained service-specific traffic, a prediction model based on a modified LSTM network is presented in to accurately predict the average traffic load, while a deep learning framework is proposed for the maximum service traffic prediction in , which can help to reduce resource over-provisioning and SLA violations. Based on historical user service requests and a known user mobility model, Sciancalepore *et al.* develop an unsupervised learning based forecasting module to predict service traffic load, with the traffic load prediction accuracy depending on the accuracy of the user mobility model. The existing preliminary studies illustrate the potential of an AI-based prediction method to accurately capture service traffic patterns. Such an accurate online service-specific traffic prediction can help to eliminate the requirement of *a priori* accurate traffic modeling in RAN slicing.

Moreover, AI-based methods can facilitate efficient resource allocation in RAN slicing. An online AI-based resource allocation decision process has the potential to achieve a low complexity after an offline training procedure, which addresses the high computational complexity challenge in the conventional model-based optimization methods. Recently, extensive research works have shown that AI-based methods can be widely applied in solving complicated resource management problems in wireless networks, such as power allocation for the interference management [7,5], resource block allocation in cloud radio access networks (CRANs) [7,6], SBS on/off scheduling in cellular networks [7,4], and computing task offloading in space-air integrated networks [7,7]. In general, the resource allocation problem is formulated as a Markov decision process (MDP), and an RL framework is developed for the MDP problem to make online decisions. As the essence of RAN slicing can be viewed as an optimization problem with the objective of maximizing network performance under constraints of satisfying QoS requirements, RL-based methods can be applied. In [4,5], an RL algorithm is presented to determine the optimal set of admitted slices in order to maximize the welfare of the infrastructure provider (e.g., 5G broker). Note that traditional RL methods, such as Q-learning, suffer from the curse of dimensionality, which are only suitable for RAN slicing problems in small-scale networks. Deep RL methods incorporate deep learning networks in the RL framework can effectively address the complexity issues in large-scale networks. Chen *et al.* present a deep RL learning based scheduling strategy to minimize service latency in a

ISSN: 2233-7857 IJFGCN

sliced RAN [4,7], using a modified deep RL method for computing power allocation and task transmission scheduling. An enhanced RL method, deep deterministic policy gradient (DDPG), is proposed to dynamically slice the shared time-varying spectrum resources in indoor small cell networks [4, 6]. In addition to the centralized network resource management for RAN slicing in, decentralized RAN slicing can be formulated as a multi-tenant RAN slicing problem, in which multiple tenants (i.e., slice owners) contend for network resources from an infrastructure provider. The multi-agent RAN slicing problem aims at bidding and allocating network resources to maximize the revenue of each tenant. The multi-tenant RAN slicing problem can be modeled as a non-cooperative stochastic game and solved by a stochastic learning algorithm. A deep learning approach based on a double deep Q network can be applied for jointly allocating communication and computing resources to maximize the welfare of each tenant. The existing studies demonstrate the potential of using AI-based approaches to address the RAN slicing problem in various contexts.

On the other hand, AI-based RAN slicing faces its unique challenges, such as achieving strict QoS guarantee within the RL framework. How to satisfy the QoS constraints in the RL framework requires innovative solutions in the RAN slicing optimization. Due to limitations of the Q-value based mathematical modeling, the QoS requirements are usually integrated into the reward function by some predefined weights [7]. In such a manner, strict QoS requirements cannot be guaranteed unless appropriate weights for the QoS requirements are determined, which is difficult to achieve, especially when the QoS requirements are multi-dimensional. Most of existing solutions can only satisfy soft QoS requirements [6]. Developing an efficient RL-based RAN slicing algorithm while satisfying strict QoS requirements requires further investigation.

IV. Automated RAT Selection

In NGWNs, multiple types of RAT will coexist. Thus, proper RAT selection for each user is essential. RAT selection is closely related to user association, which associates each user with specific APs.[3] In the simple scenario of a homogeneous network, RAT selection is basically user association. However, in a general scenario with heterogeneous networks, associating a user to an AP requires both the selection of an RAT and the selection of a specific AP given the chosen RAT. In this section, we use "*RAT selection*" as a synonym of "*user association*" in the case of heterogeneous networks.

User association has been widely studied for various network scenarios, especially in the case of homogeneous networks. Existing studies focus on user association in a multi-tier cellular network [7], under particular physical-layer settings (such as MIMO [8], mmWave [6], energy harvesting [1]), and other networking environments (such as self-organizing networks [2], D2D communications [3], and UAV-to-ground communications [9]. Many performance metrics, including spectrum efficiency, energy efficiency, and energy consumption, are considered in the study of user association [4].

In the rest of this section, we focus on RAT selection in slicing based NGWNs. Firstly, we give an overview of conventional user association schemes. Then, we introduce RAT selection in network-slicing based networks. After that, we review and discuss AI-assisted RAT selection.

A. Conventional User Association Approaches

As shown in Fig. 6, conventional user association can be divided into three categories: centralized, distributed, and hybrid based on control paradigm A global controller is assumed in the case of a centralized solution to collect network-related information. The centralized method determines a network association strategy by formulating the problem based on a Markov model or through a centralized optimization. Using a Markov model, the network selection is usually formulated as a joint user association and mobile traffic offloading problem. The target is to obtain a desirable admission and offloading policy to optimize certain system-level metrics such as service blocking probability. In the centralized optimization approach a selection algorithm is executed each time when

association decisions need to be updated. The target is to optimize system-level performance such as energy efficiency, spectrum efficiency, load balancing, or system aggregated utility, subject to network resource availability and user association constraints. However, both Markov model-based and optimization-based centralized user association approaches have their own limitations. In the Markov-based approach, user mobility and handoff are generally ignored (i.e., the users are assumed to associate with the same AP until the end of a service session). Moreover, users are treated without any differentiation, i.e., no consideration of user preference and service priorities in general. On the other hand, the centralized optimization method suffers from scalability and efficiency issues. In addition, the optimality is usually achieved at the cost of signaling overhead in the information gathering and the policy enforcement stages.

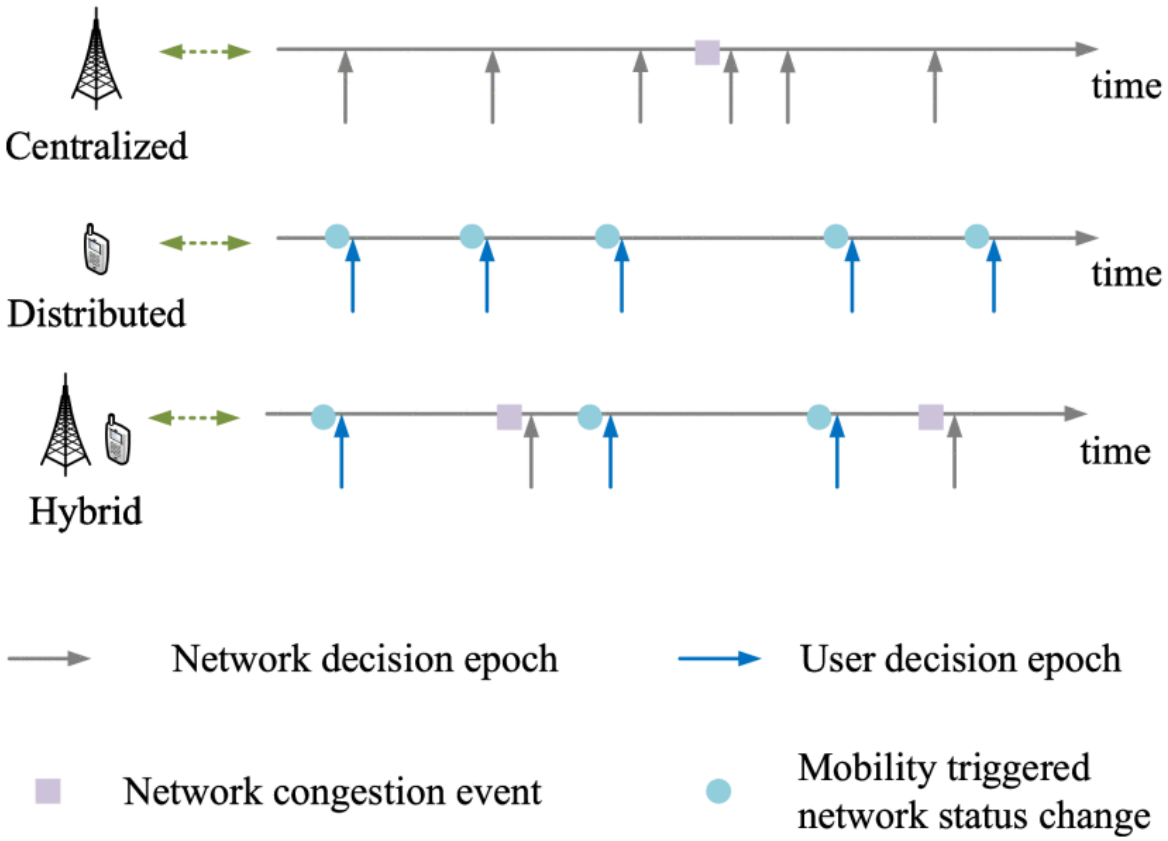

Figure 6. The classification of RAT selection based on the control paradigm: Centralized, distributed, and hybrid. For distributed user association, the case of mobility triggered network selection is illustrated here as an example.

Distributed user association has been studied using various methods, including multiple attribute decision making (MADM) , MDP, fuzzy-logic , game theory (e.g. cooperative game [1], and non-cooperative game . Under the distributed setting, network attributes are collected or estimated at the user side. The user then chooses the AP with the best performance. Compared to the centralized method, the distributed selection scheme can usually be implemented with lower complexity. Further, a decentralized approach can reduce the signaling overhead at the cost of suboptimal performance. The limitations of distributed selection schemes include that i) non-cooperative user association can lead to network load oscillation when multiple devices try to associate and disassociate with the same AP concurrently; and ii) the design of effective information exchange is necessary for distributed cooperative user association but can be very challenging.

The hybrid selection can achieve a tradeoff between network performance and signaling overhead, which can be implemented as a mixture of centralized and distributed control. In Elayoubi *et al.* solve the user association problem using a Bayesian game. Two types of players are involved, which

represent the different networks and the users. Each user selfishly maximizes their own utility without any user-level cooperation, while each network cooperates with users within its coverage by broadcasting its current status (such as the traffic load), to maximize the total utility of its users. Such a design may result in multiple Nash equilibria. Therefore, the information that the network needs to broadcast should be carefully designed, so that an equilibrium with a high efficiency can be achieved

B. RAT Selection in Slicing Based Heterogeneous Networks

Figure 7 shows an envisioned scenario of NGWNs in the presence of a SDN controller. In such networks, multiple types of RATs, multiple types of APs, multiple types of UEs with various service requirements, and multiple types of resources jointly contribute to an unprecedented level of heterogeneity. Next, we discuss the main differences of RAT selection under the slicing based NGWNs from that in conventional user association.

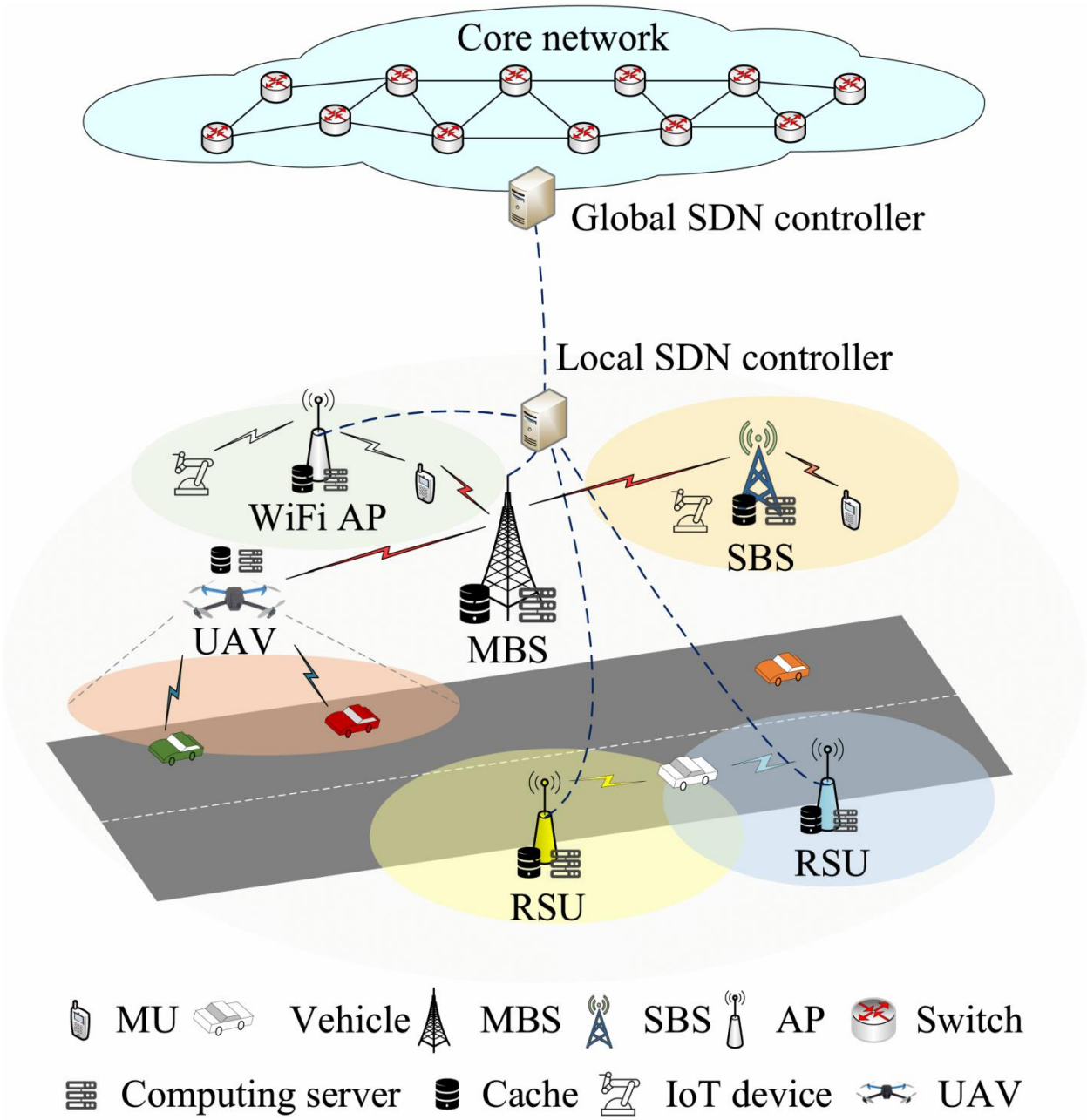

Figure 7. An illustration of RAT selection with multiple services in heterogeneous wireless networks

Firstly, from the control perspective, the capacity of the SDN controller for information collection and centralized control should be leveraged in the RAT selection. It can be seen from Fig. 7 that a global view of the network is enabled through the deployment of the SDN controller. Network status information such as current network loads, user service demands, and user distribution, as well as user status information such as user location, speed, and moving direction can be obtained by the SDN controller for centralized decision making. However, such centralized control is not scalable and can yield significant signaling overhead. Therefore, a hybrid control architecture is preferred in the NGWNs, in which users make distributed RAT selection decisions at a small timescale, while the centralized control is triggered at a large timescale.

Secondly, given the SDN/NFV enabled network slicing architecture [4], each slice is assigned with only a portion of physical resources based on its target services. Therefore, the resource availability and resource utilization level of each slice become a concern and need to be accounted for properly via the RAT selection. Further, considering network slicing, it is possible that a user is involved in multiple network slices [93]. The RAT selection in such a case requires further investigation.

Thirdly, new types of network resources are emerging, which can affect the RAT selection. Traditional network infrastructures (e.g., cellular BSs and WiFi APs) only have the communication functionality. As the networks continue to evolve, these infrastructures will support more and more caching and computing services. Such a trend leads to diverse network resources as compared to that in previous generations. In addition, an unprecedented level of network and service heterogeneity is expected and, correspondingly, the complexity of solving the RAT selection problem will increase in the NGWNs.

C. Research Challenges

Based on the preceding discussion, several research challenges related to RAT selection in the NGWNs are identified as follows.

1) Service Modeling

The dependence between service requirements and the corresponding demands for multi-dimension resources has not been modeled explicitly in conventional user associations. In the literature, users can either select the network based on a radio link quality, i.e., the always-best-connected (ABC) [4], or associate with a nearby AP that has the content of their interest in its cache [5] or a nearby AP that has high computing capability [1]. However, the network selection in NGWNs should be determined based on multi-dimensional resource availability for communication, caching, and computing, considering that different services can have totally different requirements on these three types of resources. For example, video streaming services require most attention to the communication resources (e.g., the link quality and the available bandwidth); vehicles downloading high-definition maps should connect to an AP which caches contents of their interest; for VR applications [6] (e.g., Pokémon GO), computing resources are of the foremost concern. As a result, the demand for different resources will impact the RAT selection.

Some recent works provide ideas on how to model tasks of different services. Mao *et al.* model a computing task using three parameters: offloaded task size (in bits), computation intensity (in CPU cycles per bit), and completion deadline [7]. A VR related task modeling with three-dimension resources is proposed in [8], in which cached contents are used as inputs of the computing stage. However, a general model to characterize the dependence between service requirements and multi-dimensional resources is not available yet. The complexity in developing such a service model comes from the variety of services and their diversified requirements.

ISSN: 2233-7857 IJFGCN

2) Resource Slicing

Different RATs can use different resource allocation schemes and yield different resource utilization. Therefore, resource allocation and RAT selection are mutually dependent, and joint network resource allocation and user association should be considered. Existing studies on joint computing resource allocation and user association joint caching resource allocation and user association or joint allocation of the communication, computing and caching resources and user association, do not consider network slicing. On the other hand, current works on network slicing consider focus on one type of RAT and/or only the communication resource which limits their applications in NGWNs with multiple RATs and multiple resources.

After RAN slicing in the planning stage, multiple slices are established. Within each network slice, RAT selection adjustments may be required in the scheduling stage due to user mobility, network load distribution dynamics, scheduled power-off of APs and so on. For such adjustments of RAT selection within a slice, it may be possible to extend some existing works on user association without considering network slicing, to develop an RAT selection adjustment solution.

3) User Mobility

Mobility is an essential issue in the RAT selection. A properly designed association algorithm should avoid unnecessary handoffs, since a re-association procedure incurs extra signaling and excessive execution latency. To avoid unnecessary handoffs, RAT selection can be based on predicted user mobility. Many state-of-the-art prediction algorithms have been proposed to estimate user trajectory, cell dwelling time, and other mobility-related information, using data-based and model-based mobility prediction methods. With the prediction of user mobility, a proactive network resource adjustment can be designed to achieve a timely and smooth handoff. For instance, the networks can adaptively or proactively adjust their resource allocation, using a mobility-aware computing strategy and/or caching strategy . However, these mobility-aware resource allocation strategies assume that the user association policy is known and fixed, which is inappropriate in the NGWNs. In order to efficiently utilize network resources, user association should be considered jointly with mobility-ware network resource allocation. For example, a joint user association and content placement in an edge caching scenario should account for user mobility . To jointly consider both mobility and resource allocation, the RAT selection problem is much more complex than the conventional one.

Moreover, due to user mobility, a communication or computing task may not be completed while a user is temporally connected to an AP. As a result, a task handover from the original AP is necessary. For example, when a user is moving out of the coverage of an AP and thus cannot finish downloading a content, it should connect to another AP that caches the same content if possible, to continue the downloading task. Similarly, in order to preserve service continuity in a computation task, the original task can be decomposed into several subtasks. Each subtask is offloaded to an AP with computing capacity, so that it can be finished before the user moves out of the coverage of its current AP. Therefore, the current user task completion status should be incorporated in the RAT selection in the scheduling stage.

4) Multi-Connectivity

In addition to the multi-mode capacity which allows only one RAT connection at any time, multi-connectivity/multi-homing terminals have the ability to support multiple RAT connections simultaneously Using concurrent connections for a single service has the benefit of improving service reliability . The multi-homing related RAT selection has been investigated from different perspectives. From the perspective of networks, the service operator aims to optimally allocate downlink (DL) bandwidth among multiple radio connections to support users with different services in a multi-RAT environment . In contrast, from the perspective of a single user, the goal is to enhance QoS by optimally distributing packets among multiple radio interfaces during an uplink (UL) transmission. In general, an RAT selection problem for multi-homing terminals is much more

complicated than a problem of multi-mode, since we need to determine how many connections to establish and which set of the available radio networks to connect for the user. Preliminary work on multi-homing connection does not consider network slicing and focuses only on communication resources .

D. AI-Based RAT Selection

Here, we first review optimization-based solutions, and then the learning-based solutions, followed by a discussion of the challenges in applying AI to RAT selection. Table 3 summarizes a few related works on user association in heterogeneous networks. Some works adopt optimization techniques, while others use learning-based approaches.

TABLE 3 Summary of Literature on User Association in Heterogeneous Networks

User Mode	Work	Control Paradigm	Objective	Scenario	Link	User Mobility	Resource
Multi-Mode	[21]	Centralized	Fairness	Two-tier network	DL	No	Bandwidth
	[60]	Centralized	Average drone-to-user pathloss minimization	UAV-assisted RAN	DL	No	Time slots
	[61]	Centralized	Load balancing	Vehicular network	DL	Yes	Scheduled time fraction
	[62]	Distributed	Energy efficiency maximization	MEC	UL	Yes	CPU-cycle frequency
	[63]	Distributed	Profit related utility maximization	Information-centric network	DL	No	Virtual resources & caching resources
	[64]	Centralized	Computational overhead minimization	MEC	UL& DL	No	Communication, computing & caching resources
	[65]	Centralized	Average service latency minimization	Fog-computing IoT network	UL& DL	No	Communication, computing & caching resources
	[66]	Centralized	Profits maximization	CRAN	DL	No	Cache & resource blocks
	[67]	Centralized	Load balancing	Cellular network	DL	No	Transmission power
	[68]	Centralized	Load balancing & energy efficiency maximization	mmWave communications	DL	No	Transmission power
	[69]	Centralized	Service capacity maximization	mmWave communications	DL	No	Communication, computing & caching resources
Multi-Connectivity	[70]	Centralized	Service capacity maximization	mmWave communications	DL	No	RF chains
	[71]	Centralized	Energy efficiency maximization	Multi-tier network	DL	No	Transmission power
	[72]	Centralized	Power consumption minimization	MIMO cellular network	DL	No	Transmission power

Optimization-based approaches for solving user association problems can be categorized into two classes: deterministic optimization based approaches and stochastic optimization based approaches. Both classes have the following limitations. Firstly, user association related problems, in general, are formulated as combinatorial optimization problems, which are non-convex and usually NP-hard. Therefore, applying optimization-based methods can result in significant computation latency and overhead. Even if the optimal solution can be found, the cost of finding the solution can be prohibitive as the network size or the set of service types grows due to the exponentially increasing complexity. Secondly, optimization-based approaches rely on the prior knowledge of the network (e.g., network topology, user density, mobility, channel statistics, and service requirements) and/or the assumptions made for mathematical tractability (e.g., Poisson arrivals, exponential service time, uniform user distribution, and so on). When network dynamics vary, established theoretical models may no longer be applicable and the performance of a previously obtained association solution can degrade significantly.

Different from optimization-based approaches, model-free RL provides an alternative approach for finding the optimal solution of a problem through “trial and error” in the interactions with the networking environment. According to the type of the learning agent, RL-based RAT selection can be classified into two classes. The first class chooses individual users as the learning agents. In, a distributed Q-learning-based handoff is proposed to optimize the long-term discounted rewards of users. RL can also be combined with a traditional RAT selection algorithm for performance improvement. In, RL is adopted by users to learn the optimal cell range extension bias with the global objective of minimizing the total number of devices in outage. In the scenario of edge computing, RL

can help with user association decisions to improve computing energy efficiency . The second class of RL-related works assume the APs as the learning agents. For example, the BS can be the learning agent to achieve load balancing through user association in a vehicular network .

There are technical challenges in developing ML-based approaches for solving the RAT selection problem. The underlying MDP model may not accurately capture the RAT automation problem. It is possible that only partial information is available, or there exist observation errors. In such cases, a generalized partially observable MDP (POMDP) model can be adopted. Also, deriving models and metrics to characterize the performance, or even a performance bound, of the learning algorithm is not an easy task. Most learning algorithms are evaluated only numerically. Sun *et al*. present proof on the performance bounds for their proposed learning algorithm in [6]. However, a unified framework on the convergence and performance analysis of RL is yet to be developed.

v. Mobile Edge Caching and Content Delivery

As mentioned previously, resources in NGWNs will extend beyond communication resources and include caching resources. Mobile edge caching leverages storage spaces at the network edge to cache popular contents within the RAN. As a result, mobile edge caching can help to reduce content retrieval time for users and alleviate backhaul congestion for the network [115]. As mobile edge caching is usually limited by the cache size, it is necessary to optimize caching strategies for maximal caching resource utilization. In this section, we present research challenges of mobile edge caching in conventional and network-slicing based wireless networks, respectively. The related research works are reviewed, and future research directions on AI-based mobile edge caching are discussed.

A. Research Challenges

There are two main research issues in mobile edge caching, i.e., content placement and content delivery. Content placement determines which contents to be coached at the edge, while content delivery determines how to deliver cached contents to users. The first major challenge in content placement roots from time-variant content popularity and/or an evolving content catalogue. If the content popularity could be accurately estimated, the problem of maximizing cache hit rate would be simple. However, it can be very difficult to predict the content popularity, especially when the popularity demonstrates spatial-temporal variations. The second challenge in content placement is due to the multi-tier cache system with overlapped spatial coverage in heterogeneous wireless networks. When content delivery is considered, the joint optimization of communication and caching strategy, which corresponds to a complex decision-making problem, becomes another major challenge.

NGWNs demonstrate heterogeneity in both resources and service types. In the network-slicing based architecture, the resources in RAN, including communication, caching, and computing resources, are orchestrated and sliced to support the corresponding virtual networks with QoS guarantee. An overview of caching-centric resource management in a network-slicing based architecture is illustrated in Fig. 8. After slicing the resources in the planning stage, the resources will be further scheduled in each slice to improve user service experience. In resource scheduling, in addition to content placement and content delivery, joint caching and communication resource management should be taken into account. Under this network-slicing based architecture, new challenges in content placement and content delivery are summarized as follows.

ISSN: 2233-7857 IJFGCN

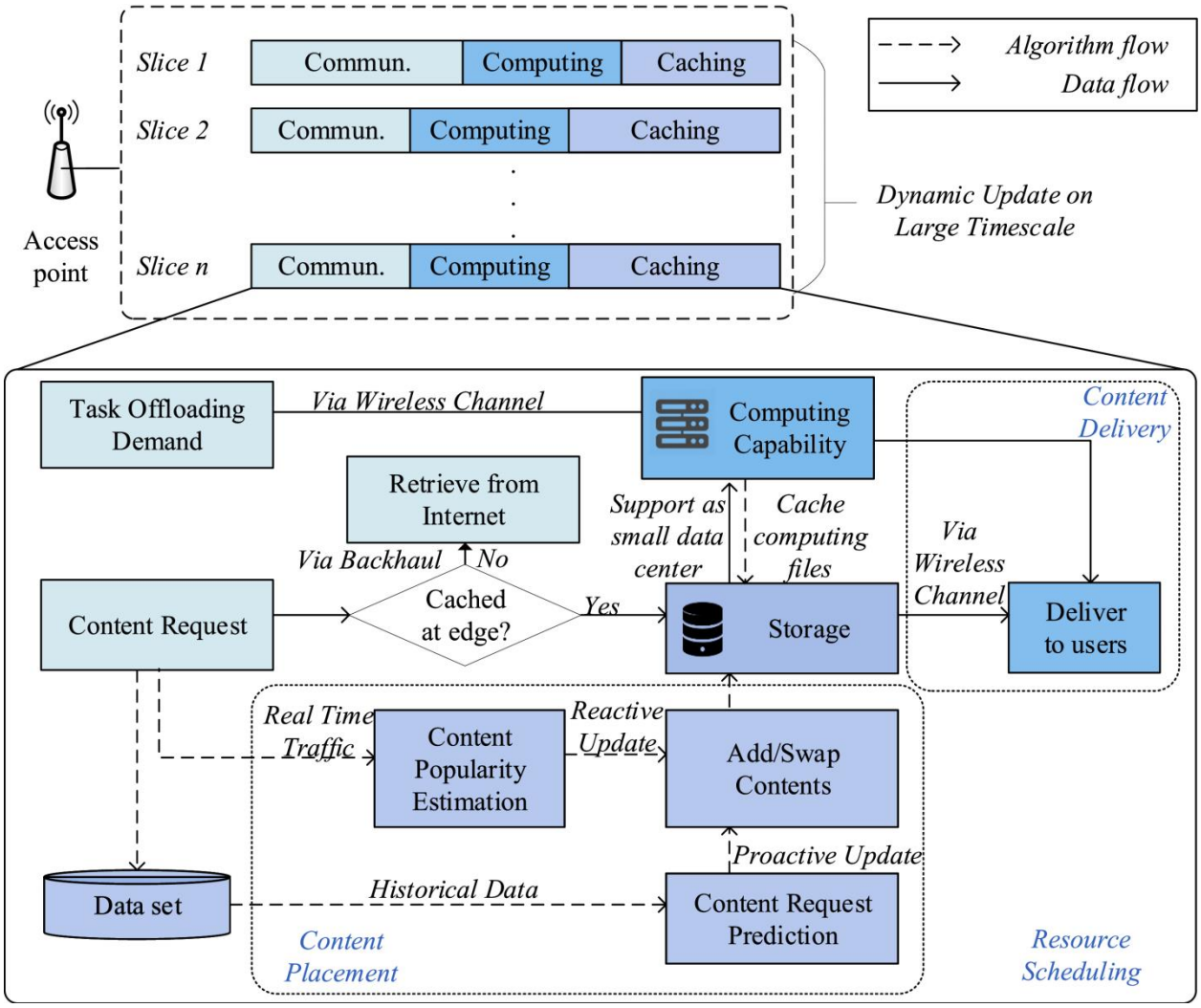

Figure 8. Overview of caching-centric resource management in a network-slicing based architecture.

- Heterogeneity among different slices: The physical cache for an edge entity can be sliced into several logical caches for serving different applications with diverse QoS requirements. Different from the conventional caching, content requests are accommodated into different slices in NGWNs. User access patterns and popular contents in different virtual networks can have distinct characteristics. For example, in IoT applications, users usually fetch contents from a static content catalogue periodically while, in mobile applications, users typically request contents from an evolving catalogue. Hence, designing a customized content placement policy to support diversified virtual networks is challenging and requires further investigation;

- Dynamic cache size: By resource virtualization, the cache size for a slice can be modified as a result of dynamic slicing on a large timescale. The cache placement policy should be updated dynamically to adapt to the variable cache size. When the cache size allocated to a slice is sufficient, contents with a large size can be cached to reduce the backhaul usage. Otherwise, popular contents with a small size should be cached to improve the cache hit rate. Thus, in addition to time-variant content popularity and evolving content catalogue, new uncertainty is introduced due to the variable cache size, which can make the content placement problem intractable;

- Multi-resource allocation: MEC is expected to be a basic element in NGWNs. The contents cached by an edge will not only include popular contents such as videos, but also include the

essential files for implementing computing functions. To improve the computing service performance, the allocation of computing, caching, and communication resources should be jointly optimized in terms of content delivery. However, the multi-resource allocation problem can be too complicated to solve in real-time using model-based approaches.

B. State-of-the-art Caching Solutions

Here, we review existing works on content placement and content delivery. For content placement, we summarize research efforts on content updating strategies at a single caching server. For content delivery, we focus on research works for joint caching and communication resource management. A summary of the literature is provided in Table 4.

TABLE 4 Summary of Literature on Caching Resource Management

Topic	Related Work	Contribution	Approach
Content Placement	[121]	Reduce content popularity estimation interval by utilizing knowledge obtained from user's interactions with a social community.	Transfer learning
	[122]	Increase cache hit rate by cooperative content popularity estimation among multiple servers.	Transfer learning
	[123]	Prefetch contents according to the mobility of user and demonstrate performance improvement on content prefetching.	Markov chain
	[124]	Track and predict time-variant content requests from users.	Neural network
	[40]	Reduce cost of downloading contents from the Internet by adding or swapping contents according their lifetime.	Reinforcement learning
Content Delivery	[125]	Analyze the coupled relation between content placement and communication resource allocation to reduce backhaul traffic.	Heuristic
	[126]	Maximize the quantity of contents delivered by a cache-enabled UAV.	Optimization
	[63]	Allocate caching and communication resources to maximize the revenue by serving end users.	Distributed optimization
	[64]	Minimize overall content access time and energy overhead by selecting the optimal spectrum for content delivery.	Hopfield neural network
	[65]	Minimize content delivery and computing latency by jointly optimize caching, computing, and communication resource.	Reinforcement learning
	[127]	Determine caching and computing offloading decisions to reduce operational cost on edge server, where the interference from other edge servers is evaluated.	Deep Q learning

1) Content Placement

Content popularity is time-varying in general and, hence, the cached contents stored at a server need to be updated dynamically. The main goal of content placement is to maximize the cache hit probability. As illustrated in Fig. 8, there are two types of caching policies to update the contents in a cache, namely the reactive caching policy and the proactive caching policy.

In a reactive caching policy, the edge node determines whether or not to cache a content after a request for that content arrives. A common assumption is that the content popularity follows a stochastic distribution, such as Zipf distribution . However, the popularity of contents varies over time, and different types of contents can exhibit a variety of popularity evolution patterns. To adapt to non-stationary traffic and content popularity, content updating policies have been proposed. For example, the least recently used (LRU) policy replaces the least recently requested content in the cache when the cache is full. To improve the cache hit rate, the content popularity can be estimated according to the content request during a period of time. However, in practice, the reactive caching policies adapt slowly to changes in content popularity .

Proactive caching policies aim to prefetch popular contents that are likely to be requested by users ahead of time. Therefore, proactive caching can mitigate backhaul usage if prefetching is scheduled during off-peak hours. Proactive caching is illustrated in the bottom part of Fig. 8, where historical content requests used for predicting popular contents are added into a data set as records. Future content requests can be predicted by exploiting the spatio-temporal association among the records in

the data set, such that the edge server can proactively cache contents for improving the cache hit rate . Another category of solutions does not directly predict requests, but formulates an MDP problem to find an optimal content placement policy which maximizes the cache hit rate in the long term . Since the content placement problem has large state-action space and unknown state transition probabilities caused by a dynamic network environment, it is difficult to solve the MDP problem by the conventional dynamic programming method. RL can be utilized to solve the MDP problem according to the reward feedback from the network environment. However, RL has some limitations in solving the content placement problem. The first limitation is the Markov property of the underlying MDP, which is assumed when RL is applied. As a result, it is difficult for RL to explore the temporal correlation in a sequence of historical user requests. Second, most works using RL for optimizing caching strategies assume that the catalogue of contents is known in advance, which can be unrealistic for the scenario in which the content catalogue changes dynamically. Therefore, in the case when new contents dynamically emerge, the RL based approaches in most existing studies, such as , cannot be applied since they cannot predict the popularity of new contents. One potential solution to handle a changing content catalogue is to add or remove contents based on their lifetime, as proposed in, so that caching decisions can be made for an evolving catalogue of contents. However, such an approach yields another challenge, i.e., estimating the lifetime of contents. In addition, existing works predict content requests according to all content requests received at the server, without considering the type of services or applications. However, such granularity is not fine enough in a network-slicing based network architecture as different slices may have different traffic and content popularity patterns.

In summary, both reactive and proactive content placement approaches aim to improve the cache hit rate. A reactive caching policy can handle a varying content catalogue or evolving content popularity via an online content update, while a proactive caching policy exploits historical user requests and predicts the content popularity offline. Since the pattern of the evolving content popularity can become more evident after the network resources are sliced based on service types, proactive caching can be a potential approach to find a customized content placement policy for virtual networks. In addition to the spatio-temporal features of user requests, other features can be excavated to further improve the performance of service-specific content request prediction, such as the QoS requirements and application types for the corresponding virtual network. In addition, existing works generally assume that all contents have an identical size, which is not practical. The content size can substantially affect caching performance due to the dynamic slicing of the physical cache. Therefore, the trade-off between the backhaul usage decrease and cache hit rate improvement needs to be studied in the context of dynamic slicing of the physical cache.

2) Content Delivery

The main goal of content delivery is to reduce content transmission time. In order to achieve this objective, except caching popular contents in the edge servers, the average communication delay between users and the edge server should be minimized. A trade-off between the transmission delay and caching service coverage is discussed in, where a cache-enabled UAV is deployed as an edge server. When the UAV is deployed at a high altitude, it can cover a large number of users and reduce content delivery time for these served users. However, the data rate of content delivery from the UAV to the users can be low, due to the high altitude of the UAV. By contrast, when the UAV is deployed at a low altitude, the data rate can be improved, but fewer users can benefit from the cached contents due to the reduced coverage of the UAV.

In a more general scenario, e.g., when there are multiple edge servers connected with each other, a dependency relation between cache and communication resources emerges across the servers. As shown in Fig. 9, in addition to determining which content to be cached, the cooperation among edge servers and the network topology should be considered in the content placement and delivery problem. The edge servers can cooperate with each other in content placement and delivery in order to improve the overall cache hit rate and, as a result, reduce the backhaul congestion. A user can access contents from both its own server and, via the relay of its server, other edge servers (shown as the

ISSN: 2233-7857 IJFGCN

virtual link in Fig. 9). The joint problem of content placement and routing among the servers can be formulated as a mixed-integer problem, which generally has high complexity. For example, a Hopfield neural network (HNN) framework can be explored to solve the problem of routing without considering resource allocation in cooperative caching. However, solving the joint resource allocation and routing problem in a scalable manner for cooperative caching with multiple edge servers remains an open research problem. As the network topology becomes more complex in NGWNs, the association between users and edge devices should be considered, while deciding content placement, in order to minimize the content delivery delay. In this case, a user can connect with multiple edge servers that cache popular contents. The trade-off between caching diversity and spectrum efficiency is investigated in, while cooperative caching and transmission is considered in the context of coordinated multi-point transmission. In addition, with network topology dynamics, the optimal content placement and delivery decision can vary significantly. The network connectivity graph is analyzed in for allocating content to multiple edge servers given the network topology and the content popularity. When the number of users increases, the optimal content placement policy becomes intractable.

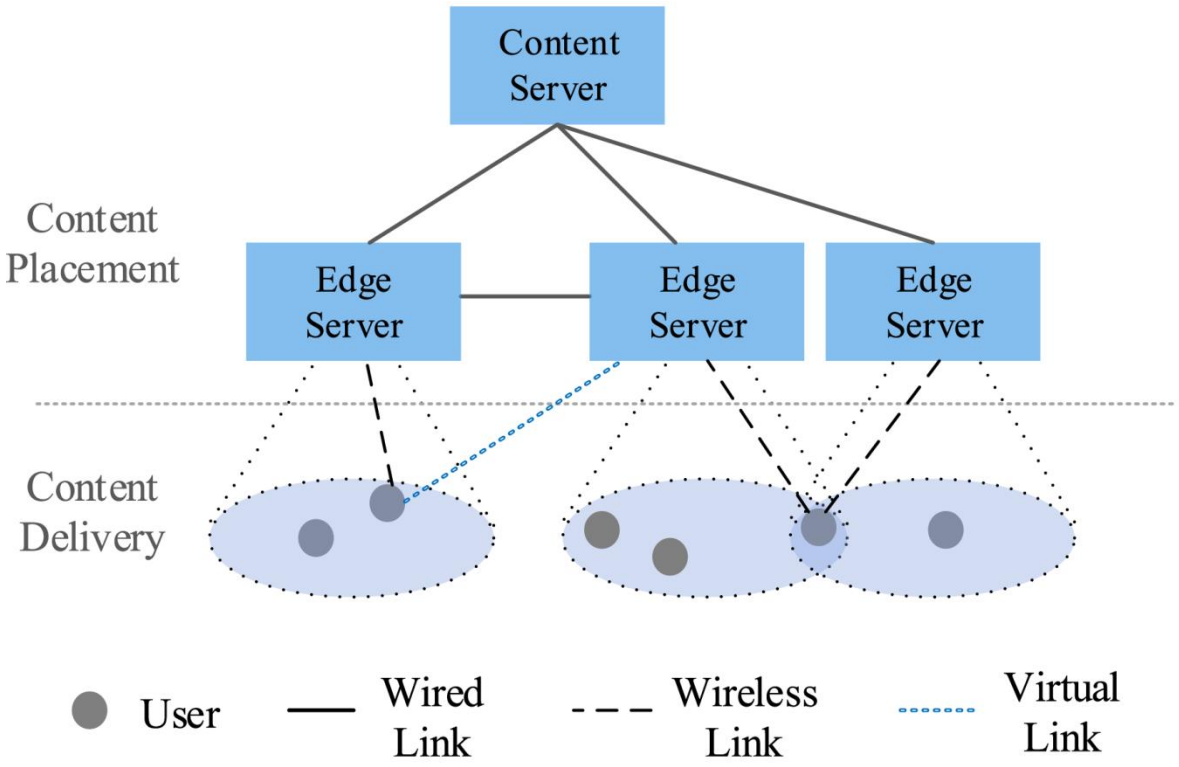

Figure 9. An illustration of the relation between caching placement and content delivery

With the emergence of MEC in 5G networks, contents stored in the cache include the files and data for implementing various computing functions. To satisfy QoS requirements of MEC, the resources (including communication, caching, and computing resources) should be jointly optimized. As mentioned, RL is known for solving complex decision-making problems and has been adopted in existing studies to jointly allocate caching, computing, and communication resources. While the resulting caching strategies obtained using RL can achieve a near-optimal performance, they cannot handle an evolving content catalogue in general. Moreover, existing works assume that computing and caching resources can be allocated independently, while computing tasks can only be executed when the corresponding files and data are stored at the edge. Such dependency can further complicate the content placement and computing offloading decision.

ISSN: 2233-7857 IJFGCN

In summary, the essence of content delivery is a multi-dimensional resource management problem. In a conventional caching scenario, communication and caching resources are jointly optimized to balance the content delivery time and the cache hit rate. In the network-slicing based architecture, computing resources at the edge are utilized to perform latency-critical tasks in virtual networks. Thus, the concept of content delivery is extended to computing offloading and execution. The multi-resource allocation problem incurs high complexity in problem solving, while the conventional optimization techniques are hard to make real-time caching and offloading decisions. Moreover, the dependency among caching, computation, and communication resources needs to be further investigated. The caching performance is not only restricted by sliced caching resources (i.e., the size of logical caches), but also constrained by sliced computing and communication resources. Last, users with high mobility can fail to download the content from or offload their tasks to the edge due to intermittent connections. Thus, user mobility prediction should be incorporated to improve the performance of content delivery.

C. Future Research Directions in AI-Based Caching

Given the aforementioned challenges, next, we discuss the potential applications of ML for caching from three aspects: content popularity prediction in proactive content placement, dynamic content placement policy adjustment, and multi-resource allocation in content delivery.

For content popularity prediction, ML based approaches, e.g., DNN, can extract features from recorded content request data to facilitate the content popularity prediction. In the network-slicing based architecture, a local SDN controller can be deployed to monitor content requests and associate the requests with user IDs, request time instants, and locations. Given a sufficiently large data set of request records collected by the controller, DNN can utilize the records for predicting the content requests in future. Compared with conventional statistical methods, such as linear regression or Kalman filter, DNN has the advantage of exploiting a large data set to make more accurate predictions. However, the performance of content popularity prediction can be degraded by many factors, such as an evolving content catalogue or time-variant content popularity. As a variant of DNN, recurrent neural network (RNN) has been widely adopted for prediction from historical data due to its ability to track time-variant patterns. Compared to the conventional neural networks, RNN applies internal memory to capture temporal correlations in the input data. Therefore, RNN has been adopted to track time-variant content popularity in the literature. In addition to the temporal correlation in the content requests, ML based approaches can be used to capture the spatial correlation in the requests. Convolution neural network (CNN) can be a potential tool for capturing such spatial correlations. Despite the various advantages, several issues need to be addressed while developing ML based approaches for caching in future communication networks. Firstly, while CNN and RNN have the potential to predict future content requests with a high accuracy based on spatio-temporal features, deploying CNN/RNN based prediction modules for content placement can lead to a high computation load. This, in turn, requires a characterization of the improvement in caching performance versus the resulting computation load. As a result, a trade-off between computation and caching performance needs to be investigated. Secondly, in the network-slicing based architecture, content requests are distributed into different virtual networks. Consequently, deploying one neural network as an open module that all slices can use is preferred in terms of complexity, but may not adapt to the specific characteristics of individual slices. By contrast, deploying one neural network for each slice allows a customized prediction module for each service or application, but can lead to prohibitive complexity.

For caching policy adjustment, RL is a potential approach to update the content placement policy in a dynamic environment with flexible cache size for each slice and time-variant content request pattern. The instantaneous cache size, cached contents, requested content, etc., can be modeled as states, and the content update can be modeled as actions. The resulting MDP model with unknown state transition probabilities can explore RL techniques to find an efficient content updating policy . However, the assumption of an underlying MDP model, and the associated Markov property in state

transitions, can be impractical. The RL based approach may not fully capture the correlations of content requests over the time domain. In addition, as mentioned in Subsection V.B, the ML algorithm should be able to handle an evolving content catalogue while updating the caching policy. Devising such an ML algorithm without incurring significant complexity, e.g., having to deploy an additional module for predicting the lifetime of all contents, remains an open and challenging problem.

For content delivery, deep RL has a potential to provide a tractable approach to coordinate and allocate multiple types of resources, including communication, computation, and caching. While the state-action space of multi-dimensional decision making in joint caching, computing, and communication resource allocation can be too large for conventional RL, deep RL adopts deep learning techniques to estimate policy and value function, and thus can handle the large state-action space from the joint allocation of multiple resources for content delivery. However, classic deep RL has limitations when dealing with constrained decision-making problems, in which the dependency among the resources exists and introduces constraints in resource allocation. For example, a user with a computing task to offload prefers an edge server that caches the data and files for this computing task. In such a case, the content placement at the edge server yields a constraint on the task offloading decision of the user. Constrained MDP can be a possible model for incorporating the constraints, while how to develop a deep RL based solution for the constrained MDP problem needs further investigation.

VI. Conclusion

In this paper, we have illustrated the network-slicing based architecture, focusing particularly on the RAN, and elaborated how AI can potentially empower this architecture for NGWNs. Through the investigation of three research problems, i.e., RAN slicing, automated RAT selection/user association, and content placement and delivery, we have demonstrated new challenges, as a result of the heterogeneity, dynamic environment, and/or strict and diversified service requirements, in network management and resource orchestration under the network-slicing based architecture. Most of these challenges cannot be addressed by directly extending existing research. Therefore, it is necessary to develop novel models, technique tools, and/or problem-solving approaches. Summarizing related research efforts, we have demonstrated the potential approaches and benefits in the application of AI for solving the three problems. Meanwhile, we have also noted the challenges of applying AI-based approaches, e.g., handling non-stationary network environment. Through the three considered problems, this paper takes an initial step towards understanding the development of models and algorithms for intelligent network management and resource orchestration in network-slicing based NGWNs.

References:

1. Z. Zhang et al., "6G wireless networks: Vision requirements architecture and key technologies", *IEEE Veh. Technol. Mag.*, vol. 14, pp. 28-41, Sep 2019.
2. L. Hobert, A. Festag, I. Llatser, L. Altomare, F. Visintainer and A. Kovacs, "Enhancements of V2X communication in support of cooperative autonomous driving", *IEEE Commun. Mag.*, vol. 53, no. 12, pp. 64-70, Dec 2015.
3. S. Jeong, W. Na, J. Kim and S. Cho, "Internet of Things for smart manufacturing system: Trust issues in resource allocation", *IEEE Internet Things J.*, vol. 5, no. 6, pp. 4418-4427, Dec 2018.
4. H. Beyranvand, M. Lévesque, M. Maier, J. A. Salehi, C. Verikoukis and D. Tipper, "Toward 5G: FiWi enhanced LTE-A HetNets with reliable low-latency fiber backhaul sharing and WiFi offloading", *IEEE/ACM Trans. Netw.*, vol. 25, no. 2, pp. 690-707, Apr 2017.
5. X. Zhu, C. Jiang, L. Yin, L. Kuang, N. Ge and J. Lu, "Cooperative multigroup multicast transmission in integrated terrestrial-satellite networks", *IEEE J. Sel. Areas Commun.*, vol. 36, no. 5, pp. 981-992, May 2018.
6. M. Peng, S. Yan, K. Zhang and C. Wang, "Fog-computing-based radio access networks: Issues and challenges", *IEEE Netw.*, vol. 30, no. 4, pp. 46-53, Jul 2016.
7. Q. Li, L. Zhao, J. Gao, H. Liang, L. Zhao and X. Tang, "SMDP-based coordinated virtual machine allocations in cloud-fog computing systems", *IEEE Internet Things J.*, vol. 5, no. 3, pp. 1977-1988, Jun 2018.

ISSN: 2233-7857 IJFGCN