Network Slicing in Fog Radio Access Networks: Issues and Challenges

Hongyu Xiang, Wenan Zhou, Mahmoud Daneshmand, and Mugen Peng

The authors present a novel architecture and related key techniques for access slicing in F-RANs. The proposed hierarchical architecture of access slicing consists of a centralized orchestration layer and a slice instance layer, which makes access slicing adaptively implementable in a convenient way.

ABSTRACT

Network slicing has been advocated by both academia and industry as a cost-efficient way to enable operators to provide networks on an as-a-service basis and meet the wide range of use cases that the fifth generation wireless network will serve. The existing works on network slicing are mainly targeted at the partition of the core network, and the prospect of network slicing in radio access networks should be jointly exploited. To solve this challenge, enhanced network slicing in F-RANs, called access slicing, is proposed. This article comprehensively presents a novel architecture and related key techniques for access slicing in F-RANs. The proposed hierarchical architecture of access slicing consists of a centralized orchestration layer and a slice instance layer, which makes access slicing adaptively implementable in a convenient way. Meanwhile, key techniques and their corresponding solutions, including radio and cache resource management as well as social-aware slicing, are presented. Open issues in terms of standardization developments and field trials are identified.

INTRODUCTION

Driven by the emerging applications of mobile Internet and the Internet of Things (IoTs), the fifth generation (5G) wireless communication systems are expected to satisfy diverse use cases and business models [1]. Unfortunately, the legacy cellular network architectures were originally designed for mobile broadband consumers, without considering the characteristics of emerging massive machine-type communications (mMTC) and ultra-reliable MTC (uMTC) [2]. To meet the diverse use cases and business models in 5G, network slicing has recently been proposed to flexibly provide software-defined networking in a cost-efficient way.

In the concept of network slicing [3], the network entities are sliced into multiple isolated network slice instances with appropriate network functions and radio access technologies. Network slicing has attracted a lot of interest from both academia and industry. In [4], the process of network slicing as a service within operators is illustrated by typical examples, and service orchestration and service level agreement mapping for quality assurance are introduced to illustrate the architecture of service management across different levels of service models. Network slice instances for specific purposes are being studied as well; for example, software-defined wireless networking enabled WLAN slices to provide fine-grained spectrum are introduced in [5], and operator slices to act as mobile virtual network operators are discussed in [6]. Meanwhile, core network (CN)-based network slicing is presented in [7], where two example network slice instances are illustrated to explain the impact of use case requirements on network slice design. In the 5G technical report given by the Third Generation Partnership Project (3GPP), support for network slicing appears as one of the key requirements [8].

Despite the evident attractive advantages, traditional CN-based network slicing also comes with its own challenges. First, the conventional network slicing solution is mainly business-driven and only addresses the needs of use cases, which does not highlight the characteristics of radio access networks (RANs) on network slicing creation. The requested network slicing may not be effective when there is a shortage of the radio resource in RANs. As a result, network slicing should consider the radio transmission impacts, and corresponding network slicing jointly considering the status of RANs should be highlighted. Second, most of the existing work on network slicing is purely based on CNs, while network slicing as an end-to-end solution should cover the specific characteristics of RANs. Especially in some cases like uMTC demanding ultra low latency, CN-based slicing with a general RAN is hard to satisfy with the requirements of 5G performance.

The fog RAN (F-RAN) has emerged as a promising 5G RAN that can satisfy diverse quality of service (QoS) requirements like high spectral efficiency, high energy efficiency, low latency, and high reliability for different service types in 5G [9]. Traditional CN-based network slicing cannot take full advantage of the edge characteristics of F-RANs to efficiently support mMTC and uMTC service types.

To cope with the aforementioned challenges for CN-based network slicing to meet, a technological framework of a novel type of network slicing called access slicing specified for F-RANs is proposed in this article. The proposed access slicing is compatible with the existing CN-based network slicing solution [8], and it has a significant difference from both CN-based network slicing

Digital Object Identifier:
10.1109/MCOM.2017.1700523

Hongyu Xiang, Wenan Zhou, and Mugen Peng (corresponding author) are with Beijing University of Posts and Telecommunications; Mahmoud Daneshmand is with the Stevens Institute of Technology.

0163-6804/17/$25.00 © 2017 IEEE

and RAN-based access slicing [10]. The proposed access slicing can be regarded as combining the advantages of CN-based and RAN-based network slicing in 5G, and it can take full advantage of F-RANs. Further, the advanced access slicing is information-aware to guarantee diverse QoS requirements, and quality of experience (QoE) requirements as functions of QoS requirements.

In this article, motivated to offer comprehensive research on access slicing in F-RANs, the corresponding architecture is first presented, in which a new management entity (e.g., access slice orchestration) is defined to fulfill the legacy function of CN-based network slicing. This new entity helps to orchestrate proper network functions for different access slices and achieves the coexistence of the multiple access slices. Meanwhile, resource management between different slices is exploited, in which the cache and radio resources are jointly designed. Furthermore, the social-aware slicing technique is presented as well, which assists to create and adjust access slice instances. In addition, the open issues and future research works are presented.

The remainder of this article is organized as follows. The new architecture for access slicing in F-RANs is introduced in the following section. Then key techniques, including resource management and social-aware slicing, are presented. Open issues are discussed, followed by conclusions in the final section.

Hierarchical Architecture for Access Slicing

The hierarchical architecture of access slicing in F-RANs is capable of providing numerous services with the desired QoS, including the typical enhanced mobile broadband (eMBB), uMTC, and mMTC defined in 5G. To fulfill the orchestration, control, and data functions in networks, two layers are defined in the proposed access slicing architecture as shown in Fig.1: a centralized orchestration layer and a slice instance layer. In the centralized orchestration layer, the access slice orchestration is used to handle the centralized network orchestration for dynamic provisioning of the slices and manage the resource between the implemented access slice instances. The slice instance layer consists of various access slice instances that provide the requested services. With the guarantee of slice isolation, each access slice instance can operate as a logically separate network and has specified control/data planes. Based on big data analysis and the identified results on software-defined access slice orchestration, access slicing is implemented to fulfill the edge computing and virtualization capabilities.

Key Components

The proposed access slicing architecture in F-RANs, as illustrated in Fig. 1, comprises two kinds of key components: access slice orchestration and access slice instance. By analyzing the information collected from mobile applications, user equipments, base station, and service platform, the access slice orchestration identifies the service type. Taking the profiles of existing slice instances into account, the access slice orchestration determines whether the needed access slice instances should be created. During the slices' runtime, the access slice orchestration performs centralized resource allocation and inter-slice resource management. Once the access slice instance is in crisis or any accident event happens, adjustment of slice instances can be triggered immediately with the help of social-aware mechanisms.

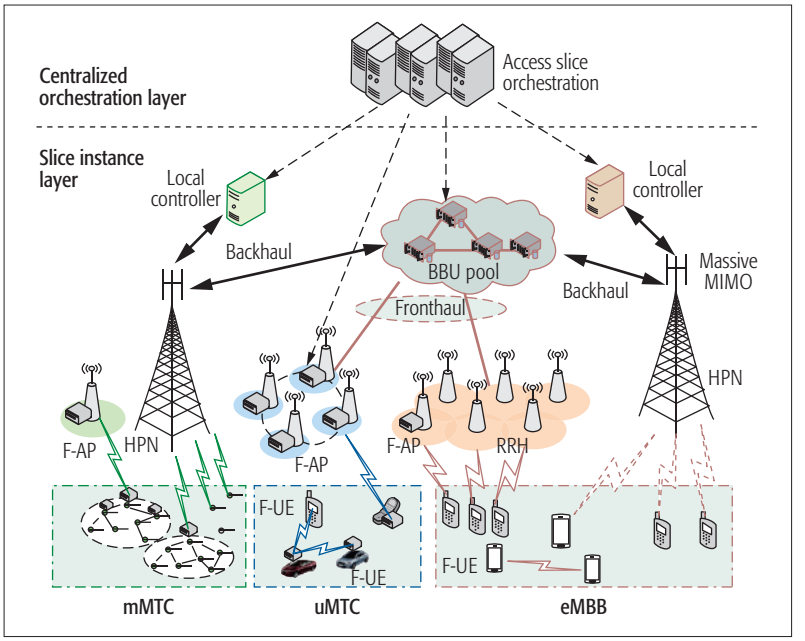

Figure 1. An illustration of access slicing in F-RANs.

Besides access slice orchestration, the access slicing architecture includes various instances, such as mMTC, uMTC, and eMBB, which is described in Fig. 1. Note that, based on the illustrated eMBB, uMTC, and eMBB instances, different service types can be handled as a combination of the three service types. Taking full advantage of the convergence of fog computing, cloud computing, and heterogeneous networking, F-RANs [11] are able to support different kinds of access slices. With a substantial amount of storage, communication, control, configuration, measurement, and management placed at the fog access points (F-APs) and fog user equipments (F-UEs), the traditional cloud computing paradigm is extended to the network edge. Hence, both the centralized cloud computing in the baseband unit (BBU) pool and the scalable fog computing composed of F-APs and F-UEs can provide collaboration radio signal processing (CRSP) and cooperative radio resource management (CRRM). CRSP and CRRM in the BBU pool are centralized and more effective, while CRSP and CRRM in the F-APs and F-UEs are closer to the edge and real-time, which makes F-RANs adaptive to dynamic traffic and the time-varying radio environment. Meanwhile, advanced 5G techniques, such as massive multiple-input multiple-output (MIMO) and non-orthogonal multiple access, can be applied directly in access slices.

eMBB Instance: To support the eMBB service with goals of high data rates and ubiquitous connection, the decomposition of the control and data planes is applied in the eMBB instance. In this instance, high-power nodes (HPNs) configured

The hierarchical architecture of access slicing in F-RANs makes access slicing adaptively implementable in a convenient way. In the centralized orchestration layer, the multi-dimensional information acquired from devices and nodes is collected, like network, device, application, and others, which may influence the provided QoS and QoE.

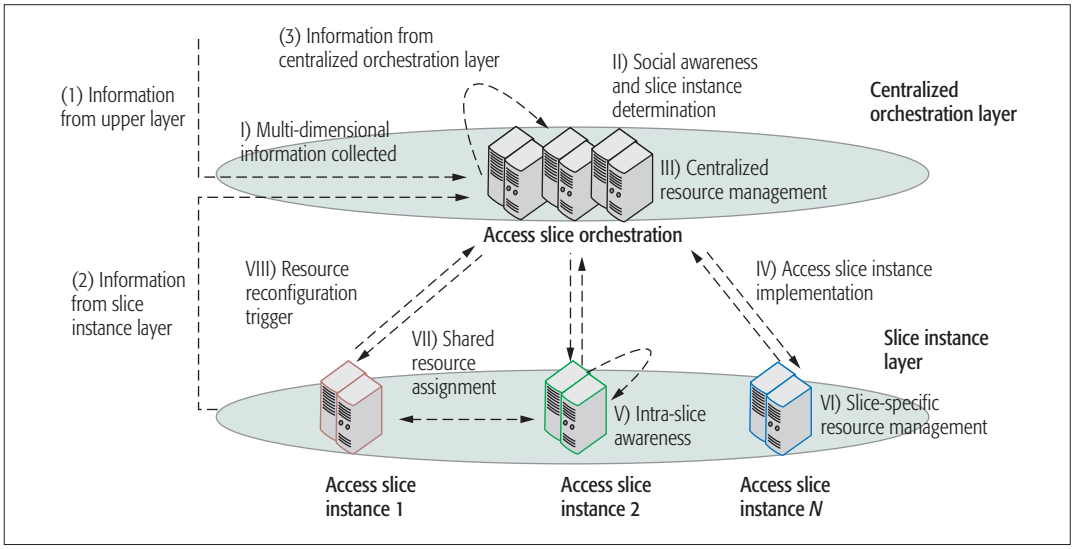

Figure 2. An illustration of the hierarchical architecture.

with completed protocol stack functions are mainly used to deliver control signaling and system broadcasting information for accessed UEs with a basic transmit data rate. Massive multiple-input multiple-output (MIMO) is configured in HPNs to provide additional diversity and multiplexing gains. In special zones like hotspots, the dense remote radio heads (RRHs) with only radio functions are randomly distributed to provide high capacity and connected to the BBU pool for achieving the large-scale centralized CRSP gains. The BBU pool, with physical layer, medium access control layer, and network layer functions, is interfaced with HPNs via backhaul, which indicates that there is coordination between the BBU pool and the HPN over the backhaul links. To mitigate the inter-tier interference between F-APs/RRHs and HPNs, cooperative processing and scheduling is implemented with the aid of the interface between the BBU pool and the HPN. Through the centralized large-scale CoMP approach, the inter-tier interference can be coordinated conveniently.

uMTC Instance: To meet the stringent latency demand of uMTC service, both the processing delay and transmit delay should be carefully constrained. To decrease the transmit latency, F-APs with cached content are deployed close to the desired UEs. To minimize the processing delay in the air interface, the protocol design in the physical layer, medium access control layer, and network layer should be redesigned. A shortened transmission time interval in the physical layer is essential to enable reduced processing delay, as well as reduced access delay and hybrid automatic repeat request Acknowledgment/negative Acknowledgment round-trip time. To address the state transition delay issue, the UE radio resource control state should also be optimized. Moreover, the device-to-device communication technique can be used. Since the radio environment is time-varied and the traffic in this instance is delay-sensitive, it is essential to approach the minimal delay via CRRM.

mMTC Instance: The realization of an mMTC instance is specified with massive connection in a dedicated area, and the clustering mechanism can be considered. As shown in Fig. 1, the adjacent UEs are formed into a mesh or tree-like topology cluster, wherein the packet traffic generated in the cluster is delivered to the F-AP or HPN via the elected cluster head. To reduce the cost caused by massive devices, the functions of traditional air interface protocols need to be re-designed elaborately. Considering the small-volume data and delay-tolerant features, a smaller resource block and a larger transmission time interval in mMTC instances are adopted. The mobile management is simplified and the handover between APs is not supported, since most devices in mMTC scenarios are immobile. The paging function can be sinked and settled in the local controller of an HPN, which avoids frequent communications between RANs and CNs. To handle the dilemma between massive UEs and scarce resource, the social-aware technique can be used to automatically detect active UEs, which assists distributed CRRM in managing the resource flexibly for both intra-clusters and inter-clusters.

Hierarchical Architecture

As shown in Fig. 2, the hierarchical architecture of access slicing in F-RANs makes access slicing adaptively implementable in a convenient way. In the centralized orchestration layer, the multi-dimensional information acquired from devices and nodes is collected, like network, device, application, and others, which may influence the provided QoS and QoE. Note that the multi-dimensional information can be divided into:

- The requested services and subscription information from the upper layer (e.g., the subscriber repository)
- The characteristics of RANs comprising UE capabilities from the slice instance layer
- The configurations of access slice instances from the centralized orchestration layer

Via the proper machine learning and data mining algorithms, some suggestions or results can be obtained to determine the required access slice instance. It is noted that not only can the requirements (e.g., QoS and QoE) of all access slice instances be fulfilled, but also the performance of the entire F-RAN can be optimized. To optimize the overall F-RAN, the access slice orchestration

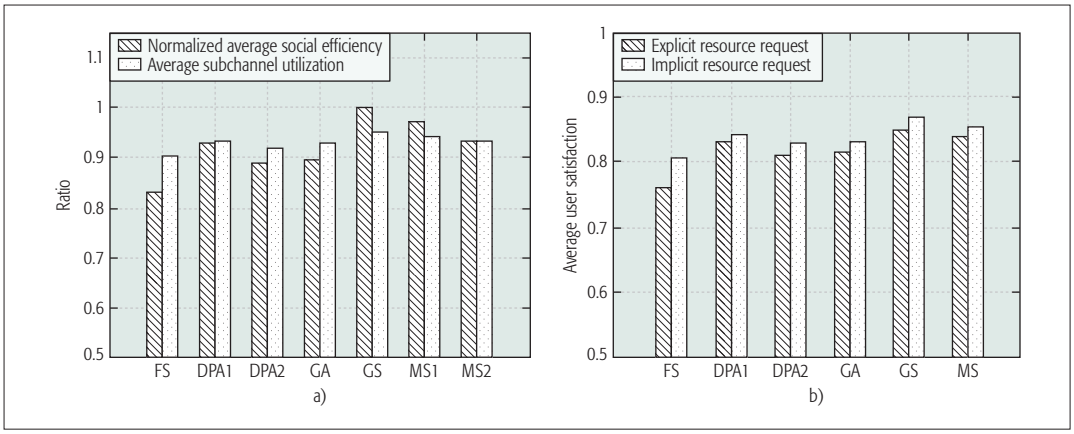

Figure 3. Performance comparisons among different algorithms applied in the proposed hierarchical auction model [12]: a) normalized average social efficiency and average subchannel utilization under different hierarchical auction mechanisms; b) average user satisfaction under different hierarchical auction mechanisms. FS: fixed sharing, DPA1: dynamic programming-based algorithm with group size 1, DPA2: dynamic programming-based algorithm with group size 5, GA: greedy algorithm, GS: general sharing, MS1: multiple seller exact solution, MS2: multiple seller approximate solution.

The proposed algorithm can achieve a relative higher average social welfare, average resource utilization, and average user satisfaction than the other baselines, including the fixed sharing scheme where the mobile virtual network operators are preassigned a fixed subset of resources.

performs the centralized resources allocation for all slice instances. Thanks to the social-aware technique, both the resource available in the whole F-RAN and the resource needed for each access slice instance can be fully acquired by the access slice orchestration. The access slice orchestration determines the proportion of resources for each slice instance. Note that the resource for each slice instance can be orthogonal or shared, which is determined by the orchestration for the aim to optimize the whole F-RAN.

In the slice instance layer, each access slice instance takes charge of the slice-specific operation and resource management based on the intra-slice awareness. Due to the introduction of cache featured F-APs and F-UEs, the resource management includes both the radio and caching resources. The caching resource management has a significant impact on the performance of different slice instances. As in the case of an eMBB instance, F-APs that have pre-cached popular content can offload part of the traffic in hotspots. On the other hand, the shortage of radio resource causes a significant constraint on the performance of slice instances. Hence, it is key to jointly optimize the radio and caching resources specified for each access slice instance. Among these different slice instances, some shared resources can be assigned. When the dedicated resource in a instance cannot achieve the desired QoS, these shared resources can be used under the condition that the incurring interference is not beyond a pre-defined threshold. If the shared resource has been totally used up, and the performance gap among different slice instances is too big, the resource reconfiguration of all slice instances will be triggered to provide a re-assigned resource allocation solution in the access slice orchestration.

Key Techniques for Access Slicing in F-RANs

The access slice instances run on the radio hardware and baseband resource pool, which exhibits more elasticity than the pure CN-based network slice solution. However, the guarantee of complete isolation between slices is challenging in access slicing. Herein, enhanced resource management is urgent to enable slice-specific operation for each isolated access slice. To realize the intelligent control and management of the proposed slicing architecture, the social-aware capability of the network is essential. By realizing the awareness of multi-dimensional information, the access slice instances can be flexibly and optimally designed.

Radio and Cache Resource Management

Efficient resource management is important to not only accommodate the dynamic demands of users in slices, but also satisfy the requirements of efficient resource allocation and isolation between slices. To enable intra-slice customization and inter-slice isolation, the potential mutual influence between slices should be taken into account when doing resource management. In [12], the resource allocation problem satisfying the requirements of efficient resource allocation, strict inter-slice isolation, and intra-slice customization is studied, in which a hierarchical combinatorial auction mechanism is derived. Based on the hierarchical auction model, a winner determination problem is formulated, and sub-efficient semi-distributed resource allocation is proposed. To evaluate the performance of the proposed algorithm, average social welfare (i.e., the sum value of all accepted bids), average resource utilization (i.e., the proportion of resources utilized), and average user satisfaction (i.e., the ratio of users whose resource requests are satisfied) are taken into account. As shown in Figs. 3a and 3b, the proposed algorithm can achieve relatively higher average social welfare, average resource utilization, and average user satisfaction than the other baselines, including the fixed sharing scheme where the mobile virtual network operators are preassigned a fixed subset of resources.

For access slicing in F-RANs, the joint resource optimization of radio and cache resources is often complex due to the tight relationship between radio and caching resources. F-APs with larger

As more and more communications occur between UEs with close relationships, the network is getting increasingly human-centric and social-aware. To exploit UEs' behaviors and interactions in the social domain, social awareness turns out to be indispensable information for the design and optimization of RANs.

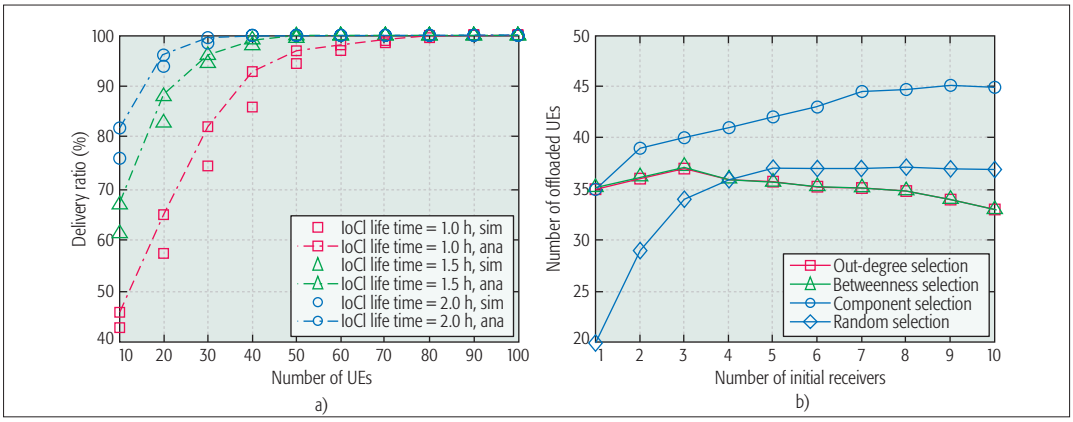

Figure 4. Performance results for the extending coverage and offloading traffic scenarios in which opportunistic communication is applied [14]: a) the average successful delivery ratio before the expiry of IoCI; b) the average number of UEs receiving IoCI under different selection schemes of the initial receiver set.

cache capacity are capable of caching more requested files and performing more sophisticated cooperation between them, so the radio resource will be reduced. On the other hand, if an AP is assigned more radio resources, fewer files need to be stored, thus saving the caching resources. To address the difference caused by the cache as a kind of resource, a framework for edge cache-based performance analysis and radio resource allocation is given in [9], where a comprehensive summarization on the recent advances of the two issues in F-RANs is presented. In particular, the optimization of spectral efficiency, energy efficiency, and latency via resource allocation is quite different from that in existing works for traditional wireless networks. The cell association to optimize spectral efficiency needs to consider the scenario in which users prefer to access F-APs when the contents they need have been locally cached. As a result, the serving F-AP cluster is determined by both the F-AP caching contents and the reference signal received power. Similarly, the energy efficiency optimization should highlight the effect of local caching because the energy consumption in the local F-AP increases, but with the benefit that the energy consumption of the backhaul is decreased. Besides, the latency optimization problem is becoming challenging due to the coexistence of various transmission modes in the F-RAN. The BBU pool can store numerous contents but only provides latency-tolerant service, while F-APs with limited cache volume can provide contents with small latency. Hence, the trade-off should be balanced between the BBU pool and F-APs with joint consideration of spectral and energy efficiency. In addition to energy efficiency, spectral efficiency, and latency, there are also other new performance metrics taken into account when doing resource management. In [13], serviceability is proposed as a key foundational criterion satisfying the IoT paradigm in the 5G era. Serviceability is defined as the ability of a network to serve UEs within desired requirements (e.g., throughput, delay, and packet loss). Given the distribution of cached content and the desired data rate constraints, a serviceability maximization problem in F-RAN is proposed, which is handled with an adaptive resource balancing scheme.

Social-Aware Slicing Techniques in F-RANs

Social awareness is an emerging approach to enable the realization of network slices in a convenient way. As more and more communications occur between UEs with close relationships, the network is getting increasingly human-centric and social-aware. To exploit UEs' behaviors and interactions in the social domain, social awareness turns out to be indispensable information for the design and optimization of RANs. By exploiting social properties of nodes and UEs in RANs, social awareness helps to provide efficient resource allocation and networking. Inspired by the attractive advantages of social awareness, a detailed survey on the cross-disciplinary research area of social-network-analysis-aided telecommunication networking is shown in [14]. In particular, two eMBB-specific application scenarios are studied to illustrate the benefits of social awareness: the extending coverage scenario and the offloading traffic scenario. In the extending coverage scenario, limited base stations are sparsely distributed in a large area. To extend the coverage, the opportunistic communications amongst the roaming UEs are exploited based on opportunistic contacts in the social-aware domain. Roaming UEs with the information of common interest (IoCI) carried can deliver the IoCI to other UEs via opportunistic links once they meet. It is demonstrated in Fig. 4a that, with the aid of opportunistic communication, the delivery ratio of the IoCI in the extending coverage scenario increases from 45 to 100 percent. Moreover, the delivery ratio is improved further as the number of UEs increases and the IoCI lifetime becomes longer, since more UEs participate in the information dissemination process and the UEs can tolerate longer latency in receiving the IoCI.

In the offloading traffic scenario, numerous UEs are occasionally densely crowded in a small area. To alleviate the burden of base stations and improve spectrum efficiency, large-scale opportunistic networks are utilized to offload some traffic from cellular networks. Based on the social contact graph of the UEs and specific selection schemes, the initial receiver set is determined and a opportunistic network is constructed. The BSs first inject some copies of the IoCI into the opportunistic network. Thereafter, the IoCI is disseminated via opportunistic contacts of UEs as

	Fog computing	Mobile edge computing	Cloudlet
Developed by	Cisco in 2011	ETSI as an industry specification in 2014	Carnegie Mellon University in 2013, later supported by various companies including Intel, Huawei, and Vodafone
Concepts	An extension of cloud computing at the edge of networks	Push cloud computing capabilities close to the RAN in 4G and 5G	The middle tier of a three-tier hierarchy: "mobile device-cloudlet-cloud"
Features	• A completely distributed, multi-layer network architecture • Any device with cache and computing capabilities can be incorporated into the network • Achieving cooperation between computing, communication, and control is characterized	A number of standardized MEC servers and applications are deployed at the edge of the network, which is capable of providing the cache, analysis, processing, and control functions	• Extra management is not needed • The implementation is based on virtual machines and the Openstack++ platform • A fairly strong computing ability is provided
Main standardization actors	Open Fog	ETSI	Open Edge Computing
Typical use cases	IoT, optimized local content distribution, data caching, and so on	Video analytics, location services, IoT, augmented reality, and so on	Face detection, voice recognition (used for traffic offloading), emergency scenarios, and so on

Table 1. Comparison of fog computing, mobile edge computing, and cloudlet.

well as via direct injections from the BSs. With the aid of opportunistic multicast scheduling in the small-scale opportunistic scenario, opportunistic communication between UEs can be established and heavy traffic is offloaded. To validate the benefits of opportunistic communication applied in the offloading traffic scenario, the contact traces of 78 UEs in a crowed area are studied based on the realistic mobility trace of INFOCOM 2006. As shown in Fig. 4b, opportunistic communication is capable of offloading as high as 58 percent of the traffic (i.e., 45 UEs) from cellular networks.

Challenging Work and Open Issues

Although access slicing in F-RANs has been initially researched, and some primeval progress has been achieved, there are still many challenges and open issues to be discussed in the future. Besides the key techniques initially researched in the aforementioned section, standard development and field trials are two other important issues.

Access Slicing Standardization

Given the relative infancy, the current standardization efforts on network slicing are still in the early stage. Lots of industry-wide studies about network slicing are ongoing and a large number of related organizations are involved, including the Next Generation Mobile Networks (NGMN) Alliance, 3GPP, and 5G Public Private Partnership (5GPPP). In NGMN, a detailed definition and applications of network slicing have been output. The network slice composing the business application layer is defined as an indispensable part of the 5G architecture [3]. In 3GPP, network slicing is identified as one of the key technologies to be developed in 5G standardization. The study of CN-based network slicing has been undertaken in the System Architecture 2 Working Group [8], where five work tasks for network slicing have been defined, and the corresponding solutions like network slice selection and identification have been discussed. The realization of network slicing in RANs has been incorporated into the future work schedule. In particular, the key principles for the support of network slicing in RANs have been identified, while the issues of resource management and isolation among slices are still under investigation. In 5GPPP, several projects related to network slicing have been activated, and the architecture enablers and resource management for network slicing related to diverse 5G services are listed in [15]. However, the current discussions on network slicing in the industry are mainly focused on the concept and requirements of network slicing, the solutions for realization of network slicing, and the potential CN-based slicing architecture. It is anticipated that more attention will be focused on access slicing since it combines the advantages of both CN-based and RAN-based slicing.

Also, there are a large number of research projects on the development of F-RANs. Numerous works have been done on the related fog computing, as well as other similar concepts like mobile edge computing and cloudlets. As illustrated in Table 1, these similar concepts are partially overlapping and complementary. All of these similar concepts are driven by an anticipated future characterized by IoT, which can be categorized into mMTC with low data rate and minimal power consumption and uMTC with ultra-low latency. Unlike mobile edge computing and cloudlets, which use dedicated servers and devices, F-RAN can utilize any device with cache and computing capabilities anywhere in the network to enhance system performance. Thus, the F-RAN offers more flexibility in the choice of devices and can be implemented in a shorter time. Moreover, the Open Fog Consortium has established the foundation for an open fog computing reference architecture. As a result, access slicing in F-RANs is a promising development, and its standardization should be exploited, regarded as a further step in the existing network slicing standards in various standard organizations.

Testbed and Field Trials

The industry has made lots of efforts on the realization and test of CN-based network slicing. During the Mobile World Congress 2016, Deutsche Telekom and Huawei demonstrated the world's first end-to-end 5G system to validate

Given the relative infancy, there are still quite a number of outstanding problems that need further investigation. Notably, it is anticipated that great attention will be focused on the progress of standardization development and trial tests on access slicing in F-RANs, which makes the commercial rollout of access slicing as early as possible.

CN-based network slicing for diverse 5G use cases, in which autonomous end-to-end network slicing, adding the dynamic and real-time slicing of the 5G RAN, CN, and the interconnecting transmission network, has been implemented, and the corresponding results have been shown. These demonstrations have shown that CN-based network slices can be automatically created in an optimized way with a cost of sub-minute time. Motivated by the CN-based network slicing testbed, it can be anticipated that access slicing in F-RANs can be tested in future 5G field trials, and the gains will be significant.

To show the performance gains achieved by the proposed access slicing, the first step is to build a powerful F-RAN testbed. Then the proposed architecture and key techniques for access slicing should be fulfilled in the implemented testbed. Notable achievements are anticipated to be gained, and preliminary results for access slicing in trial tests will be output.

Conclusion

In this article, access slicing in fog radio access networks has been proposed. Compared to separate CN-based and RAN-based network slicing, the proposed slicing combines the advantages of these two types of network slicing, taking full advantage of the edge characteristics of F-RANs. To enable the implementation of the proposed access slicing architecture, the body of key techniques are broadly divided into hierarchical resource management and multi-dimensional social-aware slicing. With the goal of understanding further intricacies of key techniques, diverse problems within these key techniques have been summarized, and corresponding solutions have been presented. Nevertheless, given its relative infancy, there are still quite a number of outstanding problems that need further investigation. Notably, it is anticipated that much attention will be focused on the progress of standardization development and trial tests on access slicing in F-RANs, which will bring about the commercial rollout of access slicing as early as possible.

Acknowledgment

The authors would like to acknowledge the support of the State Major Science and Technology Special Project under Grant 2016ZX03001020-006, and the support of the National Natural Science Foundation of China under Grant 61361166005.

References

[1] S. Chen and J. Zhao, "The Requirements, Challenges, and Technologies for 5G of Terrestrial Mobile Telecommunication," *IEEE Commun. Mag.*, vol. 52, no. 5, May 2014, pp. 36–43.
[2] ITU-R Rec. M, "[IMT.VISION]: IMT-Vision-Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond," June 2015.
[3] R. Hattachi and J. Erfanian, "5G White Paper," Feb. 2015; http://www.ngmn.org/fileadmin/ngmn/content/downloads/Technical/2015/NGMN5GWhitePaperV10.pdf, accessed Sept. 6, 2017.
[4] X. Zhang *et al.*, "Network Slicing as a Service: Enabling Enterprises' Own Software-Defined Cellular Networks," *IEEE Commun. Mag.*, vol. 54, no. 7, July 2016, pp. 146–53.
[5] W. Wang *et al.*, "A Software-Defined Wireless Network Enabled Spectrum Management Architecture," *IEEE Commun. Mag.*, vol. 54, no. 1, Jan. 2016, pp. 33–39.
[6] K. Wang *et al.*, "Virtual Resource Allocation in Software-Defined Information-Centric Cellular Networks with Device-to-Device Communications and Imperfect CSI," *IEEE Trans. Vehic. Tech.*, vol. 62, no. 12, Dec. 2016, pp. 10,011–21.
[7] M. Sama *et al.*, "Reshaping the Mobile Core Network via Function Decomposition and Network Slicing for the 5G Era," *Proc. 2016 Wireless Commun. and Networking Conf.*, Apr. 2016, pp. 1–7.
[8] 3GPP TR 23.799, "Study on Architecture for Next Generation System," Rel-14, Oct. 2016; http://www.3gpp.org.
[9] M. Peng and K. Zhang, "Recent Advances In Fog Radio Access Networks: Performance Analysis and Radio Resource Allocation," *IEEE Access*, vol. 4, Aug. 2016, pp. 5003–09.
[10] O. Sallent *et al.*, "On Radio Access Network Slicing from a Radio Resource Management Perspective," to appear, *IEEE Wireless Commun.*, 2017. DOI: 10.1109/MWC.2017.1600220WC.
[11] M. Peng *et al.*, "Fog Computing Based Radio Access Networks: Issues and Challenges," *IEEE Network*, vol. 30, no. 4, July 2016, pp. 46–53.
[12] K. Zhu and E. Hossain, "Virtualization of 5G Cellular Networks as a Hierarchical Combinatorial Auction," *IEEE Trans. Mobile Comp.*, vol. 15, no. 10, Oct. 2016, pp. 2640–54.
[13] N. Dao *et al.*, "Adaptive Resource Balancing for Serviceability Maximization in Fog Radio Access Networks," *IEEE Access*, vol. 5, June 2017, pp. 14548–59.
[14] J. Hu *et al.*, "Bridging the Social and Wireless Networking Divide: Information Dissemination in Integrated Cellular and Opportunistic Networks," *IEEE Access*, vol. 3, Sept. 2015, pp. 1809–48.
[15] P. Arnold *et al.*, "Deliverable D2.2 Draft Overall 5G RAN Design," June 2016; https://metis-ii.5g-ppp.eu/wp-content/uploads/deliverables/METIS-IID2.2V1.0.pdf, accessed Sept. 6, 2017.

Biographies

Hongyu Xiang (xianghongyu88@163.com) received his B.E. degree in telecommunication engineering from Fudan University, China, in 2013. He is currently pursuing a Ph.D. degree at the Key Laboratory of Universal Wireless Communications (Ministry of Education) at Beijing University of Posts and Telecommunications (BUPT). His research focuses on cooperative radio resource management and collaborative radio signal processing in heterogeneous cloud radio access networks and fog radio access networks.

Wenan Zhou (zhouwa@bupt.edu.cn) is currently an associate professor at the School of Computer Science, BUPT. She received her Ph.D. degree in electrical engineering from BUPT in 2002. In 2007, she furthered her study of broadband wireless communication technology at the University of California, San Diego as a visiting scholar. Her current research interests include wireless mobile communication theory, radio resource management, and QoE management in 5G.

Mahmoud Daneshmand (mdaneshm@stevens.edu)) received his Ph.D. and M.S. degrees in statistics from the University of California, Berkeley. He is currently a professor in the School of Business at Stevens Institute of Technology. He is an expert in big data analytics, data mining algorithms, machine learning, probability and stochastic processes, and statistics. He is a co-founder and chair of the Steering Committee of the *IEEE IoT Journal* and *IEEE Transactions on Big Data*.

Mugen Peng [M'05, SM'11] (pmg@bupt.edu.cn) received his Ph.D. degree from BUPT in 2005. He has been a full professor at BUPT since 2012. His main research interests focus on cooperative communication, self-organizing networking, heterogeneous networking, cloud communication, and the Internet of Things. He was a recipient of the 2014 IEEE ComSoc AP Outstanding Young Researcher Award, and the best paper awards of the *Journal of Communications Networking*, IEEE WCNC 2015, WASA 2015, GameNets 2014, and so on.