Network Slicing for 5G: Challenges and Opportunities

Network slicing for 5G provides a network as a service (NaaS) for different use cases, allowing network operators to build multiple virtual networks on a shared infrastructure. With network slicing, service providers can deploy their applications and services flexibly and quickly to accommodate diverse services' specific requirements. As an emerging technology with a number of advantages, network slicing has raised many issues for the industry and academia alike. Here, the authors discuss this technology's background and propose a framework. They also discuss remaining challenges and future research directions.

Xin Li
Wuhan University, China

Mohammed Samaka
Qatar University

H. Anthony Chan
Huawei Technologies

Deval Bhamare
Qatar University

Lav Gupta
Washington University

Chengcheng Guo
Wuhan University, China

Raj Jain
Washington University

Research in fifth-generation (5G) mobile networks has gained momentum in both academia and industry recently. These 5G mobile networks will carry a plethora of mobile devices and provide faster network connection speed. In addition, mobile data traffic in future 5G networks will experience explosive growth. According to the latest Cisco forecast,[1] by 2020, the number of devices connected to mobile networks will exceed the world's population, reaching 11.6 billion. At that time, mobile data traffic will be 30.6 exabytes per month, which is 8 times that in 2015. However, the most challenging aspect is that 5G networks will support a wide variety of use cases, as we're entering the era of the Internet of Everything (IoE). Some applications, such as ultra-high definition (UHD) video and augmented reality need high-speed, high-capacity communications, while others such as the mission-critical Internet of Things (IoT) and autonomous vehicles require ultra-low latency, ultra-reliable services.

Traditional mobile communication networks employ the one-size-fits-all approach to providing services to mobile devices, regardless of the communication requirements of vertical services. This design philosophy can't offer differentiated services. Hence, it's necessary for the research community to explore new techniques to address the challenges associated with supporting vertical industries in 5G networks.

Software-defined networking (SDN) and network functions virtualization (NFV) have been proposed as key technologies to build softwarized, virtualized, and cloudified 5G systems in

1089-7801/17/$33.00 © 2017 IEEE

recent years. SDN[2] decouples network control and data forwarding. Network control functions can run as applications independently in the logically centralized controllers. NFV[3] decouples specific network functions from dedicated and expensive hardware platforms to general-purpose commodity hardware. Network operators can implement a variety of virtual network functions (VNFs) over the standard commodity servers. In addition, Mobile Edge Computing (MEC)[4] as a key emerging technology in 5G is expected to serve low-latency communication that's one of the use cases in future 5G. It moves computing, storage, and networking resources from remote public clouds closer to the edge of the network. Thus, mobile clients can request virtual resources within the access network and experience low end-to-end delay.

The concept of *network slicing*[5] has been proposed to address the diversified service requirements in 5G under the background of the aforementioned technologies. Network slicing is an end-to-end logical network provisioned with a set of isolated virtual resources on the shared physical infrastructure. These logical networks are provided as different services to fulfill users' varying communication requirements. Network slicing provides a network-as-a-service (NaaS) model, flexibly allocating and reallocating resources according to dynamic demands, such that it can customize network slices for diverse and complex 5G communication scenarios.

Network slicing will be the fundamental feature of 5G networks. Slice-based 5G has the following significant advantages when compared with traditional networks:

- Network slicing can provide logical networks with better performance than one-size-fits-all networks.
- A network slice can scale up or down as service requirements and the number of users change.
- Network slices can isolate the network resources of one service from the others; the configurations among various slices don't affect each other. Therefore, the reliability and security of each slice can be enhanced.
- Finally, a network slice is customized according to service requirements, which can optimize the allocation and use of physical network resources.

With this in mind, next we discuss some details on network slicing for 5G networks and explore why 5G needs network slicing, as well as how to implement network slicing in 5G. For others' work on network slicing for 5G, see the related sidebar.

Network Slicing in 5G

Fifth-generation networks need to integrate multiple services with various performance requirements — such as high throughput, low latency, high reliability, high mobility, and high security — into a single physical network infrastructure, and provide each service with a customized logical network (that is, network slicing). The Third-Generation Partnership Project (3GPP) has identified network slicing as one of the key technologies to achieve the aforementioned goals in future 5G networks. Some 3GPP work items have the features of network slicing: for example, the dedicated core (DÉCOR) feature supports the operator to deploy multiple dedicated core networks by sharing a common Public Land Mobile Network (PLMN). An exhaustive description about how the next-generation wireless system will support network slicing is provided elsewhere.[6]

Network slicing refers to partitioning of one physical network into multiple virtual networks, each architected and optimized for a specific application/service. Specifically speaking, a network slice is a virtual network that's created on top of a physical network in such a way that it gives the illusion to the slice tenant of operating its own dedicated physical network. A network slice is a self-contained network with its own virtual resources, topology, traffic flow, and provisioning rules. There might be various network slices to meet the specific communication needs of different users in the future mobile network systems. For example, a massive industrial IoT slice might need a light 5G core, with no handover but a large number of connections. On the other hand, a mobile broadband slice might need a high-capacity core, full feature mobility, and low latency. Slices are logically isolated, but resources can be shared among them. Figure 1 illustrates the network slicing concept.

Related Work on Network Slicing for 5G

Network slicing makes use of the concept known as *network virtualization*. Network virtualization enables flexible and dynamic network management to address the problem of network ossification by allowing multiple heterogeneous and service-specific virtual networks to share a single substrate network.

In addition, the emerging technologies software-defined networking (SDN) and network functions virtualization (NFV) are considered as the necessary tools to implement network slicing. The following work summarizes use of SDN and NFV to implement 5G network slicing.

Csaba Simon and colleagues[1] propose a flexible 5G network architecture to support the coexistence of heterogeneous services and to quickly create services. The authors propose to use SDN and NFV in the architecture to realize the sharing of resources by different services and orchestrating resources automatically. The concept of resource slicing is similar to network slicing in the proposed architecture. Resource slices, which include virtual resources and virtual network functions, are tailored on-demand according to service categories, namely slice as a service (SlaaS).

Manuel Peuster and colleagues designed an NFV-based platform called Multi Datacenter service ChaIN Emulator (MeDICINE) for network services.[2] It allows management and orchestration (MANO) system to deploy virtual network resources for network services in a multi-domain infrastructure. The design and implementation of this platform show that NFV plays an important role in realizing network slicing.

Navid Nikaein and colleagues[3] proposed a novel slice-based 5G architecture based on SDN, NFV, and cloud computing. They designed the elements required to implement network slicing and present a validation prototype. The concept of a *network store* in this work can achieve dynamic 5G network slicing. The authors discuss building a multitenant and multiservice end-to-end 5G mobile network architecture using SDN, NFV, and network slicing in the 5G NORMA project.[4]

Mobile operators, hardware manufacturers and open source communities are also actively studying the ways to implement 5G network slicing. Ericsson and NTT DOCOMO successfully showed a dynamic 5G network slicing proof of concept on 9 June 2016.[5] Ericsson[6] discussed the decomposition schemes of radio access network (RAN) functions that can support network slicing. The schemes provide a meaningful guidance for implementing flexible and resilient RAN slices.

Very recently, Huawei (with other three enterprises) released a network slicing white paper.[7] They discussed several key technologies, such as network management system and security to achieve service-guaranteed network slicing, which will be helpful for network industries.

It's worth noting that the Open Network Automation Platform (ONAP; www.onap.org), established in February 2017, is working on a cloud-centric and SDN/NFV-based network platform, which might lay the groundwork for the implementation of 5G network slicing. Also, currently, the Wireless World Research Forum (WWRF) is launching a working group for 5G network slicing.

References

1. C. Simon et al., "5G Exchange for Inter-Domain Resource Sharing," *Proc. IEEE Int'l Symp. Local and Metropolitan Area Networks*, 2016; doi:10.1109/LANMAN.2016.7548842.
2. M. Peuster, H. Karl, and S. Van Rossem, "MeDICINE: Rapid Prototyping of Production-Ready Network Services in Multi-PoP Environments," arXiv preprint, 2016; https://arxiv.org/pdf/1606.05995.pdf.
3. N. Nikaein et al., "Network Store: Exploring Slicing in Future 5G Networks," *Proc. 10th Int'l Workshop on Mobility in the Evolving Internet Architecture*, ACM, 2015, pp. 8–13.
4. P. Rost et al., "Mobile Network Architecture Evolution Toward 5G," *IEEE Comm.*, vol. 54, no. 5, 2016, pp. 84–91.
5. *DOCOMO and Ericsson Perform Successful Proof of Concept of Dynamic 5G Network Slicing*, press release, NTT DOCOMO, 13 June 2016; https://www.nttdocomo.co.jp/english/info/media_center/pr/2016/0613_00.html.
6. E. Westerberg, "4G/5G RAN Architecture: How a Split Can Make the Difference," *Ericsson Technology Rev.*, vol. 93, no. 6, 2016; www.ericsson.com/assets/local/publications/ericsson-technology-review/docs/2016/etr-ran-architecture.pdf.
7. *5G Service-Guaranteed Network Slicing*, white paper, Huawei, Mar. 2017.

Network slicing needs to be implemented in an end-to-end manner to meet diverse service requirements. Each slice may have its own network architecture and protocols. 5G network slicing includes slicing the 5G radio access network (RAN), 5G core network, and even end-user devices. 5G RAN slicing can be implemented through logical abstraction of physical radio resources (such as spectrum) and physical hardware (such as a base station). SDN and NFV can configure the virtual network resources flexibly, which include network bandwidth, server processing capability, and network element processing capability, to build the core network slices for specific service requirements. RAN slices and core network slices can either be dedicated to a class of service users or shared between multiple classes of service users. To provide an identical slice for a specific service, a common function — the *slice-selecting function* — is needed. It's responsible for selecting RAN and core network slices to form the end-to-end slice.

From the RAN point of view, the main task of the RAN slice is to achieve the sharing and flexible management of spectrum resources. Therefore, the software-defined RAN is a promising approach to implement RAN slicing. The sharing of radio resources between different slices can be accomplished by scheduling conducted by the controller in software-defined RAN. The controller can allocate radio resources to a slice according to service request characteristics, such as traffic load. Moreover, different radio access methods for RAN slices are needed in terms of diverse service requirements. For example, besides the mobile broadband access in the traditional cellular networks, the critical communication slice in 5G needs low-latency radio access. In addition, the implementation of RAN slicing needs logical decompositions of RAN functions to figure out which functions are common functions and which functions are provided with specific hardware.[7]

5G core network slicing can provide tailored logical networks for different services or verticals in an agile and adaptable fashion. The key technologies of core network slicing are SDN and NFV. SDN technology can be used to separate the control plane and data plane of the core network so that the control plane and data plane can be deployed independently. The control plane can be centralized to facilitate management, while the data plane can be distributed. For example, the data plane in a network slice for low-latency services can be distributed on the network edge in combination with the MEC technology. NFV technology can provide the necessary VNFs for the data plane of the core network slices according to the service type. These VNFs can be scaled on-demand as the service changes dynamically.

5G network slicing brings flexibility while increasing the complexity of network management. Thus, there's a need to design automated management schemes to address such complexity. The lifecycle management is responsible for slice creation, reconfiguration, deletion, and so forth. It involves orchestrating and allocating infrastructure resources. Therefore, it's crucial to improve the 5G infrastructure's resource usage.

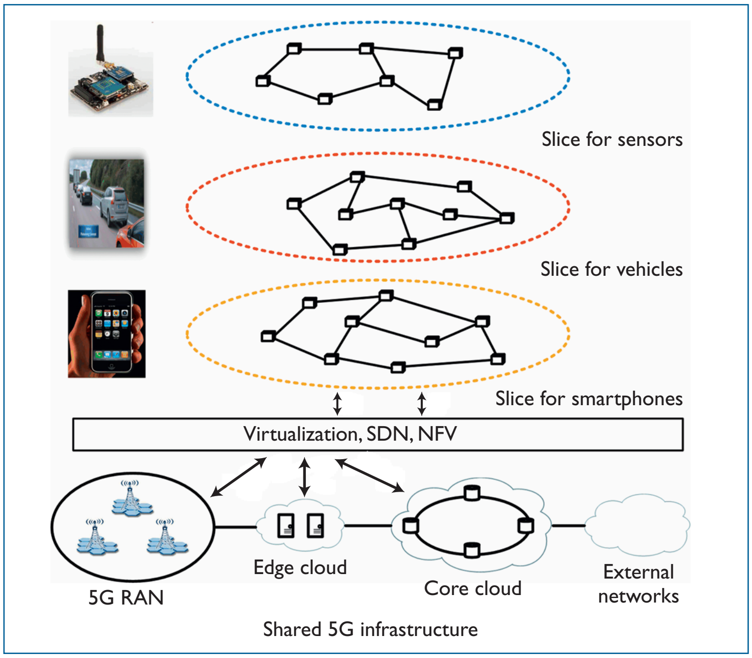

Figure 1. Conceptual illustration of network slicing. NFV stands for network functions virtualization, RAN stands for radio access network, and SDN stands for software-defined networking.

With this in mind, next we present our 5G network slicing framework based on previous work.[8]

5G Network Slicing Framework

We show our three-layer network slicing framework for 5G networks in Figure 2. The proposed framework supports flexible deployment and management of diverse network applications over one common infrastructure. The framework contains a 5G software-defined infrastructure (5G-SDI) layer, virtual resource layer, application and service layer, and a slicing management and orchestration (MANO) functional component. We use three different colors in Figure 2 to indicate three network slices for connected vehicles, virtual reality, and mobile broadband (MBB).

The bottom layer in Figure 2 is the 5G infrastructure layer, which is composed of multi-domain 5G software-defined infrastructures with software-based control and management. The use of the software-defined approach

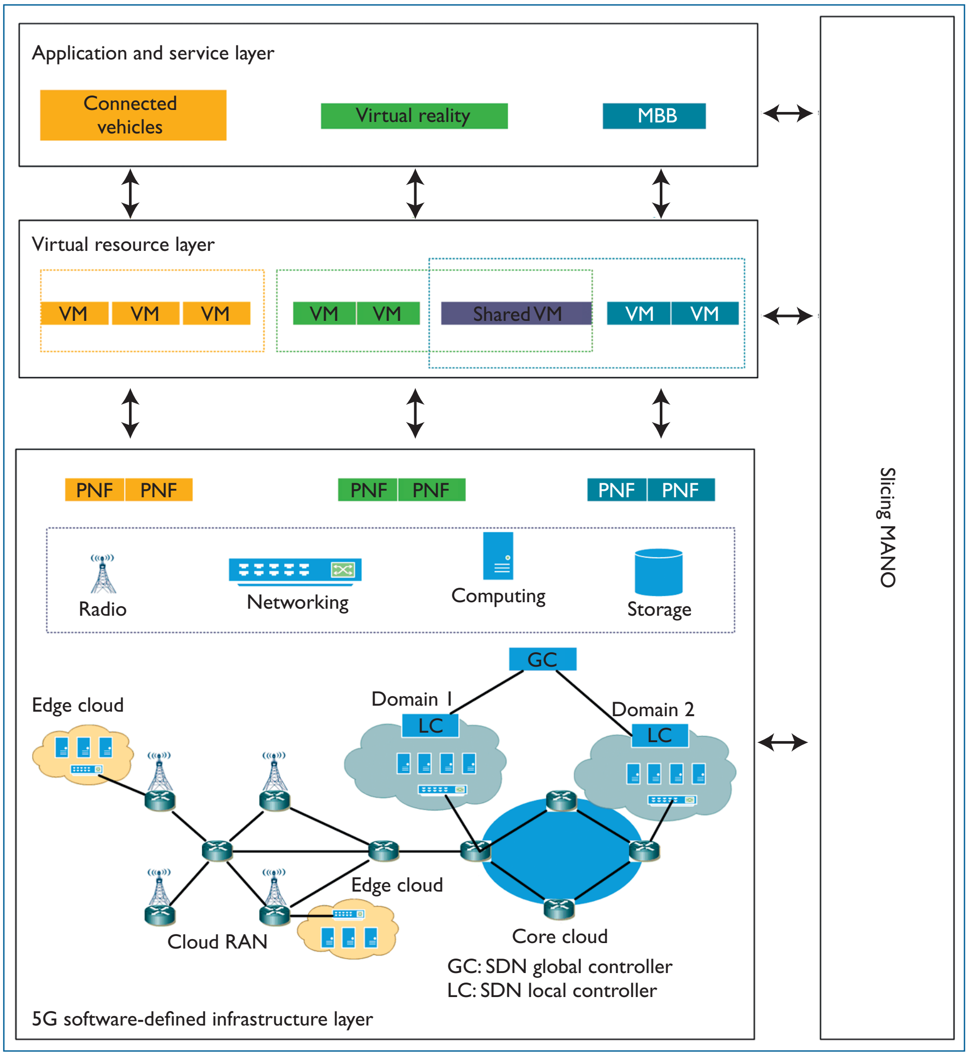

Figure 2. Fifth-generation (5G) network slicing framework. MANO stands for management and orchestration; MBB stands for mobile broadband; PNF stands for physical network function; and VM stands for virtual machine.

brings the following benefits to the proposed framework:

- it provides an abstraction of resources and allows a dynamic allocation of resources;
- the virtualization of infrastructure resources can flexibly form service-specific slices; and
- both the network provider and the slice user can manage virtual resources.

The design of the SDN controller in the framework uses the hierarchical method. Each domain has a local controller and these local controllers eventually converge to the central global controller. Some examples of the 5G-SDI controllers include Open Application Delivery Networking (OpenADN),[8] OpenDayLight, FloodLight, and OpenStack. Cloud-based technologies such as Cloud RAN and MEC can be introduced to enable a cloudified 5G infrastructure. VNFs in a low-latency slice (such as the slice for connected vehicles) can be placed in edge clouds, while VNFs in the MBB slice can be provisioned on commodity servers in the core cloud.

The virtual resource layer provides all of the virtual resources required for network slices, such as radio, computing, storage, and network bandwidth. Virtual resources and VNFs run as virtual machine (VM) instances. These resources are in fact an abstraction of the underlying physical resources managed at 5G-SDI and can be shared across slices — for example, the VM instance shared between the virtual reality slice and MBB slice in Figure 2. Moreover, these resources can be used on-demand by network slices, which enhances the network resource utilization.

Heterogeneous service instances such as connected vehicles, virtual reality, and MBB constitute the service layer. These service instances belong to multiple tenants. However, one tenant can still have several service types. The service instance informs slicing MANO of its service requirements, which is then mapped to the corresponding network slice. The dynamic nature of services mandates the network slices to be scalable at runtime. Services that have similar demands might share one network slice, and services with different service requirements can still share the slices partially.

The slicing MANO functional component is a key part of the network slicing framework.

It manages and orchestrates the physical resources, virtual resources, and network slices in the way the NFV-MANO will be built in. Slicing MANO can have subfunctions, such as the infrastructure manager, VNF manager, and slices orchestrator. The tasks of slicing MANO include

- creating and managing VM instances by using the infrastructure resources;
- mapping network functions to virtual resources and connecting network functions to create service chains; and
- managing the life cycle of network slices by interacting with the application and service layer — for example, automated creation of service-oriented slices and dynamic maintenance by monitoring service requirements and virtual resources.

In fact, the management of network slices might be performed at two levels, which are inter-slice management and intra-slice management. Inter-slice management is performed by the slice orchestrator, while intra-slice management is implemented by running a virtual manager function as part of the slice.

Challenges and Future Research Areas

Next, we discuss some important research challenges and directions for network slicing in 5G networks. They include resource sharing; dynamic creation and management; isolation among network slices; mobility management; security and wireless resource virtualization in network slicing; and algorithmic aspects of resource allocation.

Resource Sharing

The Next-Generation Mobile Networks (NGMN) Alliance has figured out that resource sharing among slice tenants is a key issue.[5] Resource sharing can be implemented by static partition sharing or elastically dynamic sharing. Because of the dynamic characteristics of network load, the dynamic resource sharing among slice tenants makes network resource usage more efficient. There are some specific issues to be solved in resource sharing. For example, radio resources can be shared among RAN slices. It requires a good radio-scheduling mechanism to allocate radio resources among these slices. In addition, computational resource sharing and other resource sharing must be considered. Although resource sharing brings benefits to infrastructure providers, it also brings more challenges to other problems such as slice isolation.

Dynamic Slice Creation and Management

One of the concerns for future 5G network operators is how to allocate network resources to maximize their benefits. In this context, lifecycle management of the network slices is a critical problem to be solved. To accommodate the maximum number of diversified service requests, 5G network operators need to deploy VNFs and allocate network resources rapidly to build network slices. In addition, they should be able to scale slices dynamically according to the varying service load. On the other hand,

Although resource sharing brings benefits to infrastructure providers, it also brings more challenges to other problems such as slice isolation.

although a network operator has the maximum control over his network slices, one slice might still need to exercise some control over itself to improve the service quality. Accordingly, the network-slicing technique should consider how to open partial permissions to each slice to configure and manage it without raising security issues. In addition, management of the network slice should be implemented in an automated fashion to avoid manual efforts and errors.

Isolation among Network Slices

Different services in 5G networks have unique requirements. As a consequence, dedicated virtual network resources are needed to guarantee the service quality at each slice. This requires slices to be highly isolated from each other. The isolation among network slices can be achieved via data plane isolation and control plane isolation. In general, the slice control function can be shared among slices, while for some services such as mission-critical communications, the slice needs its own control function. Further, the effective isolation of network slices can ensure that the failure or security attack on one slice doesn't affect other slices' operation. Hence, the slice isolation mechanism is a key challenge of implementing network slicing.

Mobility Management in Network Slicing

Mobility aspects such as seamless handover and interference management bring new challenges to network slicing. Fast mobility handover is crucial for real-time services, which has a direct influence on the quality of service. The mobility support for network slicing is optional for some specific network slices. For example, network slices serving industrial control don't need mobility management functions because of fixed position of devices. However, network slices that do need mobility management differ in their mobility requirements. For instance, the mobility requirement of the mobile broadband slice is different from the slice for automated driving services. Thus, the design of a slice-oriented mobility management protocol is imperative to tackle mobility challenges in network slicing.

Security in Network Slicing

Security in network slicing is a critical problem to be addressed because of resource sharing among slices. Network slices serving different types of services may have different levels of security policy requirements. Therefore, while designing network slicing security protocols, it's necessary to consider the impact on other slices and entire network systems. In addition, security issues become more complex when network slicing is implemented on the multidomain infrastructure. Security policy coordination mechanisms among different domain infrastructure must be designed.

Wireless Resource Virtualization

Virtualization is the most important technology to implement network slicing. Virtualization technology has evolved over the past 20 years and been applied mainly in wired networks. A lot of work has focused on core network virtualization. However, wireless communication is more complex, as wireless links are time-varying and vulnerable to interference. Some of the virtualization methods in wired network can't be directly used for their wireless counterparts. Therefore, designing new virtualization mechanisms to achieve the sharing of wireless spectrum resources and virtualization of base stations plays a vital role in implementing RAN slicing.

Algorithmic Aspects of Resource Allocation

How to efficiently allocate slice resources is a challenging problem to be addressed. The slice resource allocation problem is similar to the virtual network embedding problem. The formulation of this problem can use mathematical methods in operation research, such as integer linear programming (ILP). The algorithms to solve the problem might adopt different strategies according to the slice's size — for example, the number of virtual links and virtual nodes. The resource allocation for small-scale slices can be solved by exact algorithms, while for large-scale slices, heuristic or meta-heuristic strategies are more efficient. It's worth noting that, because of the dynamic characteristics of a 5G network, we must design practical resource allocation algorithms with the ability to reconfigure and migrate slice resources.

Network operators have begun to seek new ways to support heterogeneous services. Network slicing is emerging as a possible methodology for allowing network operators to create service instances quickly and flexibly, enabling different services to have their own logical slice instances on a shared infrastructure. Network slicing — together with SDN, NFV, MEC, and other technologies — will become a key in the service-oriented 5G network architecture.

Although network slicing has just begun to attract attention of academia and industry, we believe that it will be an indispensable technology in future 5G systems. Working on the related specification and standardization will be crucial for better execution and promotion of this technology's deployment and application.

Acknowledgments

This publication was made possible by the NPRP award (NPRP 8-634-1-131) from the Qatar National Research Fund (a member of The Qatar Foundation), Huawei Technologies, and the China Scholarship Council (201506270075). The statements made herein are solely the responsibility of the authors.

References

1. *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2015-2020*, white paper, Cisco, 2016.
2. N. McKeown et al., "OpenFlow: Enabling Innovation in Campus Networks, *J. ACM Sigcomm Computer Comm. Rev.*, 2008, vol. 38, no. 2, 2008, pp. 69–74.
3. *Network Functions Virtualisation: An Introduction, Benefits, Enablers, Challenges & Call for Action*, white paper, European Telecommunications Standards Institute (ETSI), 2012; https://portal.etsi.org/NFV/NFV_White_Paper.pdf.
4. *Executive Briefing – Mobile Edge Computing (MEC) Initiative*, white paper, ETSI, 2014; https://portal.etsi.org/portals/0/tbpages/mec/docs/mec%20executive%20brief%20v1%2028-09-14.pdf.
5. *Description of Network Slicing Concept*, white paper, Next-Generation Mobile Networks (NGMN) Alliance, 2016.
6. *Study on Architecture for Next Generation System (Release 14)*, v. 14.0.0, Third-Generation Partnership Project (3GPP) TR 23.799, Dec. 2016; https://portal.3gpp.org/ngppapp/CreateTdoc.aspx?mode=view&contributionUid=SP-160829.
7. E. Westerberg, "4G/5G RAN Architecture: How a Split Can Make the Difference," *Ericsson Technology Rev.*, vol. 93, no. 6, 2016; https://www.ericsson.com/assets/local/publications/ericsson-technology-review/docs/2016/etr-ran-architecture.pdf.
8. S. Paul and R. Jain, "OpenADN: Mobile Apps on Global Clouds Using OpenFlow and Software Defined Networking," *Proc. 2012 IEEE Globecom Workshops*, 2012, pp. 719–723.

Xin Li is a PhD student in the School of Electronic Information at Wuhan University, China, and is currently a visiting PhD student at Washington University. His research interests include 5G, network functions virtualization, and software-defined networking. Contact him at lixin_itlab@whu.edu.cn.

Mohammed Samaka is an associate professor in the Department of Computer Science and Engineering (CSE), College of Engineering at Qatar University. His research interests include wireless software architecture and technology, and mobile applications and services. Samaka has a PhD in computing from Loughborough University, England and a postgraduate diploma in computing from Dundee University, Scotland. Contact him at samaka.m@qu.edu.qa.

H. Anthony Chan is a senior researcher at Huawei Technologies, and an honorary professor at the University of Hong Kong. His research focuses on wireless Internet. Li has a PhD in physics from the University of Maryland. He is an IEEE Fellow, and a Distinguished Lecturer of the IEEE Communications Society and IEEE Reliability Society. Contact him at h.anthony.chan@huawei.com.

Deval Bhamare is a postdoctorate researcher at Qatar University. His research interests include network optimization, middleware architecture for cloud-based services, cloud security, and software-defined networks. Bhamare has a dual PhD from IITB-Monash Research Academy, a joint venture between Indian Institute of Technology, Bombay, India and Monash University, Melbourne, Australia. Contact him at devalb@qu.edu.qa.

Lav Gupta is a doctoral candidate in the Computer Science and Engineering Department at Washington University. His research focuses on telecommunications planning and deployment. Gupta has an MS in computer science from the Indian Institute of Technology, Kanpur, India. He's a senior member of IEEE. Contact him at lavgupta@wustl.edu.

Chengcheng Guo is a professor and PhD supervisor in the School of Electronics and Information, Wuhan University, China. His research interests are Internet and communication technology, wireless mesh networks, industry control networks, and real-time and reliability communications. Contact him at netccg@whu.edu.cn.

Raj Jain is the Barbara J. and Jerome R. Cox, Jr., Professor of Computer Science and Engineering at Washington University. His research focuses on congestion control and avoidance, traffic modeling, performance analysis, and error analysis. Jain has a PhD in applied mathematics and computer science from Harvard University. He's a Fellow of IEEE, ACM, and AAAS. Contact him at jain@wustl.edu.