A Network Slicing Framework for End-to-End QoS Provisioning in 5G Networks

Qiang Ye, *Member, IEEE*, Junling Li, *Student Member, IEEE*, Kaige Qu, Weihua Zhuang, *Fellow, IEEE*, Xuemin (Sherman) Shen, *Fellow, IEEE*, and Xu Li

***Abstract*—With software-defined networking (SDN) and network function virtualization (NFV) technologies, network slicing emerges as a promising solution for resource orchestration to achieve quality-of-service (QoS) isolation among customized services in 5G networks. In this article, we propose a comprehensive network slicing framework for end-to-end (E2E) QoS provisioning, with the consideration of differentiated resource types in both wireless and wired network domains. For the wireless network domain, a dynamic radio resource slicing scheme is proposed, in which the overall bandwidth resources are sliced for different BSs to maximize the network utility. The optimal bandwidth slicing ratios are dynamically adjusted based on instantaneous network load conditions. For the wired network domain, bottleneck-resource generalized processor sharing (BR-GPS) is employed as a bi-resource slicing scheme among multiple traffic flows traversing an NFV node. In addition to the property of bottleneck-resource fair allocation with high resource utilization, we show that the BR-GPS minimizes packet queueing delay for each flow at the outgoing link of the NFV node. Some open research problems regarding network slicing are discussed. A case study is presented to demonstrate the effectiveness of the proposed network slicing framework.**

I. Introduction

The fifth generation (5G) communication networks are expected to provide QoS-guaranteed end-to-end (E2E) service deliveries for a massive number of Internet-of-Things (IoT) devices (e.g., intelligent home appliances, smart sensors and actuators) supporting diversified use cases and applications, including smart homing, industrial automation, intelligent transportation, and e-health care systems. The 3rd generation partnership project (3GPP) identifies in its recent technical reports three main features for the future networking paradigm [1]:

1) **Enhanced mobile broadband:** The network deployment will be highly densified to provide a seamless communication coverage for end devices with mobility and to support high data rate services (e.g., high-definition video streaming [2] with up to Gbps peak data rate). Hence, a multi-tier hierarchical network cell deployment (i.e., small-cells underlaying macro-cells) is envisioned for an enlarged network coverage of wireless access networks. At the same time, the numbers of network routers, computational powerful servers and physical links with high transmission bandwidth are increased in the wired core network to accommodate high traffic volume and respond timely to service requests;
2) **Massive IoT:** A large number of heterogeneous IoT devices will be interconnected to support various types of services. To accommodate the massive network access and support efficient E2E packet transmissions, the network capacity needs to be boosted by further improving the utilization of both communication resources and computing resources;
3) **Critical communications:** The 5G networks will support diversified types of applications with differentiated QoS requirements. Some time-critical machine-to-machine (M2M) communications require ultra-high reliability and low latency, e.g., industrial control applications, e-health care, and remote monitoring. Therefore, QoS-oriented service customization is desired to achieve QoS isolation among different services. QoS isolation ensures that the minimum level of QoS experienced by devices (or users) belonging to one type of service is not violated when network states change, including device mobility, varying channel conditions and traffic load fluctuations, at another service type.

The distinctive characteristics of 5G networks pose challenges on the evolving network architecture for both wireless domain and wired domain. In the wireless network domain, to provide wide area network coverage and accommodate massive access from machine-type devices, current radio spectrum utilization needs to be significantly improved. Hence, multi-tier small-cell base stations (SBSs) are deployed underlaying the coverages of macro-cell BSs (MBSs) to exploit spatial multiplexing gain. However, the increasingly densified network deployment will expand both the capital and operational expenditure (CapEx and OpEx) on communication infrastructures and aggravate inter-cell interference. For E2E service deliveries, data packets from the wireless network domain are aggregated and grouped into different traffic flows according to service types, which are then forwarded through wired backhaul links to the edge routers of the core network. A traffic (service) flow refers to an aggregation of packets belonging to the same service type and traversing two end points in the core network. Each flow will traverse a sequence of servers executing specific functions and a number of physical transmission links and network routers before reaching its destination. Packets of each traffic flow consume computing resources (i.e., CPU time) for processing when traversing network servers, and occupy

Qiang Ye, Junling Li, Kaige Qu, Weihua Zhuang, and Xuemin (Sherman) Shen are with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada, N2L 3G1 (emails: {q6ye, j742li, k2qu, wzhuang, sshen}@uwaterloo.ca).

Xu Li is with Huawei Technologies Canada Inc., Ottawa, ON, Canada, K2K 3J1 (email: Xu.LiCA@huawei.com).

bandwidth resources on physical links and network routers for transmission. With increasingly diversified E2E services, traffic flows are required to pass through different sets of network functions in the core network for differentiated QoS provisioning, which leads to an increase of CapEx and OpEx for deploying more function-specific servers.

A. Network Slicing

To achieve high utilization of both communication and computing resources and minimize the infrastructure deployment cost, network function virtualization (NFV) becomes a cost-effective solution, in which different network functions are softwarized as virtual network functions (VNFs), decoupled from the physical substrate network, and placed in software-programmable commodity servers (called NFV nodes) running virtual machines (VMs). In the core network, through a virtualization layer [3] on each NFV node, physical computing resources, i.e., CPU cores for processing tasks, are abstracted as virtual CPU cores (vCPUs) and are allocated to different VMs hosting VNFs. All VNFs are centrally controlled by a virtualization controller, and can be flexibly placed onto NFV nodes in different network locations, which is also called *VNF embedding* [4]. For each service, specific sets of VNFs and the virtual links connecting them are orchestrated to form a logic service function chain (SFC), which is then embedded on the substrate network to achieve a desired cost-performance tradeoff, i.e., each VNF is operated on an NFV node and each virtual link represents a sequence of transmission links and network routers. Hence, a traffic flow traversing an embedded SFC consumes CPU resources on NFV nodes and bandwidth resources on transmission links and network routers. To further enable the programmability on virtualized resources and on routing configurations among VNFs, the virtualization controller is software-defined networking (SDN) enabled [3], which indicates that control functions on each network node are also migrated to the controller. The SDN-enabled virtualization controller can program each NFV node with an appropriate amount of computing resources, and configure an embedded routing path for packets of each flow with transmission bandwidth resources.

In the SDN-enabled NFV framework, VNFs are instantiated as software instances on NFV nodes and are flexibly orchestrated to create different SFCs embedded on the physical network for differentiated E2E service deliveries. When SFCs share a common embedded physical path, a set of network resources, including computing resources on NFV nodes and bandwidth resources on transmission links, should be properly sliced among traffic flows such that QoS isolation can be achieved. This process is called *network slicing*. In the core network, network slicing is interpreted as *bi-resource slicing*. In the wireless network domain, the radio access function on each BS is softwarized and centrally managed by the SDN-enabled virtualization controller. The controller can determine the amount of radio resources allocated to each BS to improve the overall spectrum utilization. Therefore, network slicing in the wireless domain is called *radio resource slicing*, which mainly deals with how to slice the overall radio resources for different device groups to ensure QoS isolation. Existing studies present new architectures for network slicing in either a wireless [1], [5], [6] or a core network [3]. However, limited research works focus on how to determine the sets of resources for different services to achieve a desired tradeoff between high resource utilization and QoS isolation, and how network slicing should be conducted for wireless and wired network domains, considering heterogeneous resources.

In this article, we propose a network slicing framework for both the wireless and core networks. For heterogeneous wireless access networks (HetNets), we investigate how to determine slicing ratios of radio resources at each BS. To exploit resource multiplexing gain, the amount of resources of each slice is dynamically adjusted according to changes of network conditions. In the core network, each service flow needs to traverse a specific sequence of VNFs and virtual links, representing logic SFC, to fulfill certain E2E service requrements. Different logic SFCs can be embedded on a common physical network path, sharing a set of CPU and bandwidth resources to exploit traffic multiplexing gain. Since traffic flows traversing an NFV node demonstrate bottleneck resource consumption on different resource types [7], [8], we study how bi-resources are sliced among flows passing through a common NFV node to achieve both high resource utilization and fair resource usage among flows. Specifically, we evaluate the bi-resource slicing performance in terms of improvement of packet queueing delay of each flow at the outgoing link of the NFV node. A case study is presented to evaluate the performance of the proposed network slicing framework.

II. Radio Resource Slicing for HetNets

In wireless HetNets, a multi-tier of SBSs is deployed underlaying an MBS to explore the spatial multiplexing gain of currently employed spectrum. However, the increasingly deployed SBSs both increase the CapEx and OpEx and intensify inter-cell interference. Moreover, the dynamic and unbalanced traffic load over each cell coverage makes high utilization of radio resources challenging. With SDN-enabled function softwarization, all radio resources on heterogeneous BSs are abstracted and reconfigured by the controller to create different resource slices for different BSs, which are subsequently allocated to end devices, to enhance the utilization of current spectrum and provide QoS isolation among diverse services.

A. Dynamic Radio Resource Slicing

Radio resource slicing for HetNets requires network (service)-level and device-level resource partitioning. At the network level, the abstracted resources are physically partitioned into a number of resource slices and allocated to each BS; At the device level, resources associated with each BS are further divided among end devices to fulfill differentiated QoS demands. Since devices from each service provider (SP) are scattered over the areas of different cells, the entire radio resources are logically sliced for different SPs, but are physically partitioned among end devices. Existing studies mainly focus on device-level resource slicing [1], [6], [9], where radio spectrum resources at each BS are preallocated according to specified policies and are sliced among different groups of end devices under the coverage of the BS. However, the network-level bandwidth slicing needs to be determined for maximal resource utilization.

1) Network Architecture: Consider a two-tier downlink HetNet, where an MBS, denoted by M_0, is deployed for a wide area coverage, and a set of SBSs, $\mathcal{M} = \{M_k, k = 1, 2, ..., n\}$ (n is the number of small cells) are randomly placed underlaying the coverage of the macro-cell, to support heterogeneous M2M devices (MTDs) and mobile terminals (MTs) subscribed from different SPs, as shown in Fig. 1. Since subscribers of all SPs are randomly scattered over the entire network region, we denote the set of machine-type devices along with its set cardinality, staying in the coverage of M_k $(k = 0, 1, 2, ..., n)$ and belonging to SP s $(s = 1, 2, ..., S)$, as $\mathcal{N}_{s,k}$ and $N_{s,k}$, where $\mathcal{N}_{s,0}$ and $N_{s,0}$ indicate the set and number of machine-type devices, residing only in the coverage of the MBS. Assume that all MTs generate one type of data services subscribed from SP 0 and connect to MBS M_0 to avoid frequent handover [10]. Thus, we use $\mathcal{N}_{0,0}$ and $N_{0,0}$ to indicate the set and number of MTs in the network, respectively. Since MTDs are almost stationary or likely have limited mobility, each MTD located in the coverage of an SBS can choose to associate with either its home SBS or the MBS. We use binary variable $x_{i,s,k}$ to indicate the network association pattern for MTD i from SP s located in SBS M_k ($x_{i,s,k} = 1$ if MTD i is associated with the SBS M_k; $x_{i,s,k} = 0$ if it is associated with the MBS). If $k = 0$ and $s \neq 0$, $x_{i,s,0}$ indicates the network association pattern for MTD i located only in the coverage of M_0; If $k = 0$ and $s = 0$, $x_{i,0,0}$ represents the network association pattern for MT i in the HetNet coverage. Since in both cases the device (or the MT) always associates with M_0, we have $x_{i,s,0} = 1$. Every BS has a number of transmission queues, each of which is used for downlink packet transmissions to an end device. Let λ_s denote the packet arrival rate at a transmission queue destined for an MTD (or an MT) i from SP s. For different types of services, packet arrival processes at each transmission queue behave differently. Since M2M traffic is often event-driven with burstiness, packet arrivals at a BS destined for an MTD is modeled as a Poisson process, whereas packet arrivals of a data service destined for an MT is modeled as a periodic packet arrival process.

2) Optimal Bandwidth Slicing Ratios: Bandwidth resources of the MBS and SBSs are preallocated and denoted by $B_{\rm m}$ and $B_{\rm a}$ (m for MBS and a for SBS), respectively, which are mutually orthogonal to avoid the inter-tier interference. Since SBSs can be physically separated by distances, $B_{\rm a}$ are reused at each SBS to exploit the spatial multiplexing gain. With SDN-enabled function softwarization, the spectrum bandwidths of the MBS and SBSs are abstracted as $B_{\rm v}$ $(= B_{\rm m}+B_{\rm a})$, divided into two bandwidth slices $\theta_{\rm m}B_{\rm v}$ and $\theta_{\rm a}B_{\rm v}$, and reallocated to the MBS and SBSs to improve overall resource utilization, where $\theta_{\rm m}$ and $\theta_{\rm a}$ are the slicing ratios. The bandwidth slices are then partitioned and allocated to their associated end devices. Based on a set of BS-device association patterns $\{x_{i,s,k}\}$ and a customized bandwidth allocation scheme, the fraction of bandwidths, $g_{i,s,k}$, allocated to end device i from SP s staying in M_k can be determined. Given transmit power of each BS and wireless channel conditions (including slow fading and shadowing effects, and inter-cell interference) for downlink packet transmissions, the downlink effective achievable rate, $c_{i,s,k}$ (in packet per second), at end device i from SP s staying in M_k can be obtained as a function of $\theta_{\rm m}$, $\theta_{\rm a}$, $B_{\rm v}$, and $g_{i,s,k}$. Note that the bandwidth slicing ratios, BS-device association patterns, and the fraction of bandwidths allocated to each associated MTD and MT, are updated in a large time scale to reduce the communication overhead [11]. For example, bandwidth slicing is updated when traffic load in each cell varies. Thus, $c_{i,s,k}$ is treated as a constant during each bandwidth slicing period.

The objective of bandwidth slicing is to determine the optimal slicing ratios $\theta^*_{\rm m}$ and $\theta^*_{\rm a}$ along with the set of optimal BS-device association patterns $\{x^*_{i,s,k}\}$ to maximize the overall resource utilization, under the constraints of satisfying the minimum rate requirements for devices from different SPs. Thus, an optimization problem is formulated as in (P1).

$$(\text{P1}): \max_{\substack{\theta_{\rm m},\theta_{\rm a},\\ x_{i,s,k},g_{i,s,k}}} \sum_{k=0}^{n}\sum_{s=0}^{S}\sum_{i\in\mathcal{N}_{s,k}} x_{i,s,k}\,\mathcal{U}(c_{i,s,k})$$

$$\text{s.t.} \begin{cases} \displaystyle\sum_{s=0}^{S}\sum_{i\in\mathcal{N}_{s,0}} g_{i,s,0} + \sum_{k=1}^{n}\sum_{s=1}^{S}(1-x_{i,s,k})g_{i,s,k} = 1 & \text{(1a)} \\ \displaystyle\sum_{s=1}^{S}\sum_{i\in\mathcal{N}_{s,k}} x_{i,s,k}g_{i,s,k} = 1, \quad \forall k & \text{(1b)} \end{cases}$$

where $\mathcal{U}(\cdot)$ is a concave utility function with diminish marginal utility (e.g., a logarithm function) for an end device. In (P1), the objective function is to maximize the aggregate network utility. Two basic constraints (1a) and (1b) indicate that the fractions of bandwidth resources allocated to each end device depend on the BS-device association patterns, considering equal bandwidth partitioning among devices associated with one BS. One implicit constraint of (P1) is $\theta_{\rm m} + \theta_{\rm a} = 1$, and one additional constraint guaranteeing the minimum rate requirement r_s for each device from SP s is $c_{i,s,k} \geq r_s$ (not listed in (P1) for brevity) for any i, s, k. For all desicion variables, we have $g_{i,s,k}$, $\theta_{\rm m}$, and $\theta_{\rm a}$ lie within interval $[0, 1]$, and $x_{i,s,k}$ takes on value 0 or 1, for any i, s, k. Problem (P1) can be transformed into a biconcave maximization problem, and then solved for a set of partial optimal solutions [12].

The procedure for bandwidth slicing consists of three steps, as illustrated in Fig. 1:

Step 1 – Through control links between the virtualization controller and the BSs, each BS periodically reports the network information updates to the controller, including the number of end devices $N_{s,k}$ from all SPs, traffic statistics λ_s, and long-term wireless channel conditions between a BS and an associated end device;

Step 2 – With updated network information, the controller conducts the radio resource slicing optimization described in (P1) to determine the set of optimal bandwidth slicing ratios $\theta^*_{\rm m}$ and $\theta^*_{\rm a}$ for the MBS and SBSs, and optimal set of BS-device association patterns $\{x^*_{i,s,k}\}$;

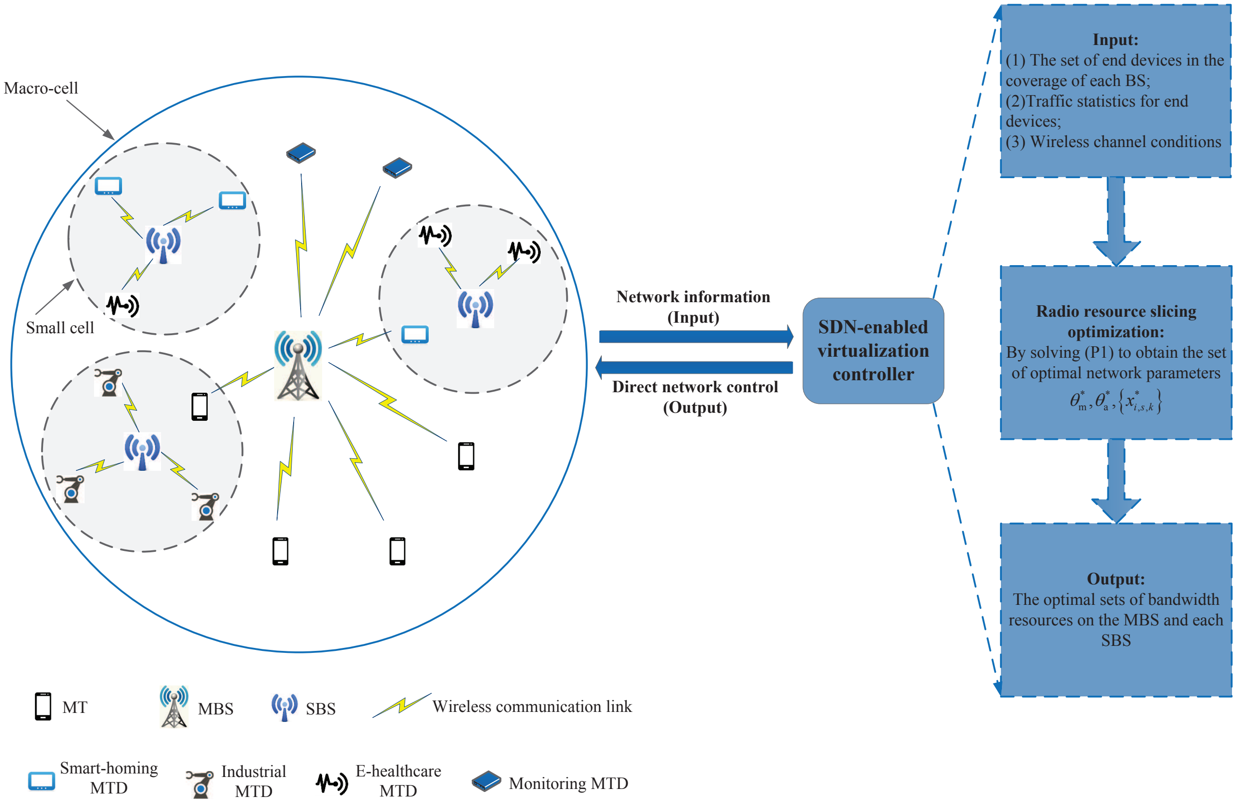

Macro-cell
Small cell
Network information
(Input)
Direct network control
(Output)
SDN-enabled virtualization controller
Input:
(1) The set of end devices in the coverage of each BS;
(2)Traffic statistics for end devices;
(3) Wireless channel conditions
Radio resource slicing optimization:
By solving (P1) to obtain the set of optimal network parameters
$\theta^*_{\rm m}, \theta^*_{\rm a}, \{x^*_{i,s,k}\}$
Output:
The optimal sets of bandwidth resources on the MBS and each SBS
MT
MBS
SBS
Wireless communication link
Smart-homing MTD
Industrial MTD
E-healthcare MTD
Monitoring MTD

Fig. 1: The dynamic radio resource slicing framework for a two-tier small-cell underlaid HetNet with the coexistence of MTDs and MTs.

Step 3 – The virtualization controller allocates the optimal sets of bandwidth resources, $\theta^*_{\rm m} B_{\rm v}$ and $\theta^*_{\rm a} B_{\rm v}$, to the MBS and SBSs, respectively, which are further partitioned into resource sub-slices for different groups of end devices subscribing services from different SPs under the coverage of each BS.

III. Bi-Resource Slicing for Core Network

A. SFC Embedding

Traffic flows aggregated from wireless networks through backhaul links represent different types of services and need to pass through different logic SFCs, consisting of a sequence of VNFs and the virtual links connecting them, to fulfill differentiated QoS requirements. QoS for E2E service deliveries refers to certain performance metrics, e.g., delay, for evaluating packets of a traffic flow passing through a pair of end points in 5G networks. With the SDN-enabled virtualization controller, *SFC embedding* places logic SFCs on selected physical network paths, with VNFs operated on NFV nodes and virtual links represented by physical transmission links and network routers. To improve resource utilization, logic SFCs traversed by multiple traffic flows can be embedded on a common physical network path sharing a set of computing resources on NFV nodes and bandwidth resources on transmission links and routers. In Fig. 2, we illustrate two traffic flows, x and y, requiring different logic SFCs, traverse one embedded physical path to fulfill E2E service requirements. Packets of flow x go through the first VNF (a firewall function) F_1 on NFV node V_1 for processing, and are transmitted on the outgoing link L_0 of V_1. They are then forwarded by a set of transmission links $\{L_1, L_2, ..., L_{n_1}\}$ and network routers $\{S_1, S_2, ..., S_{n_1}\}$ before arriving at destination VNF F_3, a domain name system (DNS) function, on the NFV node V_2. On the other hand, packets of flow y follow the same embedded physical path to traverse F_1 on V_1, and then F_2, an intrusion detection system (IDS) function operated on the same NFV node V_2 as flow x. In a general scenario, a set J of traffic flows are embedded on a common physical network path, passing through a sequence of m NFV nodes $\{V_1, V_2, ..., V_m\}$ and n_u pairs of transmission links and network routers between consecutive NFV nodes $V_u\,(u < m)$ and V_{u+1} before reaching the destination node in the core network.

B. Bottleneck Resources

As discussed, when a traffic flow traverses an NFV node, each packet of the flow consumes CPU time for packet processing and then occupies the outgoing link bandwidth resources for transmission, whereas in the wireless network domain the main function is radio access for wireless transmission, and processing is relatively insignificant. We define *resource profiles* for flow $x\,(\in J)$ traversing an NFV node as a two-dimensional time vector, $[t_{x,1}, t_{x,2}]$, indicating two time durations consumed sequentially by one packet of flow x for CPU processing and link transmission, if all CPU time and link bandwidth resources on the NFV node are allocated to the flow. We also define a two-dimensional rate profiles, $[R_{x,1}, R_{x,2}]$, as the reciprocal of the corresponding resource profiles, indicating maximum achievable rates for processing and transmitting packets. Different service flows

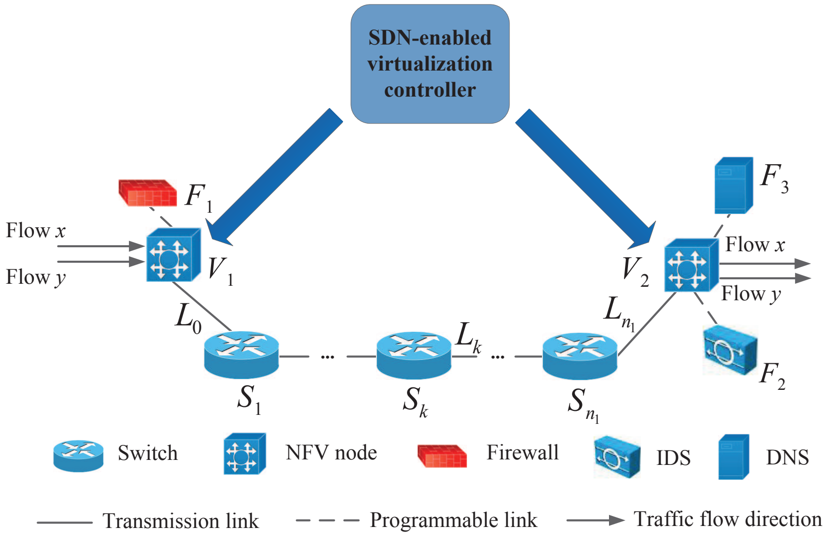


Fig. 2: Traffic flows traversing embedded SFCs in the core network.

have discrepant resource profiles for CPU processing and link transmission when passing through an NFV node. For example, a short packet with large packet header (e.g., a DNS request packet) requires more time for CPU processing than that for link transmission, whereas a long packet with small header (e.g., a video data packet) occupies more time for link transmission. We define *bottleneck resource* at an NFV node as the resource type that a packet of each traffic flow requires more time for either CPU processing or link transmission.

C. Bi-Resource Slicing

When multiple traffic flows traverse an NFV node, both CPU and link bandwidth resources need to be sliced properly and allocated to each flow to achieve high resource utilization with fair resource allocation among flows. Under the assumption that resources are infinitely divisible, Generalized processor sharing (GPS) is a benchmark fluid-flow based resource scheduling scheme in traditional communication networks supporting differentiated services [13]. With GPS, each traffic flow, say flow $x (\in J)$, multiplexing at a common GPS server, such as a network router or a transmission link, is assigned a positive weighting value ψ_x. Flow x is thus guaranteed a minimum service rate, $\frac{\psi_x}{\sum\limits_{x \in J} \psi_x} R$, if all flows at the server have backlogged packets to transmit, where R is the maximum packet service rate of the GPS server. When some of the flows have empty transmission queues, their allocated transmission rates are re-distributed among the remaining backlogged flows to exploit the traffic multiplexing gain. The GPS has properties of achieving QoS isolation among flows and improving single-resource utilization.

When the GPS is applied directly to multiple flows with bi-resource consumption at an NFV node, it is difficult to achieve high performance in both packet processing and packet transmission and to maintain a fair resource usage among flows, since traffic flows have discrepant bottleneck consumption on the two resource types. Suppose we have two equally-weighted flows x and y traversing firewall function F_1 at V_1 as shown in Fig. 2. The two flows have resource profiles $[t_{x,1}, t_{x,2}]$ and $[t_{y,1}, t_{y,2}]$ respectively, with bottleneck resource consumption on different resource types, i.e., $t_{x,1} > t_{x,2}$ and $t_{y,1} < t_{y,2}$. The following resource slicing policies can be considered:

1) Bi-resource GPS: When both flows are backlogged, the fractions of CPU and bandwidth resources allocated to flow x and flow y are equalized (applying GPS on both resource types), i.e., $f_{x,i} = f_{y,i} = \frac{1}{2}\,(i = 1, 2)$, where $f_{x,i} = \frac{r_{x,i}}{R_{x,i}}$ and $f_{y,i} = \frac{r_{y,i}}{R_{y,i}}$, and $r_{x,i}$ and $r_{y,i}$ denote the allocated packet processing or packet transmission rate to flow x and flow y, respectively. However, because of the discrepancy of the resource profiles, the equal allocation on both CPU and bandwidth resources between the two flows result in unbalanced service rates for packet processing and packet transmission. For flow x, some of the link bandwidth resources are wasted since $r_{x,1} < r_{x,2}$; For flow y, packets are accumulated for link transmission since the allocated processing rate is larger than the transmission rate, i.e., $r_{y,1} > r_{y,2}$, leading to a large packet queueing delay;

2) Single-resource GPS with equalized service rates (applying GPS on one resource type): Consider total packet delay which is the duration from the time instant that a packet of a flow reaches its processing queue for CPU processing at an NFV node till the instant the packet is transmitted through the node's outgoing link. To reduce the total packet delay for both backlogged flows at the NFV node, a basic principle is to allocate the fractions of CPU and bandwidth resources for each flow in proportion to its resource profiles, i.e., $\frac{f_{x,1}}{f_{x,2}} = \frac{t_{x,1}}{t_{x,2}}$, such that the allocated processing and transmission rates can be equalized. However, if we apply GPS on one type of resources while the other type of resources are allocated accordingly for equalized packet processing and transmission rates, the resource usage on the other type is unbalanced between the two flows due to the discrepancy of resource profiles. The characteristics and limitations for both resource slicing policies are summarized in Table I.

To achieve a low packet delay and to maintain a fair allocation on both types of resources, we employ a *bottleneck-*

TABLE I: An evaluation of resource slicing policies

Evaluation Resource slicing policies	Characteristics	Limitations
Bi-resource GPS	Apply GPS on CPU and bandwidth resources: A fair allocation on both resource types	Unbalanced services rates for packet processing and packet transmission among multiple flows
Single-resource GPS with equalized service rates	Apply GPS on one of the resources; The other type of resources are allocated for equalized processing and transmission rates	The usage on the other type of resources among multiple flows is unbalanced

resource GPS (BR-GPS) scheme [8], which combines bottleneck resource fairness [14] with GPS, for bi-resource slicing among multiple flows traversing an NFV node. With BR-GPS, the fractions of bottleneck resource allocation for each backlogged flow are equalized, and the non-bottleneck resources are allocated in proportion to the resource profiles of each flow to guarantee equalized packet processing and transmission rates. When some of the flows have no packets to transmit, their allocated resources are re-distributed among other backlogged flows (one of the properties of GPS). For the preceding example of two backlogged flows x and y at the NFV node V_1, with BR-GPS, the fraction of CPU resources allocated to flow x equalizes the fraction of bandwidth resources allocated to flow y, i.e., $f_{x,1} = f_{y,2}$, and the allocation for the other resource type follow the basic principle to guarantee $r_{x,1} = r_{x,2}$ and $r_{y,1} = r_{y,2}$.

Since each flow requires more resources on its bottleneck resource type, BR-GPS equalizes the bottleneck resource shares among backlogged flows, providing a fair allocation on the more demanding resource type. More importantly, by equalizing the allocated processing and transmission rate, BR-GPS reduces total packet delay for each flow by minimizing the packet queueing delay at the outgoing link of an NFV node. With the properties of GPS, the BR-GPS achieves service isolation by guaranteeing each backlogged flow minimum fractions of CPU and bandwidth resources, and achieves high resource utilization by traffic multiplexing [8].

IV. Open Research Issues

There are open research issues on network slicing for a 5G network:

QoS-Aware Radio Resource Slicing – The 5G network will support diversified types of M2M services with ultra-high reliability and critical latency requirements. Moreover, traffic arrival statistics for different use cases are highly diverse with a combination of deterministic and bursty characteristics. In one constraint of $(\mathrm{P1})$, we use the minimum rate requirement for each service as a coarse QoS description. However, data services and M2M services have differentiated QoS indicators. An M2M service having bursty traffic arrivals requires every packet being transmitted within a stringent delay bound. Therefore, to properly slice the resources among BSs for fine-grained heterogeneous QoS satisfaction, it is required to determine the amount of resources for each downlink transmission from a BS to an end-device. Effective bandwidth (capacity) theory [10] is a potential approach to explore appropriate resource-QoS mapping with specific traffic modeling for each service;

Cost-Effective Radio Resource Slicing – In the proposed radio resource slicing framework, the virtualization controller can dynamically adjust the amount of bandwidth resources at each BS to maximize the overall resource utilization and network utility. However, the global network information (i.e., end-device locations, number of devices in each cell, instantaneous wireless channel conditions between a BS and an associated end device) is required by the controller to determine the optimal bandwidth slicing ratios. Therefore, each BS needs to collect and update the network information to the controller through control links, which inevitably incurs communication overhead for the bi-directional control information exchange between the controller and the BSs. If the network information is updated more frequently, the controller makes better decisions for bandwidth slicing to improve the network utility, at the cost of a higher communication overhead. Therefore, how to maximize the radio resource slicing gain by considering the communication cost is a potential research topic;

Delay-Aware SFC Embedding – We employ BR-GPS as a bi-resource slicing scheme at each NFV node in the core network, and identify its property of minimizing packet queueing delay at the outgoing link of an NFV node. However, to make the SFC embedding delay-aware, an E2E delay analysis is required for packets of multiple traffic flows from different logic SFCs traversing an embedded physical network path with BR-GPS at each NFV node. The E2E delay analysis is technically challenging. With BR-GPS as the bi-resource slicing scheme among flows, both packet processing rate and transmission rate allocated to one flow depend on the backlog status of other flows at the same NFV node. This coupling effect makes packet queueing modeling difficult for each flow passing through the NFV node, including packet processing and packet transmission at the outgoing link; Further, each flow traverses a sequence of NFV nodes, physical links and routers before reaching the destination node. However, packet arrival process for each flow at a subsequent NFV node is correlated with packet service process at its preceding NFV node. This dependency makes the modeling of tandem queues [15] not accurate for E2E packet delay analysis. Therefore, how to remove the coupling effect of service processes among different flows at one NFV node and how to model the E2E delay for packets passing through a sequence of NFV nodes need further investigation.

V. A Case Study

In this section, computer simulations are conducted to evaluate the performance of radio resource slicing and bi-resource slicing in both wireless and core networks. For the wireless network domain, a two-tier HetNet is considered, with a macro-cell of 600 m communication radius underlaid by four small cells of 200 m communication radius. An MBS and four SBSs are located in corresponding cell centers, with downlink transmit powers set to 40 dBm and 30 dBm, respectively. The distance between the MBS and each of the SBSs are set to 400 m. The preallocated spectrum bandwidth at all the BSs is 10 MHz. All MTs and MTDs are randomly scattered in the HetNet coverage, with the same number of MTDs in all the small cells. There are two SPs in the network, where MTs belong to one SP providing a data service and all MTDs are subscribed to an M2M service. Downlink packet arrivals at each transmission queue of a BS destined for an MTD are modeled as a Poisson process with rate of 5 packet/s with packet size of 2000 bits, and packet arrivals destined for an MT are periodic with rate of 20 packet/s with packet size of 9000 bits. For the core network, we consider two service flows, x and y, representing two logic SFCs from F_1 to F_3 and from F_1 to F_2 respectively, traverse one embedded physical network path, as shown in Fig. 2. The packet arrival rate at V_1 of flow x is 150 packet/s with packet size of 4000 bits for DNS function. We vary the packet arrival rate of flow y from 150 packet/s to 350 packet/s with packet size of 16000 bits for video-conferencing. The rate profiles for flows x and y traversing firewall function F_1 on V_1 are $[1000, 2000]$ packet/s and $[750, 500]$ packet/s, respectively.

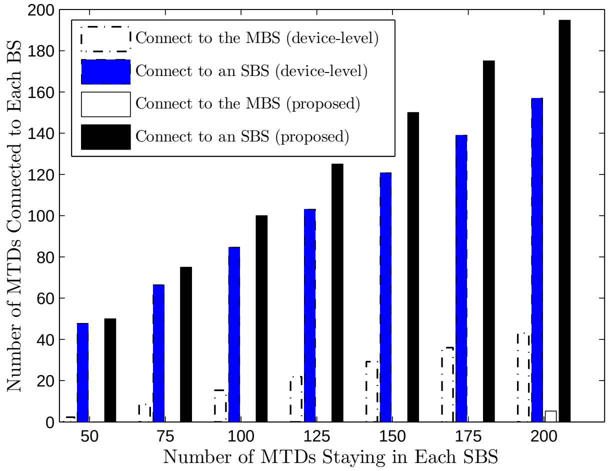


Fig. 3: BS-device association patterns for different resource slicing schemes.

Fig. 3 and Fig. 4 show the optimal BS-device association patterns and the optimal bandwidth slicing ratio on each SBS for different resource slicing schemes, where the number of MTs and MTDs connected to the MBS is 100. We can see that for the device-level bandwidth slicing scheme [9] where the amount of bandwidth resources on each BS are preallocated, more and more MTDs located in an SBS are offloaded to the MBS with an increased device number. Therefore, each end device needs to frequently change its network association pattern in network load dynamics, causing an increased communication overhead for wireless connection re-association. In contrast, for the proposed bandwidth slicing framework, the bandwidth resources on each BS are dynamically adjusted according to network load conditions and end devices maintain stable network associations with the BSs, which significantly reduce the wireless connection re-association cost. In Fig. 4, it is observed that the optimal bandwidth slicing ratio on each SBS is adapted to instantaneous network load to improve overall network resource utilization, whereas the bandwidth resources on each BS are fixed for the device-level slicing scheme.

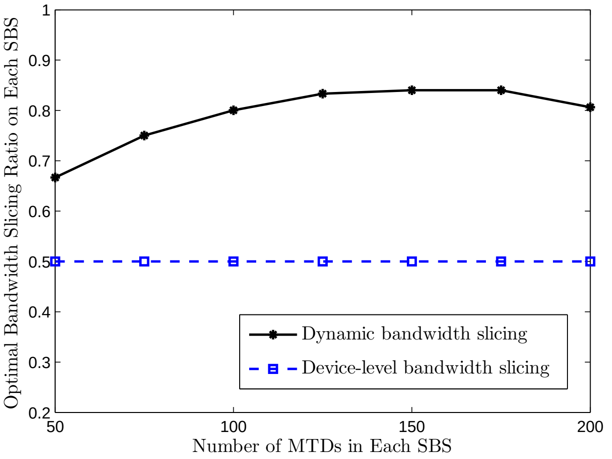


Fig. 4: Optimal bandwidth slicing ratios under different network load conditions.

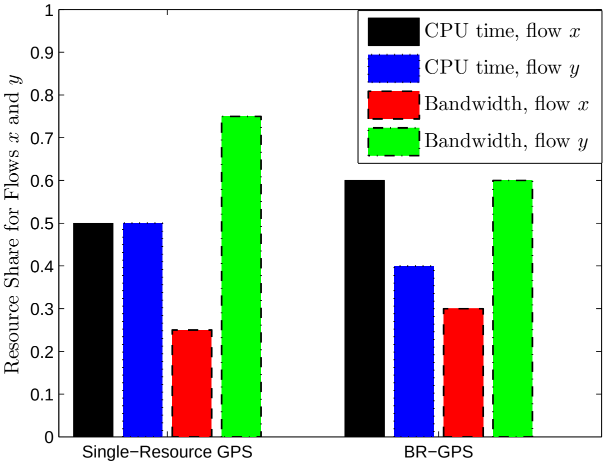


Fig. 5: Fractions of allocated resources to flows x and y under different resource slicing schemes.

For the bi-resource slicing at NFV node V_1 where the bottleneck resource types for flow x and flow y are CPU and bandwidth resources respectively, Fig. 5 shows the resource share for both flows based on BR-GPS. The BR-GPS equalizes the fractions of allocated bottleneck resources between flows x and y. The CPU processing rate is also equalized with the link transmission rate for each flow. Therefore, the BR-GPS achieves a bottleneck-resource fair allocation with high resource utilization between the flows. In contrast, using the single-resource GPS with equalized processing and transmission rates leads to an unbalanced bandwidth resource usage between the flows. We compare the BR-GPS and the bi-resource GPS in Fig. 6 in terms of packet queueing delays for both flows at outgoing transmission link of V_1. Since the service rates of CPU processing and link transmission are equalized in the BR-GPS, the queueing delays are minimal for both flows as the packet arrival rate of flow y varies. In the bi-resource GPS scheme, the required link bandwidth resource for flow y is not satisfied, leading to an increased packet queueing delay at the outgoing link for flow y.

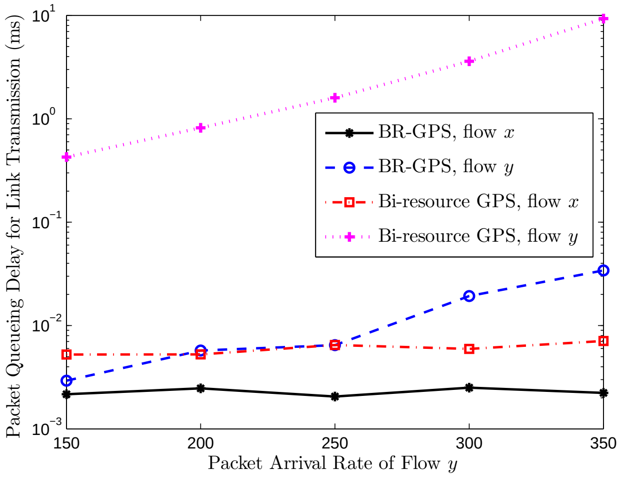


Fig. 6: Packet queueing delay for link transmission under different resource slicing schemes.

VI. Conclusion

In this article, we present a network slicing framework for both wireless and wired domains in a 5G network. Through SDN-enabled NFV technology, a dynamic radio resource slicing scheme is proposed for a HetNet, in which radio spectrum resources are partitioned into resource slices and allocated to heterogeneous BSs. The amount of resources for each BS are dynamically adjusted according to instantaneous network load conditions for improving the overall resource utilization over the device-level slicing scheme where bandwidth resources on each BS are preallocated. A network utility maximization problem is formulated to determine the set of optimal bandwidth slicing ratios between the macro-cell and small cells. For the core network, the BR-GPS is used for bi-resource slicing to achieve bottleneck-resource fairness among multiple flows traversing each NFV node of embedded SFCs. With BR-GPS, packet queueing delays for multiple flows at the outgoing link of an NFV node are reduced. Some potential research issues regarding network slicing are then discussed. Simulation results in a case study demonstrate the effectiveness of the proposed network slicing framework for the 5G network.

References

[1] A. Ksentini and N. Nikaein, "Toward enforcing network slicing on RAN: Flexibility and resources abstraction," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 102–108, Jun. 2017.

[2] M. Mu, M. Broadbent, A. Farshad, N. Hart, D. Hutchison, Q. Ni, and N. Race, "A scalable user fairness model for adaptive video streaming over SDN-assisted future networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 8, pp. 2168–2184, Aug. 2016.

[3] J. Ordonez-Lucena, P. Ameigeiras, D. Lopez, J. J. Ramos-Munoz, J. Lorca, and J. Folgueira, "Network slicing for 5G with SDN/NFV: Concepts, architectures, and challenges," *IEEE Commun. Mag.*, vol. 55, no. 5, pp. 80–87, May 2017.

[4] N. M. M. K. Chowdhury, M. R. Rahman, and R. Boutaba, "Virtual network embedding with coordinated node and link mapping," in *Proc. IEEE INFOCOM' 09*, 2009, pp. 783–791.

[5] S. Vassilaras, L. Gkatzikis, N. Liakopoulos, I. N. Stiakogiannakis, M. Qi, L. Shi, L. Liu, M. Debbah, and G. S. Paschos, "The algorithmic aspects of network slicing," *IEEE Commun. Mag.*, vol. 55, no. 8, pp. 112–119, Aug. 2017.

[6] O. Sallent, J. Perez-Romero, R. Ferrus, and R. Agusti, "On radio access network slicing from a radio resource management perspective," *IEEE Wireless Commun.*, vol. 24, no. 5, pp. 166–174, Oct. 2017.

[7] A. Ghodsi, V. Sekar, M. Zaharia, and I. Stoica, "Multi-resource fair queueing for packet processing," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 42, no. 4, pp. 1–12, Aug. 2012.

[8] W. Wang, B. Liang, and B. Li, "Multi-resource generalized processor sharing for packet processing," in *Proc. ACM IWQoS*, 2013, pp. 1–10.

[9] C. Liang, F. R. Yu, H. Yao, and Z. Han, "Virtual resource allocation in information-centric wireless networks with virtualization," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 9902–9914, Dec. 2016.

[10] A. Aijaz, M. Tshangini, M. R. Nakhai, X. Chu, and A. H. Aghvami, "Energy-efficient uplink resource allocation in LTE networks with M2M/H2H co-existence under statistical QoS guarantees," *IEEE Trans. Commun.*, vol. 62, no. 7, pp. 2353–2365, Jul. 2014.

[11] L. Li, M. Pal, and Y. R. Yang, "Proportional fairness in multi-rate wireless LANs," in *Proc. IEEE INFOCOM' 08*, 2008.

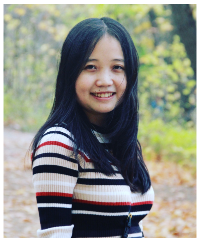
[12] J. Gorski, F. Pfeuffer, and K. Klamroth, "Biconvex sets and optimization with biconvex functions: a survey and extensions," *Math. Method. Oper. Res.*, vol. 66, no. 3, pp. 373–407, Dec. 2007.

[13] A. K. Parekh and R. G. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: The single-node case," *IEEE/ACM Trans. Netw.*, vol. 1, no. 3, pp. 344–357, Jun. 1993.

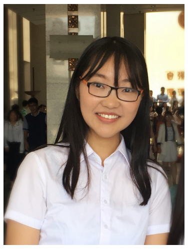
[14] A. Ghodsi, M. Zaharia, B. Hindman, A. Konwinski, S. Shenker, and I. Stoica, "Dominant resource fairness: Fair allocation of multiple resource types," in *Proc. ACM NSDI' 11*, 2011, pp. 24–37.

[15] D. P. Bertsekas, R. G. Gallager, and P. Humblet, *Data networks*. Englewood Cliffs, NJ, USA: Prentice-hall, 1987, vol. 2.

Qiang Ye (S'16-M'17) received the B.S. degree in network engineering and M.S. degree in communication and information system from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2009 and 2012, respectively, and the Ph.D. degree in electrical and computer engineering from the University of Waterloo, Waterloo, ON, Canada, in 2016. He has been a Post-Doctoral Fellow with the Department of Electrical and Computer Engineering, University of Waterloo, since 2016. His current research interests include SDN and NFV, network slicing for 5G networks, VNF chain embedding and end-to-end performance analysis, medium access control and performance optimization for mobile ad hoc networks and Internet of Things.



Junling Li (S'18) received the B.S. degree from Tianjin University, Tianjin, China, in 2013 and received the M.S. degree from Beijing University of Posts and Telecommunications, Beijing, China, in 2016. She is currently working toward the Ph.D. degree with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada. Her interests include software defined networking, network function virtualization, and vehicular networks.



Kaige Qu received the B.S. degree from Shandong University, Jinan, China, in 2013 and received dual M.S. degrees from Tsinghua University, Beijing, China and University of Leuven (KU Leuven), Leuven, Belgium, in 2016. She is currently working toward the Ph.D. degree with the Department of Electrical and Computer Engineering, University of Waterloo. Her interests include resource allocation and traffic engineering in software defined 5G networks with network function virtualization.

Weihua Zhuang (M'93-SM'01-F'08) has been with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada, since 1993, where she is a Professor and a Tier I Canada Research Chair in Wireless Communication Networks. She is the recipient of 2017 Technical Recognition Award from IEEE Communications Society Ad Hoc & Sensor Networks Technical Committee, one of 2017 ten N2Women (Stars in Computer Networking and Communications), and a co-recipient of several best paper awards from IEEE conferences. Dr. Zhuang was the Editor-in-Chief of IEEE Transactions on Vehicular Technology (2007-2013), Technical Program Chair/Co-Chair of IEEE VTC Fall 2017 and Fall 2016, and the Technical Program Symposia Chair of the IEEE GLOBECOM 2011. She is a Fellow of the IEEE, the Royal Society of Canada, the Canadian Academy of Engineering, and the Engineering Institute of Canada. Dr. Zhuang is an elected member in the Board of Governors and VP Publications of the IEEE Vehicular Technology Society. She was an IEEE Communications Society Distinguished Lecturer (2008-2011).

Xuemin (Sherman) Shen (M'97-SM'02-F'09) received Ph.D. degrees (1990) from Rutgers University, New Jersey (USA). Dr. Shen is a University Professor, Department of Electrical and Computer Engineering, University of Waterloo, Canada. His research focuses on resource management in interconnected wireless/wired networks, wireless network security, social networks, smart grid, and vehicular ad hoc and sensor networks. Dr. Shen served as the Technical Program Committee Chair/Co-Chair for IEEE Globecom'16, Infocom'14, IEEE VTC'10 Fall, and Globecom07, the Symposia Chair for IEEE ICC'10. He also serves as the Editor-in-Chief for IEEE Internet of Things Journal, Peer-to-Peer Networking and Application, and IET Communications; a Founding Area Editor for IEEE Transactions on Wireless Communications. Dr. Shen received the Excellent Graduate Supervision Award in 2006, and the Outstanding Performance Award in 2004, 2007, 2010, and 2014 from the University of Waterloo, the Premiers Research Excellence Award (PREA) in 2003 from the Province of Ontario, Canada, the Distinguished Performance Award in 2002 and 2007 from the Faculty of Engineering, University of Waterloo, the Joseph LoCicero Award and the Education Award 2017 from the IEEE Communications Society. Dr. Shen is a registered Professional Engineer of Ontario, Canada, an IEEE Fellow, an Engineering Institute of Canada Fellow, a Canadian Academy of Engineering Fellow, a Royal Society of Canada Fellow, and a Distinguished Lecturer of IEEE Vehicular Technology Society and Communications Society.

Xu Li is a staff researcher at Huawei Technologies Inc., Canada. He received a Ph.D. (2008) degree from Carleton University, an M.Sc. (2005) degree from the University of Ottawa, and a B.Sc. (1998) degree from Jilin University, China, all in computer science. Prior to joining Huawei, he worked as a research scientist (with tenure) at Inria, France. His current research interests are focused in 5G system design and standardization, along with 90+ refereed scientific publications, 40+ 3GPP standard proposals and 50+ patents and patent filings. He is/was on the editorial boards of the IEEE Communications Magazine, the IEEE Transactions on Parallel and Distributed Systems, among others. He was a TPC co-chair of IEEE VTC 2017 (fall) LTE, 5G and Wireless Networks Track, IEEE Globecom 2013 Ad Hoc and Sensor Networking Symposium. He was a recipient of NSERC PDF awards, IEEE ICNC 2015 best paper award, and a number of other awards.