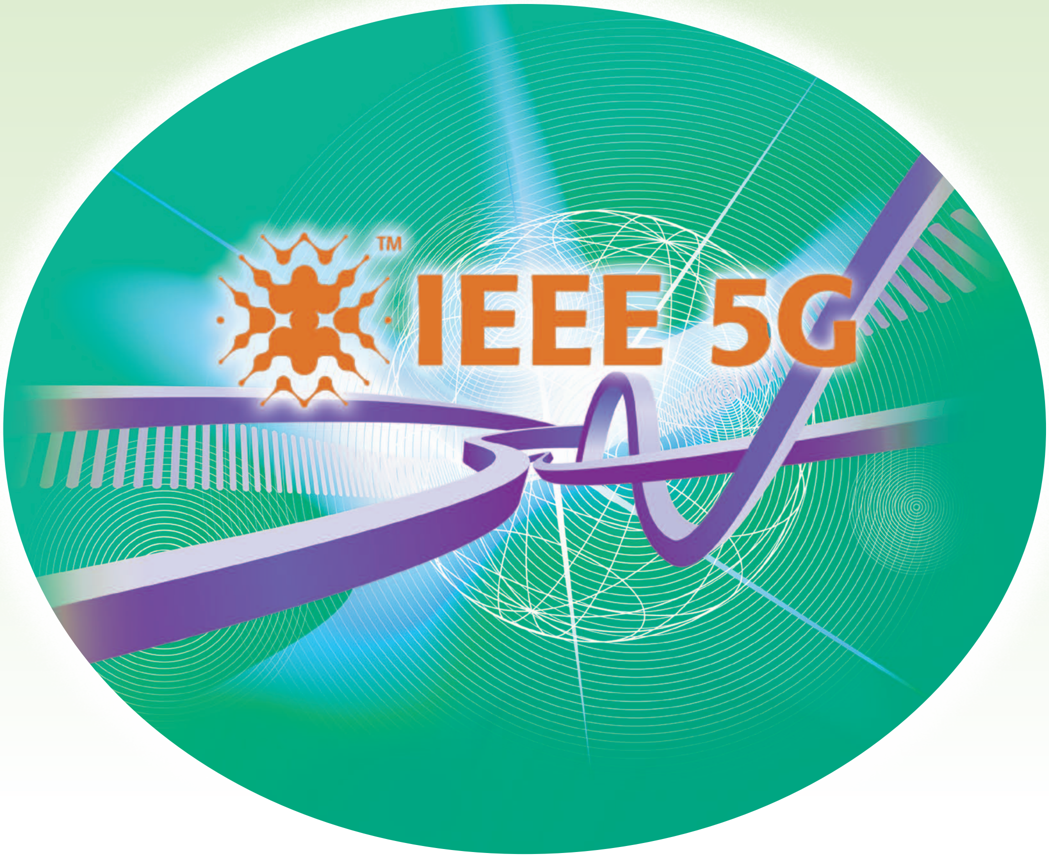
End-to-End Quality of Service in 5G Networks



©ISTOCKPHOTO.COM/DIRTYDOG CREATIVE

Examining the Effectiveness of a Network Slicing Framework

Qiang Ye, Junling Li, Kaige Qu, Weihua Zhuang,
Xuemin (Sherman) Shen, and Xu Li

With software-defined networking (SDN) and network function virtualization (NFV) technologies, network slicing is a promising solution for resource orchestration to achieve quality-of-service (QoS) isolation in customized services in fifth-generation (5G) networks. In this article, we propose a comprehensive network slicing framework for end-to-end (E2E) QoS provisioning, with differentiated resource types in both wireless and wired network domains considered.

Digital Object Identifier 10.1109/MVT.2018.2809473
Date of publication: 20 April 2018

For the wireless network domain, a dynamic radio resource slicing scheme is proposed in which the overall bandwidth resources are sliced for different base stations (BSs) to maximize network utility. The optimal bandwidth slicing ratios are dynamically adjusted based

1556-6072/18©2018IEEE

on instantaneous network load conditions. For the wired network domain, bottleneck-resource generalized processor sharing (BR-GPS) is used as a biresource slicing scheme in multiple traffic flows traversing in an NFV node. In addition to the property of BR fair allocation with high resource utilization, the BR-GPS minimizes the packet queuing delay for each flow at the outgoing link of the NFV node. This article discusses open research problems regarding network slicing and presents a case study demonstrating the effectiveness of the proposed network slicing framework.

5G Communication Networks

5G communication networks are expected to guarantee QoS E2E service deliveries for Internet of Things (IoT) devices. These devices might include intelligent home appliances, smart sensors, and actuators supporting diversified use cases; the applications include smart homing, industrial automation, intelligent transportation, and e-health-care systems. In recent technical reports, the Third-Generation Partnership Project identifies three main features for the future networking paradigm [1].

1) *Enhanced mobile broadband:* The network deployment will be highly densified to provide 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 gigabyte per second peak data rate). A multitier hierarchical network cell deployment (i.e., small cells underlaying macrocells) is envisioned for an enlarged network coverage of wireless access networks. The number of network routers, computational powerful servers, and physical links with high transmission bandwidth is increased in the wired core network to accommodate high traffic volume and respond to service requests in a timely manner.
2) *Massive IoT:* A large number of heterogeneous IoT devices can be interconnected to support various types of services. To accommodate the massive network access and support efficient E2E packet transmissions, the network capacity must be boosted by further improving the utilization of both communication and computing resources.
3) *Critical communications:* The 5G networks will support diversified applications with differentiated QoS requirements. Some time-critical machine-to-machine (M2M) communications require ultrahigh 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, which ensures that the minimum level of QoS experienced by the devices (or users) belonging to one type of service is not violated when network states change (e.g., device mobility, varying channel conditions, and traffic load fluctuations) at another service type.

5G networks pose challenges to the evolving architecture for both wireless and wired network domains. In the wireless network domain, to provide wide area network coverage and accommodate access from machine-type devices, current radio spectrum utilization needs to be significantly improved. Hence, multitier small-cell BSs (SBSs) are deployed that underlay the coverage of macrocell BSs (MBSs) to exploit any spatial multiplexing gain. However, the increasingly densified network deployment will expand both the capital expenditure (CapEx) and the operational expenditure (OpEx) of the communication infrastructures and aggravate intercell 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, a number of physical transmission links, and network routers before reaching its destination. Packets of each traffic flow consume computing resources, i.e., central processing unit (CPU) time, when traversing network servers and occupy 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, leading to increased CapEx and OpEx for deploying more function-specific servers.

Network Slicing

NFV can be used to achieve high utilization of both communication and computing resources and minimize the infrastructure deployment cost. NFV is then a cost-effective solution in which network functions become software solutions, i.e., virtual network functions (VNFs) are decoupled from the physical substrate network and placed in software-programmable commodity servers called *NFV nodes* that run 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 (vCPU) cores and allocated to different VMs hosting VNFs. All VNFs are centrally controlled by a virtualization controller and can be flexibly placed on NFV nodes in different network locations, a process known as *VNF embedding*.

For each service, specific sets of VNFs and the virtual links connecting them form a logic service function chain (SFC), which is then embedded on the substrate network to achieve a desired cost-performance tradeoff. Each VNF is operated on an NFV node, and each virtual link represents a sequence of transmission links and network routers [4], [5]. 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 controller is 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 appear 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 those that compute on NFV nodes and bandwidth resources on transmission links, should be properly sliced among traffic flows so that QoS isolation can be achieved. This process is called *network slicing*. In the core network, network slicing is interpreted as biresource 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 determines the amount of radio resources allocated to each BS to improve the overall spectrum utilization. Network slicing in the wireless domain is called *radio resource slicing*, which determines how to slice the overall radio resources for different device groups to ensure QoS isolation.

Several studies present new architectures for network slicing in either a wireless [1], [6], [7] or a core network [3]. However, there is limited research on how to determine the resources for different services to achieve the desired tradeoff between high resource utilization and QoS isolation and on how network slicing should be conducted for wireless and wired network domains considering heterogeneous resources.

We propose a network slicing framework for both wireless and core networks. For heterogeneous wireless access networks (HetNets), we investigate how to determine the slicing ratios of radio resources at each BS. To exploit the resource multiplexing gain, the number of resources of each slice is dynamically adjusted according to changes in network conditions. In the core network, each service flow must traverse a specific sequence of VNFs and virtual links, representing the logic SFC, to fulfill certain E2E service requirements. Different logic SFCs can be embedded on a common physical network path and share a set of CPU and bandwidth resources to exploit the traffic multiplexing gain.

Since traffic flows traversing an NFV node demonstrate bottleneck-resource consumption on different resource types [8], [9], we study how biresources are sliced among flows passing through a common NFV node to achieve both high resource utilization and fair resource usage. We evaluate how the biresource slicing performance improves the packet queuing 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.

Radio Resource Slicing for HetNets

In wireless HetNets, a multitier of SBSs is deployed, underlaying an MBS to explore the spatial multiplexing gain of the current spectrum. However, the SBSs increase the CapEx and OpEx and intensify intercell interference. Moreover, the dynamic and unbalanced traffic load in 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. These slices are then allocated to end devices to enhance the utilization of the current spectrum and provide QoS isolation among diverse services.

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 different cell areas, all radio resources are logically sliced for different SPs but are physically partitioned among end devices. Existing studies focus mainly on device-level resource slicing [1], [7], [10], 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.

Network Architecture

Consider a two-tier downlink HetNet, where an MBS, denoted by M_0, is deployed for a wide area coverage, and SBSs, $\mathcal{M} = \{M_k, k = 1, 2, \ldots, n\}$ (n is the number of small cells), are randomly placed underlaying the coverage of the macrocell to support heterogeneous M2M devices (MTDs) and mobile terminals (MTs) subscribed from different SPs, as shown in Figure 1.

Since all SP subscribers are randomly scattered through the network region, we denote the set of machine-type devices along with its set cardinality, staying in the coverage of $M_k (k = 0, 1, 2, \ldots, n)$ and belonging to SP $s (s = 1, 2, \ldots, S)$, with $\mathcal{N}_{s,k}$ and $N_{s,k}$, where $\mathcal{N}_{s,0}$ and $N_{s,0}$ indicate the set and number of machine-type devices,

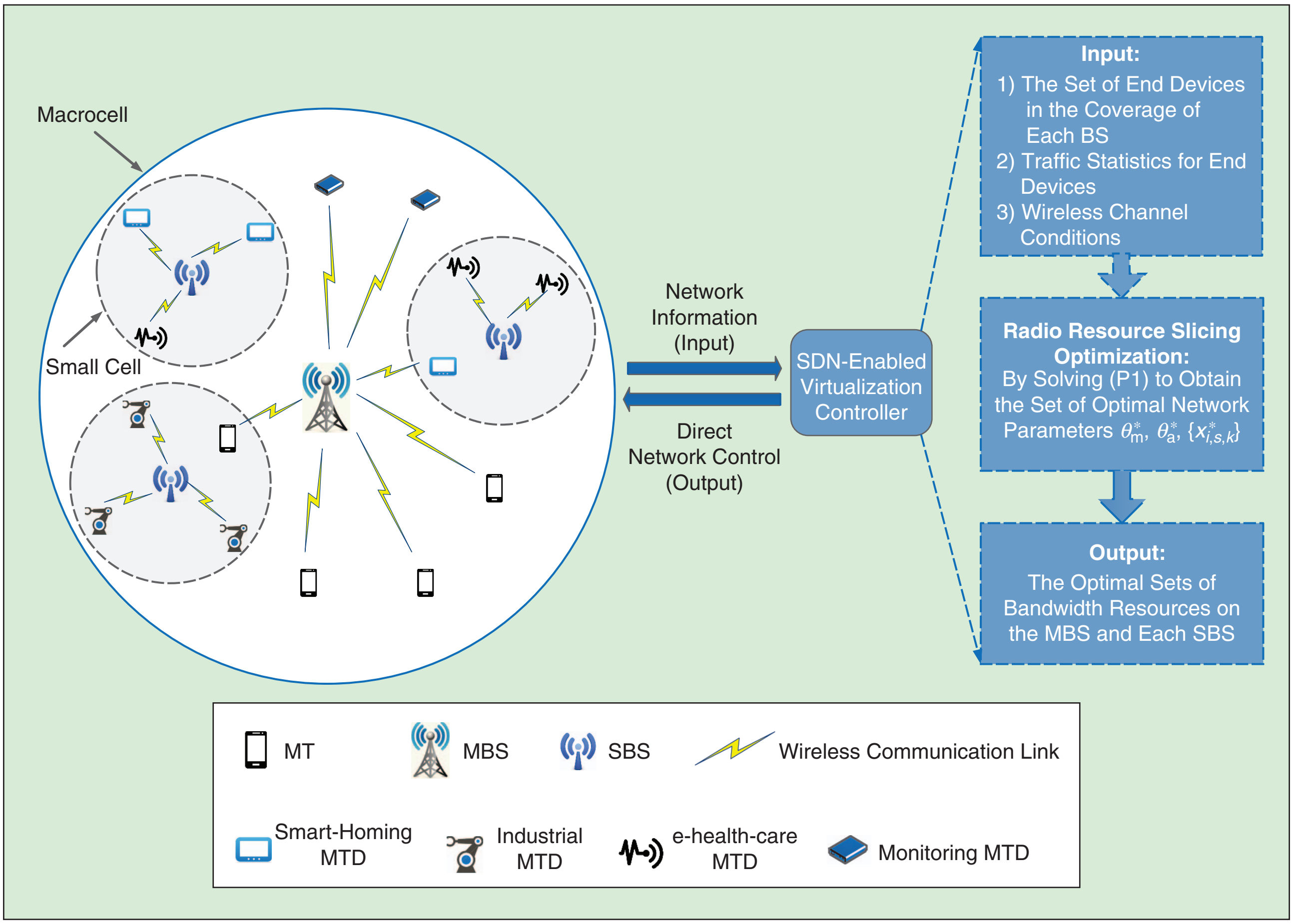

Macrocell
Small Cell
Network Information (Input)
Direct Network Control (Output)
SDN-Enabled Virtualization Controller
Input:
1) The Set of End Devices in the Coverage of Each BS
2) Traffic Statistics for End Devices
3) Wireless Channel Conditions
Radio Resource Slicing Optimization:
By Solving (P1) to Obtain the Set of Optimal Network Parameters θ_m^*, θ_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-health-care MTD
Monitoring MTD

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

residing only in the coverage of the MBS. It should be assumed that all MTs generate one type of data service subscribed from SP 0 and connect to MBS M_0 to avoid frequent handover [11]. 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 may be stationary or 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.

In both cases, the device (or the MT) associates with M_0, giving us $x_{i,s,0} = 1$. Every BS has a number of transmission queues, with each 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, the packet arrival processes at each transmission queue behave differently. Since M2M traffic is often event driven with the capacity to burst, packet arrival at a BS destined for an MTD is modeled as a Poisson process, whereas packet arrivals at a data service destined for an MT are modeled as a periodic packet arrival process.

Optimal Bandwidth Slicing Ratios

Bandwidth resources of the MBS and the SBS are preallocated and denoted by B_m and B_a (m for MBS and a for SBS), respectively, which are mutually orthogonal to avoid the intertier interference. Since SBSs can be physically separated by distance, B_a is reused at each SBS to exploit the spatial multiplexing gain. With SDN-enabled function softwarization, the spectrum bandwidths of the MBS and SBS are abstracted as $B_v (= B_m + B_a)$, divided into two bandwidth slices, $\theta_m B_v$ and $\theta_a B_v$, and reallocated to the MBS and SBS to improve overall resource utilization, where θ_m and θ_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 the SP s staying in M_k can be determined.

Given the transmit power of each BS and the wireless channel conditions (including slow fading, shadowing effects, and intercell 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 $\theta_{\mathrm{m}}, \theta_{\mathrm{a}}, B_{\mathrm{v}}$, and $g_{i,s,k}$. Note that the bandwidth slicing ratios, BS-device association patterns, and fraction of bandwidths allocated to each associated MTD and MT are updated in a large time scale to reduce the communication overhead [12]. For example, bandwidth slicing is updated when traffic loads in each cell. 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 θ_{m}^{*} and θ_{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

$$\text{(P1)}: \max_{\substack{\theta_{\mathrm{m}}, \theta_{\mathrm{a}} \\ x_{i,s,k},\, g_{i,s,k},}} \sum_{k=0}^{n} \sum_{s=0}^{S} \sum_{i \in \mathcal{N}_{s,k}} x_{i,s,k} \mathcal{U}(c_{i,s,k}), \tag{1}$$

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

where $\mathcal{U}(\cdot)$ is a concave utility function with diminished marginal utility (e.g., a logarithm function) for an end device. In (P1), the objective function is to maximize the aggregated 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.

Constraints of (P1) are $\theta_{\mathrm{m}} + \theta_{\mathrm{a}} = 1$ and 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 decision variables, $g_{i,s,k}, \theta_{\mathrm{m}}$ and θ_{a} lie within interval [0,1], and $x_{i,s,k}$ takes on a value of 0 or 1, for any i, s, k. Equation (P1) can be transformed into a biconcave maximization problem and solved for a set of partial optimal solutions [13].

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

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.
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 θ_{m}^{*} and θ_{a}^{*} for both the MBS and SBS and the optimal set of BS–device association patterns $\{x_{i,s,k}^{*}\}$.
3) The virtualization controller allocates the optimal sets of bandwidth resources, $\theta_{\mathrm{m}}^{*} B_{\mathrm{v}}$ and $\theta_{\mathrm{a}}^{*} B_{\mathrm{v}}$, to the MBS and SBS, respectively, which are further partitioned into resource subslices for various groups of end devices subscribing services from different SPs under the coverage of each BS.

Biresource Slicing for Core Network

SFC Embedding

Traffic flows from wireless networks through backhaul links represent different types of services and need to pass through logic SFCs, comprising 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 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 that shares a set of computing resources on NFV nodes and bandwidth resources on transmission links and routers. In Figure 2, two traffic flows, x and y, require different logic SFCs that 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 $\{L_1, L_2, \ldots, L_{n_1}\}$ and network routers $\{S_1, S_2, \ldots, S_{n_1}\}$ 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. Generally, a set J of traffic flow is embedded on a common physical network path, passing through a sequence of m NFV nodes $\{V_1, V_2, \ldots, V_m\}$ 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.

Bottleneck Resources

When a traffic flow traverses an NFV node, each packet of the flow consumes CPU time for packet processing

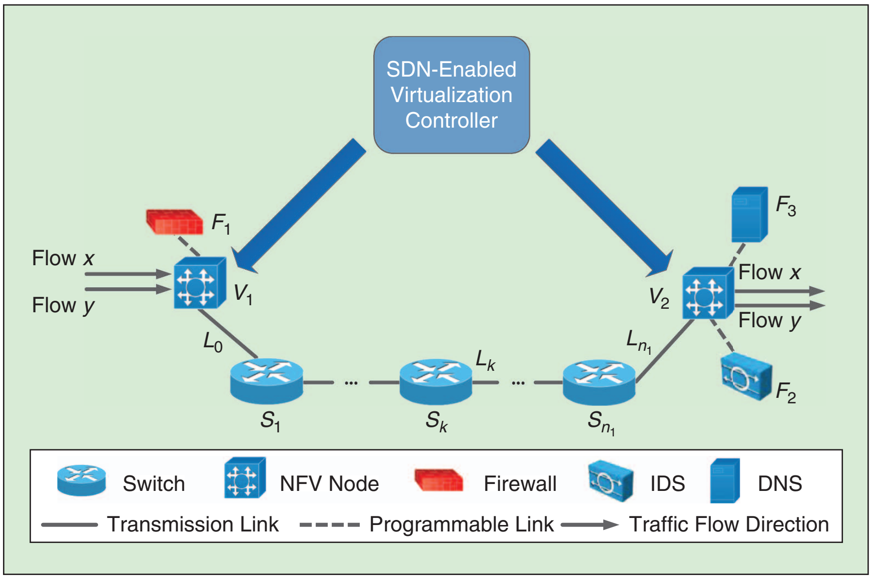


FIGURE 2 The traffic flows traverse the embedded SFCs in the core network.

and occupies any outgoing link bandwidth resources for transmission. In the wireless network domain, the main function is radio access for wireless transmission, making processing relatively insignificant. We define resource profiles for flow $x(\in J)$ traversing an NFV node as a two-dimensional (2-D) 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 2-D rate profile, $[R_{x,1}, R_{x,2}]$, as the reciprocal of the corresponding resource profiles, indicating maximum achievable rates for processing and transmitting packets. Different service flows have discrepant resource profiles for CPU processing and link transmission when passing through an NFV node. For example, a short packet with a large packet header (e.g., a DNS request packet) requires more time for CPU processing than for link transmission; a long packet with a 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 a packet of each traffic flow requires for either CPU processing or link transmission.

Biresource Slicing

When multiple traffic flows traverse an NFV node, both the CPU and link bandwidth resources must 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 allocation scheme in traditional communication networks that support differentiated services [14]. With GPS, each traffic flow, e.g., 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, $\left(\psi_x / \sum_{x \in J} \psi_x\right) R$, 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 flows have empty transmission queues, their allocated transmission rates are redistributed among the remaining backlogged flows to exploit the traffic multiplexing gain. The GPS has properties to achieve QoS isolation among flows and improve single-resource utilization.

When the GPS is applied directly to multiple flows with biresource 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. This is because 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 Figure 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, $t_{x,1} > t_{x,2}$ and $t_{y,1} < t_{y,2}$. The following resource slicing policies can be considered:

1) *Biresource GPS:* When both flows are backlogged, the fractions of CPU and bandwidth resources allocated to flows x and y are equalized, applying GPS on both resource types, i.e., $f_{x,i} = f_{y,i} = \frac{1}{2} (i = 1, 2)$, where $f_{x,i} = r_{x,i} / R_{x,i}$, $f_{y,i} = r_{y,i} / R_{y,i}$, $r_{x,i}$, 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 to both the CPU and bandwidth resources between the two flows results 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, $r_{y,1} > r_{y,2}$, leading to a large packet queuing delay.
2) *Single-resource GPS with equalized service rates:* Consider the total packet delay, which is the duration from the time a packet of a flow reaches its processing queue for CPU processing at an NFV node until 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, allocate the

fractions of CPU and bandwidth resources for each flow in proportion to its resource profiles, i.e., $(f_{x,1}/f_{x,2}) = (t_{x,1}/t_{x,2})$, such that the allocated processing and transmission rates can be equalized. However, if we apply GPS on one type of resource while the other is allocated accordingly for equalized packet processing and transmission rates, the resource usage on the other type is unbalanced between the two flows because of the discrepancy of resource profiles. The characteristics and limitations for both resource slicing policies are summarized in Table 1.

TABLE 1 *An evaluation of resource slicing policies.*

Resource Slicing Policies/ Evaluation	*Characteristics*	*Limitations*
Biresource GPS	Apply GPS on CPU and bandwidth resources; a fair allocation on both resource types	Unbalanced service rates for packet processing and packet transmission among multiple flows
Single-resource GPS with equalized service rates	Apply GPS on one of the resources: other types of resources are allocated for equalized processing and transmission rates	Usage on other types of resources among multiple flows is unbalanced

To achieve a low packet delay and maintain a fair allocation on both types of resources, we use a BR-GPS scheme [9], which combines bottleneck-resource fairness [15] with GPS for biresource 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 nonbottleneck 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 redistributed 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 equals 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 follows 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 to the more demanding resource type. More importantly, by equalizing the allocated processing and the transmission rate, BR-GPS reduces total packet delay for each flow by minimizing the packet queuing delay at the NFV node's outgoing link. With the properties of GPS, the BR-GPS can achieve service isolation by guaranteeing each backlogged flow minimum fractions of CPU and bandwidth resources and can also attain high resource utilization by traffic multiplexing [9].

Open Research Issues

Open research issues on network slicing for 5G networks still exist.

QoS-Aware Radio Resource Slicing

The 5G network will support diversified types of M2M services with ultrahigh reliability and critical latency requirements. Moreover, traffic arrival statistics for different use cases are diverse, with a combination of deterministic and bursty characteristics. In one constraint of (P1), we use the minimum rate requirement for each service as a coarse QoS description. However, data services and M2M services both have differentiated QoS indicators. An M2M service with bursty traffic arrivals requires that every packet be transmitted within a stringent delay bound. Therefore, to properly slice the resources among BSs for fine-grained heterogeneous QoS satisfaction, the number of resources for each downlink transmission from a BS to an end device must be known. Effective bandwidth capacity theory [11] 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 overall resource utilization and network utility. However, the global network information (i.e., end-device locations, number of devices in each cell, and instantaneous wireless channel conditions between a BS and an associated end device) is required by the controller to determine optimal bandwidth slicing ratios. Therefore, each BS needs to collect and update the network information to the controller through control links, which inevitably incurs overhead costs for the bidirectional control information exchange between the controller and the BSs. If the network information is updated more frequently, the controller can make better decisions for bandwidth slicing to improve the network utility at the cost of 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 use BR-GPS as a biresource slicing scheme at each NFV node in the core network and identify its property of minimizing packet queuing delay at the outgoing link of an NFV node. However, to make the SFC embedding delay-aware, an E2E delay analysis is required for

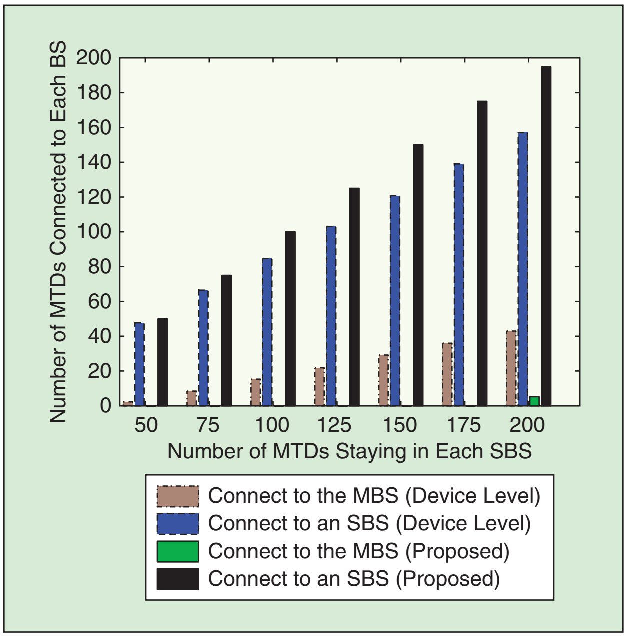


FIGURE 3 The BS-device association patterns occur for different resource slicing schemes.

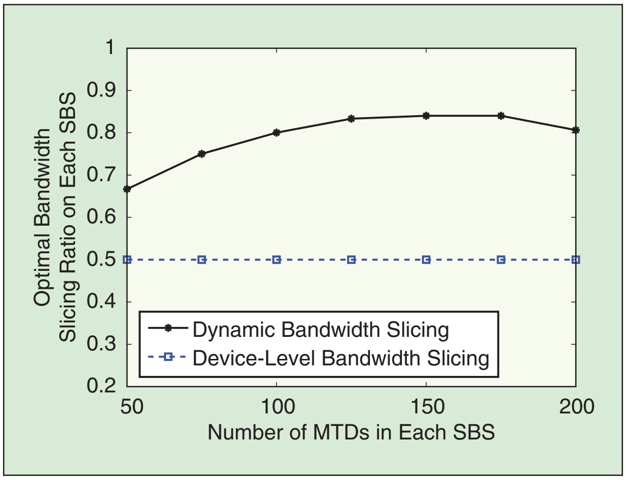


FIGURE 4 The optimal bandwidth slicing ratios under different network load conditions.

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 biresource slicing scheme among flows, both packet processing and transmission rates allocated to one flow depend on the backlog status of other flows at the same NFV node. This coupling effect makes packet queuing modeling difficult for each flow passing through the NFV node, including packet processing and packet transmission at the outgoing link; each flow traverses a sequence of NFV nodes, physical links, and routers before reaching the destination node. However, the packet arrival process for each flow at a subsequent NFV node is correlated with the packet service process at its preceding NFV node. This dependency makes the modeling of tandem queues [16] inaccurate for E2E packet delay analysis. 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.

A Case Study

Computer simulations are conducted to evaluate the performance of radio resource slicing and biresource slicing in both the wireless and core networks. For the wireless network domain, a two-tier HetNet is considered, with a macrocell of 600-m communication radius and an underlay of 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 is set to 400 m. The preallocated spectrum bandwidth at all BSs is 10 MHz. All MTs and MTDs are randomly scattered in the HetNet coverage, with the same number of MTDs in all of 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 a rate of five packets, with a packet size of 2,000 B. Packet arrivals destined for an MT are periodic with a rate of 20 packets, with a packet size of 9,000 B. 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, that traverse one embedded physical network path, as shown in Figure 2.

The packet arrival rate at V_1 of flow x is 150 packets, with a packet size of 4,000 B for the DNS function. We vary the packet arrival rate of flow y from 150 to 350 packets, with a packet size of 16,000 B for video conferencing. The rate profiles for flows x and y traversing the firewall function F_1 on V_1 are $[1{,}000, 2{,}000]$ packets and $[750, 500]$ packets, respectively.

Figures 3 and 4 show the optimal BS-device association patterns and the optimal bandwidth slicing ratio on each SBS for different resource slicing schemes, where 100 MTs and MTDs are connected to the MBS. For the device-level bandwidth slicing scheme [10], where the number of bandwidth resources on each BS is preallocated, 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 increased communication overhead for wireless connection reassociation. 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

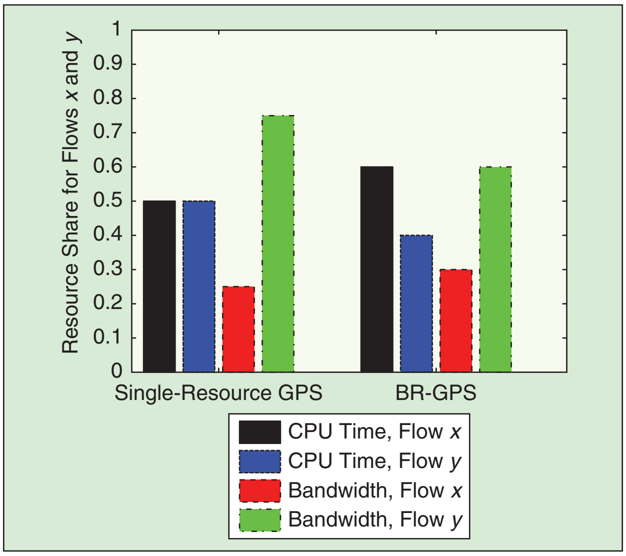


FIGURE 5 The fractions of allocated resources to flows x and y under different resource slicing schemes.

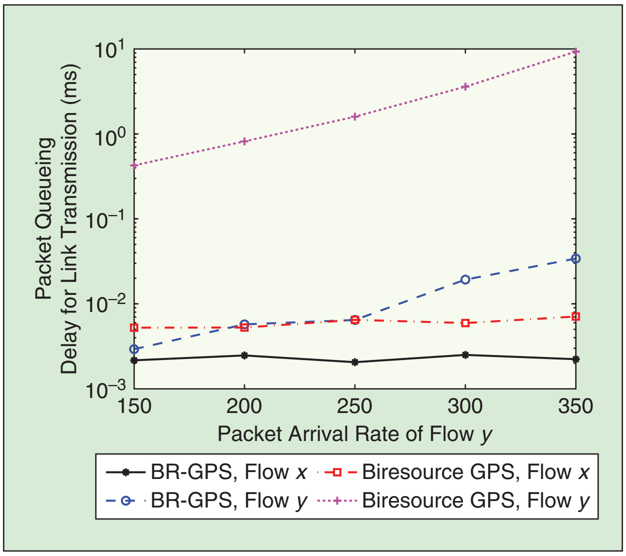


FIGURE 6 The packet queuing delay for the link transmission under different resource slicing schemes.

network associations with the BSs, which significantly reduces the wireless connection reassociation cost. In Figure 4, it is observed that the optimal bandwidth slicing ratio on each SBS is adapted to an instantaneous network load to improve the overall network resource utilization, whereas the bandwidth resources on each BS are fixed for the device-level slicing scheme.

For the biresource slicing at NFV node V_1, where the bottleneck-resource types for flows x and y are CPU and bandwidth resources, respectively, Figure 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 biresource GPS in Figure 6 in terms of packet queuing delays for both flows at the outgoing transmission link V_1. Since the service rates of CPU processing and link transmission are equalized in the BR-GPS, the queuing delays are minimal for both flows because the packet arrival rate of flow y varies. In the biresource GPS scheme, the required link bandwidth resource for flow y is not satisfied, leading to an increased packet queuing delay at the outgoing link for flow y.

Conclusions

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 number of resources for each BS is dynamically adjusted according to the instantaneous network load conditions for improving the overall resource utilization in 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 macrocell and small cells. For the core network, the BR-GPS is used for biresource slicing to achieve bottleneck-resource fairness among multiple flows traversing each NFV node of the embedded SFCs. With the BR-GPS, packet queuing 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.

Author Information

Qiang Ye (q6ye@uwaterloo.ca) received his B.S. degree in network engineering and his M.S. degree in communication and information systems from the Nanjing University of Posts and Telecommunications, China, in 2009 and 2012, respectively, and his Ph.D. degree in electrical and computer engineering from the University of Waterloo, Ontario, Canada, in 2016. He has been a postdoctoral fellow with the Department of Electrical and Computer Engineering, University of Waterloo, since 2016. His current research interests include software-defined networking and network function virtualization, network slicing for fifth-generation networks, virtual network function chain embedding and end-to-end performance

analysis, medium access control and performance optimization for mobile ad hoc networks, and the Internet of Things. He is a Member of the IEEE.

Junling Li (j742li@uwaterloo.ca) received her B.S. degree from Tianjin University, China, in 2013 and her M.S. degree from Beijing University of Posts and Telecommunications, China, in 2016. She is currently working toward her Ph.D. degree with the Department of Electrical and Computer Engineering, University of Waterloo, Ontario, Canada. Her research interests include software-defined networking, network function virtualization, and vehicular networks. She is a Student Member of the IEEE.

Kaige Qu (k2qu@uwaterloo.ca) received her B.S. degree in communication engineering from Shandong University, Jinan, China, in 2013 and her dual M.S. degrees from Tsinghua University, Beijing, China, and the University of Leuven, Belgium, in 2016, both in electrical engineering. She is currently working toward her Ph.D. degree with the Department of Electrical and Computer Engineering, University of Waterloo, Ontario, Canada. Her research interests include resource allocation and traffic engineering in software-defined fifth-generation networks with network function virtualization.

Weihua Zhuang (wzhuang@uwaterloo.ca) is a professor and a Tier I Canada research chair in wireless communication networks with the Department of Electrical and Computer Engineering, University of Waterloo, Ontario, Canada. She is currently an elected member of the Board of Governors and vice president of publications of the IEEE Vehicular Technology Society. She received the 2017 Technical Recognition Award from the IEEE Communications Society Ad Hoc and Sensor Networks Technical Committee, was one of 2017's ten Networking Networking Women (Stars in Computer Networking and Communications), a corecipient of several Best Paper Awards at IEEE conferences, and editor-in-chief of *IEEE Transactions on Vehicular Technology* from 2007 to 2013. She is a Fellow of the IEEE.

Xuemin (Sherman) Shen (sshen@uwaterloo.ca) received his Ph.D. degree from Rutgers University, New Jersey, in 1990. He is currently a professor in the Department of Electrical and Computer Engineering, University of Waterloo, Ontario, Canada, and the editor-in-chief of *IEEE Internet of Things Journal*, *Peer-to-Peer Networking and Applications*, and *Institution of Engineering and Technology Communications* as well as a founding area editor of *IEEE Transactions on Wireless Communications*. He received the Outstanding Performance Award in 2004, 2007, 2010, and 2014 from the University of Waterloo; the Distinguished Performance Award in 2002 and 2007 from the Faculty of Engineering, University of Waterloo; and the 2017 Joseph LoCicero and Education Awards from the IEEE Communications Society. His current research interests include resource management in interconnected wireless/wired networks, wireless network security, social networks, smart grid, and vehicular ad hoc and sensor networks. He is a Fellow of the IEEE.

Xu Li (Xu.LiCA@huawei.com) received his B.Sc. degree from Jilin University, China, in 1998, his M.Sc. degree from the University of Ottawa, Canada, in 2005, and his Ph.D. degree from Carleton University, Ottawa, Ontario, Canada, in 2008, all in computer science. He was a tenured research scientist at Inria, Rocquencourt, France. He is a staff researcher with Huawei Technologies Incorporated, Ottawa, Canada. He is currently or has served on the editorial boards of *IEEE Communications Magazine* and *IEEE Transactions on Parallel and Distributed Systems*. He has more than 90 refereed scientific publications, 40 Third-Generation Partnership Project standard proposals, and 50 patents and patent filings. Among others, he received the Natural Sciences and Engineering Research Council of Canada postdoctoral fellowship and the Best Paper Award from the IEEE Communications and Information Security Symposium in 2015. His current research interests include fifth-generation system design and standardization.

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, June 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*, 2009, pp. 783–791.

[5] J. Li, N. Zhang, Q. Ye, W. Shi, W. Zhuang, and X. Shen, "Joint resource allocation and online virtual network embedding for 5G networks," in *Proc. IEEE Global Communications Conf. (GLOBECOM)*, 2017, pp. 1–6.

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

[11] 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, July 2014.

[12] L. Li, M. Pal, and Y. R. Yang, "Proportional fairness in multi-rate wireless LANs," in *Proc. IEEE Infocom*, 2008, pp. 1678–1686.

[13] 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.

[14] 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, June 1993.

[15] 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 8th USENIX Symp. Networked Systems Design and Implementation*, 2011, pp. 24–37.

[16] D. P. Bertsekas, R. G. Gallager, and P. Humblet, *Data Networks*, vol. 2. Englewood Cliffs, NJ: Prentice Hall, 1987.

VT