This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2018.2859740, IEEE Transactions on Vehicular Technology

Dynamic Radio Resource Slicing for a Two-Tier Heterogeneous Wireless Network

Qiang Ye, *Member, IEEE*, Weihua Zhuang, *Fellow, IEEE*, Shan Zhang, *Member, IEEE*, A-Long Jin, Xuemin (Sherman) Shen, *Fellow, IEEE*, and Xu Li

***Abstract*—In this paper, a dynamic radio resource slicing framework is presented for a two-tier heterogeneous wireless network (HetNet). Through software-defined networking (SDN)-enabled wireless network function virtualization (NFV), radio spectrum resources of heterogeneous wireless networks are partitioned into different bandwidth slices for different base stations (BSs). This framework facilitates spectrum sharing among heterogeneous BSs and achieves differentiated quality-of-service (QoS) provisioning for data service and machine-to-machine (M2M) service in the presence of network load dynamics. To determine the set of optimal bandwidth slicing ratios and optimal BS-device association patterns, a network utility maximization problem is formulated with the consideration of different traffic statistics and QoS requirements, location distribution for end devices, varying device locations, load conditions in each cell, and inter-cell interference. For tractability, the optimization problem is transformed to a biconcave maximization problem. An alternative concave search (ACS) algorithm is then designed to solve for a set of partial optimal solutions. Simulation results verify the convergence property and display low complexity of the ACS algorithm. It is demonstrated that the proposed radio resource slicing framework outperforms the two other resource slicing schemes in terms of low communication overhead, high spectrum utilization, and high aggregate network utility.**

***Index Terms*—5G, heterogeneous wireless networks, NFV, SDN, radio resource slicing, spectrum sharing, data and M2M services, resource utilization, differentiated QoS guarantee.**

I. Introduction

The fifth generation (5G) wireless networks are envisioned to interconnect a massive number of miscellaneous end devices (e.g., smartphones, remote monitoring sensors, and home appliances) generating both mobile broadband data and machine-to-machine (M2M) services/applications (e.g., video conferencing, remote monitoring, and smart homing), to realize the ubiquitous Internet-of-Things (IoT) architecture [1]–[3]. However, the distinctive features of 5G wireless networks, with inherent radio spectrum scarcity, pose technical challenges on the evolving inter-networking paradigm. First, efficient spectrum exploitation is required in response to a surge in network traffic volume and densification of end devices (especially machine-type devices). How to improve current radio resource utilization to accommodate a large expansion of network load is a pivotal and challenging research issue. Multi-tier cell deployment (i.e., a macro-cell underlaid by several tiers of small-cells) is a potential solution to improve the spectrum efficiency, by exploring spatial multiplexing on currently employed spectrum [4]; Moreover, the heterogeneity of both service type and device type necessitates the exploration of unlicensed bands and underutilized spectrum through different wireless access technologies [5]. For example, long-term evolution (LTE) devices can utilize WiFi unlicensed bands to support delay-insensitive services [5].

However, spectrum exploitations face technical challenges. In a hierarchical multi-tier network architecture, physical base stations (BSs), i.e., macro-cell and small-cell BSs (MBSs and SBSs), are mostly owned by different infrastructure providers (InPs) [6], and radio resources are often preallocated at each BS [4]. Spectrum sharing among MBSs and SBSs are limited, due to distributed communication overhead and regulations on heterogeneous InPs [7]. Therefore, to improve the spectrum efficiency and boost network capacity, more and more SBSs are deployed underlaying the MBSs, which substantially increases capital and operational expenses (CapEx and OpEx) on network infrastructures [6], [7] and, at the same time, raises the inter-cell interference level.

Hence, to facilitate spectrum sharing among heterogeneous infrastructures without adding more network deployment cost, network function virtualization (NFV) becomes a promising solution [8], [9]. NFV is a newly-emerging approach originally used in core networks in the Internet, where a set of network/service functions (e.g., firewalls, load balancing, wide area network (WAN) optimizers) are decoupled from the physical hardware and run as software instances in virtual machines [9], [10]. This decoupling of service plane and physical plane removes the heterogeneity of physical infrastructures and facilitates service customization in a software-oriented way. In the wireless domain, network functions in each BS are a composition of radio access and processing functions for establishing wireless connections and allocating radio resources for each associated end device. In wireless NFV, radio access and processing functions run as software instances in heterogeneous BSs (MBSs and SBSs) [11] and are managed by a central controller [7], [12]. Note that the central controller is software defined networking (SDN) enabled [12], [13], which has direct control (programmability) on all BSs and the associated radio resources [14]. With SDN-enabled function softwarization, radio resources of heterogeneous BSs can be reallocated by the central controller to improve the overall resource utilization. This process is referred to as *radio*

Qiang Ye, Weihua Zhuang, Shan Zhang, A-Long Jin, and Xuemin (Sherman) Shen are with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada, N2L 3G1 (emails: {q6ye, wzhuang, s372zhan, along.jin, sshen}@uwaterloo.ca).

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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2018.2859740, IEEE Transactions on Vehicular Technology

resource slicing [15].

Radio resource slicing in the SDN-based wireless NFV framework have three major benefits: (1) Spectrum sharing among heterogeneous wireless infrastructures is achieved in a software-oriented way instead of physically deploying more SBSs with increased CapEx and OpEx; (2) The central controller has global information over the physical network, which facilitates the resource sharing without distributed information exchange; (3) The coexistence of heterogeneous services requires *QoS isolation*[1] among different groups of end devices belonging to different service types. Resource slicing is a promising approach to achieve the QoS isolation by creating resource slices for different service groups.

Most existing radio resource slicing schemes focus on the device level, to allocate a fixed amount of radio resources of a BS to different device groups in the BS coverage. However, the network-level resource sharing and resource slicing among heterogeneous BSs need further studies. As a matter of fact, an essence of network function virtualization is to enable radio resource abstraction for the purpose of managing the resources among heterogeneous wireless networks. A few studies address spectrum-level resource sharing among LTE BSs without explicitly differentiating traffic statistics and QoS descriptions among diversified services (i.e., mobile broadband data service and M2M service) [18], [19]. Therefore, to satisfy differentiated QoS requirements from heterogeneous services, traffic statistics for each specific type of service should be modeled and considered in the resource slicing. In this paper, we develop a comprehensive radio bandwidth slicing framework for a two-tier heterogeneous wireless network (HetNet) to facilitate spectrum sharing among BSs and achieve QoS isolation for different service types. The contributions of this paper are three-folded:

1) We consider the coexistence of machine-type devices (MTDs) and mobile user devices (MUs), supporting M2M service and data service, respectively. An optimization framework for spectrum bandwidth slicing is developed to maximize the aggregate network utility, with the consideration of location distribution for MTDs and MUs, traffic statistics and differentiated QoS demands, traffic load in each cell, varying distances between BSs and devices, and inter-cell interference. The outputs of the optimization problem are BS-device association patterns and optimal bandwidth slicing ratios;
2) For tractability, the original optimization problem is transformed to a biconcave maximization problem under certain approximations. Then, an alternative concave search (ACS) algorithm is designed to iteratively solve the transformed problem. We prove that the ACS algorithm converges to a set of partial optimal solutions;
3) Extensive simulation results provide insights for our proposed bandwidth slicing framework. First, the optimal bandwidth slicing ratios are independent of end device location changes when each device moves within its associated cell coverage area, which leads to low communication overhead for global network information exchange between each network cell and the central controller; Second, with our proposed bandwidth slicing framework, frequent BS-device re-association can be avoided in network dynamics; Third, the ACS algorithm is both robust and lightweight. In performance comparison with two other resource slicing schemes, the proposed framework reduces communication overhead, achieves high capacity in each cell, and provides high network utility.

We organize the remainder of this paper in following sections. Existing resource slicing schemes are summarized in Section II. We describe the system model in Section III. In Section IV, a network utility maximization problem is formulated under the constraints of differentiated QoS provisioning for heterogeneous services. The original problem is transformed to a tractable biconcave maximization problem in Section V, and an ACS algorithm is then proposed to solve the transformed problem to obtain a set of optimal bandwidth slicing ratios and optimal BS-device association patterns. In Section VI, extensive simulation results demonstrate the performance of the proposed bandwidth slicing framework. Finally, conclusions are drawn in Section VII. Important symbols are listed in Table I.

II. Related Work

Due to its advantages in reducing infrastructure deployment cost and improving spectrum utilization, resource slicing in wireless networks has drawn attention from researchers. In [9], radio resource slices are created for different service providers (SPs) to obtain the requested service capacities and QoS isolation for different groups of end users. In [6], a resource allocation scheme is presented for a two-tier HetNet, where spectrum resources on each BS are sliced for different user groups belonging to different SPs to maximize the aggregate network utility. In [20], a resource-block (RB) slicing scheme is proposed for a single-cell LTE network supporting M2M communications. The RBs used in the random access phase in the LTE system are allocated to various categories of MTDs for differentiated QoS provisioning. Resource slicing for a multi-operator radio access network is investigated in [21], where different operators have different shares of the network infrastructure and the operators with larger shares are expected to receive more resources. Specifically, a BS-user association and resource allocation problem is jointly formulated to maximize a weighted sum of all operators' utilities without service differentiation. Some fine-grained flow-level resource abstraction frameworks are proposed in literature to customize flow scheduling policies for different applications [17], [22]. For instance, in [17], a service flow is defined as a flow of packets transmitted between a BS and a user, either uplink or downlink, with specific QoS parameter settings. A two-level wireless resource abstraction framework is then developed to improve the resource utilization. At the service level, the uplink and downlink flows of a BS are grouped to form slices for differentiated QoS provisioning; At the flow-level, flow scheduling policies are customized within each slice for packet transmissions at each time instant.

[1]QoS isolation refers to that any change in network state for one type of service, including end device mobility, channel dynamics, and traffic load fluctuations, should not violate the minimum QoS performance experienced by devices belonging to another service type [16], [17].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2018.2859740, IEEE Transactions on Vehicular Technology

TABLE I: Important symbols

Symbol	Definition
B_m	The MBS
$c_{i,m}/c_{i,k}$	Achievable rate at MTD i or MU i from B_m/S_k
c^{min}	Minimum achievable rate at each MTD for a bounded delay violation probability
$D_{i,j}$	Total transmission delay for a machine-type packet from BS j to MTD i
D_{max}	Delay bound for machine-type traffic
$f_{i,m}/f_{i,k}$	Fraction of bandwidth resources allocated to MTD i or MU i from B_m/S_k
$G_{i,m}/G_{i,k}$	Average channel gain between B_m/S_k and MTD i or MU i
L_a/L_d	Machine-type/Data packet size
$\mathcal{M}_k/M_k$	Set/Number of category II MUs in the coverage of S_k
$\mathcal{N}_a/N_a$	Set/Number of category I MTDs
$\mathcal{N}_k/N_k$	Set/Number of category II MTDs in the coverage of S_k
N_{m}	Iteration limit
$\mathcal{N}_u/N_u$	Set/Number of category I MUs
P_m/P_k	Transmit power on B_m/S_k
$r_{i,m}/r_{i,k}$	Spectrum efficiency at MTD i or MU i from B_m/S_k
S_k	k th $(k = 1, 2, ..., n)$ SBS under the coverage of B_m
W_v	Aggregate bandwidth resources
$x_{i,m}/x_{i,k}$	Association pattern indicator for category II MTD i or MU i with B_m/S_k
λ_a/λ_d	Poisson/Periodic average arrival rate for machine-type traffic/data traffic
β_m/β_s	Bandwidth slicing ratio for B_m/S_k
ε	Maximum delay bound violation probability
σ^2	Average background noise power

Existing studies mainly focus on the device-level resource slicing, in which the preallocated radio resources at each BS are sliced among different service groups within the BS coverage (e.g., in [6] and [20]). Radio access networks (RAN) sharing and network-level spectrum sharing are studied in [18], without an explicit characterization on differentiated traffic statistics and QoS descriptions for heterogeneous services. The spectrum sharing is triggered when traffic overload happens in one of the coexisting virtual networks.

In the following, we present a comprehensive radio resource slicing framework to facilitate resource sharing among heterogeneous BSs, with differentiated QoS provisioning for both data and M2M services, taking account of BS-device association patterns, and instantaneous traffic load conditions for each cell.

III. System Model

We consider a two-tier downlink HetNet, where a single macro-cell with a BS at the center in the first tier is underlaid by n small cells (with BSs placed at cell centers) in the second tier, as shown in Fig. 1. There are two types of end devices distributed in the HetNet, i.e., MTDs with delay-sensitive machine-type traffic requiring high transmission reliability, and MUs generating data traffic and demanding high throughput. MUs are with low-to-moderate mobility, and MTDs are more or less stationary [16], [23]. The MBS, denoted by B_m, with high transmit power is deployed for a wide-area communication coverage, including both control signaling and data transmissions. The n SBSs, $\mathcal{S} = \{S_1, S_2, ..., S_n\}$, with lower transmit power and smaller coverage, are placed at hotspot locations within the MBS's coverage area to support the increasing communication demands. We assume that the MBS and SBSs are directly connected to edge routers of the core network via wired backhaul links. Since the HetNet is expected to accommodate a much larger number of M2M devices than MUs, the SBSs are specifically deployed to

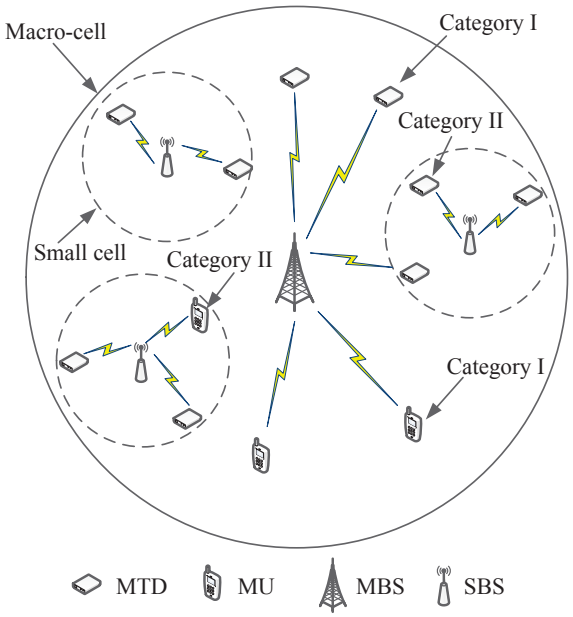


Fig. 1: A two-tier macro-cell overlaid network with the coexistence of MTDs and MUs.

enhance the network capacity in response to the increasing M2M traffic volume [24]. In most of the cases, MUs in the coverage of the macro-cell are connected to the MBS [4], [24]. Occasionally, when some of the MUs move into the coverage of an SBS, they select to connect to either the homing SBS or the MBS. There are two categories of MTDs and MUs: category I and category II. A category I device is within the macro-cell coverage but outside the coverages of all small cells, and is associated with the MBS for information packet transmissions; A category II device stays within the coverages of both the macro-cell and one of the small cells, which chooses to be associated with either the MBS or the SBS depending on the network conditions, e.g., cell loads, channel

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2018.2859740, IEEE Transactions on Vehicular Technology

conditions and network utility.

The set of category I MUs and the set of category I MTDs along with their set cardinalities are denoted by $\mathcal{N}_u$ and N_u, and $\mathcal{N}_a$ and N_a, respectively, while the sets of category II MUs and MTDs and their set cardinalities in the coverage of $S_k\,(k=1,2,...,n)$ are denoted by $\mathcal{M}_k$ and M_k and $\mathcal{N}_k$ and N_k, among which some of the devices can be associated with the MBS. There are a number of transmission queues at each BS, each of which is used for downlink transmissions between the BS and an associated device. We use link-layer packetized traffic to model arrivals of both data traffic and M2M traffic for explicit QoS characterization [25]. Data packets destined for an MU arrive at a transmission queue in the MBS periodically with rate λ_d packet/s, and each data packet has a constant size of L_d bits, whereas machine-type packets for an MTD arrive at the MBS or an SBS in an event-driven manner with a much lower packet arrival rate and a smaller packet size [16]. As suggested in [23], [26], machine-type packet arrivals at each transmission queue in a BS are modeled as a Poisson process with average rate of λ_a packet/s and constant packet size L_a bits. Due to the random nature of machine-type packet arrivals, the QoS can be guaranteed in a statistical way by applying the effective bandwidth theory [25], [27] (to be discussed in Section IV). All packets are transmitted through a wireless propagation channel to reach the destination device.

A. Communication Model

Suppose that two sets of radio resources, W_m and W_s, are initially allocated to the MBS and each SBS, respectively, and are mutually orthogonal to avoid inter-tier interference. Since each SBS maintains a small communication coverage with low transmit power, the resources W_s can be reused among all SBSs under an acceptable inter-cell interference level to improve the spectrum utilization. Transmit power for B_m and $S_k\,(k=1,2,...,n)$, denoted by P_m and $P_k\,(k=1,2,...,n)$, are pre-determined and remain constant during each resource slicing period. Resource slicing is updated when traffic load of each cell fluctuates [18]. A cell traffic load is described as a combination of number of devices admitted in the cell and traffic arrival statistics for each device. Using Shannon capacity formula, the spectrum efficiency at device i from B_m and $S_k\,(k=1,2,...,n)$ are expressed, respectively, as

$$r_{i,m}=\log_2\left(1+\frac{P_m G_{i,m}}{\sigma^2}\right),\ i\in\{\mathcal{N}_u,\mathcal{N}_a,\mathcal{N}_k,\mathcal{M}_k\},\ k=\{1,2,...,n\} \tag{1}$$

and

$$r_{i,k}=\log_2\left(1+\frac{P_k G_{i,k}}{\sum\limits_{j=\{1,2,...,n\},j\neq k} P_j G_{i,j}+\sigma^2}\right),\ i\in\mathcal{N}_k\cup\mathcal{M}_k,\ k=\{1,2,...,n\} \tag{2}$$

where $G_{i,m}$ and $G_{i,k}$ are squared average channel gains between B_m and S_k and device i, σ^2 denotes average background noise power. Inter-cell interference among small cells is included in (2). In (1) and (2), $r_{i,m}$ and $r_{i,k}$ are measured in a large time scale and are termed as *spectrum efficiency* [28], to capture long-term wireless channel conditions, which include the path loss effect. The achievable rate at each MTD and MU can be obtained based on BS-device association patterns and the received fraction of bandwidth resources from B_m or S_k.

Note that resource slicing among heterogeneous BSs, including BS-device association and bandwidth allocation for end devices, is updated in a large time scale compared with channel coherence time to reduce communication overhead [28]. Therefore, the received signal-to-noise (SNR) and signal-to-interference-plus-noise (SINR) in (1) and (2) are average SNR and SINR over each resource slicing period [18], in which channel fast fading effects are averaged out.

B. Dynamic Bandwidth Slicing Framework

As stated in Section I, in the two-tier HetNet supporting heterogeneous services, the preallocated radio resources at each BS can be inefficient due to unbalanced end device distribution, varying device locations, and increasing network load on each cell [21]. Consequently, some of the BSs are overloaded, while the resources on other light-loaded BSs can be underutilized. Additionally, a massive number of MTDs accessing the network can possibly cause QoS violation for existing users if resources are not efficiently managed between data service and M2M service [24], [29]. Therefore, proper resource slicing is an effective solution 1) to facilitate radio resource sharing among heterogeneous wireless infrastructures (MBSs and SBSs), and 2) to ensure QoS isolation among diverse services. Such resource slicing framework is required to balance the trade-off between resource sharing and QoS isolation. On one hand, resources are partitioned into different slices and allocated to M2M service and data service for differentiated QoS satisfaction; On the other hand, the amount of resources for each slice should be dynamically adjusted according to variations of network conditions to improve resource utilization. Therefore, a dynamic resource slicing framework needs to be developed with the consideration of both network conditions (i.e., end-device location distribution, cell loads, inter-cell interference levels, and BS-device association patterns) and service-level characteristics (i.e., traffic statistics and QoS requirements for heterogeneous services). Moreover, since MUs are with low-to-moderate mobility and MTDs are more or less stationary, resources for each slice are expected to be updated in a large time scale to reduce the communication overhead for network information exchange between the central controller and the BSs.

With wireless NFV, the controller has central control over all bandwidth resources from heterogeneous BSs, the amount of which is denoted by $W_v\,(W_v=W_m+W_s)$. Note that we consider radio spectrum bandwidth as the resource provisioning [9]. Resource slicing is realized in two steps: In the first step, the total bandwidth resources are sliced among the MBS and SBSs, as shown in Fig. 2. The central controller needs to specify how bandwidth resources are allocated among BSs to achieve maximal resource utilization and, at the same time, guarantee QoS isolation between M2M and data services. We define β_m and β_s (with $\beta_m+\beta_s$ =1) as *slicing ratios*, indicating the shares of bandwidth resources (out of W_v)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2018.2859740, IEEE Transactions on Vehicular Technology

allocated to the MBS and SBSs, respectively. To determine the optimal set of slicing ratios, $\{\beta_m^*, \beta_s^*\}$, we formulate a comprehensive optimization framework to maximize the aggregate network utility under the constraints of satisfying the differentiated QoS requirements for both types of services, by taking into account the network conditions and service-level characteristics (see details in Section IV). The slicing ratios are adjusted between bandwidth slices in response to the network traffic load dynamics to improve the overall resource utilization. In the second step, since the MBS and SBSs support both data and M2M services, the bandwidth slices are further split to two sub-slices and allocated to the group of MUs and the group of MTDs associated with corresponding BSs, with the bandwidth slice splitting ratios denoted by $\beta_{m,1}^*$ and $\beta_{m,2}^*$, and $\beta_{s,1}^*$ and $\beta_{s,2}^*$, respectively, as shown in Fig 2. Therefore, from the service-level standpoint, we define slicing ratios α_1^* $(= \beta_{m,1}^* + \beta_{s,1}^*)$ and α_2^* $(= \beta_{m,2}^* + \beta_{s,2}^*)$, as the fractions of bandwidth resources allocated to the data service and the M2M service, respectively.

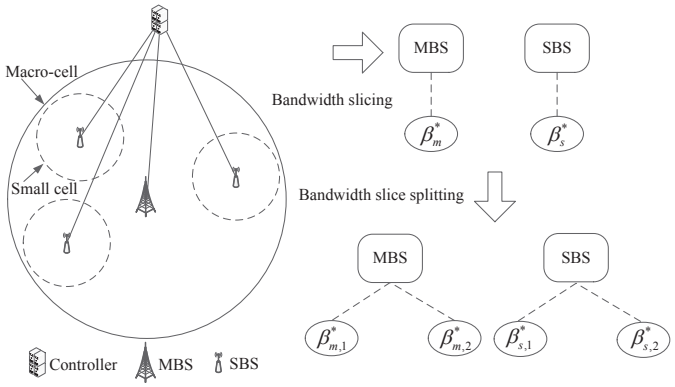


Fig. 2: Radio resource slicing and splitting.

IV. Problem Formulation

In the proposed bandwidth slicing framework, the challenging issue is how to determine the set of optimal slicing ratios, $\{\beta_m^*, \beta_s^*\}$, to 1) maximize the aggregate network utility, and 2) satisfy diverse QoS requirements between M2M and data services.

We consider the following logarithmic functions, as utility functions for an MU or an MTD (either category I or category II) associated with B_m and the utility for a category II MU or MTD associated with S_k, respectively:

$$\log\left(c_{i,m}\right) = \log\left(W_v \beta_m f_{i,m} r_{i,m}\right), i \in \mathcal{N}_u \cup \mathcal{N}_a \cup \mathcal{N}_k \cup \mathcal{M}_k, \quad k = \{1, 2, ..., n\} \tag{3}$$

$$\log\left(c_{i,k}\right) = \log\left(W_v \beta_s f_{i,k} r_{i,k}\right), \; i \in \mathcal{N}_k \cup \mathcal{M}_k, \quad k = \{1, 2, ..., n\}. \tag{4}$$

In (3) and (4), $c_{i,m}$ denotes the achievable rate at device i associated with B_m, and $c_{i,k}$ the achievable rate at category II device i from S_k; $f_{i,m}$ is the fraction of bandwidth resources (out of $W_v \beta_m$) allocated to device i from B_m, and $f_{i,k}$ the fraction of bandwidth resources (out of $W_v \beta_s$) allocated to category II device i from S_k. The logarithmic utility function is a concave function having diminishing marginal utility, which facilitates network load balancing in BS-device association and achieves certain fairness in resource allocation among end devices [6], [18], [21].

As stated in Section III-A, the downlink achievable rates, $c_{i,m}$ and $c_{i,k}$, are constant during each bandwidth slicing period, and data packets, destined for an MU, arrive at one of the transmission queues in B_m periodically. Thus, the throughput requirement for each MU can be satisfied deterministically if enough resources are allocated for each downlink data transmission. However, due to the stochastic traffic arrival of the event-driven M2M service, the downlink total transmission delay from B_m or S_k to an MTD should be guaranteed statistically for maximal resource utilization. The delay is the duration from the instant a machine-type packet arrives at a BS transmission queue to the instant the intended MTD receives the packet. We apply the effective bandwidth theory to derive the minimum service rate for each MTD to probabilistically guarantee a packet transmission delay bound.

The effective bandwidth for a machine-type traffic source, with a QoS exponent, φ_M, is expressed as

$$\varrho\left(\varphi_M\right) = \lim_{t \to \infty} \frac{1}{t} \frac{1}{\varphi_M} \log E\left[e^{\varphi_M A(t)}\right] \tag{5}$$

where $A(t)$ denotes the number of machine-type packet arrivals over $[0, t)$, and $E[\cdot]$ denotes the operation of expectation. Since $A(t)$ is modeled as a Poisson process with the rate of λ_a packet/s, (5) is further derived as [27]

$$\varrho\left(\varphi_M\right) = \lambda_a \frac{e^{\varphi_M} - 1}{\varphi_M}. \tag{6}$$

On the other hand, by applying the large deviation theory [30], [31], the probability of downlink total transmission delay $D_{i,j}$ for a machine-type packet from BS j to MTD i exceeding a delay bound D_{max} is approximated as

$$Pr\{D_{i,j} \geq D_{max}\} \approx e^{-\varphi_M p_{i,j} D_{max}} \leq \varepsilon \tag{7}$$

where $i \in \mathcal{N}_a \cup \mathcal{N}_k$ if $j = m$ for MBS B_m, and $i \in \mathcal{N}_k$ if $j = k$ for SBS S_k $(k = 1, 2, ..., n)$, ε denotes a delay bound violation probability threshold, $p_{i,j} = \frac{c_{i,j}}{L_a}$ is the achievable rate at MTD i from BS j in terms of number of packets per second. From (7), the minimum achievable rate $p^{(min)}$ of $p_{i,j}$ is given by

$$p^{(min)} = -\frac{\log \varepsilon}{\varphi_M D_{max}}. \tag{8}$$

According to the effective bandwidth theory [30], $p^{(min)}$ should equal the effective bandwidth $\varrho\left(\varphi_M\right)$ to guarantee the delay bound violation probability at most ε. Therefore, by substituting (8) into (6) and after some algebraic manipulation, we have

$$c^{(min)} = -\frac{L_a \log \varepsilon}{D_{max} \log\left(1 - \frac{\log \varepsilon}{\lambda_a D_{max}}\right)} \tag{9}$$

where $c^{(min)} = p^{(min)} L_a$.

Next, we formulate an aggregate network utility maximization problem (P1), under the constraints of differentiated QoS guarantee, BS-device association patterns and bandwidth allocation for each device:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2018.2859740, IEEE Transactions on Vehicular Technology

$$(\mathrm{P1}):$$

$$\max_{\substack{\beta_m,\beta_s,\\ x_{i,j},f_{i,j}}} \sum_{i\in\mathcal{N}_u\cup\mathcal{N}_a}\log(c_{i,m})+\sum_{k=1}^{n}\sum_{i\in\mathcal{N}_k\cup\mathcal{M}_k}\sum_{j\in\{m,k\}}x_{i,j}\log(c_{i,j})$$

$$\text{s.t.}\begin{cases} c_{i,m}\geq\lambda_d L_d, & i\in\mathcal{N}_u \quad (10a)\\ c_{i,m}\geq c^{(min)}, & i\in\mathcal{N}_a \quad (10b)\\ x_{i,j}\left[c_{i,j}-c^{(min)}\right]\geq 0, & i\in\mathcal{N}_k, j\in\{m,k\} \quad (10c)\\ x_{i,j}\left[c_{i,j}-\lambda_d L_d\right]\geq 0, & i\in\mathcal{M}_k, j\in\{m,k\} \quad (10d)\\ \sum\limits_{j\in\{m,k\}}x_{i,j}=1, & i\in\mathcal{N}_k\cup\mathcal{M}_k \quad (10e)\\ x_{i,j}\in\{0,1\}, & i\in\mathcal{N}_k\cup\mathcal{M}_k, j\in\{m,k\} \quad (10f)\\ \sum\limits_{i\in\mathcal{N}_u\cup\mathcal{N}_a}f_{i,m}+\sum\limits_{k=1}^{n}\sum\limits_{i\in\mathcal{N}_k\cup\mathcal{M}_k}x_{i,m}f_{i,m}=1 \quad (10g)\\ \sum\limits_{i\in\mathcal{N}_k\cup\mathcal{M}_k}x_{i,k}f_{i,k}=1 \quad (10h)\\ f_{i,m}\in(0,1), & i\in\mathcal{N}_u\cup\mathcal{N}_a\cup\mathcal{N}_k\cup\mathcal{M}_k \quad (10i)\\ f_{i,k}\in(0,1), & i\in\mathcal{N}_k\cup\mathcal{M}_k \quad (10j)\\ \beta_m+\beta_s=1 \quad (10k)\\ \beta_m,\beta_s\in[0,1]. \quad (10l)\end{cases}$$

In (P1), the objective function is the aggregate network utility, which is the summation of utilities achieved by all devices. A category I device is associated with B_m, whereas a category II device chooses to connect to either B_m or S_k. Hence, a binary variable, $x_{i,j}$ ($i\in\mathcal{N}_k\cup\mathcal{M}_k, j\in\{m,k\}, k\in\{1,2,...,n\}$), is introduced to indicate the association pattern for category II device i with B_m or S_k. Device i is associated with B_m, if $x_{i,m}=1$ and $x_{i,k}=0$; Otherwise, it is associated with S_k, if $x_{i,m}=0$ and $x_{i,k}=1$.

In (P1), constraints (10a) and (10d) indicate that the service rate $c_{i,j}$ for any MU i is not less than the periodic traffic arrival rate destined for the device at the base station. Constraints (10b) and (10c) ensure that the achievable rate for both category I and category II MTDs is not less than the effective bandwidth of the M2M traffic source. Constraint (10e) and (10f) indicate that a category II device is associated with either the MBS or its home SBS during each bandwidth allocation period. Constraints (10g) and (10h) demonstrate the requirements on bandwidth allocation for MTDs and MUs from different BSs. Therefore, by maximizing the aggregate network utility with QoS guarantee, the optimal set of slicing ratios β_m^* and β_s^*, BS-device association indicators $x_{i,j}^*$, and fractions of bandwidth resources $f_{i,m}^*$ and $f_{i,k}^*$ allocated to MTDs and MUs from B_m and S_k can be determined.

Since (P1) is a joint BS-device association and resource allocation problem, the fraction of bandwidth resources allocated to each device depends on all BS-device association patterns and resource slicing ratios among BSs, and this coupling makes the problem difficult to solve. For tractability, given BS-device association patterns $x_{i,j}$ and resource slicing ratios $\{\beta_m,\beta_s\}$, we first determine the optimal fractions of bandwidth resources $f_{i,m}^*$ and $f_{i,k}^*$ allocated to device i from B_m and S_k, in a function of $x_{i,j}$, to maximize the aggregate network utility.

A. Optimal Bandwidth Allocation to MUs and MTDs

We simplify (P1) by expressing $f_{i,j}$ as a function of $x_{i,j}$ to reduce the number of decision variables. Given β_m, β_s, and $x_{i,j}$, the objective function of (P1) can be expressed as a summation of $u_m^{(1)}(f_{i,m})$ and $\sum_{k=1}^{n}u_k^{(1)}(f_{i,k})$. Hence, $u_m^{(1)}(f_{i,m})$ is a function of $f_{i,m}$, indicating the aggregate utility of MUs and MTDs associated with B_m, given by

$$\begin{aligned}u_m^{(1)}(f_{i,m})=&\sum_{i\in\mathcal{N}_u\cup\mathcal{N}_a}\log\left(W_v\beta_m f_{i,m}r_{i,m}\right)\\&+\sum_{k=1}^{n}\sum_{i\in\mathcal{N}_k'}\log\left(W_v\beta_m f_{i,m}r_{i,m}\right)\end{aligned}\quad(11)$$

where $\mathcal{N}_k'=\{l\in\mathcal{N}_k\cup\mathcal{M}_k|x_{l,m}=1\}$, $u_k^{(1)}(f_{i,k})$ is a function of $f_{i,k}$, denoting the aggregate utility of category II devices associating with S_k, given by

$$u_k^{(1)}(f_{i,k})=\sum_{i\in\overline{\mathcal{N}_k'}}\log\left(W_v\beta_s f_{i,k}r_{i,k}\right)\quad(12)$$

with $\overline{\mathcal{N}_k'}=\{l\in\mathcal{N}_k\cup\mathcal{M}_k|x_{l,k}=1\}$.

Hence, (P1) can be written as (P1′):

$$(\mathrm{P1}'):\max_{f_{i,m},f_{i,k}}u_m^{(1)}(f_{i,m})+\sum_{k=1}^{n}u_k^{(1)}(f_{i,k})$$

$$\text{s.t.}\begin{cases}\sum\limits_{i\in\mathcal{N}_u\cup\mathcal{N}_a\cup\mathcal{N}_k'}f_{i,m}=1 \quad (13a)\\ \sum\limits_{i\in\overline{\mathcal{N}_k'}}f_{i,k}=1 \quad (13b)\\ f_{i,m}\in(0,1), \quad i\in\mathcal{N}_u\cup\mathcal{N}_a\cup\mathcal{N}_k' \quad (13c)\\ f_{i,k}\in(0,1), \quad i\in\overline{\mathcal{N}_k'}. \quad (13d)\end{cases}$$

From (P1′), $\{f_{i,m}\}$ and $\{f_{i,k}\}$ are two independent sets of decision variables with uncoupled constraints. Thus, (P1′) is further decomposed to two subproblems (S1P1′) and (S2P1′):

$$(\mathrm{S1P1}'):\max_{f_{i,m}}u_m^{(1)}(f_{i,m})$$

$$\text{s.t.}\begin{cases}\sum\limits_{i\in\mathcal{N}_u\cup\mathcal{N}_a\cup\mathcal{N}_k'}f_{i,m}=1 \quad (14a)\\ f_{i,m}\in(0,1), \quad i\in\mathcal{N}_u\cup\mathcal{N}_a\cup\mathcal{N}_k' \quad (14b)\end{cases}$$

$$(\mathrm{S2P1}'):\max_{f_{i,k}}\sum_{k=1}^{n}u_k^{(1)}(f_{i,k})$$

$$\text{s.t.}\begin{cases}\sum\limits_{i\in\overline{\mathcal{N}_k'}}f_{i,k}=1 \quad (15a)\\ f_{i,k}\in(0,1), \quad i\in\overline{\mathcal{N}_k'}. \quad (15b)\end{cases}$$

Proposition 1. *The solutions for* (S1P1′) *and* (S2P1′) *are*

$$f_{i,m}^*=\frac{1}{N_u+N_a+\sum\limits_{k=1}^{n}\sum\limits_{l\in\mathcal{N}_k\cup\mathcal{M}_k}x_{l,m}}\triangleq f_m^* \quad (16)$$

and

$$f_{i,k}^*=\frac{1}{\sum\limits_{l\in\mathcal{N}_k\cup\mathcal{M}_k}x_{l,k}}\triangleq f_k^*. \quad (17)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2018.2859740, IEEE Transactions on Vehicular Technology

The proof of Proposition 1 is provided in Appendix A. Proposition 1 indicates that the optimal fractions of bandwidth resources allocated to MUs and MTDs from the associated BSs are equal bandwidth partitioning.

By substituting f_m^* and f_k^* into (P1), (P1) is reformulated as (P2) with the reduced number of decision variables:

$$(\mathrm{P2}): \max_{\substack{\beta_m,\beta_s,\\ \mathbf{X}_m,\mathbf{X}_k}} u_m^{(2)}(\beta_m,\mathbf{X}_m)+\sum_{k=1}^{n} u_k^{(2)}(\beta_s,\mathbf{X}_k)$$

$$\text{s.t.}\begin{cases} W_v\beta_m f_m^* r_{i,m}\geq \lambda_d L_d, & i\in\mathcal{N}_u \quad (18a)\\ W_v\beta_m f_m^* r_{i,m}\geq c^{(min)}, & i\in\mathcal{N}_a \quad (18b)\\ x_{i,m}\left[W_v\beta_m f_m^* r_{i,m}-c^{(min)}\right]\geq 0, & i\in\mathcal{N}_k \quad (18c)\\ x_{i,m}\left[W_v\beta_m f_m^* r_{i,m}-\lambda_d L_d\right]\geq 0, & i\in\mathcal{M}_k \quad (18d)\\ x_{i,k}\left[W_v\beta_s f_k^* r_{i,k}-c^{(min)}\right]\geq 0, & i\in\mathcal{N}_k \quad (18e)\\ x_{i,k}\left[W_v\beta_s f_k^* r_{i,k}-\lambda_d L_d\right]\geq 0, & i\in\mathcal{M}_k \quad (18f)\\ \sum\limits_{j\in\{m,k\}} x_{i,j}=1, & i\in\mathcal{N}_k\cup\mathcal{M}_k \quad (18g)\\ x_{i,j}\in\{0,1\}, & i\in\mathcal{N}_k\cup\mathcal{M}_k, j\in\{m,k\} \quad (18h)\\ \beta_m+\beta_s=1 & \quad (18i)\\ \beta_m,\beta_s\in[0,1] & \quad (18j)\end{cases}$$

where $\mathbf{X}_m=\{x_{i,m}|i\in\mathcal{N}_k\cup\mathcal{M}_k, k\in\{1,2,...,n\}\}$, $\mathbf{X}_k=\{x_{i,k}|i\in\mathcal{N}_k\cup\mathcal{M}_k\}$, $u_m^{(2)}(\beta_m,\mathbf{X}_m)$ and $u_k^{(2)}(\beta_s,\mathbf{X}_k)$ are expressed as

$$u_m^{(2)}(\beta_m,\mathbf{X}_m)=\sum_{i\in\mathcal{N}_u\cup\mathcal{N}_a}\log\left(W_v\beta_m f_m^* r_{i,m}\right)+\sum_{k=1}^{n}\sum_{i\in\mathcal{N}_k\cup\mathcal{M}_k} x_{i,m}\log\left(W_v\beta_m f_m^* r_{i,m}\right) \quad (19)$$

and

$$u_k^{(2)}(\beta_s,\mathbf{X}_k)=\sum_{i\in\mathcal{N}_k\cup\mathcal{M}_k} x_{i,k}\log\left(W_v\beta_s f_k^* r_{i,k}\right). \quad (20)$$

Since the two sets of decision variables $\{\beta_m,\mathbf{X}_m\}$ and $\{\beta_s,\mathbf{X}_k\}$ are coupled by constraints (18g) and (18i), (P2) cannot be decoupled in the same way as (P1′). The simplified problem (P2) is a mixed-integer combinatorial problem with the binary variable set $\{x_{i,j}\}$, which is difficult to solve. Therefore, in the next section, we transform (P2) to a tractable form for optimal solutions.

V. Problem Transformation and Partial Optimal Solutions

To make (P2) tractable, we first relax the binary variables $\{x_{i,j}\}$ in (P2) to real-valued variables $\{\widetilde{x_{i,j}}\}$ within the range $[0,1]$. Variables $\{\widetilde{x_{i,j}}\}$ represent the fraction of time that device i is associated with B_m or S_k during each bandwidth slicing period [6]. With the variable relaxation, the objective function of (P2) becomes a summation of $u_m^{(2)}(\beta_m,\widetilde{\mathbf{X}_m})$ and $\sum_{k=1}^{n} u_k^{(2)}\left(\beta_s,\widetilde{\mathbf{X}_k}\right)$, where $\widetilde{\mathbf{X}_m}=\{\widetilde{x_{i,m}}|i\in\mathcal{N}_k\cup\mathcal{M}_k, k\in\{1,2,...,n\}\}$ and $\widetilde{\mathbf{X}_k}=\{\widetilde{x_{i,k}}|i\in\mathcal{N}_k\cup\mathcal{M}_k\}$. In Proposition 2, we state the bi-concavity property of $u_m^{(2)}(\beta_m,\widetilde{\mathbf{X}_m})$ and $\sum_{k=1}^{n} u_k^{(2)}\left(\beta_s,\widetilde{\mathbf{X}_k}\right)$, based on Definition 1 and Definition 2.

Definition 1. *Suppose set* Y *can be expressed as the Cartesian product of two subsets* $\mathrm{A}\in\mathbf{R^m}$ *and* $\mathrm{B}\in\mathbf{R^n}$*, i.e.,* $\mathrm{Y}=\mathrm{A}\times\mathrm{B}$*. Then,* Y *is called a biconvex set on* $\mathrm{A}\times\mathrm{B}$*, if* A *is a convex subset for any given* $b\in\mathrm{B}$*, and* B *is also a convex subset for any given* $a\in\mathrm{A}$*.*

Definition 2. *Function* $\mathcal{F}:\mathrm{Y}\rightarrow\mathbf{R}$ *is defined on a biconvex set* $\mathrm{Y}=\mathrm{A}\times\mathrm{B}$*, where* $\mathrm{A}\in\mathbf{R^m}$ *and* $\mathrm{B}\in\mathbf{R^n}$*. Then,* $\mathcal{F}(\mathrm{A},\mathrm{B})$ *is called a biconcave (biconvex) function if it is a concave (convex) function on subset* A *for any given* $b\in\mathrm{B}$*, and it is also a concave (convex) function on subset* B *for any given* $a\in\mathrm{A}$*.*

Proposition 2. *Both* $u_m^{(2)}(\beta_m,\widetilde{\mathbf{X}_m})$ *and the summation* $\sum_{k=1}^{n} u_k^{(2)}\left(\beta_s,\widetilde{\mathbf{X}_k}\right)$ *are (strictly) biconcave functions on the biconvex decision variable set* $\{\beta_m,\beta_s\}\times\{\widetilde{\mathbf{X}_m},\widetilde{\mathbf{X}_k},k\in\{1,2,...,n\}\}$*.*

The proof of Proposition 2 is given in Appendix B.

With the variable relaxation, (P2) is transformed to (P3):

$$(\mathrm{P3}): \max_{\substack{\beta_m,\beta_s,\\ \widetilde{\mathbf{X}_m},\widetilde{\mathbf{X}_k}}} u_m^{(2)}(\beta_m,\widetilde{\mathbf{X}_m})+\sum_{k=1}^{n} u_k^{(2)}(\beta_s,\widetilde{\mathbf{X}_k})$$

$$\text{s.t.}\begin{cases} W_v\beta_m \widetilde{f_m^*} r_{i,m}-\lambda_d L_d\geq 0, & i\in\mathcal{N}_u \quad (21a)\\ W_v\beta_m \widetilde{f_m^*} r_{i,m}-c^{(min)}\geq 0, & i\in\mathcal{N}_a \quad (21b)\\ \widetilde{x_{i,m}}\left[W_v\beta_m \widetilde{f_m^*} r_{i,m}-c^{(min)}\right]\geq 0, & i\in\mathcal{N}_k \quad (21c)\\ \widetilde{x_{i,m}}\left[W_v\beta_m \widetilde{f_m^*} r_{i,m}-\lambda_d L_d\right]\geq 0, & i\in\mathcal{M}_k \quad (21d)\\ \widetilde{x_{i,k}}\left[W_v\beta_s \widetilde{f_k^*} r_{i,k}-c^{(min)}\right]\geq 0, & i\in\mathcal{N}_k \quad (21e)\\ \widetilde{x_{i,k}}\left[W_v\beta_s \widetilde{f_k^*} r_{i,k}-\lambda_d L_d\right]\geq 0, & i\in\mathcal{M}_k \quad (21f)\\ \sum\limits_{j\in\{m,k\}} \widetilde{x_{i,j}}=1, & i\in\mathcal{N}_k\cup\mathcal{M}_k \quad (21g)\\ \widetilde{x_{i,j}}\in[0,1], & i\in\mathcal{N}_k\cup\mathcal{M}_k, j\in\{m,k\} \quad (21h)\\ \beta_m+\beta_s=1 & \quad (21i)\\ \beta_m,\beta_s\in[0,1]. & \quad (21j)\end{cases}$$

In (P3), the objective function is a nonnegative sum of two (strictly) biconcave functions, which is also (strictly) biconcave [32]. Note that $\widetilde{f_m^*}$ and $\widetilde{f_k^*}$ are f_m^* and f_k^* with $x_{i,m}$ and $x_{i,k}$ substituted by $\widetilde{x_{i,m}}$ and $\widetilde{x_{i,k}}$. Moreover, if all the constraint functions in (P3) are written in a standard form, constraints (21a) and (21b) represent linear inequality constraint functions, and constraints (21g) and (21i) represent affine equality constraint functions, with respect to the set of decision variables. However, constraints (21c) - (21f) are non-convex constraint functions. Constraints (21c) and (21d) actually indicate that if any $i\in\mathcal{N}_k\cup\mathcal{M}_k$ ($k\in\{1,2,...,n\}$) is associated with B_m, the following inequalities should be

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2018.2859740, IEEE Transactions on Vehicular Technology

satisfied,

$$\sum_{k=1}^{n}\sum_{l\in\mathcal{N}_k\cup\mathcal{M}_k}\widetilde{x_{l,m}} \le \frac{W_v\beta_m r_{i,m}}{c^{(min)}} - N_u - N_a, i\in\mathcal{N}_k, \quad (22)$$

and

$$\sum_{k=1}^{n}\sum_{l\in\mathcal{N}_k\cup\mathcal{M}_k}\widetilde{x_{l,m}} \le \frac{W_v\beta_m r_{i,m}}{\lambda_d L_d} - N_u - N_a, i\in\mathcal{M}_k. \quad (23)$$

If device $i\in\mathcal{N}_k\cup\mathcal{M}_k$ is associated with S_k, constraints (21e) and (21f) are equivalent to

$$\sum_{l\in\mathcal{N}_k\cup\mathcal{M}_k}\widetilde{x_{l,k}} \le \frac{W_v\beta_s r_{i,k}}{c^{(min)}}, \quad i\in\mathcal{N}_k, \quad (24)$$

and

$$\sum_{l\in\mathcal{N}_k\cup\mathcal{M}_k}\widetilde{x_{l,k}} \le \frac{W_v\beta_s r_{i,k}}{\lambda_d L_d}, \quad i\in\mathcal{M}_k. \quad (25)$$

Therefore, to make (P3) tractable, we simplify (P3) to (P3′), by substituting (21c) - (21f) with (22) - (25), respectively:

$$(\mathrm{P3}'): \max_{\substack{\beta_m,\beta_s,\\ \widetilde{\mathbf{X}_m},\widetilde{\mathbf{X}_k}}} u_m^{(2)}(\beta_m,\widetilde{\mathbf{X}_m}) + \sum_{k=1}^{n} u_k^{(2)}(\beta_s,\widetilde{\mathbf{X}_k})$$
$$\text{s.t. (21a), (21b), (21g) - (21j), (22) - (25).}$$

Compared to the set of constraints (21a) - (21d) in (P3), constraints (21a), (21b), (22) and (23) in (P3′) provide the lowest upper bound on the number of category II devices that can be associated with B_m. This lowest upper bound is accurate because the communication distance between the MBS and any device located in the coverage of an SBS is much longer than the location differences among MTDs and MUs in the same SBS; thus, the differences of $r_{i,m}$ among the end devices are relatively small. Similarly, compared with constraints (21e) and (21f) in (P3), constraints (24) and (25) in (P3′) indicate the lowest upper bound on the number of category II devices that can be associated with $S_k, k\in\{1,2,...,n\}$. In fact, without changing the optimal solutions for (P3), the simplified constraints (22) - (25) in (P3′) indicate a set of conservative limits on maximum numbers of category II MTDs and MUs that can be associated with B_m and $S_k, k\in\{1,2,...,n\}$.

In a standard form, (P3′) is a biconcave maximization problem due to the biconcave objective function and the set of biconvex constraint functions with respect to the biconvex decision variable set $\{\beta_m,\beta_s\}\times\{\widetilde{\mathbf{X}_m},\widetilde{\mathbf{X}_k}\}$ [32]. To solve (P3′), an *alternative concave search* (ACS) algorithm is designed to explore the bi-concavity of the problem. The procedure of the ACS algorithm is summarized in Algorithm 1. As stated in Proposition 3, due to some properties of (P3′), Algorithm 1 converges to a set of *partial optimal solutions*. The definition of a partial optimal solution for (P3′) is given in Corollary 1, based on Proposition 3 and Theorem 4.7 in [32].

Proposition 3. *Algorithm 1 converges, due to the following properties for* (P3′)*: (1) Both* $\{\beta_m,\beta_s\}$ *and* $\{\widetilde{\mathbf{X}_m},\widetilde{\mathbf{X}_k}\}$ *are closed sets, and the objective function of* (P3′) *is continuous on its domain; (2) Given the set of accumulation points*[2]*,* $\{\beta_m^{(t)},\beta_s^{(t)},\widetilde{\mathbf{X}_m}^{(t)},\widetilde{\mathbf{X}_k}^{(t)}\}$*, at the beginning of* t *th iteration, the optimal solutions at the end of* t *th iteration (at the beginning of* $(t+1)$ *th iteration), i.e.,* $\{\beta_m^{(t+1)},\beta_s^{(t+1)},\widetilde{\mathbf{X}_m}^{(t+1)},\widetilde{\mathbf{X}_k}^{(t+1)}\}$*, are unique solutions.*

The proof of Proposition 3 is given in Appendix C.

Corollary 1. *Algorithm 1 converges to a set of optimal solutions, called partial optimums* $\{\beta_m^*,\beta_s^*,\widetilde{\mathbf{X}_m}^*,\widetilde{\mathbf{X}_k}^*\}$ *where* $\widetilde{\mathbf{X}_m}^* = \{\widetilde{x_{i,m}}^* | i\in\mathcal{N}_k\cup\mathcal{M}_k, k\in\{1,2,...,n\}\}$ *and* $\widetilde{\mathbf{X}_k}^* = \{\widetilde{x_{i,k}}^* | i\in\mathcal{N}_k\cup\mathcal{M}_k\}$*, which satisfy*

$$u_m^{(2)}(\beta_m^*,\widetilde{\mathbf{X}_m}^*) + \sum_{k=1}^{n} u_k^{(2)}(\beta_s^*,\widetilde{\mathbf{X}_k}^*)$$
$$\ge u_m^{(2)}(\beta_m,\widetilde{\mathbf{X}_m}^*) + \sum_{k=1}^{n} u_k^{(2)}(\beta_s,\widetilde{\mathbf{X}_k}^*), \quad \forall\beta_m,\beta_s\in[0,1] \quad (26)$$

and

$$u_m^{(2)}(\beta_m^*,\widetilde{\mathbf{X}_m}^*) + \sum_{k=1}^{n} u_k^{(2)}(\beta_s^*,\widetilde{\mathbf{X}_k}^*)$$
$$\ge u_m^{(2)}(\beta_m^*,\widetilde{\mathbf{X}_m}) + \sum_{k=1}^{n} u_k^{(2)}(\beta_s^*,\widetilde{\mathbf{X}_k}), \quad \forall\widetilde{x_{i,m}},\widetilde{x_{i,k}}\in[0,1]. \quad (27)$$

The main logical flow for Algorithm 1 is to iteratively solve for optimal bandwidth slicing ratios and optimal BS-device association patterns, $\{\beta_m^*,\beta_s^*,\widetilde{\mathbf{X}_m}^*,\widetilde{\mathbf{X}_k}^*\}$. In each iteration, given a set of optimal values of β_m and β_s from the previous iteration, (P3′) is solved for a set of optimal BS-device association patterns $\{\widetilde{\mathbf{X}_m}^\dagger,\widetilde{\mathbf{X}_k}^\dagger\}$ and then, given $\{\widetilde{\mathbf{X}_m}^\dagger,\widetilde{\mathbf{X}_k}^\dagger\}$, (P3′) is solved again for an updated set of optimal bandwidth slicing ratios $\{\beta_m^\dagger,\beta_s^\dagger\}$. At this point, the current iteration ends, and the stopping criterion for Algorithm 1 is checked, i.e., whether the difference between the objective function values at the end of current iteration and at the end of previous iteration is smaller than a predefined threshold (set as a very small value). If the stopping criterion is met, the set of optimal solutions for current iteration converge to the final optimal solution set $\{\beta_m^*,\beta_s^*,\widetilde{\mathbf{X}_m}^*,\widetilde{\mathbf{X}_k}^*\}$. Otherwise, the next round of iteration starts, following the same procedure, until the algorithm converges. For each pair of optimal solutions $\{\widetilde{x_{i,m}}^*,\widetilde{x_{i,k}}^*\}$, we let the larger one equal 1 and the smaller one equal 0 to obtain the optimal solutions $\{x_{i,m}^*,x_{i,k}^*\}$ and ensure every end device is associated with one BS during each resource slicing period. Simulation results in Section VI-B show accuracy of the variable relaxation for solving (P2). Based on the optimal solutions from Algorithm 1, the optimal bandwidth slicing ratios, α_1^* and α_2^*, for data service and M2M

[2]An accumulation point set for the ACS algorithm denotes the set of optimal solutions at the beginning of t th (for any $t>0$) iteration.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2018.2859740, IEEE Transactions on Vehicular Technology

Algorithm 1: The ACS algorithm for solving (P3′)

Input : Input parameters for (P3′), stopping criterion δ, iteration limit N_{m}, a candidate set $\mathcal{C}$ of initial values for $\{\beta_m, \beta_s\}$.

Output: Optimal bandwidth slicing ratios, $\{\beta_m^*, \beta_s^*\}$; Optimal BS-device association pattern, $\{\widetilde{\mathbf{X}_m}^*, \widetilde{\mathbf{X}_k}^*\}$.

Step 1: Select a pair of initial values for $\{\beta_m, \beta_s\}$ from $\mathcal{C}$, denoted by $\{\beta_m^{(t)}, \beta_s^{(t)}\}$ where $t = 0$; Let $\mathcal{U}^{(t)}$ denote the maximum objective function value, with optimal decision variables $\{\beta_m^{(t)}, \beta_s^{(t)}, \widetilde{\mathbf{X}_m}^{(t)}, \widetilde{\mathbf{X}_k}^{(t)}\}$, at the beginning of t th iteration;

Step 2: $\mathcal{U}^{(0)} \leftarrow 0$;

do

  $\{\widetilde{\mathbf{X}_m}^{\dagger}, \widetilde{\mathbf{X}_k}^{\dagger}\} \leftarrow$ solving (P3′) given $\{\beta_m^{(t)}, \beta_s^{(t)}\}$;

  if *No feasible solutions for* (P3′) **then**

    Go to **Step 1** until no feasible solutions found with initial values in $\mathcal{C}$;

    Stop and no optimal solutions under current network conditions;

  else

    $\{\widetilde{\mathbf{X}_m}^{(t+1)}, \widetilde{\mathbf{X}_k}^{(t+1)}\} \leftarrow \{\widetilde{\mathbf{X}_m}^{\dagger}, \widetilde{\mathbf{X}_k}^{\dagger}\}$;

    $\{\beta_m^{\dagger}, \beta_s^{\dagger}\} \leftarrow$ solving (P3′) given $\{\widetilde{\mathbf{X}_m}^{(t+1)}, \widetilde{\mathbf{X}_k}^{(t+1)}\}$;

    if *No feasible solutions for* (P3′) **then**

      Go to **Step 1** until no feasible solutions found with initial values in $\mathcal{C}$;

      Stop and no optimal solutions under current network conditions;

    else

      $\{\beta_m^{(t+1)}, \beta_s^{(t+1)}\} \leftarrow \{\beta_m^{\dagger}, \beta_s^{\dagger}\}$;

      Obtain maximum objective function value $\mathcal{U}^{(t+1)}$ at the end of t th iteration, with $\{\beta_m^{(t+1)}, \beta_s^{(t+1)}, \widetilde{\mathbf{X}_m}^{(t+1)}, \widetilde{\mathbf{X}_k}^{(t+1)}\}$.

    end

    $t \leftarrow t + 1$;

  end

while $||\mathcal{U}^{(t)} - \mathcal{U}^{(t-1)}|| \geq \delta$ **or** N_{m} is not reached;

service (suppose the numbers of category II devices in each small cell are equal) are obtained by

$$\begin{aligned}\alpha_1^* &= \beta_{m,1}^* + \beta_{s,1}^* \\ &= \beta_m^* f_m^* \left(N_u + \sum_{k=1}^{n} \sum_{l \in \mathcal{M}_k} x_{l,m}^* \right) + \beta_s^* f_k^* \sum_{l \in \mathcal{M}_k} x_{l,k}^* \end{aligned} \tag{28}$$

and

$$\alpha_2^* = \beta_{m,2}^* + \beta_{s,2}^* = 1 - \alpha_1^*. \tag{29}$$

The computational complexity of Algorithm 1 is calculated as follows: In t th iteration, given the optimal set of bandwidth slicing ratios, $\{\beta_m^{(t)}, \beta_s^{(t)}\}$, the convex optimization problem (P3′) is solved for the optimal BS-device association patterns, $\{\widetilde{\mathbf{X}_m}^{(t+1)}, \widetilde{\mathbf{X}_k}^{(t+1)}\}$, where the number of decision variables is $2\sum_{k=1}^{n}(N_k + M_k)$; Then, given $\{\widetilde{\mathbf{X}_m}^{(t+1)}, \widetilde{\mathbf{X}_k}^{(t+1)}\}$, (P3′) with 2 decision variables is solved again for the optimal bandwidth slicing ratios, $\{\beta_m^{(t+1)}, \beta_s^{(t+1)}\}$ at the end of t th iteration. Therefore, in each iteration, both convex optimization problems are solved sequentially by using interior-point methods [33], [34], and thus the time complexity upper bound of Algorithm 1 is $\mathcal{O}\left[N_{\mathrm{m}}\left(\sum_{k=1}^{n}(N_k + M_k)\right)^4\right]$, where n is the number of small cells within the macro-cell.

VI. Simulation Results

Simulation results are presented in this section which demonstrate the effectiveness of our proposed bandwidth slicing framework. All the simulations are carried out using MATLAB. In the two-tier wireless HetNet with 1 MBS-centered macro-cell underlaid by 4 SBS-centered small cells, locations of the MBS and SBSs are fixed, and the distance between the MBS and each of the SBSs are set as 400 m. Suppose the MBS is located at the orgin of a Cartesian coordinate system, and the 4 SBSs are at the respective coordinates of $(\pm 200\sqrt{2}\,\mathrm{m}, \pm 200\sqrt{2}\,\mathrm{m})$. The downlink transmit power on the MBS is set to 40 dBm with the communication coverage radius of 600 m, whereas each SBS has identical transmit power of 30 dBm with the coverage radius of 200 m [4]. All category I devices are randomly distributed in the coverage of the MBS (Category I MUs and MTDs are outside the coverages of SBSs.); Category II devices are randomly deployed within the coverages of SBSs, and each small cell is set an equal number of category II MTDs and category II MUs, denoted by N_s and M_s, respectively. In each small cell, M_s is set to 10. For propagation models, we use $L_m(z) = -30 - 35\log_{10}(z)$ and $L_s(z) = -40 - 35\log_{10}(z)$ to describe the downlink channel gains for the macro-cell and each small cell, respectively, including the path loss effect, where z is the distance between a BS and a device. The periodic data packet arrival rate λ_d at a transmission queue of the MBS is 20 packet/s, and the average rate λ_a of machine-type packet arrivals (following a Poisson process) at each transmission queue of either the MBS or an SBS is 5 packet/s. Each simulation result is obtained upon averaging over 50 location distribution samples of MUs and MTDs. Other important system parameters for simulations are summarized in Table II.

Through extensive simulations, we first demonstrate the robustness of the proposed bandwidth slicing framework, where a set of optimal bandwidth slicing ratios are obtained with low computational complexity and are dynamically adjusted with lightweight communication overhead. Then, we compare the proposed bandwidth slicing framework with an SINR-maximization (SINR-max) based network-level resource slicing scheme mentioned in [6], in which radio resources are shared among heterogeneous BSs and each device is associated with the BS providing highest downlink SINR, and a device-level resource slicing scheme [6], i.e., bandwidth resources are preallocated to each BS, and are then sliced for different device groups in the coverage region of the BS.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2018.2859740, IEEE Transactions on Vehicular Technology

TABLE II: System parameters

Parameters	Values
Aggregate bandwidth resources (W_v)	20 MHz [6]
Background noise power (σ^2)	-104 dBm
Data packet size (L_d)	9000 bits
Machine-type packet size (L_a)	2000 bits [35]
Machine-type packet delay bound (D_{max})	100 ms [35]
Delay bound violation probability (ε)	10^{-3} [35]
Stopping criterion (δ)	0.01
Iteration limit (N_{m})	1000 rounds

A. Optimal Bandwidth Slicing Ratios

In Figs. 3(a) and 3(b), the solutions for bandwidth slicing ratio, β_s, in each iteration of Algorithm 1 are plotted. In Fig. 3(a), given the numbers of MUs and MTDs in the macro-cell and each small cell, β_s converges to the same optimal solution, regardless of the location distribution for MUs and MTDs. The same property is observed under a different network load condition in Fig. 3(b). This insight demonstrates the robustness of the proposed resource slicing framework, upon which the set of optimal slicing ratios stay steady with end device location changes inside each cell. Therefore, the slicing ratios are adjusted in a large time scale, which significantly reduces network information exchange between each cell and the central controller for updating the slicing ratios.

Fig. 4 reflects the robustness of using the proposed ACS algorithm to solve for the optimal bandwidth slicing ratio β_s^*. It can be seen that the optimal solution β_s^* does not vary with initial slicing ratio β_s^{ini}. For service-level QoS provisioning and isolation, the optimal bandwidth slicing ratio α_2^* for all MTDs in the network is also shown in Fig. 4. From both Fig. 3 and Fig. 4, we conclude that the set of optimal bandwidth slicing ratios $\{\beta_m^*, \beta_s^*\}$ vary with the numbers of MUs and MTDs in each cell.

Fig. 5 shows the computational complexity of the proposed ACS algorithm, where the average number of iterations for solving β_s^* is evaluated with respect to different initial values for β_s and different values of N_u and N_a in the macro-cell. The average iteration number is low over a wide range of β_s^{ini}, N_u, N_a, and N_s. Based on a comparison of Fig. 4 and Fig. 5, we observe that the average number of iterations decreases when β_s^{ini} approaches β_s^*.

B. Performance Comparison

In Fig. 6, optimal bandwidth slicing ratios are compared between the proposed slicing framework and the SINR-max based network-level slicing scheme for different values of N_u, N_a, and N_s. In the SINR-max based slicing scheme, since each device is always associated with the BS providing the highest SINR, the BS-device association patterns should change upon variations of end device locations. Accordingly, radio resources need to be frequently adjusted to adapt to the changing load on each BS. Fig. 6 shows that the optimal bandwidth slicing ratio in the SINR-max based slicing scheme fluctuates with different MUs and MTDs distribution samples. In comparison, the proposed bandwidth slicing framework is more robust with network condition changes, and the slicing ratios are updated in a much larger time scale to reduce communication overhead.

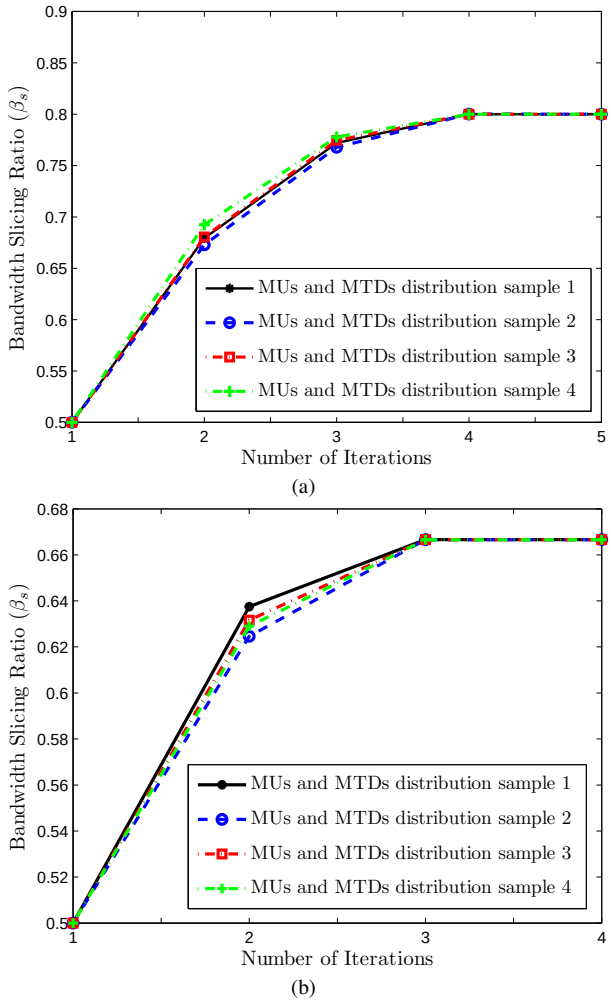


Fig. 3: Bandwidth slicing ratio (β_s) for an SBS during each iteration of Algorithm 1. (a) $N_u = N_a = 25$, $N_s = 40$, $M_s = 10$. (b) $N_u = N_a = 100$, $N_s = 90$, $M_s = 10$.

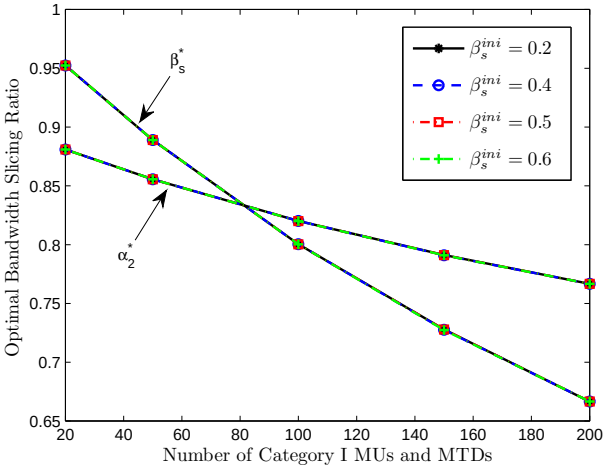


Fig. 4: Optimal bandwidth slicing ratios (β_s^* and α_2^*) ($N_u = N_a$, $N_s = 90$, $M_s = 10$).

Next, we compare the performance of the proposed resource slicing framework with the device-level resource slicing scheme in Fig. 7 to Fig. 9(b). In Fig. 7, we can see that for the device-level resource slicing scheme, with the increase of N_s,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2018.2859740, IEEE Transactions on Vehicular Technology

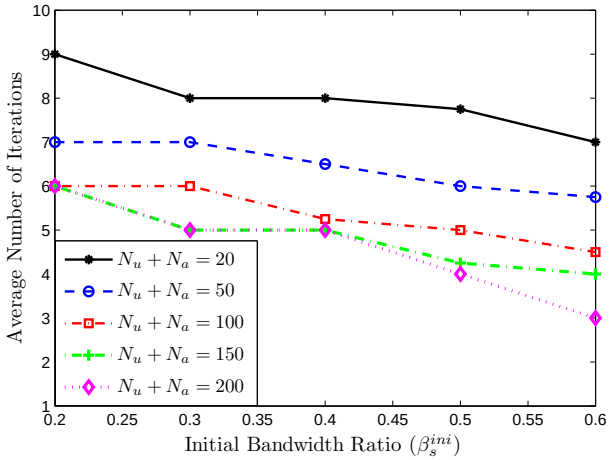


Fig. 5: Average number of iterations to solve for β_s^* using Algorithm 1 with different initial values β_s^{ini} ($N_u = N_a$, $N_s = 90$, $M_s = 10$).

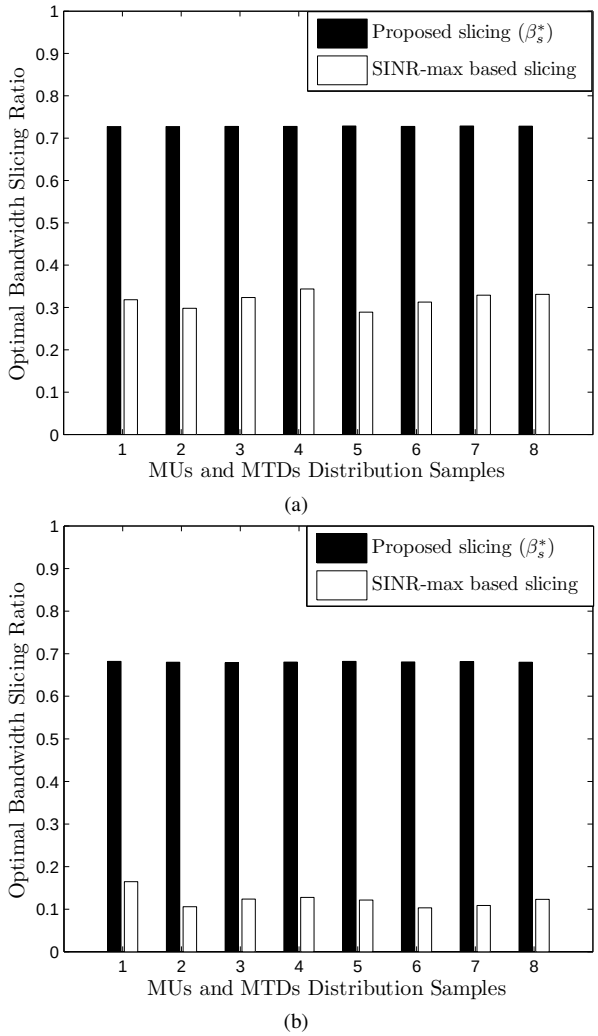


Fig. 6: Comparison of bandwidth slicing ratios between the proposed slicing framework and the SINR-max based slicing scheme. (a) $N_u = N_a = 75$, $N_s = 90$, $M_s = 10$. (b) $N_u = N_a = 100$, $N_s = 140$, $M_s = 10$.

more and more MTDs and MUs in each SBS are offloaded to the MBS to improve overall resource utilization among BSs. As a result, MTDs and MUs need to frequently change their connections with different BSs, which inevitably increases the communication overhead between end devices and BSs. In contrast, for the proposed resource slicing framework, under central control, radio resources can be dynamically adjusted among heterogeneous BSs in response to the network load changes, which makes the resource management more flexible and significantly reduces the communication cost as MTDs and MUs do not need to re-associate their connections with different BSs.

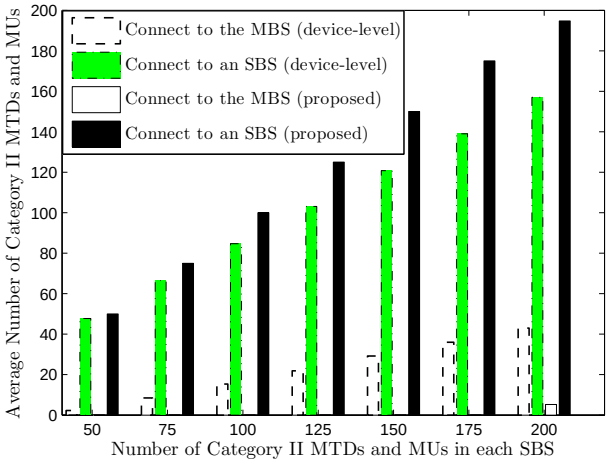


Fig. 7: Average number of category II MTDs and MUs connecting to either the MBS or their home SBSs ($N_u = N_a = 50$).

Fig. 8 shows that the optimal bandwidth slicing ratios β_s^* and α_2^* are dynamically updated with the number of MTDs, N_s, in each SBS, whereas the bandwidth resources are fixed on each BS for the device-level resource slicing scheme. Therefore, the cost of the proposed resource slicing framework is that the central controller needs to periodically obtain updated network load information from each BS for the slicing ratio adjustment. This signaling cost is relatively low since the load in each cell does not change in a small time scale.

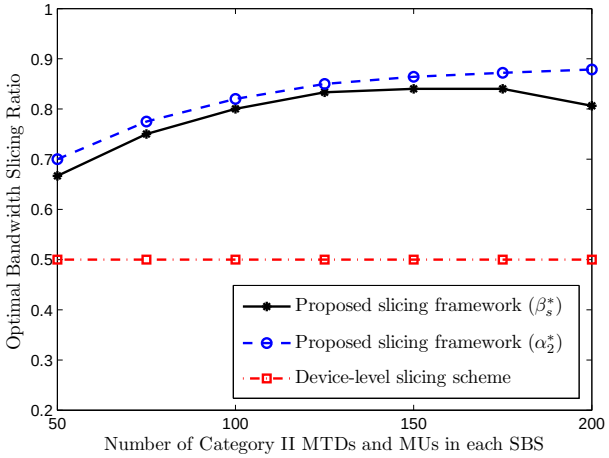
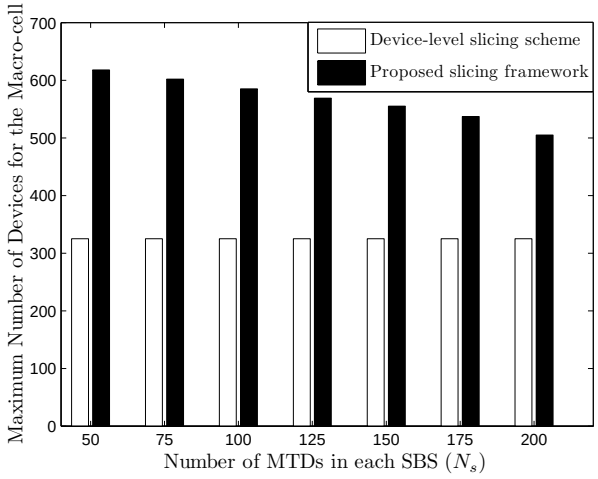


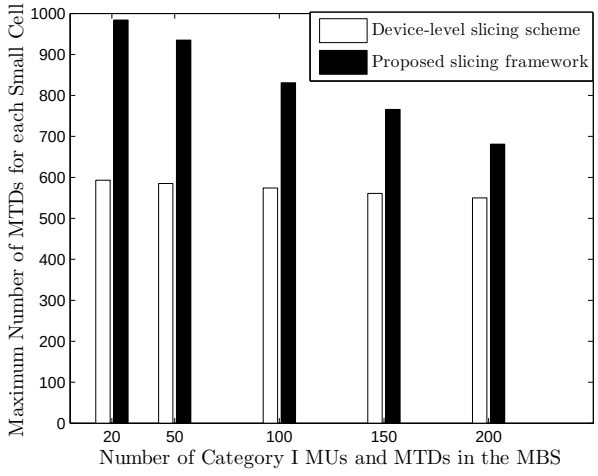
Fig. 8: Comparison of bandwidth slicing ratios between the proposed slicing framework and the device-level slicing scheme ($N_u = N_a = 50$).

Figs. 9(a) and 9(b) show that the proposed bandwidth slicing framework significantly improves the resource utilization, by

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2018.2859740, IEEE Transactions on Vehicular Technology



(a)



(b)

Fig. 9: Maximum number of admitted devices (a) for the macro-cell and (b) for each small cell.

increasing the maximum number of devices in each cell. We first evaluate the maximum number of MUs and MTDs admitted in the macro-cell under the condition that QoS requirements for 99% of MUs are satisfied. Since MUs require more bandwidth resources than MTDs, the QoS for MUs are violated first with the increase of cell load. We then evaluate the maximum number of MTDs admitted in each small cell at the premise of not violating the QoS for all MTDs (Assume that no category II MUs exist in each small cell for evaluating the maximum number of admitted devices in both the macro-cell and small cells.). In Fig. 9(a), the proposed resource slicing framework achieves a much larger number of devices admitted in the macro-cell than the device-level resource slicing scheme where the macro-cell's admission region is fixed without resource sharing under different N_s. Similar trend for the maximum number of MTDs admitted in small cells is observed in Fig. 9(b). Although some of the MTDs can be offloaded to the MBS for the device-level resource slicing scheme, the number of MTDs admitted in the macro-cell is limited due to the QoS provisioning for MUs and the utilization of bandwidth resources for the MBS.

The maximum packet loss probability (i.e., maximum delay bound variation probability) among all MTDs in the HetNet is evaluated with the variation of the number of MTDs in each SBS. It is shown in Fig. 10 that the maximum packet loss probability increases with N_s due to the reduced fractions of bandwidth resources allocated to each device. We can also see that the proposed bandwidth slicing framework guarantees the maximum packet loss probability be bounded by a threshold when the number of category II MTDs varies within the admission region of each small cell.

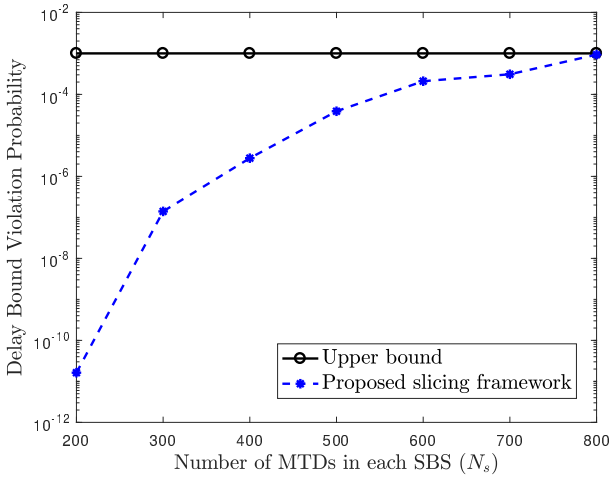


Fig. 10: Maximum packet loss probability for MTDs ($N_u = N_a = 50$).

Lastly, the aggregate network utility of the macro-cell underlaid by small cells, achieved by different resource slicing schemes, is evaluated with variations of N_u, N_a and N_s in Figs. 11(a) and 11(b). For the proposed bandwidth slicing framework, we also evaluate the effect of the approximations (i.e., the variable relaxation) made for solving (P2). It is clear that the network utility achieved by the set of fractional BS-device optimal association patterns $\{\widetilde{x_{i,m}}^*, \widetilde{x_{i,k}}^*\}$ matches closely with the one achieved by the exact binary optimal solutions $\{x^*_{i,m}, x^*_{i,k}\}$ after the rounding operation. Moreover, through bandwidth slicing among heterogeneous BSs, the overall network resource utilization, and thus the network throughput, are significantly improved. Therefore, it can be seen that our proposed resource slicing framework achieves higher network utility than both the SINR-max based network-level resource slicing scheme and the device-level resource slicing scheme.

VII. Conclusion

In this paper, we propose a dynamic radio resource slicing framework for a two-tier HetNet to determine a set of resource slicing ratios and BS-device association patterns under different network load conditions. Based on SDN-enabled radio function softwarization, spectrum bandwidth resources are centrally managed and sliced among heterogeneous BSs to improve resource utilization and achieve QoS isolation for the coexistence of data service and M2M service. To obtain a set of optimal bandwidth slicing ratios for each BS, a network utility maximization problem is formulated under the constraints of differentiated QoS guarantee for data and M2M

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2018.2859740, IEEE Transactions on Vehicular Technology

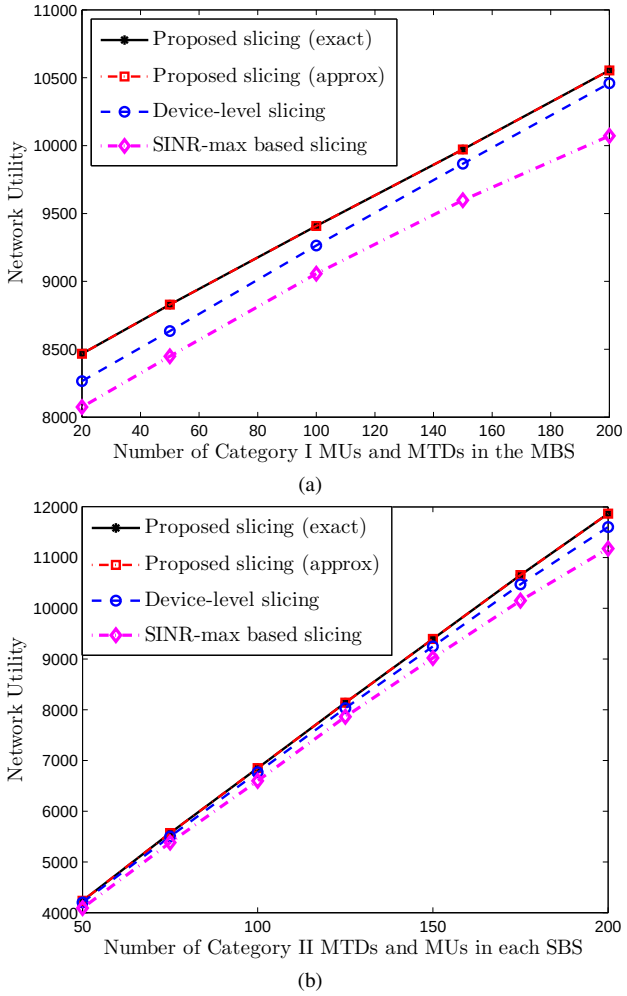


Fig. 11: Comparison of aggregate network utility (a) with respect to the number of category I MUs and MTDs ($N_u = N_a, N_s = 140, M_s = 10$) and (b) with respect to the number of category II MTDs and MUs ($N_u = N_a = 50$).

services, BS-device association patterns, and resource allocation among end devices. To reduce complexity, the original optimization problem is transformed to a tractable biconcave maximization problem. Then, an alternative concave search algorithm is designed to solve the transformed problem for a set of optimal slicing ratios and optimal BS-device association patterns. Simulation results demonstrate the robustness of the proposed algorithm due to its good convergence property and low computational complexity. In comparison with the two other resource slicing schemes, the proposed framework has lower communication overhead for updating the slicing ratios, achieves a larger device admission region for each cell, and provides higher network utility.

Appendix A. Proof of Proposition 1

For brevity, only the proof for (17) in Proposition 1 is provided. Since the bandwidth, W_s, is reused among all SBSs and the fraction of bandwidth resources allocated to device i from one SBS is independent of the fraction allocated to device q from another SBS, $(\mathrm{S2P1}')$ can be decoupled into n subproblems, each for one SBS. The subproblem for the SBS S_k $(k \in \{1, 2, ..., n\})$ is formulated as

$$(\mathrm{S2P1}'-1): \max_{f_{i,k}} u_k^{(1)}(f_{i,k})$$

$$\text{s.t.} \begin{cases} \sum\limits_{i\in\overline{\mathcal{N}_k'}} f_{i,k} = 1 & \text{(30a)} \\ f_{i,k} \in (0,1), \quad i \in \overline{\mathcal{N}_k'}. & \text{(30b)} \end{cases}$$

The objective function of $(\mathrm{S2P1}'-1)$ can be further derived as

$$\begin{aligned} u_k^{(1)}(f_{i,k}) &= \sum_{i\in\overline{\mathcal{N}_k'}} \log\left(W_v \beta_s f_{i,k} r_{i,k}\right) \\ &= \log\left(\prod_{i\in\overline{\mathcal{N}_k'}} W_v \beta_s r_{i,k}\right) + \log\left(\prod_{i\in\overline{\mathcal{N}_k'}} f_{i,k}\right). \end{aligned} \tag{31}$$

Since $r_{i,k}$ is considered constant during each bandwidth slicing period and is independent of $f_{i,k}$, $(\mathrm{S2P1}'-1)$ is equivalent to

$$(\mathrm{S2P1}'-2): \max_{f_{i,k}} \prod_{i\in\overline{\mathcal{N}_k'}} f_{i,k}$$

$$\text{s.t.} \begin{cases} \sum\limits_{i\in\overline{\mathcal{N}_k'}} f_{i,k} = 1 & \text{(32a)} \\ f_{i,k} \in (0,1), \quad i \in \overline{\mathcal{N}_k'}. & \text{(32b)} \end{cases}$$

Since geometric average is no greater than arithmetic average, we have

$$\prod_{i\in\overline{\mathcal{N}_k'}} f_{i,k} \le \left(\frac{\sum\limits_{i\in\overline{\mathcal{N}_k'}} f_{i,k}}{|\overline{\mathcal{N}_k'}|}\right)^{|\overline{\mathcal{N}_k'}|} \tag{33}$$

where the equal sign holds when $f_{i,k} = f_{l,k}, \forall\, i, l \in \overline{\mathcal{N}_k'}$, and $|\cdot|$ denotes a set cardinality. Thus, by satisfying constraints (32a) and (32b), the optimal fraction of bandwidth resources allocated to device i associated with S_k $(k \in \{1, 2, ..., n\})$ is

$$f_{i,k}^* = \frac{1}{|\overline{\mathcal{N}_k'}|} = \frac{1}{\sum\limits_{l\in\mathcal{N}_k\cup\mathcal{M}_k} x_{l,k}} \triangleq f_k^*. \tag{34}$$

Similar proof for (16) in Proposition 1 can also be made, which is omitted here.

Appendix B. Proof of Proposition 2

Given β_m, we first calculate the Hessian matrix of $u_m^{(2)}(\beta_m, \widetilde{\mathbf{X}_m})$ with respect to $\widetilde{\mathbf{X}_m}$. That is,

$$\mathbf{H}\left[u_m^{(2)}\left(\beta_m, \widetilde{\mathbf{X}_m}\right)\right] = \begin{bmatrix} -\frac{1}{h(\widetilde{\mathbf{X}_m})} & -\frac{1}{h(\widetilde{\mathbf{X}_m})} & \cdots & -\frac{1}{h(\widetilde{\mathbf{X}_m})} \\ -\frac{1}{h(\widetilde{\mathbf{X}_m})} & -\frac{1}{h(\widetilde{\mathbf{X}_m})} & \cdots & -\frac{1}{h(\widetilde{\mathbf{X}_m})} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{h(\widetilde{\mathbf{X}_m})} & -\frac{1}{h(\widetilde{\mathbf{X}_m})} & \cdots & -\frac{1}{h(\widetilde{\mathbf{X}_m})} \end{bmatrix} \tag{35}$$

where $h(\widetilde{\mathbf{X}_m}) = N_u + N_a + \sum\limits_{k=1}^{n} \sum\limits_{i=1}^{N_k+M_k} \widetilde{x_{i,m}}$, and the dimension of the matrix is $\left[\sum\limits_{k=1}^{n}(N_k+M_k)\right] \times \left[\sum\limits_{k=1}^{n}(N_k+M_k)\right]$.

For any non-zero vector $v = (v_1, v_2, ..., v_y) \in \mathbf{R}^y$, $y = \sum_{k=1}^{n} (N_k + M_k)$, we have

$$v^T \mathbf{H}\left[u_m^{(2)}\left(\beta_m, \widetilde{\mathbf{X}_m}\right)\right] v = -\frac{\left(\sum_{i=1}^{y} v_i\right)^2}{h(\widetilde{\mathbf{X}_m})} < 0. \tag{36}$$

Since the Hessian matrix $\mathbf{H}\left[u_m^{(2)}\left(\beta_m, \widetilde{\mathbf{X}_m}\right)\right]$ is negative definite, $u_m^{(2)}\left(\beta_m, \widetilde{\mathbf{X}_m}\right)$ is a (strictly) concave function in terms of $\widetilde{\mathbf{X}_m}$ for any fixed β_m. Conversely, it is obvious that $u_m^{(2)}(\beta_m, \widetilde{\mathbf{X}_m})$ is a (strictly) concave function with respect to β_m for any given $\widetilde{\mathbf{X}_m}$.

Similarly, $u_k^{(2)}\left(\beta_s, \widetilde{\mathbf{X}_k}\right)$ can also be proved as a (strictly) biconcave function. The summation $\sum_{k=1}^{n} u_k^{(2)}\left(\beta_s, \widetilde{\mathbf{X}_k}\right)$ is a nonnegative linear combination of a set of biconcave functions, which is also (strictly) biconcave [32].

APPENDIX C. PROOF OF PROPOSITION 3

Proof: Property (1) in Proposition 3 can be easily verified for (P3′). To verify the uniqueness of the set of optimal solutions, $\{\beta_m^{(t+1)}, \beta_s^{(t+1)}, \widetilde{\mathbf{X}_m}^{(t+1)}, \widetilde{\mathbf{X}_k}^{(t+1)}\}$, at the end of t th iteration, we refer to the proof of Proposition 2 that, given $\{\beta_m^{(t)}, \beta_s^{(t)}\}$, the objective function of (P3′) is a (strictly) concave function in terms of $\{\widetilde{\mathbf{X}_m}, \widetilde{\mathbf{X}_k}\}$ and, given $\{\widetilde{\mathbf{X}_m}^{(t+1)}, \widetilde{\mathbf{X}_k}^{(t+1)}\}$, the objective function of (P3′) is also a (strictly) concave function with respect to $\{\beta_m, \beta_s\}$.

REFERENCES

[1] 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.

[2] Q. Ye and W. Zhuang, "Distributed and adaptive medium access control for Internet-of-Things-enabled mobile networks," *IEEE Internet Things J.*, vol. 4, no. 2, pp. 446–460, Apr. 2017.

[3] Q. Ye and W. Zhuang, "Token-based adaptive MAC for a two-hop Internet-of-Things enabled MANET," *IEEE Internet Things J.*, vol. 4, no. 5, pp. 1739–1753, Oct. 2017.

[4] S. Zhang, J. Gong, S. Zhou, and Z. Niu, "How many small cells can be turned off via vertical offloading under a separation architecture?" *IEEE Trans. Wireless Commun.*, vol. 14, no. 10, pp. 5440–5453, Oct. 2015.

[5] N. Zhang, S. Zhang, S. Wu, J. Ren, J. W. Mark, and X. Shen, "Beyond coexistence: Traffic steering in LTE networks with unlicensed bands," *IEEE Wireless Commun.*, vol. 23, no. 6, pp. 40–46, Dec. 2016.

[6] 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.

[7] C. Liang and F. R. Yu, "Wireless network virtualization: A survey, some research issues and challenges," *IEEE Commun. Surv. Tutor.*, vol. 17, no. 1, pp. 358–380, Firstquarter, 2015.

[8] F. Bari, S. R. Chowdhury, R. Ahmed, R. Boutaba, and O. C. M. B. Duarte, "Orchestrating virtualized network functions," *IEEE Trans. Netw. Serv. Manage.*, vol. 13, no. 4, pp. 725–739, Dec. 2016.

[9] R. Riggio, A. Bradai, D. Harutyunyan, T. Rasheed, and T. Ahmed, "Scheduling wireless virtual networks functions," *IEEE Trans. Netw. Serv. Manage.*, vol. 13, no. 2, pp. 240–252, Jun. 2016.

[10] M. T. Beck and J. F. Botero, "Coordinated allocation of service function chains," in *Proc. IEEE GLOBECOM' 15*, 2015, pp. 1–6.

[11] M. Richart, J. Baliosian, J. Serrat, and J. L. Gorricho, "Resource slicing in virtual wireless networks: A survey," *IEEE Trans. Netw. Serv. Manage.*, vol. 13, no. 3, pp. 462–476, Sept. 2016.

[12] Q. Duan, N. Ansari, and M. Toy, "Software-defined network virtualization: An architectural framework for integrating SDN and NFV for service provisioning in future networks," *IEEE Netw.*, vol. 30, no. 5, pp. 10–16, Sept. 2016.

[13] B. A. A. Nunes, M. Mendonca, X. N. Nguyen, K. Obraczka, and T. Turletti, "A survey of software-defined networking: Past, present, and future of programmable networks," *IEEE Commun. Surv. Tutor.*, vol. 16, no. 3, pp. 1617–1634, Third, 2014.

[14] A. Belbekkouche, M. M. Hasan, and A. Karmouch, "Resource discovery and allocation in network virtualization," *IEEE Commun. Surv. Tutor.*, vol. 14, no. 4, pp. 1114–1128, Fourth, 2012.

[15] Q. Ye, J. Li, K. Qu, W. Zhuang, X. Shen, and X. Li, "End-to-end quality of service in 5G networks: Examining the effectiveness of a network slicing framework," *IEEE Veh. Technol. Mag.*, vol. 13, no. 2, pp. 65–74, Jun. 2018.

[16] T. P. C. de Andrade, C. A. Astudillo, and N. L. S. da Fonseca, "Allocation of control resources for machine-to-machine and human-to-human communications over LTE/LTE-A networks," *IEEE Internet Things J.*, vol. 3, no. 3, pp. 366–377, Jun. 2016.

[17] R. Kokku, R. Mahindra, H. Zhang, and S. Rangarajan, "NVS: A substrate for virtualizing wireless resources in cellular networks," *IEEE/ACM Trans. Netw.*, vol. 20, no. 5, pp. 1333–1346, Oct. 2012.

[18] A. A. Gebremariam, M. Chowdhury, A. Goldsmith, and F. Granelli, "Resource pooling via dynamic spectrum-level slicing across heterogeneous networks," in *Proc. IEEE CCNC' 17*, Jan. 2017, pp. 818–823.

[19] X. Foukas, M. K. Marina, and K. Kontovasilis, "Orion: RAN slicing for a flexible and cost-effective multi-service mobile network architecture," in *Proc. ACM MobiCom '17*, Oct. 2017, pp. 127–140.

[20] M. Li, F. R. Yu, P. Si, E. Sun, Y. Zhang, and H. Yao, "Random access and virtual resource allocation in software-defined cellular networks with machine-to-machine communications," *IEEE Trans. Veh. Technol.*, vol. 66, no. 7, pp. 6399–6414, Jul. 2017.

[21] P. Caballero, A. Banchs, G. de Veciana, and X. Costa-Prez, "Multi-tenant radio access network slicing: Statistical multiplexing of spatial loads," *IEEE/ACM Trans. Netw.*, vol. 25, no. 5, pp. 3044–3058, Oct. 2017.

[22] R. Sherwood, G. Gibb, K.-K. Yap, G. Appenzeller, M. Casado, N. McKeown, and G. Parulkar, "Flowvisor: A network virtualization layer," *OpenFlow Switch Consortium, Tech. Rep*, vol. 1, p. 132, 2009.

[23] E. Soltanmohammadi, K. Ghavami, and M. Naraghi-Pour, "A survey of traffic issues in machine-to-machine communications over LTE," *IEEE Internet Things J.*, vol. 3, no. 6, pp. 865–884, Dec. 2016.

[24] 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.

[25] D. Wu and R. Negi, "Effective capacity-based quality of service measures for wireless networks," *ACM Mobile Netw. Appl.*, vol. 11, no. 1, pp. 91–99, 2006.

[26] Y. Liu, C. Yuen, X. Cao, N. U. Hassan, and J. Chen, "Design of a scalable hybrid MAC protocol for heterogeneous M2M networks," *IEEE Internet Things J.*, vol. 1, no. 1, pp. 99–111, Feb. 2014.

[27] P. Rabinovitch, "Statistical estimation of effective bandwidth," Ph.D. dissertation, Carleton University, 2000.

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

[29] C. Y. Oh, D. Hwang, and T. J. Lee, "Joint access control and resource allocation for concurrent and massive access of M2M devices," *IEEE Trans. Wireless Commun.*, vol. 14, no. 8, pp. 4182–4192, Aug. 2015.

[30] D. Wu and R. Negi, "Effective capacity: A wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 630–643, Jul. 2003.

[31] A. Abdrabou and W. Zhuang, "Stochastic delay guarantees and statistical call admission control for IEEE 802.11 single-hop ad hoc networks," *IEEE Trans. Wireless Commun.*, vol. 7, no. 10, pp. 3972–3981, Oct. 2008.

[32] 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, 2007.

[33] S. P. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[34] A. Ben-Tal and A. Nemirovski, *Lectures on modern convex optimization: Analysis, algorithms, and engineering applications*. Society for Industrial and Applied Mathematics (SIAM), 2001.

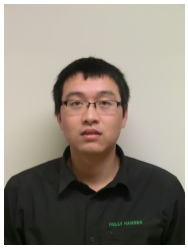
[35] "3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Feasibility Study on New Services and Markets Technology Enablers for Critical Communications; Stage 1 (Release 14) ," *3GPP TR 22.862 V14.1.0*, pp. 1–31, Sept. 2016.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2018.2859740, IEEE Transactions on Vehicular Technology

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.

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).

Shan Zhang (S13-M16) received her Ph.D. degree in electronic engineering from Tsinghua University, Beijing, China, in 2016. She is currently an assistant professor in the School of Computer Science and Engineering, Beihang University, Beijing, China. She was a post doctoral fellow in Department of Electronical and Computer Engineering, University of Waterloo, Ontario, Canada, from 2016 to 2017. Her research interests include mobile edge computing, wireless network virtualization and intelligent management. Dr. Zhang received the Best Paper Award at the Asia-Pacific Conference on Communication in 2013.



A-Long Jin A-Long Jin received his B.Eng. degree in communications engineering from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2012. He received his M.Sc. degree in computer science from University of New Brunswick, Fredericton, NB, Canada, in 2015. His research interests include software-defined networking and network function virtualization.

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.