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 developed 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 re-managed 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 with network load dynamics. To determine the set of optimal bandwidth slicing ratios and optimal BS-device (user) association patterns, a network utility maximization problem is formulated with the consideration of different traffic statistics and QoS requirements, location distribution for end devices, load conditions in each cell, wireless channel conditions 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.

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, alongjin, sshen}@uwaterloo.ca).

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

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], [2]. 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). Thus, 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' small-cells) is a potential solution to improve the spectrum efficiency, by exploring spatial multiplexing on currently employed spectrum [3]; Moreover, the heterogeneity of both service type and device type necessitates the exploration of unlicensed bands and underutilized spectrum through different wireless access technologies [4], [5]. For example, long-term evolution (LTE) devices can utilize WiFi unlicensed bands (with high multiplexing gain) to support delay-insensitive services [5].

However, spectrum exploitations have limitations. 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 on each BS [3]. 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 Internet, where a set of network/service functions (e.g., firewalls, load balancing, wide area network (WAN) optimizers) are decoupled from the physical hardware (i.e., middleboxes) 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[1] [12], which has direct control (programmability) on all BSs and the associated radio resources [14]. With SDN-enabled function softwarization, radio resources of heterogeneous BSs are reconfigured (reallocated) by the central controller to improve the overall resource utilization. This process is called *radio resource slicing*.

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 (i.e., end device locations and density, channel conditions, traffic statistics on each device) over the physical network, which facilitates the resource sharing without distributed information exchange; (3) The coexistence of heterogeneous services requires *QoS isolation (resource isolation)*[2] be attained 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.

However, most existing radio resource slicing schemes are device-level slicing, which indicates that the amount of preallocated (fixed) radio resources on a BS are sliced and allocated to different device (or user) groups in the BS's coverage. The network-level resource sharing and resource slicing among heterogeneous BSs are rarely considered so far. As a matter of fact, the essence of network (radio) function virtualization (softwarization) is to enable radio resource abstraction for the purpose of managing the resources among heterogeneous wireless networks. A few studies consider 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) [17], [18]. Therefore, to satisfy differentiated QoS

[1]In the SDN architecture, all control functions are decoupled from the physical network and integrated at an SDN controller to achieve direct control and flexible programmability on the substrate networks [11], [13].

[2]QoS isolation refers to that any change in network state for one type of service, including end device (user) mobility, channel dynamics, and traffic load fluctuations, should not violate the minimum QoS performance experienced by devices (users) belonging to another service type [15], [16].

requirements from heterogeneous services, traffic statistics for each specific type of service should be modeled clearly and considered in designing the optimal resource slicing policy. Therefore, in this paper, we develop a comprehensive radio resource (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 the paper are three-folded:

1) We consider the coexistence of machine-type devices (MTDs) with mobile users (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 for both machine-type and data services, traffic load in each cell, channel conditions between BSs and MTDs (MUs) with inter-cell interference. The outputs of the optimization problem are BS-device (user) association patterns and optimal bandwidth slicing ratios (i.e., fractions of radio resources associated with each BS and allocated to each type of service to achieve QoS isolation);
2) For tractability, the original optimization problem is transformed to a biconcave maximization problem based on certain approximations. Then, an alternative concave search (ACS) algorithm is designed to iteratively solve the transformed problem. Based on some properties of the 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 remain stable with end device (user) location change under the condition that each device moves within its associated cell coverage area, which indicates the slicing ratios need to be updated only in a long time scale and the associated communication overhead (i.e., global network information exchange between each network cell and the central controller [6], [19]) are significantly reduced; Second, with our proposed bandwidth slicing framework, the BS-device (user) connections are more stable, which saves the communication overhead for frequent BS-device (user) re-association with network dynamics; Third, the designed ACS algorithm is both robust and lightweight. Upon performance comparison with the 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 the 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 designed to solve the transformed problem to obtain a set of optimal bandwidth slicing ratios and optimal BS-device (user) association patterns. In Section VI, extensive simulation results are provided to show insights of the proposed bandwidth slicing framework. Finally, conclusions are drawn in Section VII. Important symbols are listed in Table I.

TABLE I: Important symbols

Symbol	Definition
$\mathcal{B}/\mathcal{S}$	The set of MBSs/SBSs in the HetNet
B_m	A tagged MBS
$c_{i,m}/c_{i,k}$	Effective achievable rate at MTD i or MU i from B_m/S_k
c^{min}	Minimum effective 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
$G_{i,m}/G_{i,k}$	Channel gain between B_m/S_k and MTD i or MU i
L_a/L_d	Machine-type/Data packet size
$\mathcal{N}_a/N_a$	Set/Number of category I MTDs
$\mathcal{N}_k/N_k$	Set/Number of category II MTDs in the coverage of S_k
$\mathcal{N}_u/N_u$	Set/Number of MUs
P_m/P_k	Transmit power on B_m/S_k
$r_{i,m}/r_{i,k}$	Effective 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 MTD 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

II. Related Work

Due to the advantages in reducing infrastructure deployment cost and improving spectrum utilization, resource slicing in wireless networks start to draw attention from researchers. In [9], radio resource slices (time resource slices) are created for different service providers (SPs) to obtain the requested service capacities (i.e., aggregated throughputs requested by different groups of end users) and QoS isolation. However, a proper resource slicing strategy needs to be designed, by considering specific service requirements, end-user distribution and traffic statistics,

and varying channel conditions, to better utilize the radio resources. 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 (within its coverage) 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 sliced and 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 (weights) 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 [16], [22]. For instance, in [16], 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.

Existing studies mainly focus on the device-level resource slicing, in which the preallocated radio resources on each BS are sliced among different device (or 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 [17], without an explicit characterization on differentiated traffic statistics and QoS descriptions for heterogeneous services. The spectrum sharing is triggered only when traffic overload happens in one of the coexisting virtual networks. In this paper, we develop a comprehensive radio resource slicing framework to facilitate resource sharing among heterogeneous BSs with the consideration of differentiated QoS provisioning for both data and M2M services, BS-device (user) association patterns, and instantaneous traffic load conditions for each cell.

III. System Model

We consider a two-tier downlink[3] HetNet, where a number of macro-cells with a set $\mathcal{B}$ of MBSs at the center of each macro-cell in the first tier are underlaid by small cells (with a set S of SBSs placed at cell centers) in the second tier. In Fig. 1, we show one tagged macro-cell underlaid by n small cells within its coverage. There are two types of end devices (users)

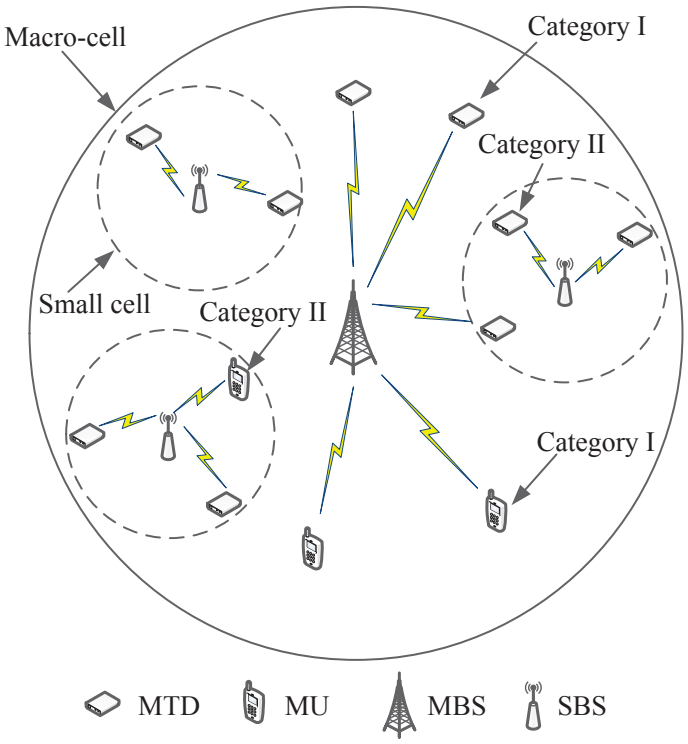

Fig. 1: A two-tier macrocell overlaid network with the coexistence of MTDs and MUs.

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 mobile, and we assume that MTDs are more or less stationary [15], [23]. For the tagged macro-cell, an MBS, denoted by $B_m\,(\in \mathcal{B})$, with high transmit power is deployed for a wide-area communication coverage, including both control signaling and real-data transmissions. The n SBSs, $\{S_1, S_2, ..., S_n\}$, with lower transmit power and shorter coverage are placed at some hotspot locations within the MBS's coverage area to increase the whole network capacity. Since the HetNet is expected to accommodate a much larger amount of M2M devices than MUs, the

[3]Radio resource slicing for an uplink HetNet requires more complicated inter-cell interference characterization and power control of end devices, which affects BS-device (user) association patterns and resource allocation for end devices, and will be investigated in our future work.

SBSs are specifically deployed to enhance the network capacity in response to the increasing M2M traffic volume [24]. In most of the cases, MUs, deployed within the coverage of the macro-cell, are connected to the MBS [3], [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 MTD (or MU) is within the macro-cell coverage but outside the coverages of all small cells, and is associated with the MBS for machine-type (or data) packet transmissions; A category II MTD (or MU) 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 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 (users) can be associated with B_m. There are a number of transmission queues at each BS, each of which is used for downlink transmissions between a BS and an associated MTD (MU). We assume data packets[4] destined for an MU arrive at a transmission queue in an MBS periodically with rate λ_d packet/s, and each data packet has fixed size of L_d bits, whereas machine-type packets for one of the MTDs arrive at the MBS or an SBS in an event-driven manner with a much lower packet arrival rate and a smaller packet size [15]. 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 fixed packet size L_a bits. Due to the stochastic 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 detail in Section IV). All packets are transmitted through the physical layer wireless propagation channel to reach to an intended MTD or MU.

A. Communication Model

Suppose that two sets of radio resources, W_m and W_s, are initially allocated to each 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

[4]In this paper, we use link-layer packetized traffic to model arrivals of both data traffic and M2M traffic for explicit QoS characterization [25].

reused among all SBSs under an acceptable inter-cell interference level to improve the spectrum utilization. The frequency reuse factor is set to 1 for each tier. Transmit power for each MBS and each SBS are preallocated and remain constant during each resource slicing period (i.e., resource slicing is updated when traffic load of each cell fluctuates [17]). The transmit power for the tagged MBS, B_m, overlaid by n SBSs, $S_k\,(k=1,2,...,n)$, are denoted by P_m and $P_k\,(k=1,2,...,n)$, respectively. Therefore, using Shannon capacity formula, the spectrum efficiency at a tagged MTD i (or MU i) from B_m and $S_k\,(k=1,2,...,n)$ are expressed, respectively, as

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

and

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

In (1) and (2), $r_{i,m}$ and $r_{i,k}$ are measured in a large time scale and are termed as *effective spectrum efficiency* [28], to capture long-term wireless channel conditions, which include path loss and shadowing effects. $G_{i,m}$ and $G_{i,k}$ are channel gains between B_m and S_k and MTD i (or MU i), and $G_{i,h}$ and $G_{i,j}$ are channel gains from an MBS and an SBS other than B_m and S_k to MTD i (or MU i). P_h and P_j are downlink transmit power at an MBS and an SBS other than B_m and S_k. σ^2 denotes average background noise power. Inter-cell interference among macro-cells and among small cells is included in (1) and (2). From (1) and (2), the effective achievable rate at each MTD and MU can be obtained based on BS-device (user) 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 (user) association and bandwidth allocation for end devices, is updated in a large time scale compared with channel dynamics to reduce communication overhead [28]. Therefore, the received signal-to-interference-plus-noise (SINR) in (1) or (2) is the average SINR over each resource slicing period [17], which is also termed as effective SINR (channel fast fading effects are also averaged out).

B. Dynamic Bandwidth Slicing Framework

As stated in Section I, in a multi-tier HetNet supporting heterogeneous services, the pre-allocated (fixed) radio resources at each BS can be inefficient due to unbalanced end device

(user) distribution, changing wireless channel conditions 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, 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. However, a proper 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/user distribution, cell loads, wireless channel conditions and interference levels, BS-device (user) association patterns) and service-level characteristics (i.e., traffic statistics and QoS requirements for heterogeneous services). Moreover, resources for each slice is expected to be updated in a large time scale to reduce the communication overhead for network information exchange between the central controller and BSs in each cell.

Resource slicing is realized in two steps: In the first step, with function softwarization, the controller has a 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]. The total bandwidth resources are then sliced among MBSs and SBSs, as shown in Fig. 2.

In the second step, the central controller needs to specify how bandwidth resources are sliced 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) allocated to MBSs 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 dynamics

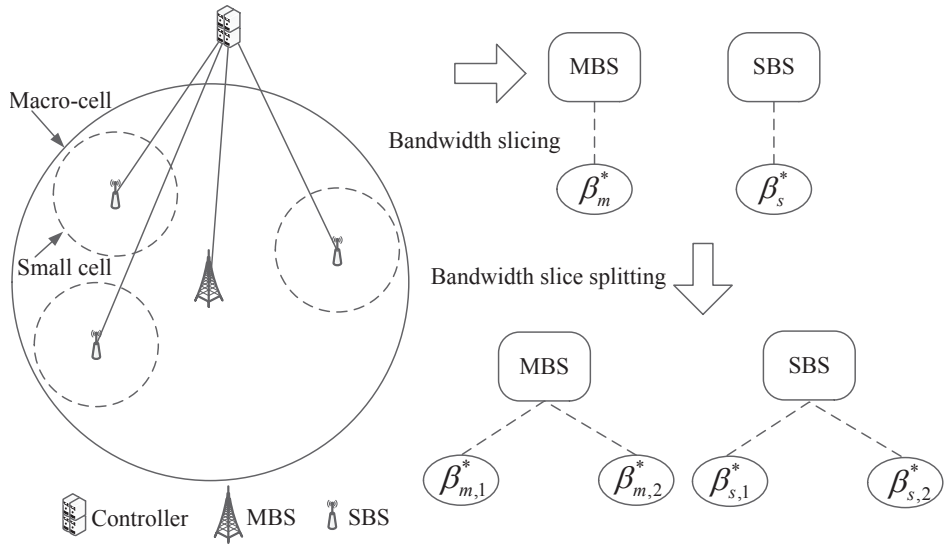

Fig. 2: Radio resource slicing and splitting.

to improve the overall resource utilization. Since MBSs 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 also 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.

IV. Problem Formulation

In the proposed bandwidth slicing framework, the challenging research 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 differentiated QoS requirements between M2M and data services.

We consider logarithmic functions, defined in (3) and (4), as utility functions to indicate the utility 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. The reason for the logarithmic utility function is that it is a concave function having diminishing marginal utility, which facilitates network load balancing in BS-device (user) association and achieves certain fairness in resource allocation among end devices [6], [17], [21].

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

and

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

In (3) and (4), $c_{i,m}$ denotes the effective achievable rate at MU i ($i \in \mathcal{N}_u \cup \mathcal{M}_k$) or MTD i ($i \in \mathcal{N}_a \cup \mathcal{N}_k$) associated with B_m, and $c_{i,k}$ the effective achievable rate at category II MTD i ($i \in \mathcal{N}_k$) or MU i ($i \in \mathcal{M}_k$) from S_k; $f_{i,m}$ is the fraction of bandwidth resources (out of $W_v \beta_m$) allocated to MU i ($i \in \mathcal{N}_u \cup \mathcal{M}_k$) or MTD i ($i \in \mathcal{N}_a \cup \mathcal{N}_k$) from B_m, and $f_{i,k}$ the fraction of bandwidth resources (out of $W_v \beta_s$) allocated to category II MTD i ($i \in \mathcal{N}_k$) or category II MU i ($i \in \mathcal{M}_k$) from S_k.

As stated in Section III-A, the downlink effective achievable rates, $c_{i,m}$ and $c_{i,k}$, are treated 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 feature of the event-driven M2M service, the downlink total transmission delay from B_m or S_k to an MTD should be guaranteed in a statistical way for maximal resource utilization. The delay is the duration from the instant a machine-type packet arrives at a transmission queue of a BS to the instant the intended MTD receives the packet. We apply the effective bandwidth theory to derive the minimum service rate (received effective achievable 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 $[0, t)$, $E[\cdot]$ denotes the operation of expectation. Since $A(t)$ is modeled as a Poisson process with the average 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 $j = m$ (m for the MBS B_m index), or $i \in \mathcal{N}_k$ if $j = k$ (k for the SBS S_k index, $k \in \{1, 2, ..., n\}$), ε denotes the maximum delay bound violation probability, $p_{i,j} = \frac{c_{i,j}}{L_a}$ is the effective achievable rate at MTD i from BS j in terms of number of packets per second. From (7), the minimum effective achievable rate $p^{(min)}$ of $p_{i,j}$ to achieve ε 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 $\varrho(\varphi_M)$ 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 $(\mathrm{P1})$, under the constraints of differentiated QoS guarantee, BS-device (user) association patterns and bandwidth allocation for each MTD and MU:

$(\mathrm{P1})$:

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

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

In $(\mathrm{P1})$, the objective function is the aggregate network utility, which is the summation of utilities achieved by each MU and MTD. An MTD $i\,(\in \mathcal{N}_a)$ or MU $i\,(\in \mathcal{N}_u)$ is associated with B_m, whereas a category II MTD $i\,(\in \mathcal{N}_k)$ or a category II MU $i\,(\in \mathcal{M}_k)$ 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 a category II MTD (or MU) i with B_m or S_k. MTD (or MU) 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 $(\mathrm{P1})$, constraints (10a) and (10d) indicate that the service rate $c_{i,j}$ for any MU i associated with B_m or S_k is not less than the periodic traffic arrival rate destined for MU i at a transmission queue of B_m, to guarantee the throughput requirement. Constraints (10b) and (10c) ensure that the achievable rate at 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 MTD or MU 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 (user) 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.

A. Optimal Bandwidth Allocation to MUs and MTDs

Inspired by [19], we first simplify $(\mathrm{P1})$ 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 $(\mathrm{P1})$ can be expressed as a summation of $u^{(1)}_m(f_{i,m})$ and $\sum_{k=1}^{n} u^{(1)}_k(f_{i,k})$. Hence, $u^{(1)}_m(f_{i,m})$ is a function of $f_{i,m}$, indicating the aggregate utility of MUs and MTDs associated with B_m, given by

$$u^{(1)}_m(f_{i,m}) = \sum_{i\in\mathcal{N}_u\cup\mathcal{N}_a} \log\left(W_v\beta_m f_{i,m} r_{i,m}\right) + \sum_{k=1}^{n}\sum_{i\in\mathcal{N}'_k} \log\left(W_v\beta_m f_{i,m} r_{i,m}\right) \tag{11}$$

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

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

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

Hence, $(\mathrm{P1})$ can be written as $(\mathrm{P1}')$:

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

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

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

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

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

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

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

Proposition 1. *The solutions for* $(\text{S1P1}')$ *and* $(\text{S2P1}')$ *are*

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

and

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

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:

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

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

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

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

and

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

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), $(\mathrm{P2})$ cannot be decoupled in the same way as $(\mathrm{P1}')$, which is difficult to solve due to the non-concavity of the problem structure. Therefore, in the next section, we transform $(\mathrm{P2})$ to a tractable form for optimal solutions.

V. Problem Transformation and Partial Optimal Solutions

To make $(\mathrm{P2})$ tractable, we first relax the binary variables $\{x_{i,j}\}$ in $(\mathrm{P2})$ to real-valued variables $\{\widetilde{x_{i,j}}\}$ within the range $[0,1]$. The variables $\{\widetilde{x_{i,j}}\}$ represent the fraction of time that MTD (or MU) i is associated with B_m or S_k during each bandwidth slicing period [6]. With the variable relaxation, the objective function of $(\mathrm{P2})$ becomes a summation of $u_m^{(2)}(\beta_m, \widetilde{\mathbf{X}_m})$ and $\sum\limits_{k=1}^{n} u_k^{(2)}\left(\beta_s, \widetilde{\mathbf{X}_k}\right)$, 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* $\mathcal{F} : \mathrm{Y} \to \mathbf{R}$ *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, $(\mathrm{P2})$ is transformed to $(\mathrm{P3})$:

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

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

In $(\mathrm{P3})$, the objective function is a nonnegative sum of two (strictly) biconcave functions,

which is also (strictly) biconcave [32]. $\widetilde{f_m^*}$ 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 $(\mathrm{P3})$ 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 still non-convex constraint functions. Constraints (21c) and (21d) actually indicate that, for any $i \in \mathcal{N}_k \cup \mathcal{M}_k \, (k \in \{1, 2, ..., n\})$, if it is associated with B_m, the following inequalities should be satisfied,

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

and

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

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

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

and

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

Therefore, to make $(\mathrm{P3})$ tractable, we simplify $(\mathrm{P3})$ to $(\mathrm{P3}')$, 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 $(\mathrm{P3})$, constraints (21a), (21b), (22) and (23) in $(\mathrm{P3}')$ provide the lowest upper bound on the number of category II MTDs and MUs that can be associated with B_m (This lowest upper bound is accurate because the communication distance between the MBS and any MTD or MU located in the coverage of an SBS is much longer than the location differences among MTDs and MUs in the same SBS and thus the differences of $r_{i,m}$ among the end devices are relatively small.). Similarly, compared with constraints (21e) and (21f) in $(\mathrm{P3})$, constraints (24) and (25) in $(\mathrm{P3}')$ indicate the lowest upper bound on the number

of category II MTDs and MUs that can be associated with $S_k, k \in \{1, 2, ..., n\}$. In fact, without changing the optimal solutions for $(\mathrm{P3})$, the simplified constraints (22) - (25) in $(\mathrm{P3}')$ indicate a set of conservative capacities on maximum numbers of category II MTDs and MUs that can be associated with B_m and $S_k, k \in \{1, 2, ..., n\}$.

If $(\mathrm{P3}')$ is written in a standard form, it 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 $(\mathrm{P3}')$, 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 $(\mathrm{P3}')$, Algorithm 1 converges to a set of *partial optimal solutions*. The definition of a partial optimal solution for $(\mathrm{P3}')$ 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* $(\mathrm{P3}')$*: (1) Both* $\{\beta_m, \beta_s\}$ *and* $\{\widetilde{\mathbf{X}_m}, \widetilde{\mathbf{X}_k}\}$ *are closed sets, and the objective function of* $(\mathrm{P3}')$ *is continuous on its domain; (2) Given the set of accumulation points*[5]*,* $\{\beta_m^{(t)}, \beta_s^{(t)}, \widetilde{\mathbf{X}_m}^{(t)}, \widetilde{\mathbf{X}_k}^{(t)}\}$*, 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.*

Proof: Property (1) can be easily verified for $(\mathrm{P3}')$. 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 $(\mathrm{P3}')$ 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 $(\mathrm{P3}')$ is also a (strictly) concave function with respect to $\{\beta_m, \beta_s\}$.

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*

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

[5]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.

and

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

Algorithm 1: The ACS algorithm for solving $(\mathrm{P3}')$

Input : Input parameters for $(\mathrm{P3}')$, 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 (user) 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 $(\mathrm{P3}')$ given $\{\beta_m^{(t)}, \beta_s^{(t)}\}$;

  if *No feasible solutions for* $(\mathrm{P3}')$ **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 $(\mathrm{P3}')$ given $\{\widetilde{\mathbf{X}_m}^{(t+1)}, \widetilde{\mathbf{X}_k}^{(t+1)}\}$;

    if *No feasible solutions for* $(\mathrm{P3}')$ **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;

The main logical flow for Algorithm 1 is to iteratively solve for optimal bandwidth slicing ratios and optimal BS-device (user) 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, $(\mathrm{P3}')$ is solved for a set of optimal BS-device (user) association patterns $\{\widetilde{\mathbf{X}_m}^{\dagger}, \widetilde{\mathbf{X}_k}^{\dagger}\}$ and then, given $\{\widetilde{\mathbf{X}_m}^{\dagger}, \widetilde{\mathbf{X}_k}^{\dagger}\}$,

$(\mathrm{P3}')$ 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 the stopping criterion[6] (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 $(\mathrm{P2})$. Based on the optimal solutions from Algorithm 1, the optimal bandwidth slicing ratios, α_1^* and α_2^*, for data service and M2M service (suppose the numbers of category II devices in each small cell are equal) are obtained by

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

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 $(\mathrm{P3}')$ is solved for the optimal BS-device (user) 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 $\{\widetilde{\mathbf{X}_m}^{(t+1)}, \widetilde{\mathbf{X}_k}^{(t+1)}\}$, $(\mathrm{P3}')$ 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 $\mathcal{O}\left[N_m \left(\sum_{k=1}^{n} (N_k + M_k) \right)^4 \right]$, where n is the number of small cells within the tagged macro-cell.

[6]The iteration limit is set large enough to ensure the stopping criterion be met. The convergence of Algorithm 1 for solving $(\mathrm{P3}')$ is proved in Proposition 3.

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 2 MBS-centered macro-cell underlaid by 4 SBS-centered small cells, locations of MBSs and SBSs are fixed, and the distances between MBSs and between each MBS and one of its covering SBSs are set as 1200 m and 400 m, respectively. The downlink transmit power on an 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 [3], [19]. All category I MUs and MTDs are randomly distributed in the coverage of each MBS (Category I MTDs are outside the coverages of SBSs.); Category II MTDs and MUs 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. 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 each macro-cell and each small cell, respectively, including path loss and shadowing effects, where z is the distance between a BS and a device or a user. The periodic data packet arrival rate λ_d at a transmission queue of an 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 an 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 resource slicing scheme mentioned in [19], in which radio resources are shared among heterogeneous BSs and each MTD (MU) is associated with the BS providing highest downlink SINR, and a device-level resource slicing scheme [6], i.e., bandwidth resources are preallocated (fixed) on each BS, and are then sliced for different device groups (with differentiated services) in the coverage region of the BS (no resource sharing among BSs).

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), under certain network load condition (i.e., numbers of MUs and MTDs in each 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 (user) location changes (i.e., variations of wireless channel conditions) inside each cell. Therefore, the slicing ratios are adjusted in a large time scale, which significantly reduces the communication overhead (i.e., 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^*.

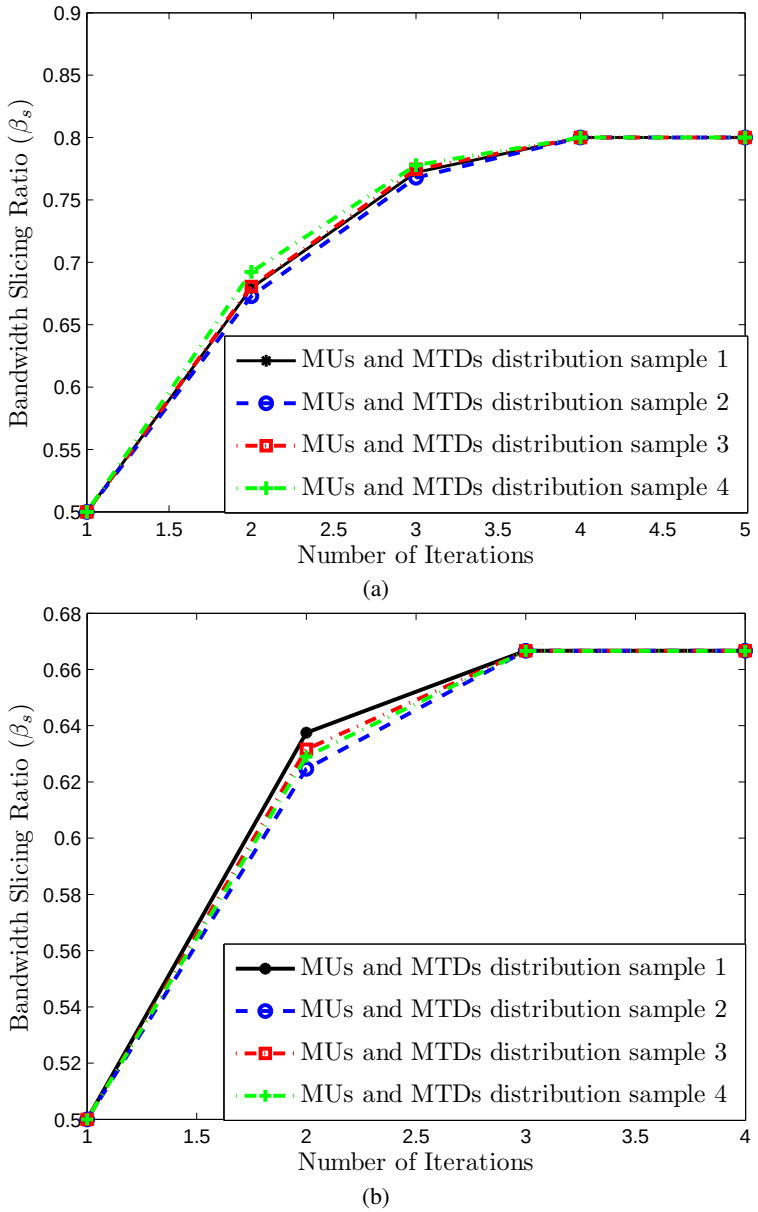

Fig. 3: Bandwidth slicing ratio (β_s) for an SBS during each iteration of Algorithm 1. (a) $N_u = 25$, $N_a = 25$, $N_s + M_s = 50$. (b) $N_u = N_a = N_s + M_s = 100$.

B. Performance Comparison

In Fig. 6, optimal bandwidth slicing ratios are compared between the proposed slicing framework and the SINR-max based slicing scheme under different network load conditions (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 (user) association patterns should change upon variations of wireless channel conditions and end device locations.

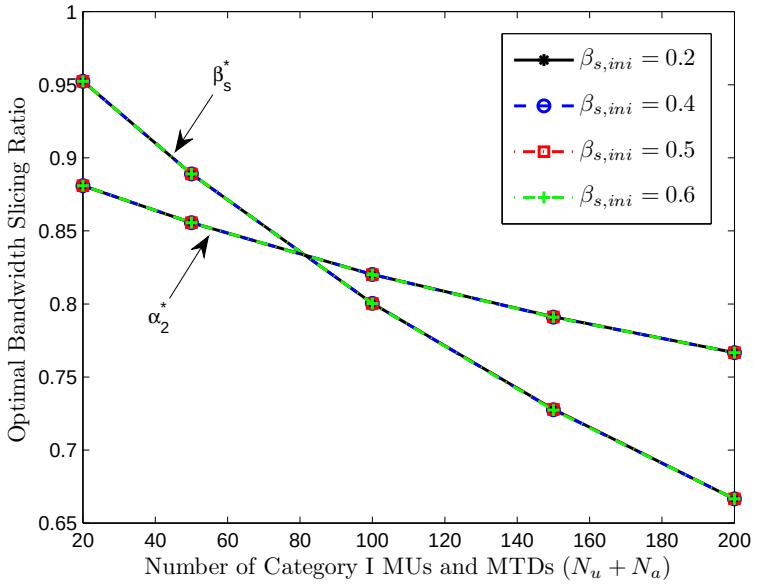

Fig. 4: Optimal bandwidth slicing ratios (β_s^* and α_2^*) ($N_s + M_s = 100$, $N_u = N_a$).

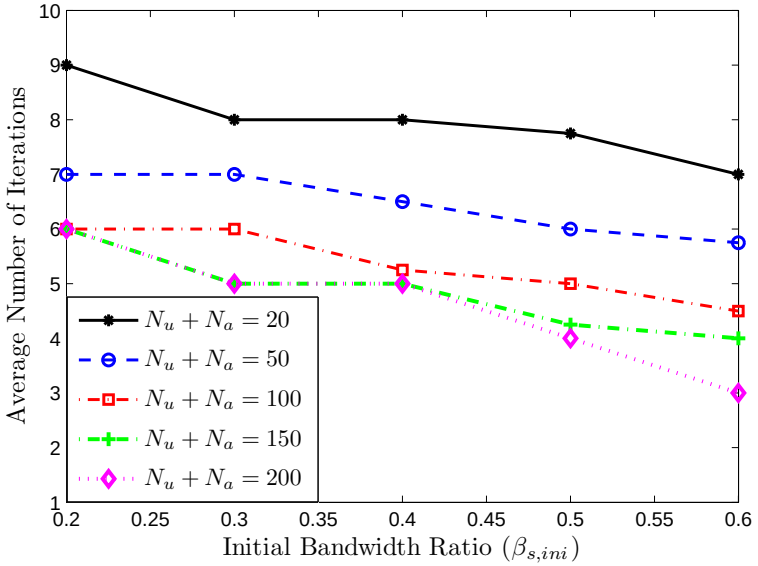

Fig. 5: Average number of iterations to solve for β_s^* using Algorithm 1 with different initial values $\beta_{s,ini}$ ($N_s + M_s = 100$, $N_u = N_a$).

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.

Next, we compare the performance of the proposed resource slicing framework with the

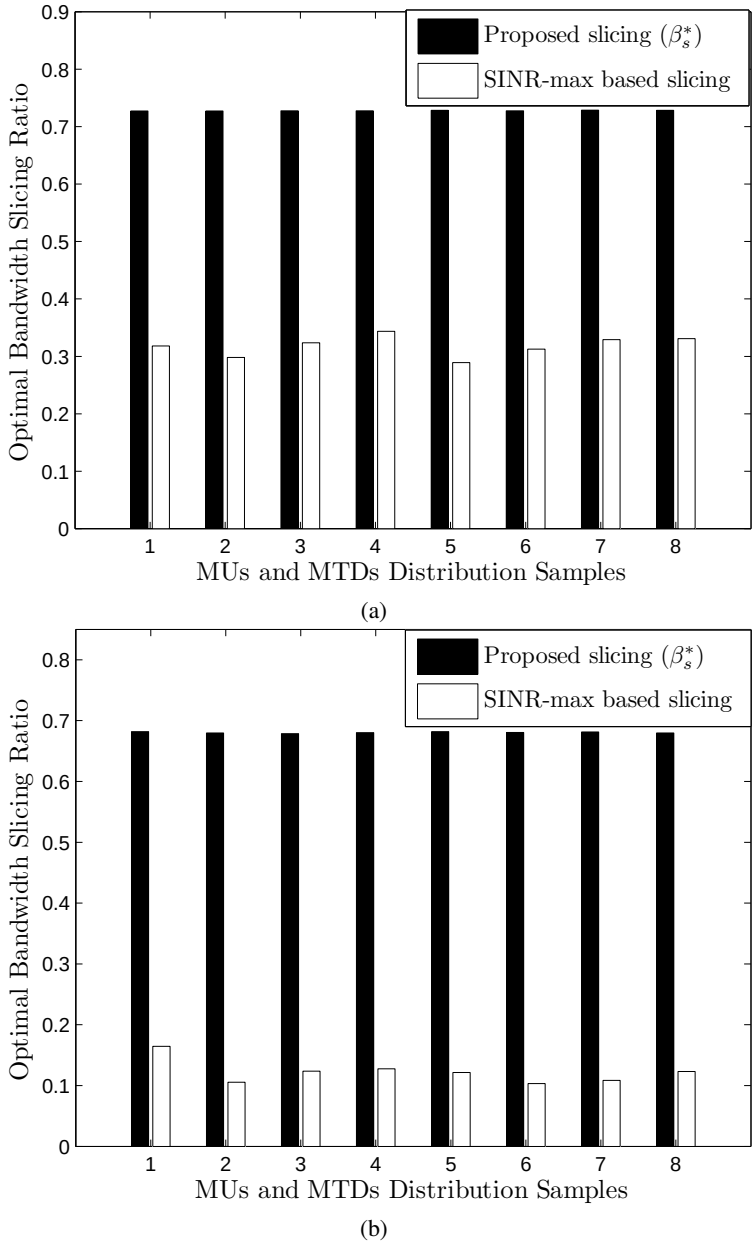

Fig. 6: Comparison of bandwidth slicing ratios between the proposed slicing framework and the SINR-max based slicing scheme. (a) $N_u = 75$, $N_a = 75$, $N_s + M_s = 100$. (b) $N_u = 100$, $N_a = 100$, $N_s + M_s = 150$.

device-level resource slicing scheme (no resource sharing among heterogeneous BSs) 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, 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.

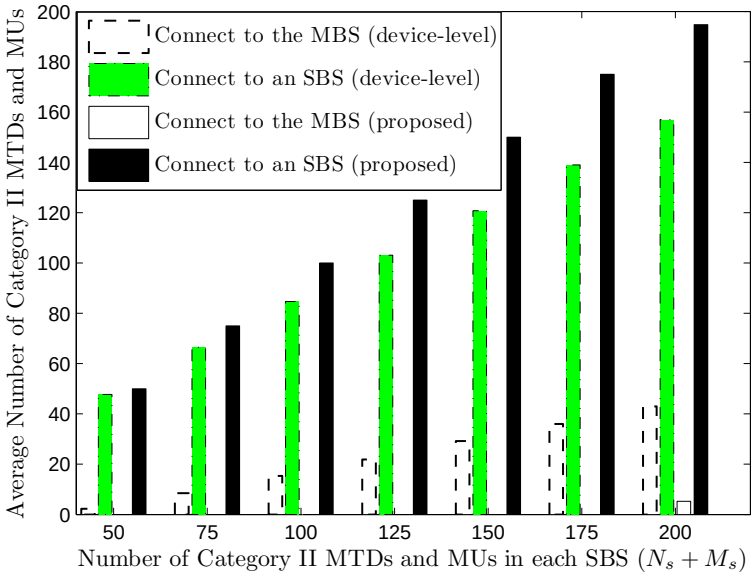

Fig. 7: Average number of category II MTDs and MUs connecting to either the MBS or their home SBSs ($N_u = N_a = 50$).

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.

Figs. 9(a) and 9(b) show that the proposed bandwidth slicing framework significantly improves the resource utilization, by increasing the capacity for MUs and MTDs in each cell. The macro-cell capacity is defined as the maximum number of MUs and MTDs admitted in the 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. The small cell capacity is defined as the maximum MTDs admitted in the cell at the premise of not violating the QoS for all MTDs (Assume no category II MUs exist in each small cell for evaluating the capacities for both macro-cells and small cells). In Fig. 9(a), the proposed resource slicing framework achieves much larger macro-cell capacity than the device-

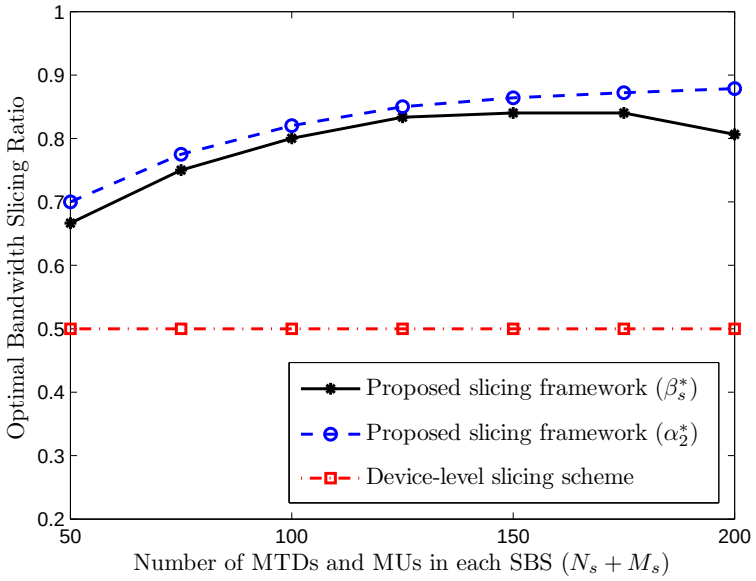

Fig. 8: Comparison of bandwidth slicing ratios between the proposed slicing framework and the device-level slicing scheme ($N_u = N_a = 50$).

level resource slicing scheme where the macro-cell capacity is fixed without resource sharing under different N_s. Similar trend for small cell capacity is observed in Fig. 9(b). Although some of the MTDs can be offloaded to an 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.

Lastly, the aggregate network utility of a tagged 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. 10(a) and 10(b). For the proposed bandwidth slicing framework, we also evaluate the effect of the approximations made for solving $(\mathrm{P2})$ (i.e., the variable relaxation). It is clear that the network utility achieved with approximations matches closely with the one without approximations. Moreover, through bandwidth slicing among heterogeneous BSs, the overall network resource utilization is significantly improved. Therefore, it can be seen that our proposed resource slicing framework achieves higher network utility than both the SINR-max based 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 (user) association patterns under

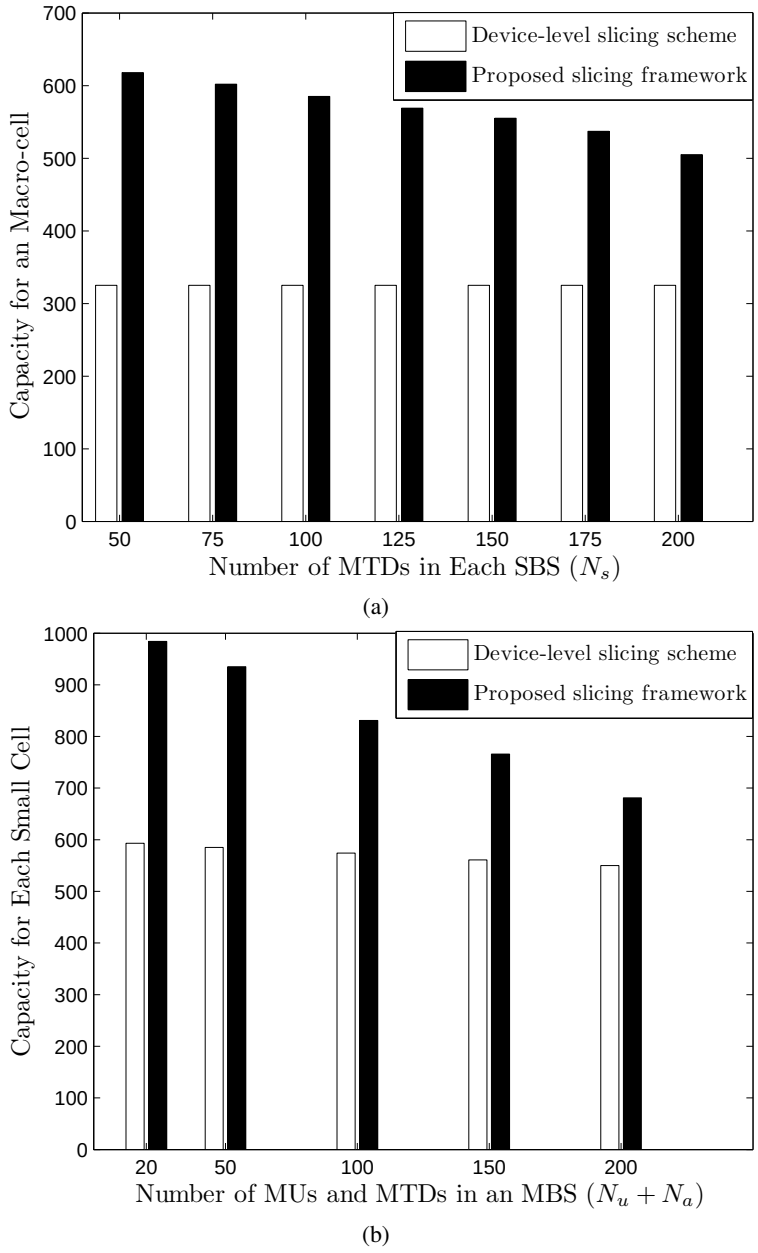

Fig. 9: Capacity (a) for an MBS and (b) for each SBS.

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 services, BS-device (user) association patterns, and resource allocation among end

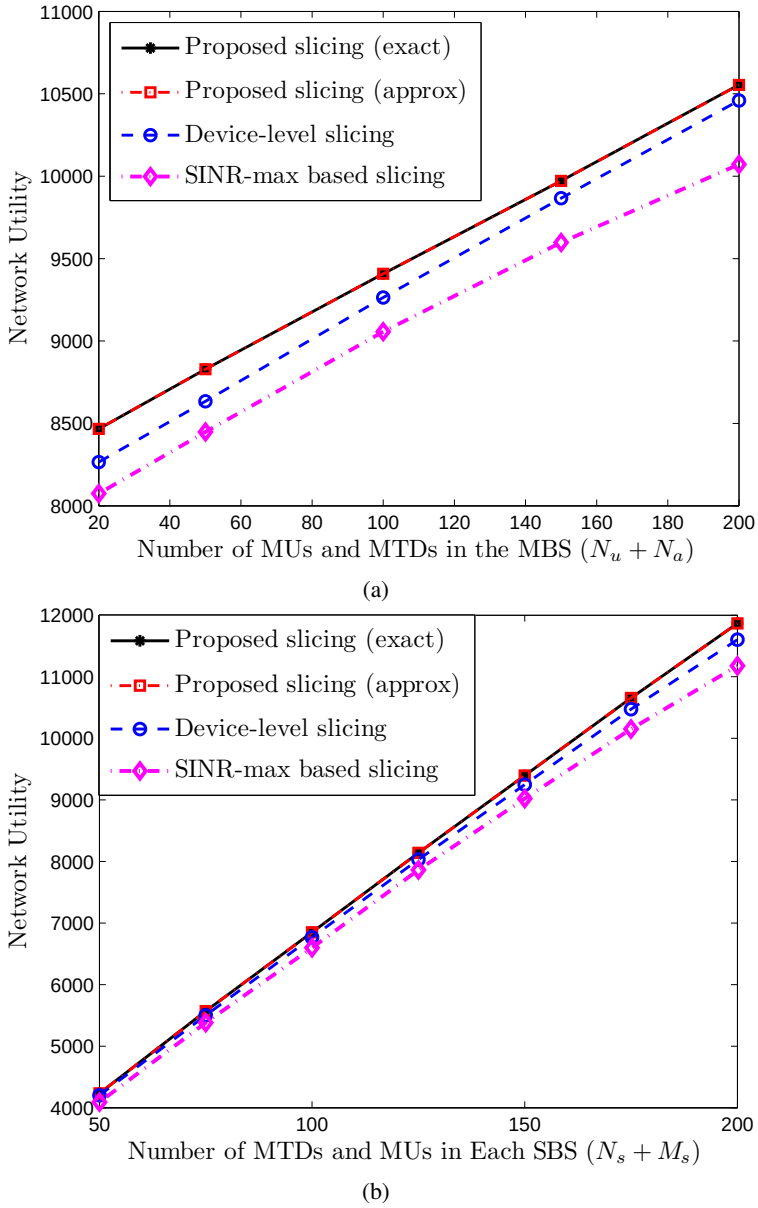

Fig. 10: Comparison of aggregate network utility (a) with respect to the number of category I MUs and MTDs ($N_s + M_s = 150, N_u = N_a$) and (b) with respect to the number of category II MTDs and MUs ($N_u = N_a = 50$).

devices. To reduce complexity, the original optimization problem is transformed to a tractable biconcave maximization problem. Then, an alternative concave search (ACS) algorithm is designed to solve the transformed problem for a set of optimal slicing ratios and optimal BS-device (user) association patterns. Simulation results demonstrate the robustness of the proposed ACS algorithm due to its good convergence property and low computational complexity. In comparison with the two other resource slicing schemes, the proposed framework has a lower communication overhead for updating the slicing ratios, achieves higher capacity in 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 MTD (or MU) i from one SBS is independent of the fraction allocated to MTD (or MU) 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

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

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

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

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

Similar to [19], since geometric average is no greater than arithmetic average, we have

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

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 MTD (or MU) 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 $v = (v_1, v_2, ..., v_y) \in \mathbf{R}^y$, $y = \sum\limits_{k=1}^{n}(N_k+M_k)$, we have

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

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_s. 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 $\sum\limits_{k=1}^{n} u_k^{(2)}\left(\beta_s, \widetilde{\mathbf{X}_k}\right)$ is a nonnegative linear combination of a set of biconcave functions, which is also (strictly) biconcave [32].

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] 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.

[4] W. Wu, Q. Shen, M. Wang, and X. Shen, "Performance analysis of IEEE 802.11.ad downlink hybrid beamforming," in *Proc. IEEE ICC' 17*, May 2017, pp. 1–6.

[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] 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.

[16] 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.

[17] 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.

[18] 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.

[19] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2706–2716, Jun. 2013.

[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 (M2M) communications," *IEEE Trans. Veh. Technol.*, 2016, to appear. DOI: 10.1109/TVT.2016.2633525.

[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.*, 2017, to appear, doi: 10.1109/TNET.2017.2720668.

[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.