Service Provisioning Framework for RAN Slicing: User Admissibility, Slice Association and Bandwidth Allocation

Yao Sun, Shuang Qin, *Member, IEEE*, Gang Feng, *Senior Member, IEEE*, Lei Zhang, *Senior Member, IEEE*, and Muhammad Ali Imran, *Senior Member, IEEE*

Abstract—Network slicing (NS) has been identified as one of the most promising architectural technologies for future mobile network systems to meet the extremely diversified service requirements of users. In radio access networks (RAN) slicing, service provisioning for slice users becomes much more complicated than that in traditional mobile networks, as the constraints of both user physical association with base station (BS) and logical association with NS should be considered. In other words, the user-BS-NS three layer association relationship should be addressed in provisioning tailored service for diversified use cases with various quality of service (QoS) requirements. Therefore, service provisioning in RAN slicing becomes an essential yet challenging issue for 5G and beyond systems. In this paper, we propose a unified framework for service provisioning in RAN slicing with aim of maximizing resource utilization while guaranteeing QoS of users. The framework consists of two steps. The first step is to identify a set of slice users whose QoS can be satisfied simultaneously; while the second step performs joint slice association and bandwidth allocation with aim to minimize bandwidth consumption. Numerical results show that in typical scenarios, our proposed service provisioning framework can achieve significant performance gain in terms of the number of serving users and wireless bandwidth utilization compared with traditional schemes.

Index Terms—Service provisioning, slice association, bandwidth allocation, radio access network slicing

1 Introduction

It is foreseen that in future mobile communication systems, networks will be abstracted into network slicing (NS), which enables design, deployment, customization, and optimization of isolated virtual sub-networks/slices on a common physical network infrastructure [1], [2]. In order to provide tailored service thus to meet specific quality of service (QoS) requirements in different communication scenarios (e.g., enhanced Mobile Broadband, massive Machine Type Communication, and Ultra Reliable Low Latency Communication) [3], [4], the radio access networks (RANs) should be also sliced for provisioning tailored services. This new architecture is called as RAN slicing, and it can dramatically improve the flexibility and efficiency of networks, and thus enhance network capabilities in terms of connectivity, latency, transmission rate, etc.

To take advantage of the aforementioned benefits of NS, service provisioning scheme (including user access control, slice association and resource allocation) is of paramount importance to be investigated in RAN slicing. An appropriate service provisioning scheme can improve both network and user performance by flexibly allocating inter-/intra NS resources. Moreover, driven by the explosive growth of wireless data traffic, improving resource utilization is crucial for future communication systems [3]. Therefore, from both user QoS and network resource utilization perspectives, it is imperative to develop efficient service provisioning schemes for RAN slicing.

In RAN slicing, the service provisioning mechanism is fundamentally different from that in conventional mobile networks due to the introduction of NS. First, from service model aspect, NS-based networks provide guaranteed QoS for all serving users [5] instead of the traditional *best effort* model [6]. Second, from network architecture aspect, slices are logically virtualized and isolated over a shared physical network. Hence, both physical and virtual resource constraints need to be considered to create a function chain for a specific service. Third, from user association aspect, user equipment (UE) should be associated with an NS via a specific base station (BS), thus forming a UE-BS-NS three-layer association relationship. Not all BSs are able to provide the specific slice/service due to either functionality missing or limited resources. Hence, a joint optimization of NS and BS selection for a UE with specific QoS requirements should be addressed. Due to the aforementioned differences, applying traditional service provisioning schemes to RAN slicing may lead to low resource utilization, poor QoS provisioning and/or frequent NS re-configurations. Therefore, designing new service provisioning schemes dedicated to RAN slicing to optimize network performance becomes an essential yet challenging issue.

Unfortunately, while most recent related work focuses on resource management in NSs to accomplish optimal NS

- *Y. Sun, S. Qin, and G. Feng are with the National Key Lab on Communications, University of Electronic Science and Technology of China, Chengdu 610051, China. E-mail: {sunyao, blueqs, fenggang}@uestc.edu.cn.*
- *L. Zhang and M.A. Imran are with the School of Engineering, University of Glasgow, G12 8QQ Glasgow, United Kingdom. E-mail: {Lei.Zhang, Muhammad.Imran}@glasgow.ac.uk.*

Manuscript received 12 June 2019; revised 18 Jan. 2020; accepted 31 May 2020. Date of publication 8 June 2020; date of current version 3 Nov. 2021. (Corresponding author: Shuang Qin.)
Digital Object Identifier no. 10.1109/TMC.2020.3000657

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

deployment [7], [8], [9], very little attention is paid to service provisioning scheme for RAN slicing. The most relevant work is on UE access control in NS-based networks. The authors of [5] and [10] point out that UE access control is a key issue in RAN slicing, but no further optimization is studied in their work. The authors of [11] and [12] focus on the optimization of user association without consideration of resource allocation to users. Indeed, resource allocation is a closely coupled issue with service provisioning in RAN slicing, as it affects QoS of users as well as network resource utilization.

In this paper, we propose a unified service provisioning framework for RAN slicing with aim to maximize bandwidth utilization while guaranteeing QoS of users. Due to the limited resource, the QoS of all users may not be satisfied simultaneously. Thus, we should first select admissible users and then conduct slice association and resource allocation for these admissible users. Accordingly, the proposed framework performs the two functions by two steps. The fist step is to identify a set of users whose QoS can be satisfied, and the second step is to jointly performs slice association and bandwidth allocation (SABA) for the admissible users identified in the first step. Specifically, we design two policies in the first step for optimizing the QoS and the number of admissible users respectively. In the second step, we propose network-centric SABA policy trying to achieve the global optimality of bandwidth consumption while the UE-centric SABA policy to achieve a sub-optimal solution with a low computational complexity. Numerical results show that in typical scenarios, our proposed service provisioning framework can significantly outperform traditional schemes in terms of the number of admissible users and wireless bandwidth consumption.

In the rest of the paper, we begin with an overview of related work in Section 2. We present system model and service provisioning problem formulation in Sections 3 and 4 respectively. In Section 5, two algorithms are proposed to identify the admissibility of users, and then the SABA problem for the admissible users is solved in Section 6. We present numerical results in Section 7 and finally conclude this paper in Section 8.

2 Related Work

We present related work on service provisioning for traditional heterogeneous cellular networks (HetNets) and RAN slicing separately.

2.1 Service Provisioning for Traditional HetNets

In HetNets, where macro base stations (MBSs) are overlaid with lower transmit power small base stations (SBSs), service provisioning becomes challenging due to the significant difference of transmit power between MBS and SBS. In recent years, many user association and bandwidth allocation policies have been proposed to optimize the instantaneous or long-term network performance in terms of load balance among BSs, system throughput, and fairness among users.

Existing user service provisioning policies are usually focused on instantaneous network performance by leveraging game theory [13], [14] and convex optimization [15], [16]. The authors of [13] propose an auction-based algorithm to achieve load balance between MBS and SBSs. In [14], the authors formulate the user association problem as a noncooperative game, and then propose a distributed algorithm to solve the problem. They analyze the convergence and Pareto-efficiency of this algorithm. In work [15] and [16], the authors formulate the joint user association and wireless resource allocation problem as a mixed linear programming, and propose a distributed algorithm to achieve load balance [15] and minimization of the global outage probability [16] respectively.

Some researchers propose service provisioning policies with aim to optimize the long-term network performance by using some stochastic optimization tools such as machine learning [17], Markov decision process (MDP) [18] and multi-armed bandit (MAB) [19]. In [17], the authors propose a BS selection policy for users when handoff occurs by using reinforcement learning algorithm to reduce redundant handoffs. The authors of [18] leverage MDP model to design a hybrid user association policy where users are assisted in their decisions by broadcasting load information. The authors of [19] formulate the user association problem as a MAB model by considering user behaviors, and derive a user association policy to maximize the long-term system throughput.

2.2 Service Provisioning for RAN Slicing

Recently, most work on service provisioning for RAN slicing focuses on NS function virtualization and softwarizaion [7], [8], [9], where the optimization of resource configuration between multiple NSs as well as some NS deployment problems are investigated. Work [7] studies RAN resource splitting among multiple slices in a multi-cell network. Four different slice spitting approaches are presented and compared from various perspectives. Work [8] explores radio and core network resource allocation respectively for network slicing by using deep reinforcement learning. The authors in [9] provide a guidance to the optimization of resource configuration between multiple NSs. The authors first review the state-of-the-art research work on resource allocation, and then investigate the relationship between public and private wireless resources.

Meanwhile, some research works investigate the wireless resource allocation among different RAN slices from network perspective. Specifically, the authors in work [20] propose a framework of wireless spectrum sharing named CellSlice to achieve network slicing for both downlink and uplink. Work [21] studies the wireless resource scheduling for RAN slicing by using stochastic learning. The authors in work [22] propose a dynamic radio resource slicing framework for a two-tier heterogeneous wireless network with aim to facilitate spectrum sharing among BSs and guarantee QoS requirement for different service types. The authors of [23] exploit deep learning in conjunction with reinforcement learning to optimize the resource for multiple slices at different time-scale. The authors in [24] present a network slicing framework for both RAN and core networks to dynamically allocate network resources based on SDN controller.

Thus far, there are only a few recent investigations on service provisioning from UE perspective. Work [5] introduces the concept of end-to-end network slicing where UEs are considered as components of an NS. The authors of [5]

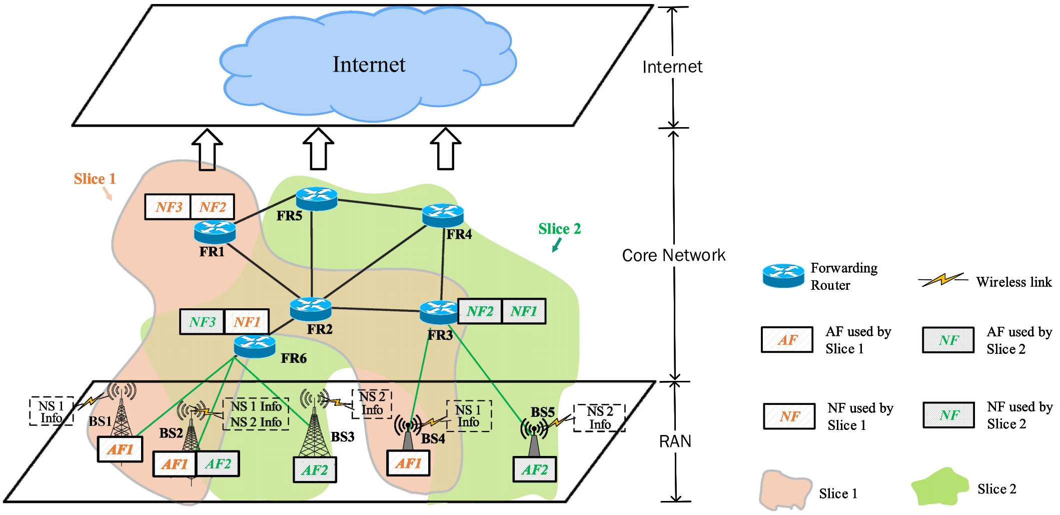


Fig. 1. Network slice-based network architecture.

and [10] point out that UE access control is one of the key issues in NS-based RAN, but no further optimization is studied in their work. The authors of work [25] elaborate the process of user association in RAN slicing, where it should be split into two steps: (i) the discovery of slices via physical BSs; (ii) the mapping of UEs to slices. The authors clearly explain the user association process, although there is no further optimization on the association mechanism. Work [11] and [12] focus on the optimization of user-slice association mechanism, and do not investigate resource allocation in RAN slicing.

3 System Model

In this section, we present the NS-based mobile network architecture, RAN slicing model, and the UE QoS model, respectively.

3.1 NS-Based Network Architecture Model

We consider an NS-based mobile network shown in Fig. 1, which consists of the core and radio access networks. Some network function (NF) modules and access function (AF) modules are deployed to form an end-to-end network slice for provisioning specific service. NFs and AFs are related to some specific logic functions, such as connection management, mobility management, security, etc., in core network and RAN respectively. The detailed descriptions of the network architecture can be found in [5]. Here we focus on the resource management aspect of the system.

From Fig. 1, we see that slices share resources in both core and radio access networks. We use the green lines to denote the backhaul links between access network and core network. There are totally 5 backhaul links in Fig. 1, and the backhaul link of BS 2 should be shared by NS 1 and 2. In the core network, the slices using the same link should share link bandwidth resources, computing resources and NFs. In RAN, the slices covering the same BS (we use BS to denote all kinds of access point through out this paper) will share wireless resources as well as AFs. The slice information is broadcast by the BSs, and not all NSs will be accessible via every BS [5]. In the example of Fig. 1, BS 2 broadcasts the information of NS 1 and NS 2, thus UEs associated with BS 2 can access to the both NSs; while UEs associated with BS 1 can access to only NS 1 due to that only AF 1 is deployed on BS 1. Note that there are mainly two differences of the UE-BS-NS model when compared with the traditional UE-BS relation. First, service type should be considered in priority for making user association in UE-BS-NS scenario. Each slice provides different QoS for the serving users, thus we should choose the slice that can satisfy the QoS requirements of the user. Second, the coverage of both BS and NS should be taken into account since not all NSs will be accessible via every BS.

3.2 RAN Slicing Model

Focusing on RAN slicing, we consider a multi-slice and multi-BS model shown in Fig. 2, where the BSs used by multiple slices are deployed in the area. Each BS can support multiple NSs with different provisioned QoS, and each NS may also be covered by multiple BSs (i.e., each slice information is broadcast by several BSs). Multiple UEs are randomly distributed in this area with different QoS requirements. They can access to a specific NS via a BS in the coverage of the NS, and thus forming a three-layer association relationship. Let $\mathcal{B}$, $\mathcal{S}$ and $\mathcal{U}$ denote the set of BSs, NSs and UEs, respectively. For a specific BS, say BS k, we use $\mathcal{S}_k$ to denote the set of NSs supported by BS k.

We identify a specific NS, say NS j, by transmission rate, delay, and the resource allocation in core and access networks. Specifically, besides the slice ID, four elements $(R_j, D_j, \Lambda_j, \vec{B}_j)$ are used to identify the jth slice, where R_j

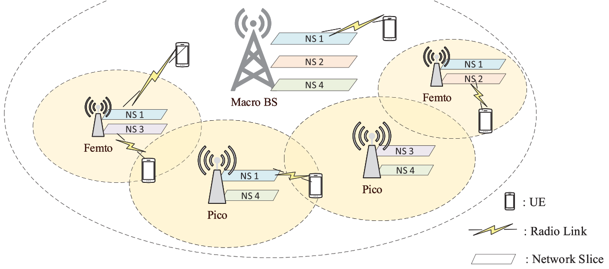


Fig. 2. Multi-slice and multi-BS RAN slicing.

and D_j respectively denotes the minimum transmission rate and the maximum delay that NS j can provide to its serving UEs, Λ_j denotes the bandwidth allocated to NS j in the core network, and $\vec{B}_j$ is a vector denoting the wireless bandwidth allocation of NS j from all BSs. Let $b_j^{(k)}$ be the kth element of vector $\vec{B}_j$ denoting the bandwidth of NS j allocated by BS k. $b_j^{(k)} = 0$ when BS k is not in the coverage of NS j.

3.3 UE QoS Model

For a specific UE, say UE n with q_n volume data to transmit, the QoS can be described by two metrics: transmission rate $\bar{r}_n$ and tolerant delay $\bar{d}_n$ [26]. Thus, NS j is admissible for UE n only if $R_j \geq \bar{r}_n$ and $D_j \leq \bar{d}_n$. We now discuss the two QoS metrics respectively. Let $r_n^{j,k}$ be the transmission rate of UE n served by NS j via BS k. For simplicity, we use Shannon theory to define the transmission rate, i.e., $r_n^{j,k} = w_n^{j,k} log_2(1 + SINR_n^k)$, where $w_n^{j,k}$ is the wireless bandwidth that BS k allocates to the UE n served by NS j, and $SINR_n^k$ is the signal-to-interference-plus-noise-ratio (SINR) between UE n and BS k. In our work, we assume that the BSs in HetNet share the same spectrum and thus the co-channel interference is considered. Similar to that in [19] we assume that the channel is flat and the transmit power of BS is allocated uniformly to each sub-channel. Hence the SINR of UE n associated with BS k can be represented as

$$SINR_n^k = \frac{P_{Tn} g_{kn}}{\sum_{j \in \mathcal{M} j \neq k} P_{Tj} g_{jn} + \sigma^2}, \ k \in \mathcal{B}, \tag{1}$$

where P_{Tn} denotes the transmit power of BS n; g_{kn} is the channel gain between UE n and BS k; and σ^2 is the noise level.

We use $d_n^{j,k} = q_n / r_n^{j,k}$ to denote the delay in RAN of UE n served by NS j via BS k. Thus, combined with the delay in core network, the end-to-end delay can be approximately calculated as $d_n^{j,k} + D_j$. Note that more sophisticated and accurate transmission rate and delay models can be used here. However, they do not affect the following derivations. This is because that the models in this work may affect the absolute value of bandwidth consumption, but do not invalidate the relative performance enhancement of our proposed policies. We will also conduct some experiments to verify this point in the simulation part.

Traditional mobile networks usually provide the *best effort* service model to users. In this model, the network allocates resources among all of the active UEs and attempts to serve all of them without making any explicit commitment on rate or any other service quality [6]. In comparison, NS-based networks need to guarantee the different and heterogeneous QoS of UEs without interference among one another [10]. Hence, the best effort service model cannot be applicable directly to RAN slicing since it cannot guarantee individual user's QoS. Next, we will formulate the service provisioning problem in RAN slicing to meet individual UE requirements.

4 Problem of Service Provisioning and Solution Framework

We first formulate the service provisioning problem for UEs with aim to minimize wireless bandwidth consumption while guaranteeing the QoS of UEs. Then we propose a unified service provisioning framework to solve the formulated problem.

4.1 Problem Formulation

The optimization problem in this work can be stated as: minimizing the wireless bandwidth consumption subject to QoS requirements, NS and BS resource constraints. We first define a binary variable $x_n^{j,k} \in \{0,1\}, \forall (n,j,k) \in \mathcal{U} \times \mathcal{S} \times \mathcal{B}$, where $x_n^{j,k} = 1$ indicates that UE n is served by NS j via BS k. We can now formulate the service provisioning problem as **P1**:

$$\mathbf{P1:} \min \sum_{n \in \mathcal{U}} \sum_{j \in \mathcal{S}} \sum_{k \in \mathcal{B}} x_n^{j,k} w_n^{j,k}, \tag{2}$$

$$\text{s.t.} \sum_{n \in \mathcal{U}} \sum_{k \in \mathcal{B}} x_n^{j,k} r_n^{j,k} \leq \Lambda_j, \ \forall j \in \mathcal{S} \tag{2-1}$$

$$\sum_{n \in \mathcal{U}} x_n^{j,k} w_n^{j,k} \leq b_j^{(k)}, \ \forall j \in \mathcal{S}, \forall k \in \mathcal{B} \tag{2-2}$$

$$\sum_{j \in \mathcal{S}} \sum_{k \in \mathcal{B}} x_n^{j,k} r_n^{j,k} \geq \bar{r}_n, \ \forall n \in \mathcal{U} \tag{2-3}$$

$$\sum_{j \in \mathcal{S}} \sum_{k \in \mathcal{B}} x_n^{j,k} R_j \geq \bar{r}_n, \ \forall n \in \mathcal{U} \tag{2-4}$$

$$\sum_{j \in \mathcal{S}} \sum_{k \in \mathcal{B}} x_n^{j,k} (d_n^{j,k} + D_j) \leq \bar{d}_n, \ \forall n \in \mathcal{U} \tag{2-5}$$

$$\sum_{j \in \mathcal{S}} \sum_{k \in \mathcal{B}} x_n^{j,k} = 1, \ \forall n \in \mathcal{U} \tag{2-6}$$

$$x_n^{j,k} \in \{0,1\}, \ \forall (n,j,k) \in \mathcal{U} \times \mathcal{S} \times \mathcal{B} \tag{2-7}$$

where $x_n^{j,k}$ and $w_n^{j,k}$ are the optimization variables. Constraint (2-1) refers to the wired link resource constraint to guarantee that the total transmission rate offered by an NS does not exceed the link capacity in the core network. Constraint (2-2) states the wireless bandwidth constraint, ensuring that the total bandwidth allocated to UEs by NS j via BS k does not exceed the maximum value $b_j^{(k)}$. Constraints (2-3), (2-4), and (2-5) guarantee the QoS (rate and delay) of UEs can be satisfied by its' serving BS and NS. Constraints (2-6) and (2-7) ensure that each UE can only access one NS via one BS at a time. Note that constraint (2-2) also ensures that UEs cannot access an NS via the BSs that cannot provide such a service. This is because that when the BS that cannot provide the required service type, we have $b_j^{(k)} = 0$. In this case, the only way to satisfy constraint (2-2) when $x_n^{j,k} \neq 0$ is to set $w_n^{j,k} = 0$ and $x_n^{j,k} = 1$. However, the setting of $w_n^{j,k} = 0$ and $x_n^{j,k} = 1$ means that the achievable rate $r_n^{j,k} = w_n^{j,k} log_2(1 + SINR_n^k)$ equals to 0. Therefore, the constraint (2-3) in this case cannot be satisfied. Hence, $\left(w_n^{j,k} = 0, x_n^{j,k} = 1\right)$ should not be a feasible solution, and $x_n^{j,k} = 0$ always holds for the BSs that cannot provide such a service.

4.2 Service Provisioning Framework in RAN Slicing

Since the formulated service provisioning problem **P1** requires to guarantee the QoS of all the UEs with limited resources, there may be no feasible solution in the case of dense UE distribution and/or high QoS requirements. Therefore, some UEs cannot be served in the network with certain amount of resources. To solve the problem we propose a

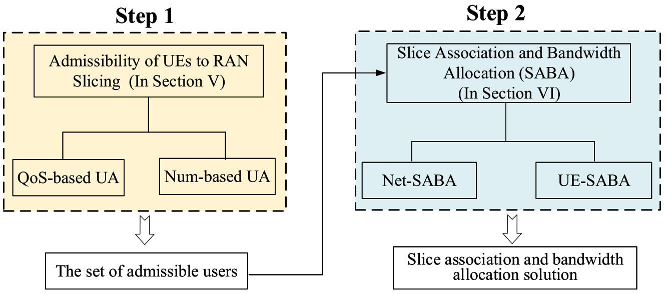

Step 1
Admissibility of UEs to RAN Slicing (In Section V)
QoS-based UA
Num-based UA
The set of admissible users
Step 2
Slice Association and Bandwidth Allocation (SABA) (In Section VI)
Net-SABA
UE-SABA
Slice association and bandwidth allocation solution

Fig. 3. Framework of service provisioning in RAN slicing.

service provisioning framework with two steps as shown in Fig. 3, where the first step guarantees the admissibility of UEs to RAN slicing by selecting suitable serving UEs whose QoS can be satisfied simultaneously for the network, and the second step is to solve a feasible problem **P1** to determine the slice association and bandwidth allocation for these serving UEs.

In detail, to select admissible UEs to be served, we first propose two policies of determining UE admissibility (UA) to RAN slicing, namely QoS-based UA and Num-based UA, to optimize the QoS and the number of serving users respectively while guaranteeing the QoS of the selected UEs. Next, we solve **P1** to determine the slice association and bandwidth allocation (SABA) for the serving UEs selected in the first step. Net-SABA policy for global optimality of bandwidth consumption and UE-SABA policy with low computational complexity are proposed respectively in this step. In the subsequent sections, we will elaborate our proposed service provisioning framework.

5 Admissibility of UEs to RAN Slicing

We first determine UE admissibility to RAN slicing to identify the users whose QoS can be satisfied simultaneously. The main idea is that if we cannot meet all UEs' QoS with limited resource, we will reject some of the UEs to give more resources to others, thus to guarantee the QoS of these admissible UEs as well as to avoid significant overall network performance degradation. For convenience, we call the UEs whose QoS can be guaranteed as admissible UEs throughout this paper. Before studying this problem, we first give the following assumption and definition.

Assumption 1. *Network slice allocates the minimal required wireless bandwidth to UEs to satisfy the UEs' QoS requirements.*

According to Assumption 1, the original constraints (2-3), (2-4), and (2-5) regarding UEs QoS become equality constraints, and the admissible UEs obtained under Assumption 1 are still admissible for the original problem. Moreover, as we use this assumption, the allocated bandwidth $w_n^{j,k}$ is not the optimization variable in this section, and we will optimize $w_n^{j,k}$ in the next section by solving **P1** for those admissible UEs.

Definition 1. *Subset $\mathcal{A}$ is an admissible UE set (AUS) if problem **P1** is feasible when $\mathcal{U}$ is replaced by $\mathcal{A}$.*

Definition 1 describes the admissibility of a UE subset. Hence, for a specific AUS, the network can simultaneously guarantee the QoS of all the UEs in this AUS. However, it is not a sufficient condition to achieve good overall network performance. For example, it is meaningless to choose an AUS which contains only one UE. Therefore, we need to design an approach to select admissible UEs as well as achieve good overall network performance. We will thus develop two policies of determining UE admissibility under Assumption 1, namely QoS-based UA and Num-based UA, to optimize the QoS and the number of admissible UEs respectively.

5.1 QoS-Based UA Policy

We first consider UE admissibility from QoS perspective. As not all UEs could be admissible, there is a gap between the achieved QoS and the required QoS of UEs (we use QoS degradation to represent this gap). Our main idea is to select the UEs whose QoS can be satisfied simultaneously while minimizing the overall QoS degradation. To this end, we formulate a new optimization problem based on **P1**, and then design QoS-based UA Policy according to the solution to the new problem. Let us illustrate the process in detail.

First, we introduce two elastic variables $\check{r}_n$ and $\check{d}_n$ for UE n to describe the rate and delay degradation respectively. We restrict that $0 \leq \check{r}_n \leq \bar{r}_n$ and $0 \leq \check{d}_n \leq \bar{D}$, where $\bar{D}$ is a very large parameter. Therefore, the rate and delay requirement of UE n can be referred to as $\bar{r}_n - \check{r}_n$ and $\bar{d}_n + \check{d}_n$ respectively. UE n is admissible when $\check{r}_n = 0$ and $\check{d}_n = 0$, i.e., the QoS can be satisfied. Then we give the following definition to describe a subset of UEs whose QoS can be simultaneously satisfied while the QoS degradation of others is the minimum.

Definition 2. *Subset $\mathcal{A}$ is a QoS-admissible UE set (QoS-AUS) if $\mathcal{A}$ is an AUS with the minimum achievable value of $\sum_{n\in\mathcal{U}\setminus\mathcal{A}}\left(\frac{\check{r}_n}{\bar{r}_n}+\frac{\check{d}_n}{\bar{d}_n}\right)$.*

Here we use the normalized degradation of transmission rate (i.e., $\check{r}_n/\bar{r}_n$) and delay ($\check{d}_n/\bar{d}_n$). This definition describes both the feasibility and the QoS performance of a UE subset. In the following, we design QoS-based UA policy to find such UE set QoS-AUS in Definition 2. By introducing elastic variables $\check{r}_n$ and $\check{d}_n$, we formulate problem **P2** as follows:

$$\textbf{P2}: \min \sum_{n\in\mathcal{U}}\left(\frac{\check{r}_n}{\bar{r}_n}+\frac{\check{d}_n}{\bar{d}_n}\right), \tag{3}$$

$$\text{s.t.} \sum_{n\in\mathcal{U}}\sum_{k\in\mathcal{B}} x_n^{j,k} r_n^{j,k} \leq \Lambda_j,\ \forall j\in\mathcal{S} \tag{3-1}$$

$$\sum_{n\in\mathcal{U}} x_n^{j,k} w_n^{j,k} \leq b_j^{(k)},\ \forall j\in\mathcal{S}, \forall k\in\mathcal{B} \tag{3-2}$$

$$\sum_{j\in\mathcal{S}}\sum_{k\in\mathcal{B}} x_n^{j,k} r_n^{j,k} = \bar{r}_n - \check{r}_n,\ \forall n\in\mathcal{U} \tag{3-3}$$

$$\sum_{j\in\mathcal{S}}\sum_{k\in\mathcal{B}} x_n^{j,k} R_j = \bar{r}_n - \check{r}_n,\ \forall n\in\mathcal{U} \tag{3-4}$$

$$\sum_{j\in\mathcal{S}}\sum_{k\in\mathcal{B}} x_n^{j,k} (d_n^{j,k} + D_j) = \bar{d}_n + \check{d}_n,\ \forall n\in\mathcal{U} \tag{3-5}$$

$$\sum_{j\in\mathcal{S}}\sum_{k\in\mathcal{B}} x_n^{j,k} = 1,\ \forall n\in\mathcal{U} \tag{3-6}$$

$$x_n^{j,k} \in \{0,1\},\ \forall (n,j,k)\in\mathcal{U}\times\mathcal{S}\times\mathcal{B} \tag{3-7}$$

$$0 \le \check{r}_n \le \bar{r}_n,\ \forall n \in \mathcal{U} \qquad (3-8)$$

$$0 \le \check{d}_n \le \bar{D},\ \forall n \in \mathcal{U}. \qquad (3-9)$$

In **P2**, the objective is to minimize the normalized QoS degradation of all UEs. Compared with the constraints in **P1**, the only difference is using equalities in constraints (3-3), (3-4), and (3-5) to replace the inequalities in (2-3), (2-4), and (2-5) by introducing elastic variables. In **P2**, the optimization variables are binary indicators $x_n^{j,k}$ as well as the continuous elastic variables $\check{r}_n$ and $\check{d}_n$. Hence, **P2** is a mixed integer liner programming (MILP). As we introduce the elastic variables into the MILP, the QoS of UEs can vary with the elastic variables, and thus problem **P2** is always feasible. In the following, we solve **P2** by using Lagrange decomposition theory [27].

We first introduce constraints (3-3), (3-4), and (3-5) into the optimization objective by associating Lagrange multipliers λ_n, v_n and μ_n. Let $\boldsymbol{\lambda}, \boldsymbol{v}, \boldsymbol{\mu}, \check{\boldsymbol{r}}, \check{\boldsymbol{d}}$ and $\boldsymbol{x}$ be the corresponding vectors of $\lambda_n, v_n, \mu_n, \check{r_n}, \check{d}_n$ and $x_n^{j,k}$, respectively. For **P2**, we give Lagrange dual problem **P3** with respect to constraints (3-3), (3-4), and (3-5)

$$\begin{aligned}&\mathbf{P3:}\ g(\boldsymbol{\lambda},\boldsymbol{v},\boldsymbol{\mu}) \triangleq \inf_{\check{\boldsymbol{r}},\check{\boldsymbol{d}},\boldsymbol{x}} L(\boldsymbol{\lambda},\boldsymbol{v},\boldsymbol{\mu},\check{\boldsymbol{r}},\check{\boldsymbol{d}},\boldsymbol{x})\\ &\text{s.}t.\ \text{Constraints}\ (3-1),(3-2),(3-6)-(3-9),\end{aligned} \qquad (4)$$

where Lagrangian $L(\boldsymbol{\lambda},\boldsymbol{v},\boldsymbol{\mu},\check{\boldsymbol{r}},\check{\boldsymbol{d}},\boldsymbol{x})$ is expressed as

$$\begin{aligned}&L(\boldsymbol{\lambda},\boldsymbol{v},\boldsymbol{\mu},\check{\boldsymbol{r}},\check{\boldsymbol{d}},\boldsymbol{x})\\ &=\sum_{n\in\mathcal{U}}\left(\frac{\check{r}_n}{\bar{r}_n}+\frac{\check{d}_n}{\bar{d}_n}\right)+\sum_{n\in\mathcal{U}}\lambda_n\left(\sum_{j\in\mathcal{S}}\sum_{k\in\mathcal{B}}x_n^{j,k}r_n^{j,k}-(\bar{r}_n-\check{r}_n)\right)\\ &\quad+\sum_{n\in\mathcal{U}}v_n\left(\sum_{j\in\mathcal{S}}\sum_{k\in\mathcal{B}}x_n^{j,k}R_j-(\bar{r}_n-\check{r}_n)\right)\\ &\quad+\sum_{n\in\mathcal{U}}\mu_n\left(\sum_{j\in\mathcal{S}}\sum_{k\in\mathcal{B}}x_n^{j,k}(d_n^{j,k}+D_j)-(\bar{d}_n+\check{d}_n)\right)\end{aligned} \qquad (5)$$

$$\begin{aligned}&=\sum_{n\in\mathcal{U}}\left(\frac{\check{r}_n}{\bar{r}_n}+\frac{\check{d}_n}{\bar{d}_n}\right)+\sum_{n\in\mathcal{U}}\left(\lambda_n\check{r}_n+v_n\check{r}_n-\mu_n\check{d}_n\right)\\ &\quad+\sum_{n\in\mathcal{U}}\lambda_n\left(\sum_{j\in\mathcal{S}}\sum_{k\in\mathcal{B}}x_n^{j,k}r_n^{j,k}-\bar{r}_n\right)\\ &\quad+\sum_{n\in\mathcal{U}}v_n\left(\sum_{j\in\mathcal{S}}\sum_{k\in\mathcal{B}}x_n^{j,k}R_j-\bar{r}_n\right)\\ &\quad+\sum_{n\in\mathcal{U}}\mu_n\left(\sum_{j\in\mathcal{S}}\sum_{k\in\mathcal{B}}x_n^{j,k}(d_n^{j,k}+D_j)-\bar{d}_n\right).\end{aligned} \qquad (6)$$

Here, strong duality holds. Therefore, we first solve **P3** with the fixed Lagrange multipliers $\boldsymbol{\lambda}$, $\boldsymbol{v}$ and $\boldsymbol{\mu}$, and then maximize $g(\boldsymbol{\lambda},\boldsymbol{v},\boldsymbol{\mu})$ to find the optimal solution to **P2**.

Due to the independence of $\check{\boldsymbol{r}},\check{\boldsymbol{d}}$ and $\boldsymbol{x}$, $L(\boldsymbol{\lambda},\boldsymbol{v},\boldsymbol{\mu},\check{\boldsymbol{r}},\check{\boldsymbol{d}},\boldsymbol{x})$ can be decoupled into two sub-functions $L^1(\boldsymbol{\lambda},\boldsymbol{v},\boldsymbol{\mu},\check{\boldsymbol{r}},\check{\boldsymbol{d}})$ and $L^2(\boldsymbol{\lambda},\boldsymbol{v},\boldsymbol{\mu},\boldsymbol{x})$, i.e., $L(\boldsymbol{\lambda},\boldsymbol{v},\boldsymbol{\mu},\check{\boldsymbol{r}},\check{\boldsymbol{d}},\boldsymbol{x})=L^1(\boldsymbol{\lambda},\boldsymbol{v},\boldsymbol{\mu},\check{\boldsymbol{r}},\check{\boldsymbol{d}})+L^2(\boldsymbol{\lambda},\boldsymbol{v},\boldsymbol{\mu},\boldsymbol{x})$, where

$$\begin{aligned}L^1(\boldsymbol{\lambda},\boldsymbol{v},\boldsymbol{\mu},\check{\boldsymbol{r}},\check{\boldsymbol{d}})&\triangleq\sum_{n\in\mathcal{U}}\left(\frac{\check{r}_n}{\bar{r}_n}+\frac{\check{d}_n}{\bar{d}_n}+\lambda_n\check{r}_n+v_n\check{r}_n-\mu_n\check{d}_n\right)\\ &=\sum_{n\in\mathcal{U}}\left[\left(\frac{1}{\bar{r}_n}+\lambda_n+v_n\right)\check{r}_n+\left(\frac{1}{\bar{d}_n}-\mu_n\right)\check{d}_n\right],\end{aligned} \qquad (7)$$

and

$$\begin{aligned}L^2(\boldsymbol{\lambda},\boldsymbol{v},\boldsymbol{\mu},\boldsymbol{x})&\triangleq\sum_{n\in\mathcal{U}}\lambda_n\left(\sum_{j\in\mathcal{S}}\sum_{k\in\mathcal{B}}x_n^{j,k}r_n^{j,k}-\bar{r}_n\right)\\ &\quad+\sum_{n\in\mathcal{U}}v_n\left(\sum_{j\in\mathcal{S}}\sum_{k\in\mathcal{B}}x_n^{j,k}R_j-\bar{r}_n\right)\\ &\quad+\sum_{n\in\mathcal{U}}\mu_n\left(\sum_{j\in\mathcal{S}}\sum_{k\in\mathcal{B}}x_n^{j,k}(d_n^{j,k}+D_j)-\bar{d}_n\right).\end{aligned} \qquad (8)$$

Hence, **P3** is now decomposed into two sub-problems

$$\begin{aligned}&\mathbf{P3(1)}\ \mathbf{:}g^1(\boldsymbol{\lambda},\boldsymbol{v},\boldsymbol{\mu})\triangleq\inf_{\check{\boldsymbol{r}},\check{\boldsymbol{d}}}L^1(\boldsymbol{\lambda},\boldsymbol{v},\boldsymbol{\mu},\check{\boldsymbol{r}},\check{\boldsymbol{d}})\\ &\text{s.}t.\ \text{Constraints}\ (3-1),(3-2),(3-6)-(3-9),\end{aligned} \qquad (9)$$

and

$$\begin{aligned}&\mathbf{P3(2)}\ \mathbf{:}g^2(\boldsymbol{\lambda},\boldsymbol{v},\boldsymbol{\mu})\triangleq\inf_{\boldsymbol{x}}L^2(\boldsymbol{\lambda},\boldsymbol{v},\boldsymbol{\mu},\boldsymbol{x})\\ &\text{s.}t.\ \text{Constraints}\ (3-1),(3-2),(3-6)-(3-9).\end{aligned} \qquad (10)$$

When $\boldsymbol{\lambda},\boldsymbol{v},\boldsymbol{\mu}$ are fixed, we solve **P3(1)** and **P3(2)** respectively. The detailed solution to **P3(1)** and **P3(2)** can be found in Appendices A and B respectively, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TMC.2020.3000657. By updating $\boldsymbol{\lambda},\boldsymbol{v},\boldsymbol{\mu}$, we can obtain the optimal solution to **P2** denoted by $\{\check{\boldsymbol{r}}^*,\check{\boldsymbol{d}}^*,\boldsymbol{x}^*\}$. To improve the clarity, we streamline the whole solution process of **P2** in Appendix C, available in the online supplemental material.

Then we design QoS-based UA Policy according to the solution to **P2** to find the UE set QoS-AUS in Definition 2. Let $\check{r}_n^*$ and $\check{d}_n^*$ be the optimal solution of UE n. If $\check{r}_n^*=0$ and $\check{d}_n^*=0$, it means that there is no QoS degradation of UE n. In other words, the network can provide satisfied service for this UE, and thus the UE is admissible. Based on this observation, we design QoS-based UA Policy, where the UEs with $\check{r}_n^*=0$ and $\check{d}_n^*=0$ can be accepted by the network, and others are rejected due to limited resources. Hence, the admissible set of UEs can be expressed by $\mathcal{A}_{Q-A}=\{n:\check{r}_n^*=0,\check{d}_n^*=0,n\in\mathcal{U}\}$.

Theorem 1. *Subset $\mathcal{A}_{Q-A}$ is a QoS-AUS.*

Proof. To prove Theorem 1, we should prove the feasibility and QoS degradation performance (i.e., minimum value of $\sum_{n\in\mathcal{U}\setminus\mathcal{A}_{Q-A}}\left(\frac{\check{r}_n^*}{\bar{r}_n}+\frac{\check{d}_n^*}{\bar{d}_n}\right)$) of $\mathcal{A}_{Q-A}$ respectively.

(1) feasibility: Let $x_n^{j^*k^*}$ be the optimal solution $\boldsymbol{x}^*$ of UE n to **P2**. We denote $\boldsymbol{y}$ as an $|\mathcal{A}_{Q-A}|$-dimensional vector with elements y_n, where $y_n=x_n^{j^*,k^*}$ for $n\in\mathcal{A}_{Q-A}$. As $\boldsymbol{x}^*$ is the optimal solution to **P2**, it is also a feasible solution. Moreover, $\forall n\in\mathcal{A}_{Q-A}$, we have $\check{r}_n^*=0,\check{d}_n^*=0$. Hence, it is easy to verify that $\boldsymbol{y}$ can satisfy all constraints in **P1**

when we use $\mathcal{A}_{Q-A}$ to replace $\mathcal{U}$. In other words, $\mathcal{A}_{Q-A}$ is an AUS.

(2) QoS performance: According to Definition 2, we need to prove that the achievable value of $\sum_{n\in\mathcal{U}\setminus\mathcal{A}_{Q-A}}\left(\frac{\check{r}_n}{\bar{r}_n}+\frac{\check{d}_n}{d_n}\right)$ is no greater than that of any other AUS $\mathcal{H}$. Let us start from set $\mathcal{A}_{Q-A}$. According to the definition of $\boldsymbol{y}$, we have $\check{r}_n^*=0,\ \check{d}_n^*=0$ for all $n\in\mathcal{A}_{Q-A}$, and thus $\sum_{n\in\mathcal{U}\setminus\mathcal{A}_{Q-A}}\left(\frac{\check{r}_n^*}{\bar{r}_n}+\frac{\check{d}_n^*}{d_n}\right)=\sum_{n\in\mathcal{U}}\left(\frac{\check{r}_n^*}{\bar{r}_n}+\frac{\check{d}_n^*}{d_n}\right)$. For any other AUS $\mathcal{H}$, let $\check{\boldsymbol{r}}^{\mathcal{H}}$, $\check{\boldsymbol{d}}^{\mathcal{H}}$ and $\boldsymbol{x}^{\mathcal{H}}$ be the optimal solution of $\check{\boldsymbol{r}}$, $\check{\boldsymbol{d}}$ and $\boldsymbol{x}$, respectively, and thus the minimum achievable value is $\sum_{n\in\mathcal{U}}\left(\frac{\check{r}_n^{\mathcal{H}}}{\bar{r}_n}+\frac{\check{d}_n^{\mathcal{H}}}{d_n}\right)$. As $\mathcal{H}$ is an AUS, we have $\check{r}_n^{\mathcal{H}}=0,\ \check{d}_n^{\mathcal{H}}=0$ for all $n\in\mathcal{H}$, and thus $\sum_{n\in\mathcal{U}\setminus\mathcal{H}}\left(\frac{\check{r}_n^{\mathcal{H}}}{\bar{r}_n}+\frac{\check{d}_n^{\mathcal{H}}}{d_n}\right)=\sum_{n\in\mathcal{U}}\left(\frac{\check{r}_n^{\mathcal{H}}}{\bar{r}_n}+\frac{\check{d}_n^{\mathcal{H}}}{d_n}\right)$. As $\{\check{\boldsymbol{r}}^*,\check{\boldsymbol{d}}^*,\boldsymbol{x}^*\}$ is the optimal solution to **P2**, $\sum_{n\in\mathcal{U}}\left(\frac{\check{r}_n^*}{\bar{r}_n}+\frac{\check{d}_n^*}{d_n}\right)\le\sum_{n\in\mathcal{U}}\left(\frac{\check{r}_n^{\mathcal{H}}}{\bar{r}_n}+\frac{\check{d}_n^{\mathcal{H}}}{d_n}\right)$, and thus $\sum_{n\in\mathcal{U}\setminus\mathcal{A}_{Q-A}}\left(\frac{\check{r}_n^*}{\bar{r}_n}+\frac{\check{d}_n^*}{d_n}\right)\le\sum_{n\in\mathcal{U}\setminus\mathcal{H}}\left(\frac{\check{r}_n^{\mathcal{H}}}{\bar{r}_n}+\frac{\check{d}_n^{\mathcal{H}}}{d_n}\right)$.

Therefore, according to the above proof of feasibility and QoS degradation performance, we can conclude that subset $\mathcal{A}_{Q-A}$ is a QoS-AUS. □

According to Theorem 1, QoS-based UA Policy guarantees both the feasibility to **P1** and the QoS degradation performance. Moreover, this policy also guides network operators to re-allocate bandwidth in the NS re-configuration phase thus to satisfy QoS of all the UEs in set $\mathcal{U}$ with the minimum bandwidth consumption. However, network slice reconfiguration is beyond the scope of this work.

QoS-based UA Policy is focused on network QoS while the performance in term of the number of admissible UEs cannot be guaranteed. In the next subsection, we will design another policy, Num-based UA policy, of determining UE admissibility to maximize the number of admissible UEs.

5.2 Num-Based UA Policy

In the proposed QoS-based UA policy, we find that some UEs with only unsatisfied rate or delay requirement (i.e., $\check{r}_n^*>0,\check{d}_n^*=0$ or $\check{d}_n^*>0,\check{r}_n^*=0$) should be rejected. This means that some unviolated constraints are deleted in **P1**, implying that the network may have some spare resources to admit more UEs. Hence, from the number of admissible UEs viewpoint, the performance of QoS-based UA policy may not be good. Moreover, the number of admissible UEs is also one of the key performance measures of service provisioning for slices. Therefore, we propose Num-based UA policy to further optimize the number of admissible UEs.

By analyzing the optimal solution to **P2**, we find that the smaller $\check{r}_n^*$ or $\check{d}_n^*$ is, the more likely the rate or delay of UE n can be satisfied. Based on this observation, we develop Num-based UA policy to maximize the number of admissible UE, and we denote by $\mathcal{A}_{N-A}$ the obtained UE subset. The basic idea of this policy is trying to add the UEs with small value of $\check{r}_n^*$ and $\check{d}_n^*$ into set $\mathcal{A}_{N-A}$.

First of all, the UEs with $\check{r}_n^*=0$ and $\check{d}_n^*=0$ are definitely admissible to the network. Hence, $\mathcal{A}_{Q-A}\subseteq\mathcal{A}_{N-A}$. We then try to find more admissible UEs from set $\mathcal{U}\setminus\mathcal{A}_{Q-A}$, and add these UEs into $\mathcal{A}_{N-A}$. The details of Num-based UA policy are summarized as Algorithm 1.

In the initialization stage, we add the definitely admissible UEs (i.e., the UEs with $\check{r}_n^*=0$ and $\check{d}_n^*=0$). Then in the search stage, we check the feasibility of other UEs one by one. The smaller the value of $\left(\frac{\check{r}_n^*}{\bar{r}_n}+\frac{\check{d}_n^*}{d_n}\right)$ is, the more likely the UE is admissible. Hence, we check the UEs with the smallest value of $\left(\frac{\check{r}_n^*}{\bar{r}_n}+\frac{\check{d}_n^*}{d_n}\right)$ first. To reduce the computational complexity, once a UE is infeasible for the network, we terminate the check, and then obtain the set $\mathcal{A}_{N-A}$. Therefore, this policy needs to solve **P2** at most $|\mathcal{U}|$ times in the worst case.

Algorithm 1. Algorithm of Num-Based UA Policy

```
Input: problem P2 formulated in (2).
Output: set of admissible UEs $\mathcal{A}_{N-A}$.
Initialization Stage:
1: $\mathcal{A}_{N-A}=\emptyset$, $\mathcal{A}_{temp}=\emptyset$
2: obtain the optimal solution $\check{\boldsymbol{r}}^*$, $\check{\boldsymbol{d}}^*$ and $\boldsymbol{x}^*$ by solving P2
3: add all UEs with $\check{r}_n^*=0$ and $\check{d}_n^*=0$ into $\mathcal{A}_{N-A}$
Search Stage:
4: find UE $i$: $\min_{i\in\mathcal{U}\setminus\mathcal{A}_{N-A}}\left(\frac{\check{r}_i^*}{\bar{r}_i}+\frac{\check{d}_i^*}{d_i}\right)$
5: $\mathcal{A}_{temp}=\{\mathcal{A}_{N-A}\cup$ UE $i\}$
6: obtain the optimal solution $\check{\boldsymbol{r}}^*$, $\check{\boldsymbol{d}}^*$ and $\boldsymbol{x}^*$ by solving P2 with
   respect to $\mathcal{A}_{temp}$
7: if $\sum_{n\in\mathcal{A}_{temp}}\left(\frac{\check{r}_n^*}{\bar{r}_n}+\frac{\check{d}_n^*}{d_n}\right)=0$ then
8:   $\mathcal{A}_{N-A}=\mathcal{A}_{temp}$
9:   Go back to line 4
10: else
11:   break
12: end if
13: output $\mathcal{A}_{N-A}$
```

6 Slice Association and Bandwidth Allocation (SABA) Scheme

After determining UE admissibility, we obtain two admissible UE subsets, $\mathcal{A}_{Q-A}$ and $\mathcal{A}_{N-A}$, in which the UEs' QoS can be satisfied. We now focus on slice association and bandwidth allocation (SABA) for these admissible UEs by solving problem **P1**. Note that **P1** is feasible with respect to the two admissible UE sets after conducting UE admission policies.

6.1 Network-Centric SABA Policy Net-SABA

We first develop the network-centric SABA policy Net-SABA with aim to reach the global optimality of bandwidth consumption. For convenience, we use $y_n^{j,k}=w_n^{j,k}/b_j^{(k)}$ to replace $w_n^{j,k}$, and thus $y_n^{j,k}\in[0,1]$. In addition, we define a mapping function $\phi(n,j,k)$ to determine a unique integer between 1 and $|\boldsymbol{x}|$ when n,j,k are given, where $|\boldsymbol{x}|$ is the number of the elements of all $x_n^{j,k}$. Then, we transform variables $x_n^{j,k}$ and $y_n^{j,k}$ into $\tilde{x}_{\phi(n,j,k)}$ and $\tilde{y}_{\phi(n,j,k)}$, i.e., $x_n^{j,k}=\tilde{x}_{\phi(n,j,k)}$ and $y_n^{j,k}=\tilde{y}_{\phi(n,j,k)}$. Let $\phi_{(n)}^{-1}$, $\phi_{(j)}^{-1}$ and $\phi_{(k)}^{-1}$ be the inverse function of n, j and k, respectively. In this way, for a given integer between 1 and $|x|$, we use $\phi_{(n)}^{-1}$, $\phi_{(j)}^{-1}$ and $\phi_{(k)}^{-1}$ to find the value of n, j and k, respectively. Let $\widetilde{\boldsymbol{x}}$ and $\widetilde{\boldsymbol{y}}$ be the set of $\tilde{x}_{\phi(n,j,k)}$ and $\tilde{y}_{\phi(n,j,k)}$ respectively. In the following, we solve **P1** with respect to $\widetilde{\boldsymbol{x}}$ and $\widetilde{\boldsymbol{y}}$.

Note that it is hard to directly find the optimal solution to **P1** due to the binarity of $\widetilde{\boldsymbol{x}}$. To tackle this problem, we first relax the feasible region of $\widetilde{\boldsymbol{x}}$ and $\widetilde{\boldsymbol{y}}$ to a convex set, and then solve **P1** subject to the relaxed convex feasible region. Let Z be the original feasible region of **P1**, and thus

$Z = \{(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}}) : subject\ to$ constraints $(2-1)-(2-7)\}$. The *convex hull* of a set Z, denoted by $conv(Z)$ is the smallest convex set that contains Z [27]. Using the similar idea of [28], we give $conv(Z)$ in the following.

Define the polynomial factors of degree d as $F_d(J_1, J_2) = [\Pi_{i\in J_1}\tilde{x}_i][\Pi_{j\in J_2}(1-\tilde{x}_j)]$, where $J_1, J_2 \subseteq \{1, 2, , \ldots, |\boldsymbol{x}|\} \equiv \mathcal{J}$, $J_1 \cap J_2 = \emptyset$ and $|J_1 \cup J_2| = d$. To linearize the cross-product terms of $\widetilde{\boldsymbol{x}}$ and $\widetilde{\boldsymbol{y}}$, we define $u_J = \Pi_{i\in J}\tilde{x}_i$ and $v_{J,m} = \tilde{y}_m\Pi_{i\in J}\tilde{x}_i$ for $m = 1, \ldots, |\boldsymbol{x}|$, where $u_\emptyset = 1$ and $v_{\emptyset,m} = \tilde{y}_m$, for $m = 1, \ldots, |\boldsymbol{x}|$. We denote by $f_d(J_1, J_2)$ and $f_d^m(J_1, J_2)$ the linearized forms of polynomial expressions $F_d(J_1, J_2)$ and $\tilde{y}_m F_d(J_1, J_2)$ respectively. For convenience, let $\tilde{b}_n^{j,k} \equiv b_j^{(k)} log_2(1 + SINR_n^k)$, and $\phi \equiv \phi(n, j, k)$.

For constraints (2-1)−(2-3), and (2-5), we use constraints (11)−(14) to relax them, and for constraints (2-4), (2-6), $\tilde{x}_\phi \in \{0, 1\}$ and $\tilde{y}_\phi \in [0, 1]$, we then give (15)−(18) to relax them

$$\begin{aligned}\Lambda_j f_d(J_1, J_2) - &\sum_{\phi\in\mathcal{J}-(J_1\cup J_2)} \tilde{b}_n^{j,k} f_{d+1}^{\phi^{-1}_{(n,k)}}(J_1+\phi, J_2) \\ &- \sum_{\phi\in J_1} \tilde{b}_n^{j,k} f_d^{\phi^{-1}_{(n,k)}}(J_1, J_2) \\ &\geq 0, \forall j \in \mathcal{S}, and\ (J_1, J_2)\ of\ order\ d,\end{aligned} \tag{11}$$

$$\begin{aligned}f_d(J_1, J_2) - &\sum_{\phi\in\mathcal{J}-(J_1\cup J_2)} f_{d+1}^{\phi^{-1}_{(n)}}(J_1+\phi, J_2) \\ &- \sum_{\phi\in J_1} f_d^{\phi^{-1}_{(n)}}(J_1, J_2) \\ &\geq 0, \forall j \in \mathcal{S}, k \in \mathcal{B}, and\ (J_1, J_2)\ of\ order\ d,\end{aligned} \tag{12}$$

$$\begin{aligned}&\sum_{\phi\in J_1} \tilde{b}_n^{j,k} f_d^{\phi^{-1}_{(n,k)}}(J_1, J_2) - \sum_{\phi\in\mathcal{J}-(J_1\cup J_2)} \tilde{b}_n^{j,k} f_{d+1}^{\phi^{-1}_{(n,k)}}(J_1+\phi, J_2) \\ &- \bar{r}_n f_d(J_1, J_2) \geq 0, \forall n \in \mathcal{U}, and\ (J_1, J_2)\ of\ order\ d,\end{aligned} \tag{13}$$

$$\begin{aligned}&\sum_{\phi\in J_1}\left[q_{\phi^{-1}_{(n)}} f_d(J_1, J_2) + D_{\phi^{-1}_{(j)}} \tilde{b}_n^{j,k} f_d^{\phi^{-1}_{(n)}}(J_1, J_2)\right] \\ &- \sum_{\phi\in\mathcal{J}-(J_1\cup J_2)}\left[q_{\phi^{-1}_{(n)}} f_{d+1}(J_1+\phi, J_2) + D_{\phi^{-1}_{(j)}} \tilde{b}_n^{j,k} f_{d+1}^{\phi_{(n)}^{-1}}(J_1+\phi, J_2)\right] \\ &+ \sum_{\phi}\left[\bar{d}_{\phi^{-1}_{(n)}} \tilde{b}_n^{j,k} f_d^{\phi^{-1}_{(n)}}(J_1, J_2)\right] \geq 0, \forall n \in \mathcal{U}, and\ (J_1, J_2)\ of\ order\ d,\end{aligned} \tag{14}$$

$$\sum_{j\in\mathcal{S}}\sum_{k\in\mathcal{B}} \tilde{x}_\phi R_j - \bar{r}_n \geq 0,\ \forall n \in \mathcal{U} \tag{15}$$

$$\sum_{j\in\mathcal{S}}\sum_{k\in\mathcal{B}} \tilde{x}_\phi - 1 \geq 0,\ \forall n \in \mathcal{U} \tag{16}$$

$$f_{D_1}(J_1, J_2) \geq 0, for\ (J_1, J_2)\ of\ order\ D_1 = min\{d, |\boldsymbol{x}|\} \tag{17}$$

$$\begin{aligned}&f_{D_2}(J_1, J_2) \geq f_{D_2}^m(J_1, J_2) \geq 0, for\ m = 1, \ldots, |\boldsymbol{x}|, \\ &and\ (J_1, J_2)\ of\ order\ D_2 = min\{d+1, |\boldsymbol{x}|\}.\end{aligned} \tag{18}$$

By using these relaxed constraints, we obtain a convex relaxation Z_d of the original feasible region, where d is the degree of the relaxation, and $Z_d = \{(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}}, \boldsymbol{u}, \boldsymbol{v}) : subject\ to$ constraints $(9)-(16)\}$. Then focusing on variables $\widetilde{\boldsymbol{x}}$ and $\widetilde{\boldsymbol{y}}$ the d-degree convex relaxation of Z can be expressed as $Z_{Pd} = \{(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}}) : subject\ to\ (\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}}, \boldsymbol{u}, \boldsymbol{v}) \in Z_d\}$. In fact, for all degrees $0 \leq d \leq |\boldsymbol{x}|$, Z_{Pd} is a convex relaxation of the feasible region Z. The larger the degree d is, the tighter the relaxation Z_{Pd} is, and the higher computational complexity is incurred [28].

Theorem 2. *$Z_{P|\boldsymbol{x}|}$ is the convex hull of* ***P1****, i.e., $Z_{P|\boldsymbol{x}|} = conv(Z)$.*

Proof. By using Theorem 3.5, Extension 1 and Extension 2 in [28], we can easily obtain Theorem 2. □

After the relaxation, the feasible region of **P1** becomes a convex set with linear constraints. Hence, **P1** becomes a linear programming which can be solved easily. Based on the solution to the relaxed **P1**, we design Net-SABA policy. In Net-SABA policy, we associate UEs with NSs and BSs according to $\widetilde{\boldsymbol{x}}^*$, and allocate bandwidth according to $\widetilde{\boldsymbol{y}}^*$, where $\{\widetilde{\boldsymbol{x}}^*, \widetilde{\boldsymbol{y}}^*\}$ is the optimal solution to the relaxed problem.

In Net-SABA policy, the major computational complexity lies in the part of solving the relaxed **P1** problem, which is a linear programming with $\mathcal{O}(n^4)$ computational complexity in the worst case [29], where n is the number of variables. For the d-degree relaxed problem, the number of variables can be approximately deemed as $|\boldsymbol{x}|^d$. Hence, the computational complexity of Net-SABA policy is $\mathcal{O}(|\boldsymbol{x}|^{4d})$ when d-degree relaxation is used.

Net-SABA policy requires global network information (the association and bandwidth allocation of all UEs) and high computational complexity. Thus, Net-SABA is not suitable for delay sensitive users. In the following, we will design an efficient UE-centric SABA policy to reduce individual UE bandwidth consumption with low computational complexity.

6.2 UE-Centric SABA Policy UE-SABA

UE-SABA policy consists of two steps, obtaining initial solution and searching better solution. In the first step, let $(\boldsymbol{x}^{(0)}, \boldsymbol{w}^{(0)})$ be the optimal solution obtained by Num-based UA policy. Hence, it is also the feasible solution to **P1**. We use $(\boldsymbol{x}^{(0)}, \boldsymbol{w}^{(0)})$ as the initial solution.

Then in the second step, we try to find a better solution for each UE with the fixed associations of other UEs. Specifically, let $\boldsymbol{x}^{(s)}$ and $\boldsymbol{w}^{(s)}$ respectively be the slice association and the corresponding bandwidth allocation after s searching steps. In the $(s+1)$th searching step, for a specific UE $n \in \mathcal{A}_{N-A}$, we first fix the associations and bandwidth allocations of others, i.e., $\boldsymbol{x}^{(s+1)} = \boldsymbol{x}^{(s)}$ and $\boldsymbol{w}^{(s+1)} = \boldsymbol{w}^{(s)}$ except for the nth element. Then we optimize the nth element $x_n^{j,k}(s+1)$ and $w_n^{j,k}(s+1)$. We find the set $\mathcal{H}_n = \{(x_n^{j,k}, w_n^{j,k}) : subject\ to$ constraints $(2-3)-(2-5)$, and $w_n^{j,k} \leq w_n^{j,k}(s) - \epsilon\}$, where $w_n^{j,k}(s)$ is the bandwidth allocation of UE n at the sth step, and ϵ is an arbitrary positive parameter. If $\mathcal{H}_n = \emptyset$, we obtain $x_n^{j,k}(s+1) = x_n^{j,k}(s)$ and $w_n^{j,k}(s+1) = w_n^{j,k}(s)$. If $\mathcal{H}_n \neq \emptyset$, we find the pair $\left(x_n^{j,k}(s+1), w_n^{j,k}(s+1)\right)$ in $\mathcal{H}_n$ that satisfies: 1) **P1** is feasible respect to $\boldsymbol{x}^{(s+1)}$ and $\boldsymbol{w}^{(s+1)}$, and 2) $w_n^{j,k}(s+1)$ is the smallest one among all the pairs which satisfy condition 1). If all the pairs in $\mathcal{H}_n$ are infeasible for **P1**, we have $x_n^{j,k}(s+1) = x_n^{j,k}(s)$ and $w_n^{j,k}(s+1) = w_n^{j,k}(s)$. In this way, we obtain $x_n^{j,k}(s+1)$ and $w_n^{j,k}(s+1)$, and thus $\boldsymbol{x}^{(s+1)}$ and $\boldsymbol{w}^{(s+1)}$.

Therefore, the $(s+1)$th searching step is finished. The searching termination criteria is set as that the association and bandwidth allocation of all the UEs are unchanged.

Theorem 3. *UE-SABA policy converges in finite searching steps.*

Proof. Denote by $f^{(s)}$ the objective value of **P1** after s searching steps, and $f^{(s)}$ is bounded. Assuming that Theorem 3 is false, then there exist $\boldsymbol{x}^{(s+1)}$ and $\boldsymbol{w}^{(s+1)}$ that $f^{(s+1)} \leq f^{(s)} - \epsilon, \forall s \in Z^+$. Hence, after M searching steps, we have $f^{(s+M)} \leq f^{(s+M-1)} - \epsilon \leq \cdots \leq f^{(s)} - M\epsilon$. As $f^{(s)}$ is bounded, $f^{(s)} - M\epsilon$ can be less than f^* when M is large enough, where f^* is the optimal objective value of **P1**. Hence, we have $f^{(s+M)} < f^*$, which is a contradiction. Therefore, the assumption is not true, and Theorem 3 holds. □

Besides convergence property, let us analyze the bandwidth consumption performance of UE-SABA policy. Pareto optimality (or Pareto efficiency) is a concept to measure the performance of a resource allocation policy, which is stated as: a policy is Pareto optimal if it is impossible to make some individuals better off without making some other individuals worse off [30]. In the following, we prove that the proposed UE-SABA achieves ϵ-Pareto optimality which is weaker than Pareto optimality. The definition of ϵ-Pareto optimality is given as follows.

Definition 3. *A bandwidth allocation $\boldsymbol{w}$ is ϵ-Pareto optimal if for any feasible allocation $\boldsymbol{w}'$ we have if $\exists\ n_0 \in \mathcal{U}$ that $w'_{n_0} < w_{n_0} - \epsilon$, then $\exists\ n_1 \in \mathcal{U}$ and $n_1 \neq n_0$ that $w'_{n_1} > w_{n_1}$.*

Theorem 4. *UE-SABA policy converges to an ϵ-Pareto optimal solution.*

Proof. Let $(\boldsymbol{x}^*, \boldsymbol{w}^*)$ denote a converged solution obtained by UE-SABA policy. Assuming that $\boldsymbol{w}^*$ is not an ϵ-Pareto optimal solution, then there exists a feasible bandwidth allocation solution $\boldsymbol{w}'$ that $\exists\ n_0 \in \mathcal{U}$ that $w'_{n_0} < w^*_{n_0} - \epsilon$, and $\forall n \in \mathcal{U}, n \neq n_0$ that $w'_n \leq w^*_n$. Hence, for the solution $\boldsymbol{w}^*$, the searching termination of UE n_0 is not satisfied. However, as $\boldsymbol{w}^*$ is a converged solution obtained from UE-SABA policy, the searching termination of all UEs should be satisfied, which is a contradiction. Therefore, $\boldsymbol{w}^*$ is an ϵ-Pareto optimal solution. □

Compared with Net-SABA policy, we can see that UE-SABA does not require global network information at searching steps. Moreover, the computational complexity of UE-SABA is much lower than that of Net-SABA. UE-SABA policy only needs to find the smallest value of $w_n^{j,k}(s)$ at each searching step rather than solving an optimization problem. However, UE-SABA policy cannot guarantee the global optimality of total network bandwidth consumption. Nevertheless, UE-SABA provides an effective solution to the service provisioning problem especially for the delay sensitive users.

6.3 Mobility Management

In the section, we first illustrate the impact of user mobility on slice association and bandwidth allocation, and then propose a mobility management scheme. The signalling overhead could be extremely high if we frequently perform our proposed slice association and bandwidth allocation framework. This is because that the proposed framework is centralized and it requires the central controller to collect the information (including SINR, QoS requirement, available bandwidth) of all users and NSs. Also, the framework makes handoff decisions and bandwidth re-allocation for all users simultaneously, which is not directly applicable to mobile users. To address this issue, we develop the following mobility management scheme.

Similar to that in work [31], we give the handoff trigger condition for UE n as

$$\forall t_0 \in [t - \tau_n, t], r_n(t_0) < \gamma_n^{min}, \tag{19}$$

where $r_n(t_0)$ is the achievable transmission rate of UE n at time t_0. This condition states that UE n cannot achieve the minimum rate requirement γ_n^{min} in the last τ_n time. Note that the handoff trigger condition is different for the UEs with different service type. Once a handoff occurs, the UE should choose the target NS and BS to keep connected while moving. Since the target NS and BS selection for handoff users is not the focus of this work, we propose a simple method for selecting NS and BS when a handoff occurs. The handoff UE first chooses an NS that can provide the required service type, and then associates with the BS covered by the NS with sufficient wireless bandwidth. Note that an optimal handoff scheme is discussed in our related work [31].

7 Performance Evaluation

In this section, we evaluate the performance of the proposed service provisioning framework. To the best of the authors' knowledge, there is no explicit solution to the service provisioning to UEs for RAN slicing. Inspired by 3GPP standard [32] and the related work [25], two modified service provisioning schemes, NS-prior association (NSA) and BS-prior association (BSA), are used as our benchmark for performance evaluation of our proposed framework. Specifically, NSA scheme first finds the NS that satisfies the QoS requirement of the UE, and then finds the BS covered by this NS with sufficient bandwidth. BSA scheme first finds the BS with the maximum SINR for the UE, and then finds the NS deployed in this BS with satisfied QoS guarantee. In both NSA and BSA schemes, if such a pair of NS and BS is found, the UE is admissible and associated with the NS and BS. The bandwidth allocation policy for NSA and BSA is to allocate the minimal required bandwidth to UEs to satisfy the QoS requirement. Hence, NSA and BSA mechanisms contain both UE admissibility and SABA schemes, denoted as NSA-UA, BSA-UA and NSA-SABA, BSA-SABA respectively.

We consider a network which consists of a macro BS (MBS) located at the central of a circular area with a radius of 500 m and multiple pico BSs (PBS), femto BSs (FBS), NSs and UEs. The UEs and BSs are randomly distributed in the considered area. The number of UEs and BSs are seen as parameters varying in each simulation experiment. Each NS randomly covers 4 BSs, and provides different transmission rate and delay performance. Thus, the number of supported NSs in each BS is a random variable determined by the NS coverage. Each UE generates a type of service with different

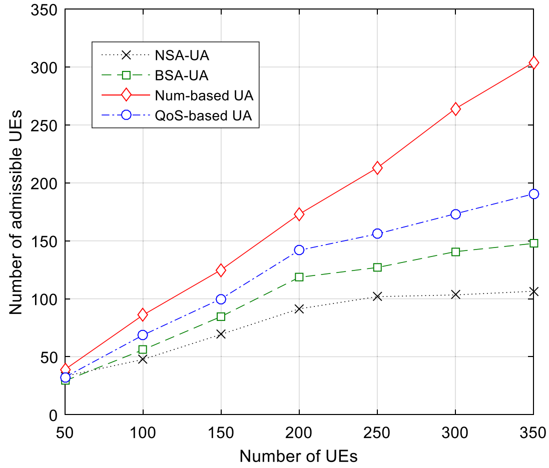


Fig. 4. Comparisons of the number of admissible UEs with different UE density. (The number of NSs is 20, and the number of BSs is 21.).

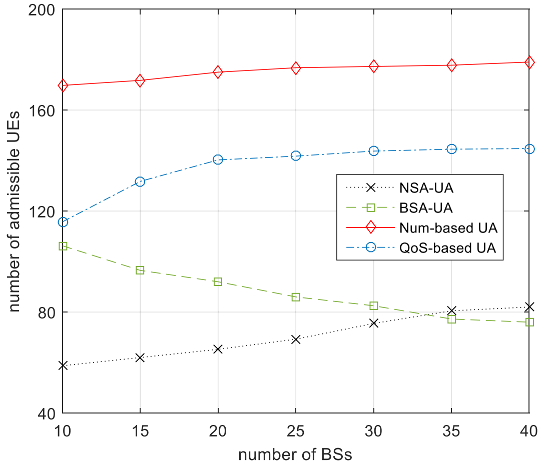


Fig. 6. Comparisons of the number of admissible UEs versus number of BSs. (The number of UEs is 200, and the number of NSs 20.).

rate and delay requirements. The transmit power of MBS, PBS and FBS is set to 46, 30 and 20 dBm, respectively. We use $L(d) = 34 + 40log(d)$ and $L(d) = 37 + 30log(d)$ to model the pass loss for the MBS/PBSs and FBSs respectively [15]. All the BSs share 20 MHz bandwidth.

7.1 Performance of UE Admission Policies

In the first experiment, we compare the number of admissible UEs of the four UE admission policies QoS-based UA, Num-based UA, NSA-UA and BSA-UA. In this experiment, we fix the number of NSs and BSs as 20 and 21 (including one MBS) respectively. Fig. 4 shows the number of admissible UEs for the four UE admission policies with different UE densities. From this figure, we can see that the number of admissible UEs of QoS-based UA and Num-based UA is always higher than that of the other two traditional schemes which do not consider the characteristics of NS. Specifically, when the number of UEs is 200, the admissible number of UEs for Num-based UA, QoS-based UA, BSA-UA and NSA-UA is 173, 142, 118 and 92, respectively. These results show that the proposed Num-based UA policy can serve 47 and 88 percent more UEs when compared with NS-Selection and BSA-UA respectively.

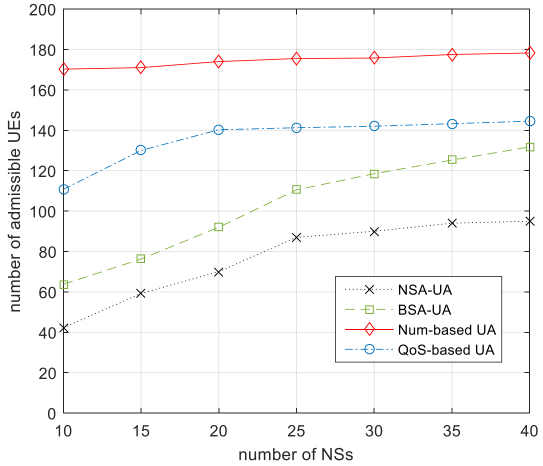
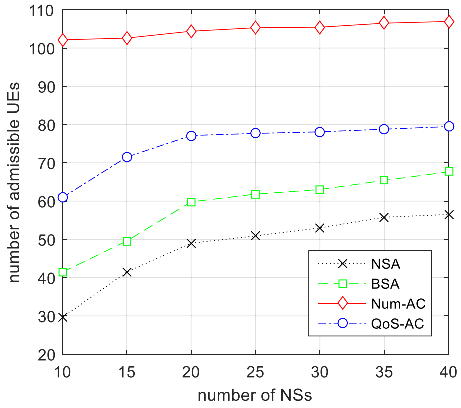


Fig. 5. Comparisons of the number of admissible UEs versus number of NSs. (The number of UEs 200, and the number of BSs 21.).

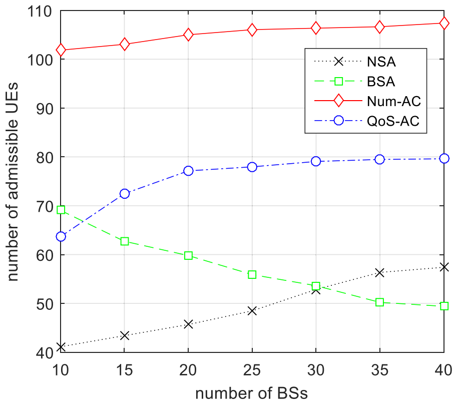
In the second experiment, we evaluate the number of admissible UEs of the four UE admission policies for varying number of NSs while using fixed number of UEs 200. Fig. 5 shows the number of admissible UEs for the four schemes as a function of number of NSs. From this figure, we can see that at the beginning (the number of NSs is lower than 25) the number of admissible UEs of all four policies increases rapidly with the number of NSs. This is because that the more NSs deployed the more association choices for UEs, and thus the more UEs can be admitted. However, when the number of NSs is larger than 25, the number of admissible UEs of all the four schemes increases slowly or even stays unchanged due to the limited resources in both core and access networks. Thus, the results of Fig. 5 keep consistent with the fact that simply increasing the number of slices would not necessarily increase the number of QoS-satisfied users. We also find that the number of admissible UEs of Num-based UA and QoS-based UA is always significantly higher than that of NSA-UA and BSA-UA under all NS density circumstances.

We next investigate the relationship between the number of admissible UEs and the number of BSs with the same parameters as those in the first experiment, and fix the number of UEs to 200. Fig. 6 shows the number of admissible UEs for the four schemes as a function of number of BSs. From this figure, we can see that the number of admissible UEs of Num-based UA, QoS-based UA and NSA-UA monotonically increases with the number of BSs, while that of BSA-UA decreases which is due to the decreasing number of NSs deployed in each BS. Moreover, the number of admissible UEs of Num-based UA and QoS-based UA is always much higher than that of the other two traditional UE admission policies. These results clearly demonstrate the performance gain of proposed Num-based UA and QoS-based UA schemes in terms of the number of admissible UEs.

Next, we investigate the performance of the proposed UE admissibility policies with the consideration of queueing delay. In this experiment, we evaluate the number of admissible UEs of the four UE admission policies for varying number of NSs and NS respectively. All the parameters stay the same with those in the second and third experiments



(a) number of admissible UEs vs number of NSs



(b) number of admissible UEs vs number of BSs

Fig. 7. Comparisons of the number of admissible UEs.(with queueing delay.).

except the introduction of queueing delay. Fig. 7 shows the comparisons of the number of admissible UEs for the four schemes when queueing delay is introduced. From both Figs. 7a and 7b, we can see that the performance gain of our proposed policies Num-based UA and QoS-based UA still be valid. Moreover, we find that the number of admissible users of all four policies could be lower than that in Figs. 5 and 6 due the the existence of queueing delay.

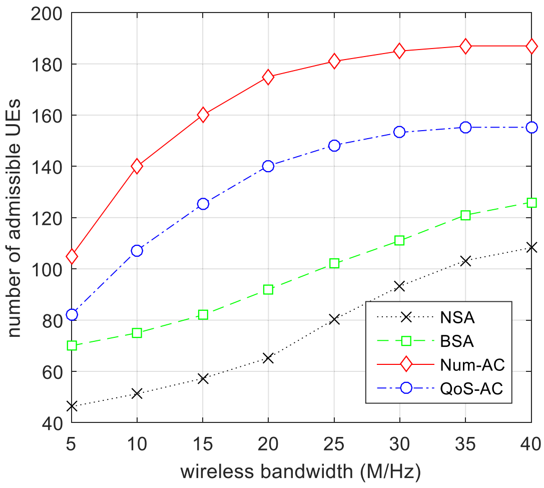


Fig. 8. Comparisons of the number of admissible UEs versus amount of wireless bandwidth.

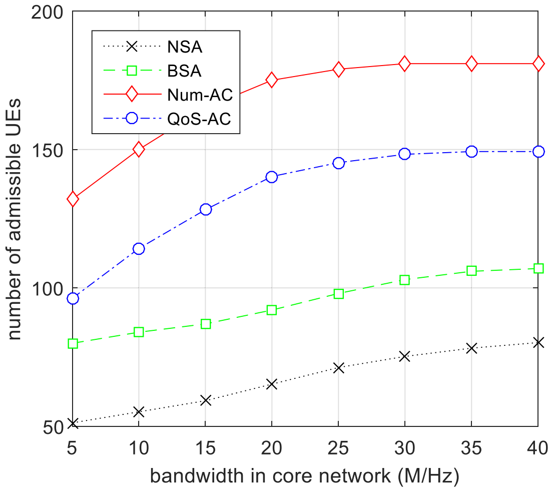


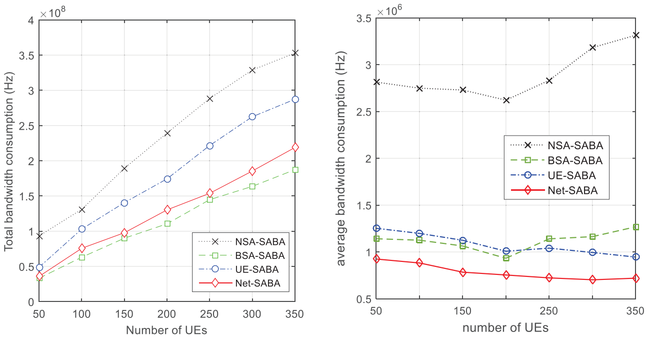
Fig. 9. Comparisons of the number of admissible UEs versus amount of bandwidth in core network.

In addition, we evaluate the number of admissible UEs of the four UE admission policies for varying amount of total wireless bandwidth. All the parameters stay the same with those in the first experiment while using fixed number of UEs 200. Fig. 8 shows the comparisons of the number of admissible UEs for the four schemes with different amount of wireless bandwidth. From this figure, we can see that the performance gain of our proposed policies Num-based UA and QoS-based UA increase with the amount of wireless bandwidth at the beginning. Moreover, the number of admissible users of Num-based UA and QoS-based UA policies increases slowly when the amount of wireless bandwidth is larger than 30 MHz due to the limitation of resource in core network.

Then, we investigate how the resource in core network can affect the proposed UE admission policies. In this experiment, we explore the number of admissible UEs of the four UE admission policies for varying amount of bandwidth in core network. All the parameters stay the same with those in the first experiment while using fixed number of UEs 200. Fig. 9 shows the comparisons of the number of admissible UEs for the four policies. From this figure, we can see that the number of admissible users of all the four policies increase with the amount of bandwidth in core network. Moreover, the number of admissible users of Num-based UA and QoS-based UA policies increases slowly when the amount of bandwidth is larger than 20 MHz because of the limited wireless spectrum.

7.2 Performance of SABA Schemes

Next, we evaluate the performance of the proposed SABA schemes. In this experiment, we compare the bandwidth consumption of the four SABA schemes (Net-SABA, UE-SABA, NSA-SABA and BSA-SABA) with the same system settings and parameters as those in the first experiment, i.e., we fix the number of NSs and BSs as 20 and 21 respectively. We compare the total and average UE bandwidth consumption, where the total bandwidth consumption is the sum bandwidth consumption for all admissible UEs, and the average UE bandwidth consumption is defined as the total bandwidth consumption divided by the number of admissible UEs. Fig. 10 shows the bandwidth consumption for the

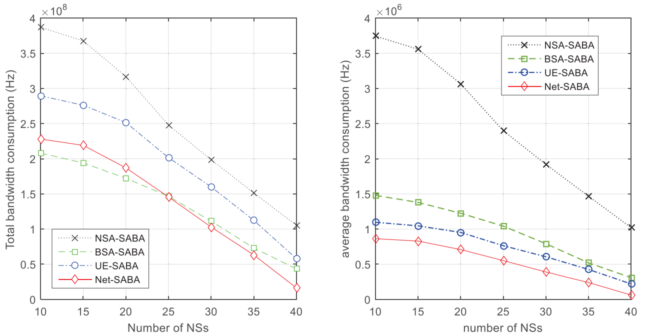


(a) Total bandwidth consumption (b) Average bandwidth consumption

Fig. 10. Relationship between bandwidth consumption and the number of UEs. (The number of NSs 20, and the number of BSs 21.).

four schemes as a function of the number of UEs. From Fig. 10a, we can see that the total bandwidth consumption of the traditional scheme BSA-SABA is always the smallest. This is because that UEs always access the BS with the maximum SINR value in BSA-SABA policy. The total bandwidth consumption of Net-SABA and UE-SABA is higher than that of Net-SABA but significantly lower than that of NSA-SABA. Moreover, we find that the difference of total bandwidth consumption between Net-SABA and BSA-SABA is relatively small (for example, 7 percent for 150 UEs), implying that much more UEs (for example, 51 percent for 150 UEs) can be served with a small compromise on total bandwidth consumption. Note that from Section 3.1, we can see that the number of admissible UEs of BSA-SABA and NSA-SABA is significantly smaller than that of the two proposed schemes. Therefore, we evaluate the average UE bandwidth consumption shown in Fig. 10b. From this figure, we can see that the average bandwidth consumption of Net-SABA is always the lowest, and when the number of UEs is larger than 200, the bandwidth consumption of UE-SABA is lower than that of BSA-SABA. For example, when the number of UEs is 300, the average UE bandwidth consumption for Net-SABA, UE-SABA, BSA-SABA and NSA-SABA is approximately 7.0×10^5, 1.0×10^6, 1.2×10^6, and 3.2×10^6 Hz. These results show that Net-SABA and UE-SABA can serve more UEs with the lower average UE bandwidth consumption.

Then, we examine the bandwidth consumption of the four SABA schemes for varying number of NSs while using fixed number of UEs 200. Fig. 8 shows the bandwidth consumption for the four schemes with different number of NSs. From Fig. 11a, we can see that the total bandwidth consumption of all four schemes monotonically decreases with the number of NSs. The more NSs deployed, the better association choice can be made, and thus the less bandwidth consumption. When the number of NSs is greater than 25, the total bandwidth consumption of Net-SABA is the lowest. Fig. 11b shows the average UE bandwidth consumption of the four SABA schemes. We can see that the average UE bandwidth consumption of Net-SABA and UE-SABA scheme is much lower than that of the other two traditional schemes.



(a) Total bandwidth consumption (b) Average bandwidth consumption

Fig. 11. Relationship between bandwidth consumption and the number of NSs. (The number of UEs 200, and the number of BSs 21.).

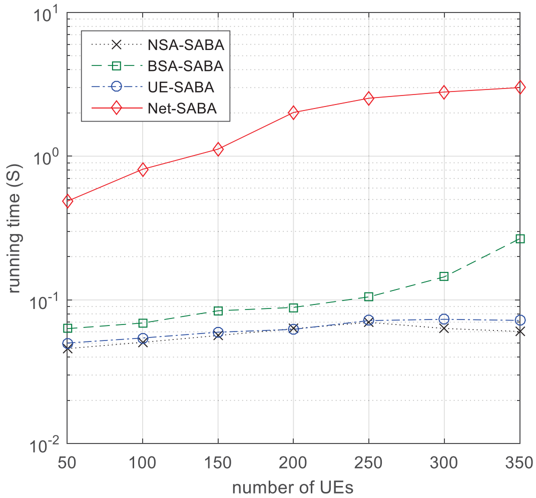


Fig. 12. Comparisons of the running time versus number of UEs. (The number of NSs 20, and the number of BSs 21.).

In the last experiment, we evaluate the running time of the four SABA schemes with the same system settings and parameters as the first experiment. Running time directly reflects the computational complexity for a SABA scheme. Our numerical computations are implemented with MATLAB codes and carried out on a PC equipped with an Intel-i5 4 core 3.2 GHz processor and 4G RAM. Fig. 12 shows the running time of the four schemes with different number of UEs. From this figure, we can see that the running time of Net-SABA is always the highest due to the part of solving an LP in Net-SABA. Moreover, we find that the proposed efficient policy UE-SABA achieves similar running time with that of the traditional scheme NSA-SABA.

8 Conclusion

In this paper, we have investigated service provisioning for UEs in RAN slicing. We have proposed a unified framework for user access control and bandwidth allocation to minimize bandwidth consumption while guaranteeing QoS of users. Numerical results demonstrate the significant performance gain of our proposed framework in terms of the number of admissible UEs and bandwidth consumption when compared with the traditional mechanisms in typical scenarios. This work illustrated the importance of service provisioning for users in RAN slicing, and gave a guidance of designing the optimal service provisioning framework.

ACKNOWLEDGMENTS

The work was supported by the National Science Foundation of China under Grant number 61971099, China Postdoctoral Science Foundation under Grant 2019M663476, and U.K. Engineering and Physical Sciences Research Council under Grant EP/S02476X/1.

REFERENCES

[1] Ericsson, "5G systems," *Ericsson White Paper*, Jan. 2015.

[2] L. Zhang, A. Ijaz, J. Mao, P. Xiao, and R. Tafazolli, "Multi-service signal multiplexing and isolation for physical-layer network slicing (PNS)," in *Proc. IEEE 86th Veh. Technol. Conf.*, 2017, pp. 1–6.

[3] H. Zhang, N. Liu, X. Chu, K. Long, A.-H. Aghvami, and V. C. Leung, "Network slicing based 5G and future mobile networks: Mobility, resource management, and challenges," *IEEE Commun. Mag.*, vol. 55, no. 8, pp. 138–145, Aug. 2017.

[4] Y. Sun, L. Zhang, G. Feng, B. Yang, B. Cao, and M. A. Imran, "Blockchain-enabled wireless Internet of Things: Performance analysis and optimal communication node deployment," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 5791–5802, Jun. 2019.

[5] X. An *et al.*, "On end to end network slicing for 5G communication systems," *Trans. Emerg. Telecommun. Technol.*, vol. 28, no. 4, 2017, Art. no. e3058.

[6] D. D. Clark and W. Fang, "Explicit allocation of best-effort packet delivery service," *IEEE/ACM Trans. Netw.*, vol. 6, no. 4, pp. 362–373, Aug. 1998.

[7] O. Sallent, J. Pérez-Romero, R. Ferrús, and R. Agustí, "On radio access network slicing from a radio rersource management perpective," *IEEE Wireless Commun.*, vol. 24, no. 5, pp. 166–174, Oct. 2017.

[8] R. Li *et al.*, "Deep reinforcement learning for resource management in network slicing," *IEEE Access*, vol. 6, pp. 74429–74441, 2018.

[9] B. Cao, Y. Li, C. Wang, G. Feng, S. Qin, and Y. Zhou, "Resource allocation in software defined wireless networks," *IEEE Netw.*, vol. 31, no. 1, pp. 44–51, Jan./Feb. 2017.

[10] A. Nakao *et al.*, "End-to-end network slicing for 5G mobile networks," *J. Inf. Process.*, vol. 25, pp. 153–163, 2017.

[11] X. Hu, R. Chai, G. Jiang, and H. Li, "A joint utility optimization based virtual AP and network slice selection scheme for SDWNs," in *Proc. 10th Int. Conf. Commun. Netw. China*, 2016, pp. 448–453.

[12] G. Zhao, S. Qin, G. Feng, and Y. Sun, "Network slice selection in softwarization-based mobile networks," *Trans. Emerg. Telecommun. Technol.*, vol. 31, no. 1, 2020, Art. no. e3617.

[13] W. Wang, X. Wu, L. Xie, and S. Lu, "Femto-matching : Efficient traffic offloading in heterogeneous cellular networks," in *Proc. IEEE Conf. Comput. Commun.*, 2015, pp. 325–333.

[14] E. Aryafar, A. Keshavarz-Haddad, M. Wang, and M. Chiang, "RAT selection games in HetNets," in *Proc. IEEE Conf. Comput. Commun.*, 2013, pp. 998–1006.

[15] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and re 3jJ. G. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2706–2716, Jun. 2013.

[16] H. Boostanimehr and V. K. Bhargava, "Unified and distributed QoS-driven cell association algorithms in heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 3, pp. 1650–1662, Mar. 2015.

[17] Y. Sun, G. Feng, S. Qin, Y.-C. Liang, and T.-S. P. Yum, "The SMART handoff policy for millimeter wave heterogeneous cellular networks," *IEEE Trans. Mobile Comput.*, vol. 17, no. 6, pp. 1456–1468, Jun. 2018.

[18] S. E. Elayoubi, A. Eitan, M. Haddad, and Z. Altman, "A hybrid decision approach for the association problem in heterogeneous networks," in *Proc. IEEE INFOCOM*, 2010, pp. 1–5.

[19] Y. Sun, G. Feng, S. Qin, and S. Sun, "Cell association with user behavior awareness in heterogeneous cellular networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 4589–4601, May 2018.

[20] R. Kokku, R. Mahindra, H. Zhang, and S. Rangarajan, "CellSlice: Cellular wireless resource slicing for active RAN sharing," in *Proc. 5th Int. Conf. Commun. Syst. Netw.*, 2013, pp. 1–10.

[21] X. Chen, Z. Han, H. Zhang, G. Xue, Y. Xiao, and M. Bennis, "Wireless resource scheduling in virtualized radio access networks using stochastic learning," *IEEE Trans. Mobile Comput.*, vol. 17, no. 4, pp. 961–974, Apr. 2018.

[22] Q. Ye, W. Zhuang, S. Zhang, A.-L. Jin, X. Shen, and X. Li, "Dynamic radio resource slicing for a two-tier heterogeneous wireless network," *IEEE Trans. Veh. Technol.*, vol. 67, no. 10, pp. 9896–9910, Oct. 2018.

[23] M. Yan, G. Feng, J. Zhou, Y. Sun, and Y.-C. Liang, "Intelligent resource scheduling for 5G radio access network slicing," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 7691–7703, Aug. 2019.

[24] Q. Ye, J. Li, K. Qu, W. Zhuang, X. S. 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.

[25] M. K. Marina, "Orion: RAN slicing for a flexible and cost-Effective multi-service mobile network architecture," in *Proc. 23rd Annu. Int. Conf. Mobile Comput. Netw.*, 2017, pp. 127–140.

[26] C. Liang and F. R. Yu, "Wireless network virtualization: A survey, some research issues and challenges," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 358–380, First Quarter 2015.

[27] S. P. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[28] H. D. Sherali and W. P. Adams, "A hierarchy of relaxations and convex hull characterizations for mixed-integer zero-one programming problems," *Discrete Appl. Math.*, vol. 52, no. 1, pp. 83–106, 1994.

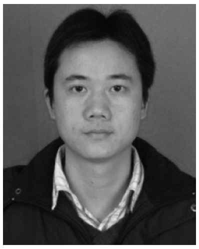
[29] N. Karmarkar, "A new polynomial-time algorithm for linear programming," *Combinatorica*, vol. 4, no. 4, pp. 373–395, 1984.

[30] A. Mas-Colell, M. D. Whinston, and J. R. Green, *Microeconomic Theory*. London, U.K.: Oxford Univ. Press, 1995.

[31] Y. Sun, G. Feng, L. Zhang, P. V. Klaine, M. A. Iinran, and Y.-C. Liang, "Distributed learning based handoff mechanism for radio access network slicing with data sharing," in *Proc. IEEE Int. Conf. Commun.*, 2019, pp. 1–6.

[32] Access, Evolved Universal Terrestrial Radio. "Radio resource control (RRC)," Protocol specification (Release 10) 290, 2013.

Yao Sun received the BS degree in mathematical sciences, and the PhD degree in communication and information system from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2014 and 2019, respectively. He has published widely in wireless networking research area, and received the IEEE ComSoc TAOS Best Paper Award in 2019. His research interests include intelligent access control, handoff, and resource management in mobile networks.



Shuang Qin (Member, IEEE) received the BS degree in electronic information science and technology, and the PhD degree in communication and information system from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2006 and 2012, respectively. He is currently an associate professor with the National Key Laboratory of Science and Technology on Communications, UESTC. His research interests include cooperative communication in wireless networks, data transmission in opportunistic networks, and green communication in heterogeneous networks.

Gang Feng (Senior Member, IEEE) received the BEng and MEng degrees in electronic engineering from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 1986 and 1989, respectively, and the PhD degrees in information engineering from the Chinese University of Hong Kong, Hong Kong, in 1998. He joined the School of Electric and Electronic Engineering, Nanyang Technological University in December 2000 as an assistant professor and became an associate professor in October 2005. At present, he is a professor with the National Laboratory of Communications, UESTC. He has extensive research experience and has published widely in wireless networking research. A number of his papers have been highly cited. He has received the IEEE ComSoc TAOS Best Paper Award and ICC best paper award in 2019. His research interests include next generation mobile networks, mobile cloud computing, AI-enabled wireless networking, etc.

Lei Zhang (Senior Member, IEEE) received the PhD degree from the University of Sheffield, Sheffield, United Kingdom. He is now a lecturer with the University of Glasgow, United Kingdom. His research interests broadly lie in the communications and array signal processing, including radio access network slicing (RAN slicing), wireless blockchain networks, new air interface design, Internet of Things (IoT), V2X, multi-antenna signal processing, massive MIMO systems, etc. He is holding 16 US/UK/EU/China granted patents on wireless communications. He also holds a visiting position in 5GIC at the University of Surrey. He is an associate editor of the *IEEE Access*.

Muhammad Ali Imran (Senior Member, IEEE) is a professor of Wireless Communication Systems with research interests in self organised networks, wireless networked control systems, and the wireless sensor systems. He heads the Communications, Sensing and Imaging CSI Research Group, University of Glasgow. He is an affiliate professor with the University of Oklahoma and a visiting professor with 5G Innovation Centre, University of Surrey, United Kingdom. He has more than 20 years of combined academic and industry experience with several leading roles in multi-million pounds funded projects. He has filed 15 patents; has authored/co-authored more than 400 journal and conference publications; was editor of three books and author of more than 20 book chapters; has successfully supervised more than 40 postgraduate students at doctoral level. He has been a consultant to international projects and local companies in the area of self-organised networks. He is a fellow of the IET and a senior fellow of the HEA.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/csdl.