IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 7, 2018, accepted September 12, 2018, date of publication September 20, 2018, date of current version October 17, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2871506

Multi-Service Resource Allocation in Future Network With Wireless Virtualization

LETIAN LI[1], NA DENG[2], WEILONG REN[1], BAOHUA KOU[3], WUYANG ZHOU[1], (Member, IEEE), AND SHUI YU[4], (Senior Member, IEEE)

[1]Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei 230026, China
[2]School of Information and Communication Engineering, Dalian University of Technology, Dalian 116024, China
[3]China Telecom Satellite Communications, Beijing 100190, China
[4]School of Software, University of Technology Sydney, Sydney, NSW 2007, Australia

Corresponding author: Wuyang Zhou (wyzhou@ustc.edu.cn)

This work was supported in part by the Key Program of National Natural Science Foundation of China under Grant 61631018 and in part by the National Natural Science Foundation of China under Grant 61701071.

ABSTRACT Future network is envisioned to be a multi-service network which can support various types of terminal devices with diverse quality of service requirements. As one of the key technologies, wireless virtualization establishes different virtual networks dependent on different application scenarios and user requirements through flexibly slicing and sharing wireless resources in future networks. In this paper, we first propose a service-centric wireless virtualization model to slice network according to service types. In this model, how to share and slice wireless resource is one of the fundamental issues to be addressed. Therefore, we formulate and solve a multi-service resource allocation problem to realize spectrum virtualization. Different from the existing strategies, we decouple the multi-service resource allocation problem in the proposed virtualization model to make it easier to solve. Specifically, it is solved in two stages: inter-slice resource allocation and intra-slice resource scheduling. In the first stage, we formulate the inter-slice resource allocation as a discrete optimization problem and propose a heuristic algorithm to get sub-optimal solution of this NP-hard problem. In the second stage, we modify several existing scheduling algorithms suitable for scheduling users of several specific services. Numerical results show the superiority of the proposed scheduling algorithms over the existing ones when applied to schedule specific services. Moreover, proposed resource allocation scheme is verified to meet the properties of virtualization and solves the multi-service resource allocation problem well.

INDEX TERMS Multi-service, resource allocation, user scheduling, wireless virtualization.

I. INTRODUCTION

A. MOTIVATION

Resource allocation is a complicated problem in future networks where services will become more diverse because of the introduction of machine type applications [1]. Recently, as the widespread popularity of smartphones, tablets and other mobile devices, many user-oriented multimedia applications, like streaming video, online gaming and mobile video conference have become important parts of customer services. These human-centric applications will coexist with machine type applications in future networks [2]. Specifically, machine type applications, including security, gaming, remote management and control, industrial wireless automation, distributed/mobile computing, health monitoring and ambient assisted living [3], are extremely diverse. While bringing convenience to people, machine type communications (MTC) also provide many challenges to wireless resource allocation. Since MTC services have characteristics of massive devices and short packets, it is difficult to manage them in traditional cellular systems which are designed for human-type communications [4]. Therefore, traditional resource allocation schemes are not fully suitable for future networks where various types of services coexist and the MTC services become an important part. It is crucial to introduce novel methods or technologies to solve this problem.

Wireless virtualization, an essential part of the future networking paradigm [5], enables the coexistence of multiple isolated logical networks on the same substrate network. In network virtualization environment, complicated physical network can be separated into several simple virtual networks which makes the multi-service resource allocation

problem simpler. Consequently, wireless virtualization is a better way to address the multi-service resource allocation problem. However, many challenges remain to be addressed before we apply wireless virtualization to solve multi-service resource allocation problem. It should be noted that in virtualized network, traditional internet service providers (ISPs) are decoupled into two independent entities [6]: infrastructure providers (InPs) and service providers (SPs). Specifically, InPs manage the physical infrastructure, while SPs offer different end-to-end services through virtual networks which are created by Mobile Virtual Network Operators (MVNOs) through leasing resources from InPs. This evolution of business model exerts influence on resource allocation problems. Besides, good isolation among slices have to be maintained to guarantee that virtual networks do not influence each other. Therefore, in this paper, we focus on how to apply wireless virtualization in the multi-service network and how to allocate resources among different users in virtualized future networks which are still open issues.

B. RELATED WORK

Resource allocation in either multi-service or virtualized network has been studied in recent years. However, few works have jointly considered these two issues and studied multi-service resource allocation problem in virtualized cellular networks.

1) MULTI-SERVICE RESOURCE ALLOCATION

Due to the diverse quality of service (QoS) requirements of various types of services, multi-service resource allocation is absolutely vital for energy efficiency, spectrum efficiency, QoS and quality of experience (QoE) provisioning. Therefore, many researchers have focused on this problem.

Classical scheduling algorithms like Round Robin (RR), Maximum Carrier to Interference (MAX C/I) and Proportional Fairness (PF) do not consider different QoS requirements. Different from these, Modified Largest Weighted Delay First (M-LWDF) and Exponential Rule (EXP-RULE), which take delay and packet loss ratio into consideration, were proposed to schedule users in multi-service network. Furthermore, many modifications [7]–[10] to these scheduling algorithms were done to get better performance. For instance, real-time (RT) and non-real-time (NRT) services [7] are scheduled differently in a uniform and centralized way. Moreover, Ali and Zeeshan [8] allocated resources among different services and each service is scheduled separately with their allocated resources. However, these authors adopted the same algorithm to schedule different services which do not take full advantage of their different characteristics.

As we can see, all the works mentioned above did not consider MTC services which are one of the fundamental parts of future networks. As the features of MTC services are quite different from human type services, these existing methods can hardly schedule them together and achieve good performance. Moreover, previous works did not try to schedule different services in a distributed way.

2) RESOURCE ALLOCATION IN VIRTUALIZED CELLULAR NETWORKS

Due to broadcast nature and stochastic fluctuation of wireless links, there are still many challenges to be addressed to realize wireless resource virtualization. Consequently, some researchers have focused on this to promote further study of wireless virtualization.

The existing works can be divided into two categories on the basis of whether their objectives are system performance or economic profit. The resource virtualization solutions that belong to the first category are studied in [11]–[14]. Generally, in this category of works, InP is the central scheduler and directly allocates radio resources to users of different MVNOs. In the second category of works, such as [15]–[19], MVNOs are also involved and the resource allocation problem becomes a hierarchical problem. To solve this problem, Zhu and Hossain [15] designed a hierarchical combinatorial auction mechanism, while Ho *et al.* [17], Fu and Kozat [18] applied game theory and Kazmi *et al.* [19] proposed a hierarchical matching game based scheme.

These works did realize wireless resource virtualization successfully. However, they did not consider how to satisfy various QoS requirements of different services. Consequently, these works can hardly be used to realize multi-service resource allocation in the future network. Different from them, in our work, different types of services (including MTC services) and wireless virtualization are both taken into consideration. In our proposed virtualization model, different services are scheduled in different virtual networks with different scheduling algorithms to get better performance. These scheduling algorithms are specially designed according to service features.

C. CONTRIBUTIONS

In this paper, we focus on the multi-service resource allocation in future networks with wireless virtualization. Generally, the multi-service resource allocation is a complicated optimization problem which should not only optimize system performance but also satisfy various QoS requirements. In our proposed service-centric wireless virtualization model, we decouple the multi-service resource allocation into two simpler problems instead of solving it directly. In such a way, the complexity of the problem is greatly reduced. In addition, it is easy to be extended to general situations with much more services due to the intrinsic characteristics of wireless virtualization. We summarize the contributions of this paper as follows.

- We focus on multi-service resource allocation in virtualized future network which is seldomly considered in previous works. MTC services, which have a tremendous impact on resource allocation scheme, are also considered in this paper.

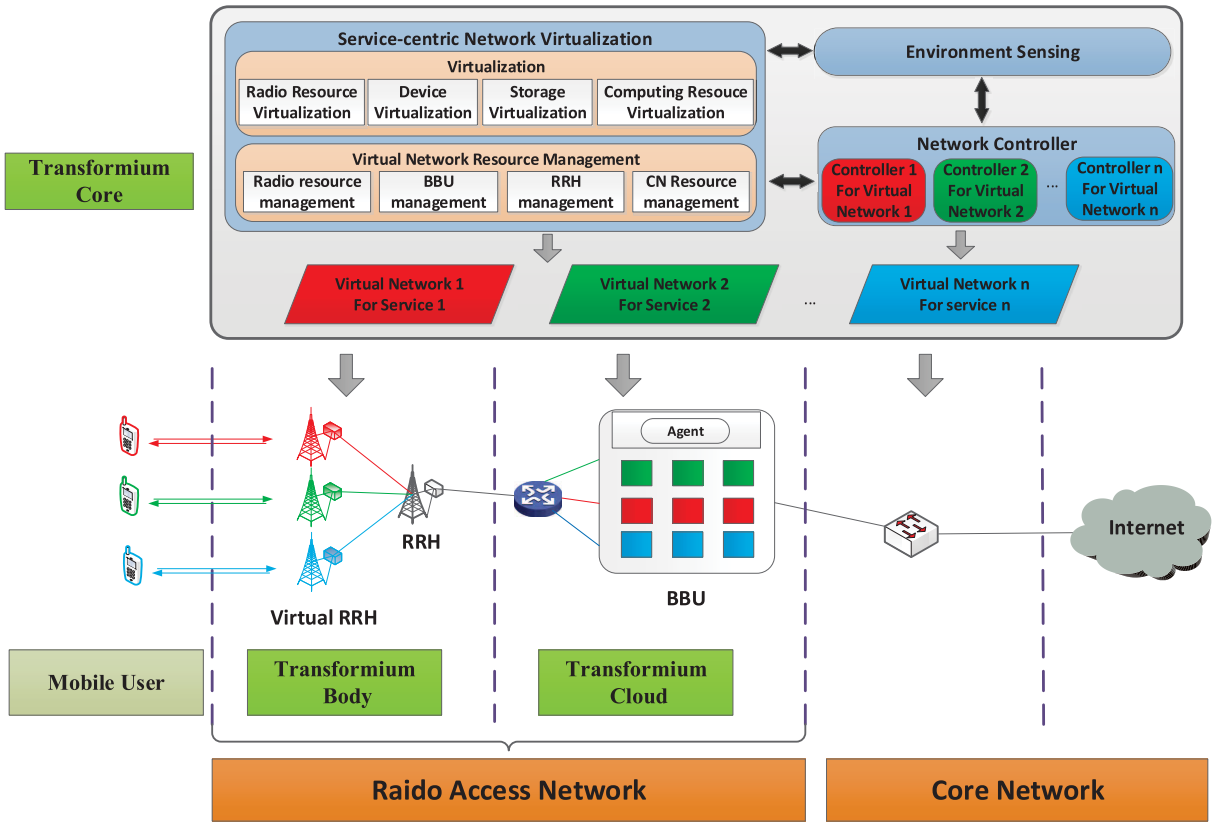


FIGURE 1. Proposed architecture of future networks.

- We propose a service-centric wireless virtualization model where physical network is sliced according to service types. In such a model, services are classified into several types and each type is served in the virtual network specially designed for them. Therefore, the resource allocation problem is decoupled into two simpler problems which are inter-slice resource allocation and intra-slice resource scheduling.
- To solve this two problems, a utility function is designed to represent satisfaction degree of a slice first. Next, we formulate inter-slice resource allocation problem based on the utility function. Then a heuristic algorithm is proposed to get a sub-optimal solution. Moreover, several scheduling algorithms for specific services are devised to solve the intra-slice resource scheduling problem.
- Simulations are implemented to analyze the performance of proposed algorithms. Numerical results show that proposed scheduling algorithms are more suitable for specific services compared with the existing ones. In addition, the proposed scheme meets the properties of virtualization and outperforms traditional scheduling algorithms.

The remainder of this paper is organized as follows. System model is introduced in Section II. Problem formulation is described in Section III. Next, we present a heuristic algorithm and several specially designed scheduling algorithms in Section IV. Simulation results and performance analysis are described in Section V. Finally, we conclude the paper in Section VI.

II. SYSTEM MODEL

To satisfy the ever increasing service types and traffic volume, current network architecture should be rethought to improve efficiency and flexibility. Therefore, a **transformium**[1] **network architecture** is proposed for future networks, which can be arbitrarily transformed for adaptation to the environmental changes of networks. As shown in Fig. 1, the proposed architecture consists of three major parts: transformium core, body and cloud, where the transformium core is logically centralized control plane and the transformium body and cloud are data plane. Network virtualization is employed in the control plane to enable the transforming ability. Network virtualization module has two functions, i.e., abstraction and virtualization of substrate resources and virtual network resource management. More detailed introduction to transformium network architecture can be seen in our previous

[1]Inspired by the movie "Transformers", we introduced the concept "Transformium". It can transform into any thing you want. We believe that future network should also has the transforming ability. Therefore, we name the proposed network architecture "Transformium Network Architecture". Detailed explanation can be seen in our pervious work [20].

work [20]. Since different services coexist in the future network, we propose a service-centric virtualization model to realize wireless resource virtualization in transfomium networks. In this paper, we focus on how to realize the multi-service resource allocation in the proposed virtualization model.

In the following, we first present the proposed virtualization model in detail. Then we display the architecture of virtualized Base Station (BS) which is devised based on LTE protocol stack. Finally, the scheme of network virtualization and communication model for a cellular network are also discussed.

A. VIRTUALIZATION MODEL

In the network virtualization environment, infrastructures are decoupled from the services it provides. As a result, ISPs are decoupled into InPs and SPs in wireless virtualization. Specifically, InPs own the infrastructures and wireless network resources including radio access networks (RANs), backhaul networks and data centers. MVNOs allocate these resources to each virtual network to realize virtualization, whereas SPs provide certain services (e.g., video, FTP or machine type services) through these virtual networks. Specifically, in the proposed transformium network architecture, transformium body and cloud are both owned by InPs, whereas the functions of MVNOs are realized in transformium core.

As described in [20], in the proposed service-centric virtualization model, we classify services into several types and construct a slice for each type of service. A "slice" [21] is a virtual network which is just like a slice of the substrate network. In other words, each type of service is served in a virtual network specially designed for it. Specifically, from the scheduling algorithm up to the protocol stack a virtual network can be devised based on the service features to get better system performance and user experience.

Due to the features of massive devices and short packets, MTC services are greatly different from traditional services. Therefore, services can be roughly divided into two parts which are human-type and MTC services. Like many previous works, services in human-type communication are classified into RT and NRT services. Referring to [22], MTC services are classified into four types: Low Priority, High Priority, Scheduled and Emergency. For simplicity, we use MTC-LP, MTC-HP, MTC-S, MTC-E to denote these four types of MTC services, respectively. In our virtualization model, we construct a virtual network for each type of service and devise special scheduling algorithm. Although more detailed classification can achieve slightly better system performance, it will bring greater challenge to slice management and scheduling algorithm design. As we focus on changes of multi-service resource allocation brought by wireless virtualization, the simple classification described above is adopted in this paper.

Considering a case that one SP provides more than one type of services, this SP will provide its services through several slices, and QoS requirements of different types of services will be guaranteed in their corresponding slices. Although our resource allocation scheme can be applied in such case, it will bring much complexity to the descriptions and notations in the paper. As a result, for convenience, we assume that each SP just provides one type of service in the following parts.

B. VIRTUALIZED BASE STATION

Based on the virtualization model described above, the virtualized BS model is presented in Fig. 2. Physical infrastructure owned by InP is placed in the bottom layer on the top of which is hypervisor. When the system starts, hypervisor is responsible for collecting user information, like QoS requirements, channel conditions, and signal to noise ratio (SNR). Then resources including spectrum, storage and computing are allocated by the hypervisor to create different slices based on these information. After slices are created, hypervisor will collect the information from each slice to maintain their operations. It can be seen that the function of MVNO is basically realized by hypervisor. On the top layer, different virtual networks are built for their corresponding types of services. Services are classified into different types and then enter into their corresponding virtual networks for differentiated management. As each virtual network serves merely one type of service, these virtual networks can be customized according to the service features. For example, scheduling algorithm of MAC scheduler can be specially devised to take charge of intra-slice resource managing for specific services. Each virtual base station sends their transmission rate requirements, user channel conditions and other related information to the hypervisor as the basis of resource allocation.

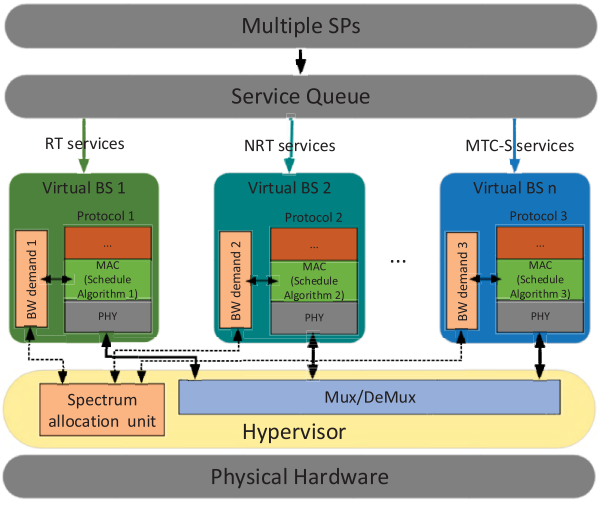


FIGURE 2. Virtualized BS model.

C. NETWORK MODEL

Since the focus of this paper is multi-service resource allocation in virtualized future networks, we consider the downlink resource allocation in a single cell with just one BS.

Resources are represented in the unit of resource block (RB) and we use $i \in I = \{1, 2, \ldots, R_{total}\}$ to denote the total resource of the BS. The BS and spectrum are owned and managed by a single InP which provides its network infrastructure as a service to MVNO. We assume that there are m SPs $j \in J = \{1, 2, .., m\}$ and n users $p \in P = \{1, 2, \ldots, n\}$. Each SP provides a certain type of service to some users denoted by $K_j = \{1, 2, \ldots, k_j\}$. It is assumed that users who apply for different services have different QoS requirements whereas those applying for the same service have the same QoS requirement. We use R^j_{min}, R^j_{max}, d^j_{max}, δ^j_{max} to denote the minimum and maximum transmission rate requirements, maximum delay and packet loss rate requirements of users in slice j respectively. Each SP will send these requirements and channel conditions to MVNO. Then MVNO allocates specific RB to different slices according to received information. And finally each SP can schedule these resources in its corresponding slice to serve their users.

III. PROBLEM FORMULATION OF RESOURCE ALLOCATION IN VIRTUALIZED FUTURE NETWORK

In this section, we formulate the multi-service resource allocation problem in virtualized wireless network which is decoupled into inter-slice resource allocation and intra-slice resource scheduling. First, a utility function is designed to represent user satisfaction degree based on some observations. Then we use the weighted utility function as the optimization objective and formulate the inter-slice resource allocation problem. Finally, we demonstrate that intra-slice resource scheduling is a scheduling algorithm devising problem.

A. UTILITY DESIGN

In this paper, we focus on improving system performance and user satisfaction degree. To achieve a good inter-slice resource allocation, we construct a utility function to measure user satisfaction degree. Generally, user satisfaction degree depends on transmission rate, delay and packet loss ratio. We intend to consider the influence of them separately.

1) UTILITY OF RATE

As transmission rate request is the major factor that influences resources allocated to different slices, we first design a utility function of assigned transmission rate. As there is no such a recognized utility function, we choose a plausible one based on two observations. First, the utility function should be monotone increasing since the more assigned rate, the better QoS users will get. Second, to conform to the marginal utility, the increasing rate of utility will decline with the allocated resources when the demands of users in the slice is basically satisfied. Following these principles, we devise the utility of rate [8], [23] as:

$$U_r(R_j) = \frac{1}{1 + \exp(\frac{-10}{R^j_{max} - R^j_{min}}(R_j - R^j_{min}))}, \tag{1}$$

where R^j_{max}, R^j_{min} are the minimum and maximum transmission rate requests, respectively, and R_j is the assigned rate of slice j. Moreover, $\frac{10}{R^j_{max} - R^j_{min}}$ is chosen as the exponential factor to control the increasing rate of the utility function.

2) UTILITY OF DELAY

To get the average delay and packet loss ratio, we formulate the buffer as a first in first out queue. In addition, we assume the packet whose delay is larger than d^j_{max} will be dropped. Therefore, when transmission rate is not high enough, packet delay increases with time and part of packets will be dropped from a certain time (denoted by t_p). Based on these, the average packet delay and loss ratio $d_j(R_j, t)$ and $\delta_j(R_j, t)$ can be calculated as:

$$d_j(R_j, t) = \begin{cases} \frac{S_j}{R_j}, & R_j t^j_I \le S_j, \\ \frac{S_j t - R_j t_I t + 2 S_j t_I}{2 R_j t_I}, & R_j t^j_I > S_j,\ t \le t_p, \\ \frac{S_j t^j_p + d^j_{max} R_j (2t - t^j_p)}{2 R_j t}, & otherwise, \end{cases} \tag{2}$$

$$d_j(R_j, t) = \begin{cases} \frac{(S_j - R_j t_I)(t - t^j_p)}{S_j t}, & R_j t^j_I > S_j,\ t > t_p, \\ 0, & otherwise, \end{cases} \tag{3}$$

where S_j, t^j_I are the average packet size and arrival interval of slice j, respectively. In addition, t_p is expressed as:

$$t_p = \frac{d^j_{max} R_j t_I - S_j t_I}{S_j - R_j t_I}. \tag{4}$$

From the equations above, $d_j(R_j, t)$ and $\delta_j(R_j, t)$ will become the function of R_j when given a fixed t. As t is irrelevant with resource allocation, we set it to a fixed value and use the notations $d_j(R_j)$ and $\delta_j(R_j)$ instead of $d_j(R_j, t)$ and $\delta_j(R_j, t)$.

Then we devise the utility of delay based on the analysis above. As we can see, average delay increases with time and approaches its asymptote d^j_{max} when R_j is not high enough. Therefore, the utility should be 0 when $d_j(R_j)$ is equal to d^j_{max}. In addition, when $d_j(R_j)$ is much less than d^j_{max}, the utility should be 1. On the contrary, when delay increases and approaches d^j_{max}, the utility should decrease rapidly. Based on the observations above, we devise the utility of delay [24] as:

$$U_d(d_j) = 1 - \exp(1/10(d_j - d^j_{max})), \tag{5}$$

where 1/10 is chosen as the exponential factor to control the increasing rate.

3) UTILITY OF PACKET LOSS RATIO

Next, we attempt to devise the utility of packet loss ratio. Generally, the utility should be 1 when $\delta_j(R_j)$ is much less than δ^j_{max}. Furthermore, the utility should decrease rapidly when $\delta_j(R_j)$ approaches and becomes larger than δ^j_{max}. Based on

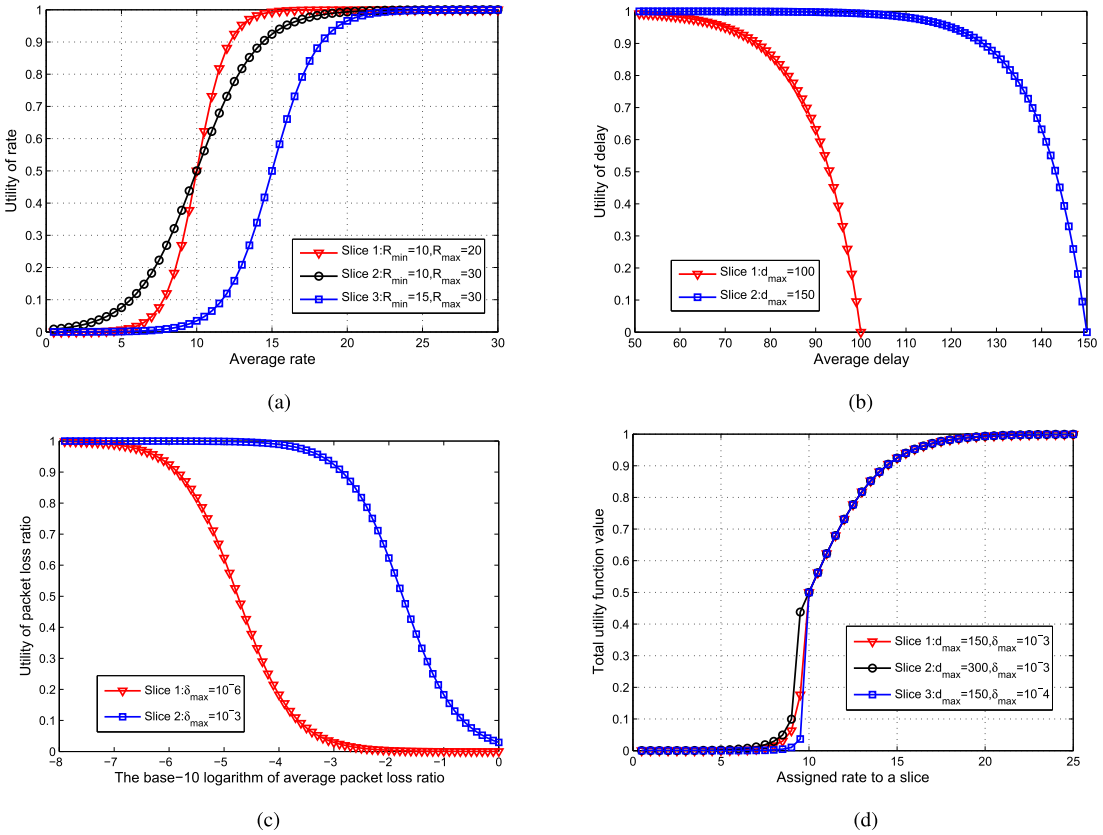


FIGURE 3. Performance of different utility functions. (a) Utility of rate. (b) Utility of delay. (c) Utility of packet loss ratio. (d) Total utility.

these observations, we devise the utility of packet loss ratio as:

$$U_p(\delta_j) = 1 - \frac{1}{1 + 10\exp(2(-log_{10}(\delta_j) + log_{10}(\delta^j_{max})))}, \tag{6}$$

where 10 and 2 are chosen to control the increasing rate.

Finally, the total utility function can be expressed as:

$$U(R_j) = U_r(R_j)U_d(d_j)U_p(\delta_j). \tag{7}$$

As d_j and δ_j are functions of R_j, the total utility can also be seen as the function of R_j.

To show the properties directly, we draw the curves of different utility functions in Fig. 3. Observing the utility of rate, we can see that it is monotone increasing. Furthermore, its derivative decreases continuously as the assigned rate increases when the rate is larger than R^j_{min}. Consequently, both properties of the utility of rate are satisfied. We think that the demand of slice j are basically satisfied when the assigned rate R_j is equal to R^j_{min}. From Fig .3(b), we can find that $U_d(d_j)$ is equal to 1 when d_j is much less than d^j_{max} and decreases rapidly when d_j becomes larger than d^j_{max}. In addition, from Fig .3(c), we can find that $U_p(\delta_j)$ is equal to 1 when δ_j is much less than δ^j_{max} and decreases rapidly when δ_j reaches and becomes larger than δ^j_{max}. Therefore, the properties of utility of delay and packet loss ratio are also satisfied. Finally, the total utility is displayed in Fig .3(d), we can see that the slice with higher delay or packet loss ratio requirement has larger utility function value when assigned the same rate.

B. PROBLEM FORMULATION

In traditional cellular networks without virtualization, resources are allocated by the centralized controller to users directly. Therefore, multi-service resource allocation in traditional cellular network is a complicated optimization problem because of various QoS requirements. Moreover, machine type applications, which are essentially different from human-type applications, bring greater challenge to the existing resource allocation schemes. Therefore, we introduce wireless virtualization to simplify the problem. In our proposed virtualization model, the resource allocation are completed in two stages: inter-slice resource allocation and intra-slice resource scheduling.

The inter-slice resource allocation is designed to allocate resources among slices. Considering the transmission rate requests of different slices, the problem is formulated as:

$$\max_{A} f(A) \tag{8}$$

$$s.t.\ a_{ij} \in \{0, 1\}, \quad \forall i, \forall j, \tag{9}$$

$$\sum_{i=1}^{m} a_{ij} \leq 1, \quad \forall i, \forall j \tag{10}$$

$$R_{min}^{j} \leq R_j \leq R_{max}^{j}, \quad \forall j, \tag{11}$$

$$R_j = \sum_{i=1}^{R_{total}} a_{ij} B \log(1 + SNR_{ij}), \quad \forall i, \forall j, \tag{12}$$

where $A = (a_{ij} : i \in I, j \in J)$ is the vector of optimization variables. We use a_{ij} to denote the RB assignment solution, where $a_{ij} = 1$ indicates that RB i is allocated to slice j and $a_{ij} = 0$ otherwise. The objective function $f(A) = \sum_{j=1}^{m} k_j U(R_j)$ presented in Eq. (8) is the weighted utility function value of all slices and k_j is the number of users in slice j. Eq. (10) assures that one RB can just be allocated to one slice. Thus the orthogonality among the resources of different slices is guaranteed. Besides, minimum and maximum transmission rate constraints are guaranteed by Eq. (11). We consider the minimum transmission rate constraints to guarantee the isolation among slices and the maximum transmission rate constraints to avoid resource waste. And finally, as shown in Eq. (12), the transmission rate of each slice is the sum of the Shannon capacity in each allocated RB. As there are multiple users in slice j, we use SNR_{ij} to denote its average signal-to-noise ratio in RB i. In this problem model, transmission rate is the result of the resource allocation algorithms and perceived by users. In other words, the units of both input and output resources are transmission rate. Therefore, our proposed scheme passes the recursive test and conforms the definition of virtualization [25].

Whereas problem in the inter-slice resource allocation stage is formulated as above, scheduling algorithms for different services need to be devised to solve the intra-slice resource scheduling problem. As slices are built according to service types in our virtualization model, the scheduler in each slice only needs to schedule single type of service with its own features. Therefore, the problem is much simpler compared with that in multi-service networks. Moreover, scheduling algorithm can be specially designed to take advantage of the service features. In addition, this will lead to a better performance of each type of service in its corresponding virtual network.

IV. PROPOSED ALGORITHMS

It is complex and sometimes infeasible to solve the multi-service resource allocation directly. Instead of that, due to the introduction of wireless virtualization, we simplify the multi-service resource allocation problem through decoupling it into inter-slice resource allocation and intra-slice resource scheduling. As we can see, the inter-slice resource allocation problem is a non-convex integer nonlinear programming problem which is NP-hard. Therefore, we propose a heuristic algorithm to get the non-negative integer solution. In addition, intra-slice resource scheduling algorithms for different types of services are also designed.

A. ALGORITHM FOR INTER-SLICE RESOURCE ALLOCATION

As discussed in [26], a simple case of non-convex mixed-integer nonlinear programs, where all variables are integer constrained, is NP-hard. Therefore, the inter-slice resource allocation problem, which is a non-convex integer nonlinear programming problem, is also NP-hard. As NP-hard problem can hardly be solved by convex optimization method [27], we resort to devising a heuristic algorithm for inter-slice resource allocation which is presented in Algorithm I.

Algorithm 1 Heuristic Algorithm

Input: m, R_{min}^{j}, R_{max}^{j}, SNR_{ij}.
Output: inter-slice resource allocation solution $A = (a_{ij})$
Initialize m, R_{min}^{j}, R_{max}^{j}, SNR_{ij}, $a_{ij} = 0$.
Construct Buffer $buffer_q$, $q = 1, 2, \ldots, m$ for each slice.
for each RB $i = 1, 2 \ldots, R_{total}$ **do**
  if average SNR of slice j on RB i SNR_{ij} is larger than other slices **then**
    put RB i into $buffer_j$.
  end if
end for
for $q = 1, 2, \ldots m$ **do**
  Sort RB sequence in $buffer_q$ according to the SNR value in descending order.
end for
Find the slice j with smallest utility function value.
if $buffer_j$ is not empty **then**
  Allocate the first RB i in $buffer_j$ to slice j ($a_{ij} = 1$).
  Remove RB i from $buffer_j$.
else
  for all RB in other $buffer_q$ (except $buffer_j$) **do**
    Find RB i with minimum value of $SNR_{iq} - SNR_{ij}$.
    Allocate RB i to slice j ($a_{ij} = 1$).
    Remove RB i from $buffer_q$.
  end for
end if
if all buffer is empty **then**
  return A as the inter-slice resource allocation solution.
end if

In the algorithm, the slice with the smallest utility function value is the first one to be allocated resources. We set such a rule for two reasons. First, this rule can guarantee the fairness among slices and thus be more likely to satisfy their transmission rate requirements. Second, as the derivative of the utility function decreases continuously as the assigned rate increases when the rate is larger than R_{min}^{j}, the allocation of each RB tends to maximize the objective of the optimization problem. Moreover, to maximize the resource utilization, each RB tends to be allocated to the slice which has the highest average SNR. In the following, we present the algorithm in detail.

At the beginning of the algorithm, some parameters are initialized and the optimization variable a_{ij} are set to be 0. The transmission rate requests and channel conditions are known

in advance by the scheduler. Then we will construct the buffer for each slice. As frequency selective fading is considered in this paper, users have different SNR values in different RBs. Based on this point, RB are classified and put into different buffers. If the average SNR of slice j on RB i is larger than other slices, RB i will be put into the buffer of slice j, i.e., $buffer_j$. In the following resource allocation process, we prefer to assign the RB in $buffer_j$ to slice j. By such a classification, each RB tends to be used by the slice with the best channel conditions. Next, for each slice, we sort the RB sequences in its buffer according to the corresponding SNR value in the descending order. As a consequence, the best RB will be used first to achieve better system performance. Finally, we will schedule the slice with the smallest utility function value which can guarantee the fairness among slices. The first RB in its buffer will be assigned if the buffer is not empty. Otherwise, we will search the buffers of other slices and find RB i in $buffer_q$ with the minimum value of $SNR_{iq} - SNR_{ij}$. This allocation scheme can minimize the loss bring by mismatch between RB and its best slice. We will return A as the inter-slice resource allocation solution when all RB are allocated. Algorithm I is running in the hypervisor as shown in Fig. 2. As described above, the hypervisor will allocate appropriate resources to different slices according to their transmission rate requirements, channel and network load conditions through Algorithm I.

B. ALGORITHMS FOR INTRA-SLICE RESOURCE SCHEDULING

Intra-slice resource scheduling can be formulated as a scheduling algorithm design problem. As different types of services are served separately in different slices, we can specially design the scheduling algorithms for certain type of service to get better performance. In the following, scheduling algorithms for different types of services are devised based on some existing ones.

1) EXISTING SCHEDULING ALGORITHMS

In the design of practical schedulers, we should jointly consider QoS requirements, channel conditions and fairness. Generally, in a scheduling algorithm, each user in the network is assigned a scheduling priority and the user with highest priority will be scheduled in each time slot. The scheduling priority in the PF algorithm is defined as

$$p_p(t) = C_p(t) \big/ \overline{R}_p(t), \tag{13}$$

where $C_p(t)$ is the channel capacity to present the maximum maximum achievable transmission rate of user p in time t, and $\overline{R}_p(t)$ is the average rate expressed as

$$\overline{R}_p(t+1) = (1 - \frac{1}{t_c})\overline{R}_p(t) + \frac{1}{t_c} R_p(t), \tag{14}$$

where $R_p(t)$ is the real transmission rate of user p in time t and t_c is the average rate update factor. As the PF scheduling algorithm does not consider the QoS requirements of users, it is not suitable for multi-service resource scheduling. Different from this, M-LWDF takes QoS requirements into consideration and defines the scheduling priority as

$$p_p(t) = -\lg(\delta^p_{max}) \frac{C_p(t)}{\overline{R}_p(t)} \frac{d_p(t)}{d^p_{max}}, \tag{15}$$

where δ^p_{max}, d_p and d^p_{max} are packet loss rate requirement, queue delay and maximum tolerable queue delay of user p, respectively. In [8], Ali and Zeeshan proposed a different scheduling algorithm called DELAY in which the scheduled user is chosen as

$$u = \arg\min_p (d^p_{max} - HOL_p(t)), \tag{16}$$

where $HOL_p(t)$ is the head of line packet delay (the difference between current time and the time it first arrives at the buffer queue) of user p.

Scheduling algorithms introduced above take time slot as the resource unit to schedule users. Actually, LTE system has consider scheduling users in a smaller granularity, i.e., RB. These algorithms can be easily extended to the situation of RB scheduling.

2) SCHEDULING ALGORITHM FOR RT/NRT SERVICES

Different from the traditional scheduling algorithms, the scheduling algorithm in the proposed virtualization model only needs to schedule users of single type of service. Therefore, scheduling algorithm design is simpler and more specialized.

To take channel condition, packet loss rate requirement and users' transmission rate into consideration, we adopt the expression similar to M-LWDF in the scheduling algorithm for RT services (SART) and that for NRT services (SANRT). Generally, RT services have lower delay and higher packet loss ratio requirements, whereas NRT services have higher delay and lower packet loss rate requirements. To satisfy different QoS requirements of RT and NRT services, we devise different utility functions of packet delay for them. In the following, we first present SART, SANRT and their utility functions of delay. Then we explain why we devise such utility functions for them.

In SART, the scheduling priority is defined as

$$p_{ip}(t) = -\lg(\delta^p_{max}) \frac{C_{ip}(t)}{\overline{R}_p(t)} U_1(d_p(t)), \tag{17}$$

where $U_1(d_p(t))$ is the utility function of delay for RT service users, defined as

$$U_1(d_p(t)) = \log_a(d_p(t)/d^p_{max} + 1), \tag{18}$$

where a ($a > 1$) is the variable parameter of U_1 function which can be adjusted to achieve better performance.

In SANRT, scheduling priority and utility function of delay are expressed as

$$p_{ip}(t) = -\lg(\delta^p_{max}) \frac{C_{ip}(t)}{\overline{R}_p(t)} U_2(d_p(t)), \tag{19}$$

$$U_2(d_p(t)) = b^{(d_p(t)/d^p_{max})}, \tag{20}$$

where b ($b > 1$) is the variable parameter of U_2 function which can be adjusted to achieve better performance.

We assume users of RT/NRT services are scheduled in RB. As a result, $p_{ip}(t)$ and $C_{ip}(t)$ in Eq. (17) and (19) are the scheduling priority and the channel capacity of user p in RB i. We assume that the packet whose delay is more than d^p_{max} will be dropped. Therefore, packet loss rate can also be controlled by the control of packet delay.

As described above, we devise U_1 and U_2 to satisfy different QoS requirements of RT and NRT services. Specifically, U_1 is devised as a logarithmic function which is a concave function. On the contrary, U_2 is devised as an exponential function which is a convex function. U_1 increases rapidly when $d_p(t)/d^p_{max}$ is small, and U_2 increases rapidly when $d_p(t)/d^p_{max}$ is large. This characteristic ensures that SART tends to schedule packets with low delay and SANRT with high delay. Furthermore, as U_2 will increase greatly as $d_p(t)/d^p_{max}$ increases, high delay packet will be scheduled first which leads to a low packet loss ratio. As a result, different QoS requirements of RT and NRT services are guaranteed through U_1 and U_2.

3) SCHEDULING ALGORITHM FOR MTC-S SERVICES

Different from traditional services in human type communications, MTC services have many distinctive features. Specifically, there are enormous amount of machine type terminals in the network, while each just transmits short and small number of packets. Besides, there is a long period between two successive data transmissions. Due to these properties, although MTC services can either be RT or NRT services, scheduling algorithms for the latter two are not appropriate for former. Therefore, the scheduling algorithm for MTC services should be specially devised.

In our virtualization model, we need to devise four scheduling algorithms for all types MTC services. As we focus on the virtualization model and resource allocation scheme, we merely devise the algorithm for the MTC-S service as an example. In the following simulations, we also just consider the MTC-S service.

In addition to the common features described, MTC-S is delay tolerant which is different from other types of MTC services. Considering these features, we devise the scheduling algorithm for MTC-S services (SAMTC). Due to the good performance of M-LWDF, we devise SAMTC based on it. Specifically, the scheduling priority of SAMTC is expressed as

$$p_{xp}(t) = -\lg(\delta^p_{max})\frac{C_{xj}(t)}{\overline{R}_p(t)}\frac{d_p(t)}{d^p_{max}}n^p_{buf}(t), \tag{21}$$

where $n^p_{buf}(t)$ is the queue length of user p in time t. Due to the features of MTC-S services described above, traffic volume in the buffer of a single user may not fully occupy the assigned resources, leading to resource wasting. To tackle this problem, we make two modifications based on M-LWDF. First, we take the queue length into consideration. Consequently, users with higher traffic volume in its buffer tend to be scheduled. Second, we propose to apply a smaller scheduling granularity. To be specific, we can choose resource element (RE) or resource element group (REG) depending on specific circumstances. In Eq. (21), we use x to represent different resources in the unit of RE or REG. As a result, the possibility of resource wasting will be reduced by these two modifications.

To schedule users with the finer granularity, we should also redesign the network to support such a scheduling scheme. Due to the isolation and customization properties of wireless virtualization, we just need to establish a virtual network which applies given scheduling scheme. However, in traditional network, it is difficult to change the substrate network. What's more, all services are served in the network not just MTC services which means any change to the network will cause the coupling effect on all services. Therefore, different from traditional network where all services have to be scheduled in the same scheduling granularity, virtualized network have the advantage of adaptive scheduling granularities.

Actually, we assume that the resource granularities applied in the proposed scheduling algorithms are adaptive. They can be adjusted according to overhead of information reporting and service features like packet size and transmission interval. If the service features change in the future, we can adjust the scheduling granularity in the meanwhile to achieve better system performance.

These proposed scheduling algorithms are running in different virtual base stations as shown in Fig. 2. For instance, SART is running in the virtual base stations for RT services. Differentiated scheduling algorithms are more likely to satisfy various QoS requirements of different services. Furthermore, this approach can be easily realized in virtualized network on account of the isolation among slices.

V. PERFORMANCE ANALYSIS

In this section, we implement intensive simulations to analyze the performance of the devised scheduling algorithms and the proposed virtualization model. First, we make a comparison between the proposed scheduling algorithms and classic ones in terms of throughput, delay and packet loss ratio for each type of service. Then we implement simulation to verify whether the proposed resource virtualization scheme satisfies the properties of virtualization like isolation, customization and high resource utilization. Finally, we also compare the overall performance of multi-service resource allocation in our virtualization model with that in traditional network. The simulation results about arguments described above will be presented and analyzed in the following.

A. SIMULATION SETUP

1) SYSTEM PARAMETERS

For simplicity, we just consider a single cell in the system. Referring to the settings in [28], system parameters used in our simulations are displayed in Table 1. The power density of thermal noise power is set to -174dBm/Hz. Users are

TABLE 1. System parameters in simulation.

Parameters	Values	Parameters	Values
Number of Cells	1	Radius of Cell	1000m
P_{macro}	46dBm	Penetration Loss	20dB
Macro Ant. Gain	14dBi	UE Ant. Gain	0 dB
User Distribution	Uniform	Velocity of User	3Km/h
Duration of Subframe	1ms	Noise Power Spectrum Density	-174 dBm/Hz

uniformly distributed in the macro cell of radius 1000m. Transmit power of macro BS is 46 dBm. The system bandwidth for simulation is 5MHz which includes 25 discrete physical resource blocks in the downlink. The path loss model used in this paper is given by $L(d) = 15.3 + 37.6\lg(d)$ where d in meter is the distance between the BS and the user. The value of t_c and t_p are 2000 timeslots and 5s respectively. According to the service features nowadays, we assume resources are scheduled in the unit of RB in SART and SANRT, and RE or REG in SAMTC. As all services have to be scheduled in the same granularity in traditional network, we assume all services are scheduled in the unit of RB in DELAY, PF and M-LWDF.

2) TRAFFIC MODEL

In our simulation, human type and machine type terminals are both included. Furthermore, we assume human type terminals apply for conversational video and FTP services which are the representatives of RT and NRT services respectively. Added by MTC-S service applied by MTC terminals, these three types of services are introduced in detail in the following:

- **RT video:** We assume RT video steaming traffic will generate packets of variable sizes periodically. Referred to the streaming video traffic in 3GPP2/TSGC.R1002, the traffic model applied in this paper is shown in Table 2. We assume that 25 frames arrive in one second and each frame consists of 8 packets. Besides, both packet size and inter-arrival time between packets follow the truncated pareto distribution [29].
- **NRT FTP:** We assume NRT FTP to be the file transmission that generates packets of fixed size periodically. We adopt a 2-state Markov (ON/OFF) model to represent the FTP traffic. The state of ON represents one file is being transmitted now and vice versa. The length of the ON and OFF periods obeys the exponential distribution with means of 1 second and 1.35 seconds, respectively. Each file consists of multiple packets with same size. For simplicity, we assume that the time interval of packet arrival is fixed.
- **MTC-S service:** MTC-S service has many applications and we just adopt a simple traffic model to embody its basic characters. These characters are massive terminals, small size packets and determined transmission frequency [30]. As a result, we assume that MTC-S service will generate a packet of 50 bytes every 50ms. The packet transmission characters of MTC-S service are reflected in such a traffic model.

TABLE 2. RT video traffic model.

Characteristics	Distribution	Parameters
Inter-arrival Time between Frames	Deterministic	40ms
Number of Packets/Frame	Deterministic	8
Packet Size	Truncated Pareto Max: 1000 bytes	$K = 600$ bytes $\alpha = 1.2$
Inter-arrival Time between Packets	Truncated Pareto Max: 10 ms	$K = 2.5$ ms $\alpha = 1.2$

In the simulation, delay requirements of these three types of services are 150, 300 and 100ms respectively. In addition, the packet loss ratio requirements are 10^{-3}, 10^{-6} and 10^{-6} respectively. Users are taken in a mixed proportion with 2.5% for RT video, 2.5% for NRT FTP, 95% for MTC-S traffic. Due to the massive terminals characters, MTC terminals account for most of users in the network.

B. SIMULATION RESULTS

1) PERFORMANCE OF PROPOSED SCHEDULING ALGORITHMS

To evaluate the performance of the proposed SART, we compare it with DELAY, PF and M-LWDF algorithms. We assume that fixed number of RBs are used to schedule RT video users whose number is increasing gradually. As shown in Fig. 4, the performances in terms of in average throughput, delay and packet loss ratio are compared among different algorithms. We can see that SART and M-LWDF have similar performance and they both outperform the other two algorithms. However, there still exist some difference between the performance of SART and M-LWDF. To be specific, SART has lower packet delay and higher packet loss ratio than M-LWDF. This is caused by the utility function U_1 in Eq. (18) which is designed to meet the requirements of RT services. In addition, the simulation results demonstrate that SART can meet the lower delay requirements of RT services at the cost of a little higher packet loss ratio.

Similar to RT services, fixed resource is scheduled by different algorithms among FTP users. Observing the results shown in Fig. 5, we can also find that SANRT has similar performance with M-LWDF. However, different from SART, SANRT achieves lower packet loss ratio and higher packet delay than M-LWDF. As mentioned above, due to different

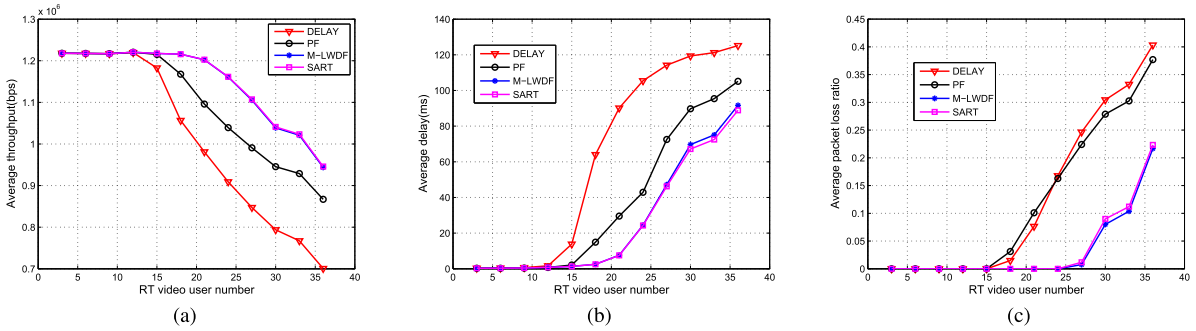


FIGURE 4. Performance comparison between different scheduling algorithms for RT video users. (a) Average throughput for RT video. (b) Average delay for RT video. (c) Average packet loss ratio for RT video.

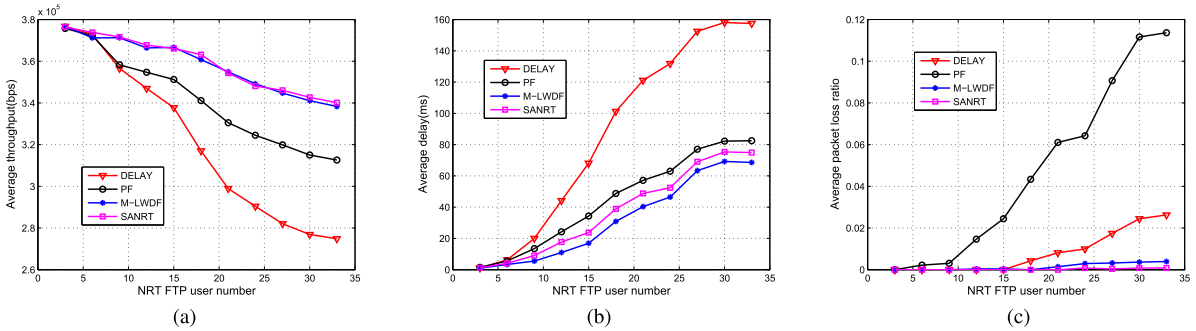


FIGURE 5. Performance comparison between different scheduling algorithms for NRT FTP users. (a) Average throughput for NRT FTP. (b) Average delay for NRT FTP. (c) Average packet loss ratio for NRT FTP.

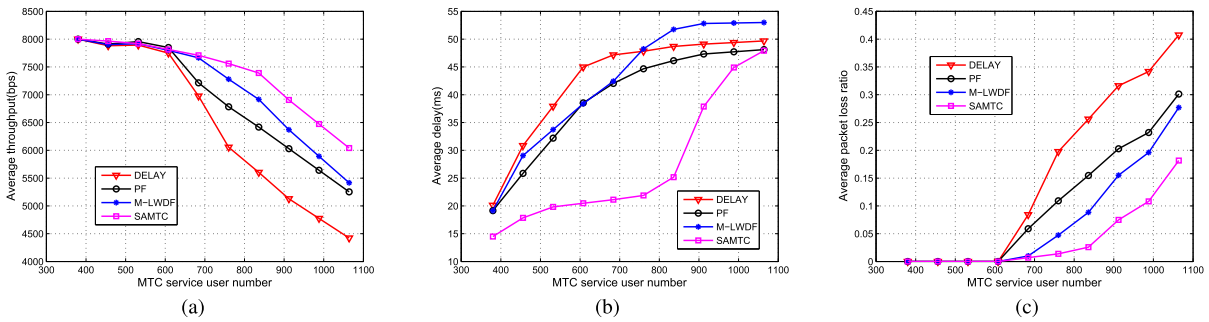


FIGURE 6. Performance comparison between different scheduling algorithms for MTC-S service users. (a) Average throughput for MTC-S. (b) Average delay for MTC-S. (c) Average packet loss ratio for MTC-S.

properties of U_1 and U_2, SART tends to schedule packets at low delay and SANRT at high delay. Moreover, in SANRT, high delay packet will be scheduled first which leads to a low packet loss ratio. Therefore, balancing the tradeoff between delay and packet loss ratio, SART prefers low delay and SANRT prefers low packet loss ratio which can satisfy different QoS requirements of RT and NRT services. In addition, parameter a in U_1 and b in U_2 can be adjusted to achieve the desired balance between delay and packet loss ratio.

To evaluate the performance of the proposed SAMTC, we compare it with DELAY, PF and M-LWDF algorithms by applying them to schedule MTC-S users with fixed number of RBs. As shown in Fig. 6, we can clearly see that SAMTC outperforms all other algorithms in all metrics. As DELAY, PF and M-LWDF just schedule one user in each RB, they cannot achieve the scheduling of all MTC-S service users which are in a large number. Moreover, packets of MTC-S users are small and the interval between two packets are long. Therefore, one MTC-S user does not have enough traffic to transmit and the resources would be wasted under the schedule of these three algorithms. On the contrary, the proposed SAMTC schedules multiple users in finer granularity than RB which avoids the two problems mentioned above. In the simulation, we schedule 4 users in each RB. Actually, finer scheduling granularity can be used to achieve better QoS performance and schedule more users.

To know the source of performance advantage of SAMTC, we also compare it with SART and SANRT which use

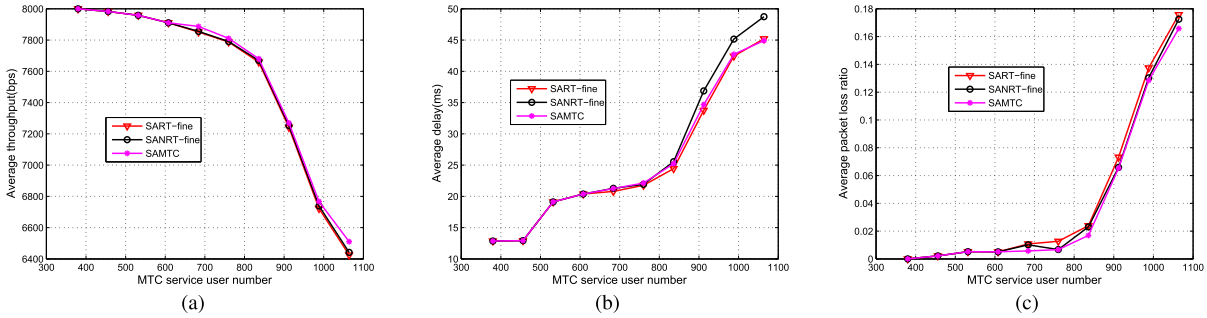


FIGURE 7. Performance comparison between different scheduling algorithms in finer granularity for MTC-S service users . (a) Average throughput for MTC-S. (b) Average delay for MTC-S. (c) Average packet loss ratio for MTC-S.

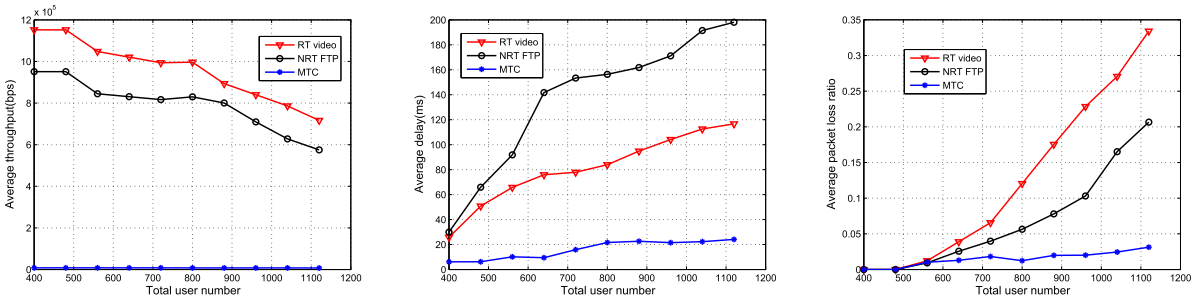


FIGURE 8. Performance of different services scheduled by proposed scheme. (a) Average throughput for multi-service. (b) Average delay for multi-service. (c) Average packet loss ratio for multi-service.

the same granularity as SAMTC. As shown in Fig. 7, SAMTC outperform the others slightly. This demonstrates that SAMTC is indeed the most appropriate for MTC-S service. Furthermore, we can also see that performance advantage is highly dependent on the finer granularity rather than SAMTC itself.

2) CUSTOMIZATION

To verify whether our virtualization model and resource allocation scheme satisfy the customization property, we implement simulation to get the performance of different services in metrics of average throughput, delay and packet loss ratio. As shown in Fig. 8, we schedule users of multiple services by our proposed scheme and obtain the performance curves of each type of service. It is easy to see that users of different services have totally different performance curves although they are scheduled in the same physical network. Specially, users of RT service have lower packet delay and users of NRT service have lower packet loss rate. With regard to users of MTC-S service, we schedule them in finer granularity which cannot be reflected by these curves. Based on the analysis above, we can draw a conclusion that virtual networks of different services are customized in scheduling algorithms.

3) ISOLATION

To verify whether our virtualization model and resource allocation scheme satisfy the isolation property, we change the condition of one virtual network and see whether other virtual networks will be affected. Specifically, the transmission rate requirements of RT video services are set as 4/3 and 5/3 times as the original one. The simulation results are shown in Fig. 9 and 10. Comparing Fig. 8 and Fig. 9, we can find that the whole changing trends in three metrics are similar. As we increase the transmission rate requirements of RT services, the RT video users get more resources and thus the performance of FTP and MTC-S users are worse. However, the difference is little in all three metrics. In other words, change of RT video slice has a negligible effect on the other two slices. Comparing Fig. 8 and Fig. 9, we can find that the delay performances are little worse but the packet loss ratio are much worse. 5/3 times transmission rate requirement greatly increases the load of network and leads to the performance degradation of all services. Therefore, in our resource virtualization scheme, isolation can be satisfied only when the network changes a little. To guarantee the absolute isolation, service contract between InP and SP needs to be considered, which is not the focus of our paper.

4) RESOURCE UTILIZATION

To verify whether our virtualization model and resource allocation scheme can achieve high resource utilization, we compare its performance with other traditional scheduling algorithms. As shown in Fig. 11, the proposed scheme achieves much higher resource utilization than other ones. In traditional schemes where all services are scheduled together, these scheduling algorithms are not suitable for

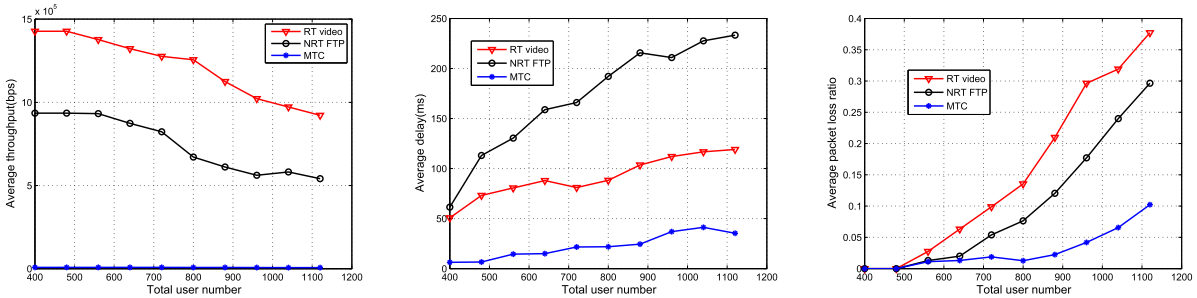


FIGURE 9. Performance of different services scheduled by proposed scheme with 4/3 times rate requirement of RT video. (a) Average throughput for multi-service. (b) Average delay for multi-service. (c) Average packet loss ratio for multi-service.

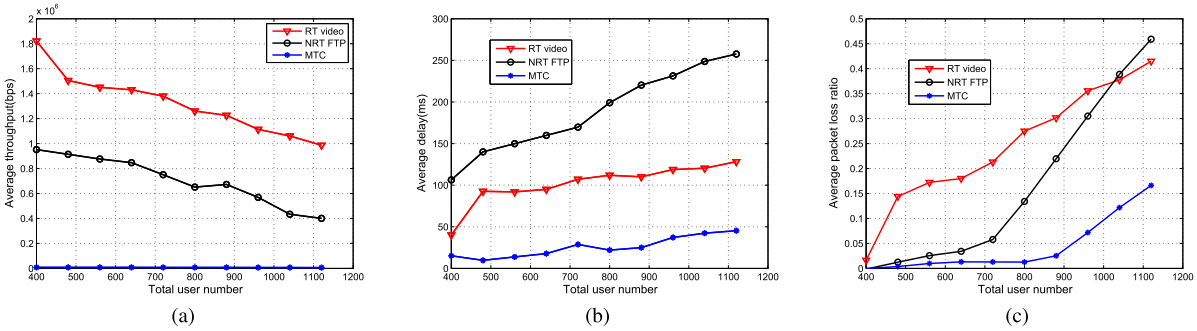


FIGURE 10. Performance of different services scheduled by proposed scheme with 5/3 times rate requirement of RT video. (a) Average throughput for multi-service. (b) Average delay for multi-service. (c) Average packet loss ratio for multi-service.

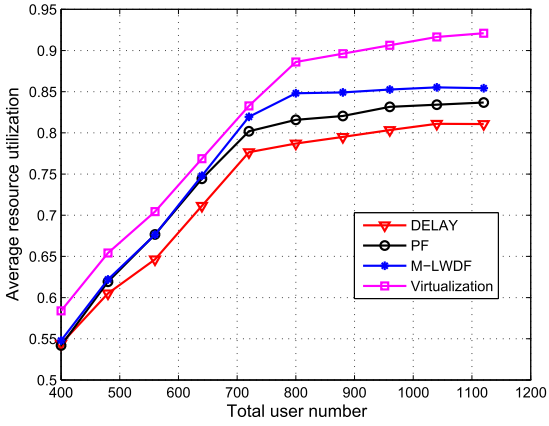


FIGURE 11. Resource utilization of different resource allocation schemes.

MTC services which have totally different traffic features with other services. Each scheduled user is allocated one whole RB which is too much for MTC user. Therefore, part of resources are wasted, leading to a low resource utilization. However, in our proposed wireless virtualization model, different types of services are scheduled separately in different customized virtual networks. In the specially designed scheduling algorithms, we used different scheduling granularities based on features of different services and thus the resources are fully utilized.

5) OVERALL PERFORMANCE ANALYSIS

Based on all the results presented above, we will analyze the overall performance of our proposed multi-service resource allocation scheme. We can see that our resource virtualization scheme can satisfy the isolation, customization, high resource utilization properties. Therefore, our proposed scheme actually realize wireless resource virtualization. As different kinds of services are served in different slices which are customized by different scheduling algorithms, their various QoS requirements can be well guaranteed. Although scheduled in the same physical network, services in different slices will not influence each other. What's more, as we choose different scheduling granularities for different services, high resource utilization can be achieved. Consequently, multi-service resource allocation problem is solved very well by the proposed scheme. It is worth noting that it is convenient to adjust resource allocation scheme when the number of services types increase in wireless virtualization. We just need to construct another virtual network and design the scheduling algorithm specially for the new type of service.

VI. CONCLUSION

In this paper, we have studied the multi-service resource allocation problem in future networks with wireless virtualization. As services will become more diverse and MTC will play an important role in future networks, traditional resource

allocation schemes are not entirely appropriate anymore. We propose a service-centric virtualization model where the network is sliced according to service types. In this model, different types of services are served in different virtual networks. Consequently, multi-service resource allocation problem is decoupled into inter-slice resource allocation and intra-slice resource scheduling. As the inter-slice resource allocation is a NP-hard problem, we design a heuristic algorithm to derive a suboptimum solution. As for intra-slice resource scheduling, we specially design the scheduling algorithms for RT, NRT and MTC-S services. Simulations are implemented to evaluate the proposed scheduling algorithms and virtualization model. Numerical results show that the proposed scheduling algorithms perform better than traditional ones in their corresponding services. Moreover, our resource allocation scheme can satisfy the isolation, customization and high resource utilization properties of virtualization. Finally, the overall performance and the performance of each kind of services show that the multi-service resource allocation problem can be well solved in our virtualization model.

With the proposed resource allocation scheme, different types of services can get appropriate spectrum resources to transmit their data. However, this is just one function of transformium core. To realize the transforming ability of RAN, the BBU and RRH management functions still need to be studied in the future work.

REFERENCES

[1] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 1617–1655, 3rd Quart., 2016.

[2] A. Osseiran *et al.*, "Scenarios for 5G mobile and wireless communications: The vision of the METIS project," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 26–35, May 2014.

[3] N. Nikaein *et al.*, "Simple traffic modeling framework for machine type communication," in *Proc. 10th Int. Symp. Wireless Commun. Syst. (ISWCS)*, Aug. 2013, pp. 1–5.

[4] C. Bockelmann *et al.*, "Massive machine-type communications in 5G: Physical and MAC-layer solutions," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 59–65, Sep. 2016.

[5] N. M. M. K. Chowdhury and R. Boutaba, "Network virtualization: State of the art and research challenges," *IEEE Commun. Mag.*, vol. 47, no. 7, pp. 20–26, Jul. 2009.

[6] 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, 1st Quart., 2015.

[7] H. Lei, L. Zhang, X. Zhang, and D. Yang, "A packet scheduling algorithm using utility function for mixed services in the downlink of OFDMA systems," in *Proc. IEEE 66th Veh. Technol. Conf.*, Sep. 2007, pp. 1664–1668.

[8] S. Ali and M. Zeeshan, "A utility based resource allocation scheme with delay scheduler for LTE service-class support," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2012, pp. 1450–1455.

[9] M. Ismail and W. Zhuang, "A distributed multi-service resource allocation algorithm in heterogeneous wireless access medium," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 2, pp. 425–432, Feb. 2012.

[10] N. U. Hassan and M. Assaad, "Dynamic resource allocation in multi-service OFDMA systems with dynamic queue control," *IEEE Trans. Commun.*, vol. 59, no. 6, pp. 1664–1674, Jun. 2011.

[11] Y. Zaki, L. Zhao, C. Goerg, and A. Timm-Giel, "LTE wireless virtualization and spectrum management," in *Proc. WMNC*, Oct. 2010, pp. 1–6.

[12] M. I. Kamel, L. B. Le, and A. Girard, "LTE wireless network virtualization: Dynamic slicing via flexible scheduling," in *Proc. IEEE Veh. Tech. Conf. (VTC)*, Sep. 2014, pp. 1–5.

[13] 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.

[14] 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. (COMSNETS)*, Jan. 2013, pp. 1–10.

[15] K. Zhu and E. Hossain, "Virtualization of 5G cellular networks as a hierarchical combinatorial auction," *IEEE Trans. Mobile Comput.*, vol. 15, no. 10, pp. 2640–2654, Oct. 2016.

[16] C. Liang and F. R. Yu, "Distributed resource allocation in virtualized wireless cellular networks based on ADMM," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Apr. 2015, pp. 360–365.

[17] T. M. Ho, N. H. Tran, S. M. A. Kazmi, and C. S. Hong, "Dynamic pricing for resource allocation in wireless network virtualization: A stackelberg game approach," in *Proc. Int. Conf. Inf. Netw. (ICOIN)*, Jan. 2017, pp. 429–434.

[18] F. Fu and U. C. Kozat, "Stochastic game for wireless network virtualization," *IEEE/ACM Trans. Netw.*, vol. 21, no. 1, pp. 84–97, Feb. 2013.

[19] S. M. A. Kazmi, N. Tran, T. Ho, and C. S. Hong, "Hierarchical matching game for service selection and resource purchasing in wireless network virtualization," *IEEE Commun. Lett.*, vol. 22, no. 1, pp. 121–124, Jan. 2018.

[20] L. Li, H. Wei, N. Deng, B. Fang, and W. Zhou, "A transforming architecture for future wireless networks: Transformium network," in *Proc. IEEE 84th Veh. Technol. Conf. (VTC-Fall)*, Sep. 2016, pp. 1–6.

[21] A. Galis and I. Chih-Lin, "Towards 5G network slicing-motivations and challenges," *IEEE 5G Tech Focus*, vol. 1, no. 1, Mar. 2017.

[22] J.-P. Cheng, C.-H. Lee, and T.-M. Lin, "Prioritized random access with dynamic access barring for RAN overload in 3GPP LTE-A networks," in *Proc. IEEE GLOBECOM Workshops (GC Wkshps)*, Dec. 2011, pp. 368–372.

[23] N. Leibowitz, B. Baum, G. Enden, and A. Karniel, "The exponential learning equation as a function of successful trials results in sigmoid performance," *J. Math. Psychol.*, vol. 54, no. 3, pp. 338–340, 2010.

[24] S. Ryu, B.-H. Ryu, H. Seo, M. Shin, and S. Park, "Wireless packet scheduling algorithm for OFDMA system based on time-utility and channel state," *Electron. Telecommun. Res. Inst. J.*, vol. 27, no. 6, pp. 777–787, 2005.

[25] J. van de Belt, H. Ahmadi, and L. E. Doyle, "Defining and surveying wireless link virtualization and wireless network virtualization," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1603–1627, 3rd Quart., 2017.

[26] S. Burer and A. N. Letchford, "Non-convex mixed-integer nonlinear programming: A survey," *Surv. Oper. Res. Manage. Sci.*, vol. 17, no. 2, pp. 97–106, 2012.

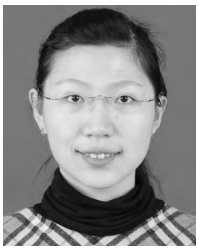
[27] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[28] D. Fooladivanda and C. Rosenberg, "Joint resource allocation and user association for heterogeneous wireless cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 1, pp. 248–257, Jan. 2013.

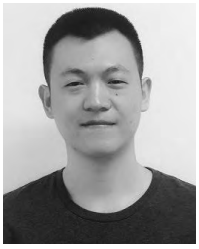
[29] I. B. Aban, M. M. Meerschaert, and A. K. Panorska, "Parameter estimation for the truncated Pareto distribution," *J. Amer. Stat. Assoc.*, vol. 101, no. 473, pp. 270–277, 2006.

[30] T. Taleb and A. Kunz, "Machine type communications in 3GPP networks: Potential, challenges, and solutions," *IEEE Commun. Mag.*, vol. 50, no. 3, pp. 178–184, Mar. 2012.

LETIAN LI received the B.S. degree in electronic engineering from the University of Science and Technology of China, Hefei, China, in 2013, where he is currently pursuing the Ph.D. degree. His research interests include wireless radio access network, wireless virtualization, radio resource management, and space information network.



NA DENG received the B.S. and Ph.D. degrees in electronic engineering from the University of Science and Technology of China, Hefei, China, in 2015 and 2010, respectively. From 2013 to 2014, she was a Visiting Student with the Prof. Martin Haenggi's Group, University of Notre Dame, Notre Dame, IN, USA. From 2015 to 2016, she was a Senior Engineer with Huawei Technologies Co., Ltd., Shanghai, China. Since 2016, she has been a Lecturer with the School of Information and Communication Engineering, Dalian University of Technology, Dalian, China. Her scientific interests include networking and wireless communications, green communications, and network design based on wireless big data.



WEILONG REN received the B.S. and Ph.D. degrees in electronics and information system from the University of Science and Technology of China (USTC), Hefei, China, in 2012 and 2017, respectively. He has been with the Personal Communication Network and Spectrum Spreading Laboratory, USTC, since 2012. He has participated in projects including Innovative Wireless Campus Experimental Networks (National Science and Technology Major Project), Research on High Frequency Networking Technologies, and Research on Transmission and Networking Technologies in Satellite Mobile Communications (National 863 Research Projects). His research interests include resource allocation, interference mitigation, and satellite communications.

BAOHUA KOU received the B.S. degree from Xi'an Jiaotong University and the Ph.D. degree from the National University of Defence Technology, Changsha, in 2007. He is currently a Senior Engineer with China Telecom Satellite Communications. His research interests include space tracking, satellite communication, and mobile communication.

WUYANG ZHOU received the B.S. and M.S. degrees from Xidian University, Xi'an, China, in 1993 and 1996, respectively, and the Ph.D. degree from the University of Science and Technology of China (USTC), Hefei, China, in 2000. He is currently a Professor with the Department of Electronic Engineering and Information Science, USTC. His research interests include mobile communication, satellite communication, and wireless networking.

SHUI YU (SM'12) is currently a Full Professor of the School of Software, University of Technology Sydney, Australia. He has published two monographs and edited two books, over 200 technical papers, including top journals and top conferences, such as IEEE TPDS, TC, TIFS, TMC, TKDE, TETC, ToN, and INFOCOM. His research interest includes security and privacy, networking, big data, and mathematical modeling. He actively serves his research communities in various roles. He initiated the research field of networking for big data in 2013. His h-index is 32. He is a member of AAAS and ACM. He is the Vice Chair of Technical Committee on Big Data of the IEEE Communication Society and a Distinguished Lecturer of the IEEE Communication Society. He has served over 70 international conferences as a member of organizing committee, such as a Publication Chair for the IEEE GLOBECOM 2015, the IEEE INFOCOM 2016 and 2017, a TPC Chair for the IEEE BigDataService 2015, and a General Chair for ACSW 2017. He is currently serving the Editorial Boards of the IEEE Communications Surveys and Tutorials, the *IEEE Communications Magazine*, the IEEE Internet of Things Journal, the IEEE Communications Letters, the IEEE Access, and the IEEE Transactions on Computational Social Systems.