IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 20, 2020, accepted June 12, 2020, date of publication June 16, 2020, date of current version June 26, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3002927

Delay-Aware Satellite-Terrestrial Backhauling for Heterogeneous Small Cell Networks

ZHE JI[1], (Member, IEEE), SUZHI CAO[2], (Member, IEEE), SHENG WU[1], (Member, IEEE), AND WENBO WANG[1], (Senior Member, IEEE)

[1]School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing 100094, China

Corresponding author: Sheng Wu (thuraya@bupt.edu.cn)

This work was supported by the Basic Scientific Research Project of Beijing University of Posts and Telecommunications under Grant 2019RC02.

ABSTRACT This paper investigates a satellite-terrestrial backhaul framework to enhance efficient data offloading for heterogeneous terminals, including delay-sensitive and delay-tolerant users. In the considered architecture, ground terminals in satellite-terrestrial small cells can access different services via the satellite-terrestrial station (STS) in each cell. The satellite offloads the requested services to corresponding STSs, and each STS provides services to terrestrial terminals via an OFDM-based downlink system. We aim to maximize the sum throughput of all small cells while integrating joint satellite backhaul power allocation and STS downlink resource allocation. The problem is firstly decomposed into two types of subproblems by decoupling the optimization of satellite backhaul capacity and downlink capacity in small cells. Then, to satisfy users' delay requirements, the downlink STS throughput is maximized over multiple slots, and we propose a two-step algorithm to schedule users during these slots. By taking advantage of a delay-violation parameter, the algorithm iteratively approaches the optimal power and subchannel solution, while guaranteeing the delay requirements. Moreover, to reduce the computational complexity, we propose a greedy-based sub-optimal scheduling algorithm where delay requirements are guaranteed by users' self-search for favorable resources, aiming at sacrificing the minimum throughput in exchange for the delay performance. Simulation results show our algorithms effectively improve the throughput performance while ensuring the delay constraints, maintaining a well-performed balance between throughput and delay performance.

INDEX TERMS Satellite-terrestrial backhauling, heterogenous services, small cells, delay awareness, low-complexity, power allocation, subchannel allocation.

I. INTRODUCTION

Playing a compelling complementary role in 5G and beyond 5G communications, satellite networks have shown a great capability to augment terrestrial services thanks to their ubiquitous coverage, data offloading, and continuous services [1]–[6]. Giambene *et al*. [3] provided an overview of an integrated satellite-5G network, and with the unbearable cost of pure terrestrial coverage, satellites will be the main approach to provide 5G services in rural and remote areas [3]–[5]. Shaat *et al*. [6] reviewed the advantages of integrated satellite-terrestrial backhaul networks and proposed link scheduling and carrier allocation strategies while the same frequency band is reused. Du *et al*. [7] considered a software-defined network (SDN) architecture and proposed an auction mechanism for spectrum sharing and traffic offloading in satellite-terrestrial networks. From the energy efficiency perspective, Ruan *et al*. [8] studied the power allocation strategies in spectrum sharing satellite-terrestrial network to maximize energy efficiency while considering the energy-spectral efficiency tradeoff.

To accommodate the increasing terrestrial multimedia traffic demands, satellite communications are expected to support heterogeneous services, including delay-sensitive and delay-tolerant services. For delay-sensitive services, the quality of service (QoS) requirements are characterized

The associate editor coordinating the review of this manuscript and approving it for publication was Angelos Antonopoulos.

by delay bounds while delay-insensitive services pursue a high throughput. To strike a proper tradeoff between system throughput and QoS requirements, efficient resource management is required for effectively matching different users with appropriate resources, such as power, subchannel, and time slots. While most existing resource allocation strategies in satellite communications focus on achieving high system throughput or improving power efficiency [9], [10], delay-aware resource allocation strategies have been investigated from different perspectives in terrestrial downlink systems [11]–[19].

Delay constraints are converted into minimum rate constraints in [11]–[13], which allows simple solutions to the optimization problem. In [11] and [12], subchannels are first allocated to delay-sensitive services under uniform power allocation, and after their minimum rate requirements are satisfied, the other subchannels will be assigned to delay-tolerant ones. In [11], the minimum rate requirement of delay-sensitive services is calculated based on the delay bound and packet arrival rate, while in [12] it is based on the deadlines and the length of all packets. In [13], instead of giving high priority to delay-sensitive services, an allocation algorithm is proposed to maximize the sum-rate of delay-tolerant services, while the rate constraints for delay-sensitive users are satisfied. Guaranteeing the alternative rate constraints makes the problem easy to solve, however, it might lead to an inefficient resource management and capacity degradation, especially when delay-sensitive users are under unfavorable channel conditions at some time slots.

Utility functions sensitive to delay constraints are exploited in [14]–[16]. The advantage of the concept of the effective capacity [17], a function of statistic delay-bound violation probabilities, is taken to satisfy delay requirements with a minimized power consumption in [14], [15]. Based on the modified largest weighted delay first (M-LWDF), subchannels are allocated based on a utility function taking the delay bound requirements of head-of-line (HOL) packets into account in [16]. Adopting utility functions concerning delay requirements is, in general, not sufficient to guarantee delay constraints. It is difficult to establish the direct relations between the effective capacity with actual delay constraints, and on the other hand, it is preferred to assign the resources considering the delay constraints of all packets other than only HOL packets.

In [18], [19], queueing theory is exploited to characterize the delay constraints of delay-sensitive users. The resource allocation is optimized while queueing theory is taken into account in modeling the queue dynamics to guarantee the average delay no larger than an upper bound. Subject to the average delay constraint, power, and subchannel allocation algorithms were proposed to maximize the total system throughput. However, the average delay constraint can not guarantee delay requirements of all packets, especially for bursty arrival packets.

In this paper, we investigate a satellite-terrestrial backhaul framework to enhance efficient data offloading for heterogeneous terminals, including delay-sensitive and delay-tolerant users. In our architecture, users are grouped into small cells according to their locations, and each cell is equipped with a satellite-terrestrial station (STS). Instead of direct satellite-terminal transmission, the satellite offloads the requested services to corresponding STS, and each STS provides services to terminals or users in its cell via an OFDM-based downlink system. Accordingly, the satellite backhaul capacity and STS downlink capacity in small cells are coupled. In other words, user scheduling and delay-aware resource allocation of all small cells are coupled since downlink capacity in small cells is upper bounded by the varying satellite backhaul capacity, which is determined by the allocated satellite power. On the other hand, resource allocation will, in turn, influence satellite power allocation to satisfy the QoS requirements of delay-sensitive services. To solve this problem, we formulate an optimization problem and aim to maximize the system throughput while integrating joint satellite backhaul power allocation and downlink resource allocation in small cells.

Different from the resource allocation in a traditional terrestrial multi-cell network, satellite backhauling leads to a large propagation delay. Without the continuous power supply, satellite resources are more critical constrained than terrestrial networks. Not all services can be delivered through satellite backhauling since the delay bounds may be expired due to the large propagation delay; moreover, a more efficient delay-aware resource allocation strategy is pursued for satellite-terrestrial data offloading while some alternative transformations of delay requirements, as mentioned above, might lead to inefficient resource consumption and capacity degradation at some slots. Therefore, instead of converting the delay requirements to inequivalent alternatives during each slot, delay constraints in our model are characterized by actual delay bounds of arriving packets, and the packet scheduling is carried out over multiple slots before their deadlines are expired.

Besides, energy efficiency is a significant concern for 5G networks, and Mesodiakaki *et al*. [20] proposed a user association strategy in a backhaul small-cell network by jointly maximizing network energy efficiency as well as spectrum efficiency, while an insightful analysis on the solution is given. Powered by solar panels, improving energy efficiency as well as balancing energy-spectral efficiency tradeoff is also important for satellite-terrestrial networks, as shown in [8]. In this paper, we focus on the throughput maximization, and energy efficiency will be considered in our following research.

The contribution of our paper are described as follows:

- We propose a scheme for data offloading in a satellite-terrestrial backhaul network where heterogeneous terminals in small cells are supported via a satellite-terrestrial station in each cell. To effectively deliver required services, a joint optimization problem is formulated to maximize the sum throughput of all small cells while integrating joint satellite backhaul power allocation and

STS downlink resource allocation. The problem is firstly decomposed into two types of subproblems by decoupling the optimization of satellite backhaul capacity and downlink capacity in small cells.

- For STS downlinks, a power and subchannel allocation problem is derived to maximize the system throughput over multiple time slots, subject to the delay constraints of delay-sensitive services. Based on a time-sharing factor, the problem is transformed into a continuous programming problem. Then, we propose a two-step algorithm to allocate resources and schedule terminals during multiple slots with low complexity by utilizing a delay-violation parameter. The system throughput is first optimized without considering the delay constraints, and the optimal allocation solution under delay constraints is then iteratively approached according to the update of the delay-violation parameter.
- To make the problem more tractable, we propose a low-complexity greedy-based sub-optimal algorithm to solve the user scheduling problem. After optimizing the system throughput without delay constraints, we formulate a scheduling problem where user scheduling is carried out to minimize the throughput loss while the delay constraints of delay-sensitive users are considered. To solve the problem, a greedy-based sub-optimal scheduling algorithm is proposed where subchannels are reallocated by users' self-search for favorable ones.

The rest of this paper is organized as follows. In Section II, we describe the system model and problem statement. Section III contains the proposed optimal allocation algorithm, and the sub-optimal algorithm is given in Section IV. The computational complexity is discussed in Section V. Simulation results are presented in Section VI. Finally, we draw a conclusion in Section VII.

II. SYSTEM MODEL AND PROBLEM STATEMENT

In this section, we introduce the satellite-terrestrial backhaul network where services are delivered to heterogeneous terminals in small cells via satellite-terrestrial stations. Then, we formulate an optimization problem to maximize the overall throughput in all small cells while integrating joint satellite backhaul power allocation and downlink resource allocation in small cells.

A. SCENARIO DESCRIPTION

Consider a satellite-terrestrial backhaul network, as shown in Fig. 1, where a satellite, K small cells within the satellite coverage, and M terrestrial terminals (interchangeably used with 'users' or 'user terminals') in each small cell. Each cell is equipped with a satellite-terrestrial station (STS), and the satellite offloads the requested services to STSs, which will deliver these services to terrestrial terminals via an OFDM-based downlink system. In our model, the terrestrial terminals or users to be served locate in remote sparsely-populated areas, such as islands, grasslands, mountainous regions, and satellite backhauling is the main approach to

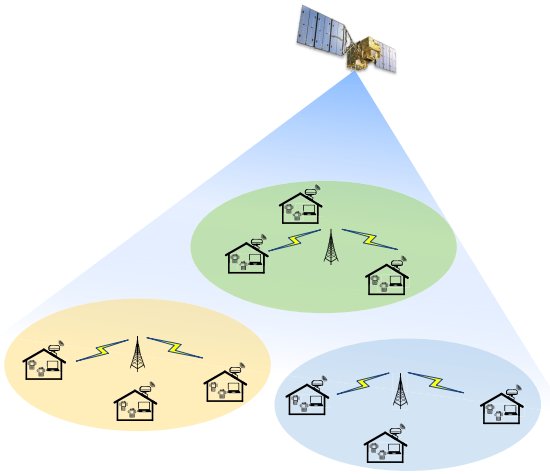

FIGURE 1. A satellite-terrestrial backhaul network.

provide multimedia services. As shown in Fig. 1, the satellite-terrestrial station in each cell transmits the required data to corresponding terrestrial terminals, which are fixed at houses and act as access points for mobile phones, computers, etc. The satellite downlink bandwidth is divided into K equivalent subbands, while each cell is allocated one band with bandwidth B_s. All small cells share the satellite transmit power P_{sat}, and the power on each cell will be determined according to their required services and channel conditions.

In each small cell, there are M_1 delay-sensitive users and $M_2 = M - M_1$ delay-tolerant users, denoted by $\mathcal{M}_1$ and $\mathcal{M}_2$, distributed randomly in each small cell. The STS in each cell transmits the traffic to users via an OFDM-based downlink system. The STS downlink systems in different small cells share the same spectrum. Since small cells locate in remote sparsely-populated areas, it is more likely two small cells locate with a certain distance between them, and we assume the interference between them can be neglected. The system is partitioned into frames of L time slots and the bandwidth of each cell B_{cell} Hz is divided into N orthogonal subchannels with bandwidth $B = B_{\text{cell}}/N$ Hz. Moreover, the power allocation in each cell will be optimized under the STS downlink maximum power constraint P_{cell}.

B. TRANSMISSION MODEL FOR STS DOWNLINKS

With varying subchannel and slot scheduling, the resource allocation for STS downlink can be represented by the binary assignment variables $\mathcal{S}_k = \{s^k_{ijm}\}_{T\times N\times M}$ in which $s^k_{ijm} = 1$ indicates subchannel j at time slot i (defined as resource block (i, j)) in small cell k is allocated to user m during an allocation and $s^k_{ijm} = 0$ otherwise. The power allocation for STS downlink can be denoted as $\mathcal{P}_k = \{p^k_{ijm}\}_{T\times N\times M}$ where p^k_{ijm} is the power from STS in small cell k allocated to its user m by the block (i, j).

We assume subchannels between STS and user terminals are independent and frequency selective [13], [18], and the channel state information (CSI) is available at the STS.

The fading coefficients of these subchannels are assumed to remain unchanged during a frame, which is a reasonable assumption for motionless user terminals since the coherence time of the channel fading is more than one second [21]. Let γ_{ijm}^k denote the channel power gain to noise power ratio in block (i, j) for user m in small cell k. The number of bits that user m in small cell k can transmit with block (i, j), can be given as

$$r_{ijm}^k = B\log_2\left(1 + P_{ijm}^k \gamma_{ijm}^k\right). \tag{1}$$

Given block allocation $\mathcal{S}_k$ and power allocation $\mathcal{P}_k$, the number of bits user m can transmit from the beginning slot to slot t in one allocation can be obtained as

$$v_{k,m}^t(\mathcal{S}_k, \mathcal{P}_k) = \sum_{i=1}^{t}\sum_{j=1}^{N} r_{ijm}^k s_{ijm}^k. \tag{2}$$

Then, the sum amount of bits the STS can transmit during an allocation can be given as

$$\sum_{m=1}^{M} v_{k,m}^T(\mathcal{S}_k, \mathcal{P}_k) = \sum_{m=1}^{M}\sum_{i=1}^{T}\sum_{j=1}^{N} r_{ijm}^k s_{ijm}^k. \tag{3}$$

C. TRANSMISSION MODEL FOR SATELLITE DOWNLINKS

The satellite operates at Ka band and let g_k denote the channel power gain between the satellite and small cell k. G_k is determined by a Welbull distribution based channel model [22]. By changing the downlink power allocated to small cells, the satellite backhaul capacity can be optimized to accommodate users' requirements in different small cells. We assume the power allocation at the satellite remains unchanged during an allocation, and thus during an allocation, the capacity of the satellite backhaul for cell k at each slot is given as

$$C_k = B_s\log_2\left(1 + P_s^k g_k/\sigma^2\right), \tag{4}$$

where σ^2 is the AWGN power at the STS receiver, and $\mathcal{P}_s = [P_s^{k,i}]_{T\times K}$

To coordinate the STS downlink capacity with the satellite backhaul capacity, the downlink capacity of each small cell is constrained to be smaller than the backhaul capacity in practice. Accordingly, their coupling is constructed as

$$\sum_{m=1}^{M}\sum_{j=1}^{N} r_{ijm}^k s_{ijm}^k \le C_k, \quad \forall m, k. \tag{5}$$

D. DELAY REQUIREMENTS FOR DELAY-SENSITIVE SERVICES

For each user terminal, there is a queue at the satellite to store the fixed sized arriving packets. Delay-sensitive services requested by terminal m in small cell k is characterized by the delay bound D_m^k, and once a packet of terminal m arrived at satellite, it is required to be delivered to the terminal in D_m^k. The delay from the satellite to terminals consists of the queueing delay at the satellite, the propagation delay from the satellite to the corresponding STS, and the queueing delay in the STS, where the propagation delay from the satellite to STSs is unified as T_d (the propagation delay from the STS to users can be neglected). Accordingly, the queueing delay should be less than $W_m^k = \lfloor D_m^k/T_s\rfloor - \lceil T_d/T_s\rceil$ slots, where T_s is the duration of a time slot.

To ensure delay constraints for delay-sensitive packets and maximize the downlink throughput in small cells, it needs the cooperation of the satellite and STSs. Firstly, the downlink allocation in small cells should be determined based on channel state information of each user terminal and the queue state information at the satellite; secondly, the packets preferred by each cell hope to be scheduled for backhaul transmission by the satellite. Therefore, the resource allocation decision in the integrated network calls for a central unit, which is the satellite in our model. The STS in each cell will send the collected channel state information to the satellite, and the satellite will determine backhaul power allocation, packet scheduling, and resource allocation in small cells. Once the STS received the packets from the satellite, it will forward them to corresponding terminals immediately as scheduled. Hence, we can assume no queueing delay in STSs with coordinated satellite and STSs. In addition, since the satellite-terrestrial station and user terminals are motionless, the coherence time of the link between them is large enough, such that the trip time to the satellite is relatively small. Accordingly, the satellite can receive accurate downlink CSI in small cells for backhaul power allocation and packet scheduling.

To ensure the delay constraints of delay-sensitive traffic which randomly arrived, the time interval T of the user scheduling should be no larger than $min\{W_m^k - 1\}, m \in \mathcal{M}_1$ slots. Before each allocation, the delay requirements of packets in delay-sensitive users' queues need to be determined. Let $\mathcal{Q}_{W_m^k-1}^t$ denote the set of packets arrived $W_m^k - 1$ slots earlier than slot t and still wait for transmission to destination m in cell k. In other words, packets in $\mathcal{Q}_{W_m^k-1}^t$ are required to be transmitted before $t + 1$. Following our idea to guarantee the delay constraints in [23], during an allocation in STS downlinks, the accumulated bits achieved by one user from slot 1 to any slot t, which is $v_{m,k}^t(\mathcal{S}_k, \mathcal{P}_k)$, should be no smaller than the bits in $\bigcup_{i=1}^{t} \mathcal{Q}_{W_m^k-1}^i$, given as

$$v_{k,m}^t(\mathcal{S}_k, \mathcal{P}_k) \ge \sum_{i=1}^{t} q_{W_m^k-1}^i, \quad \forall t, m \in \mathcal{M}_1, \tag{6}$$

where $q_{W_m^k-1}^t$ denote the number of bits in $\mathcal{Q}_{W_m^k-1}^t$ and (6) shows the delay constraints of delay-sensitive packets.

E. PROBLEM FORMULATION

Given the set of feasible resource allocation $\mathcal{S}_k$ and power allocation $\mathcal{P}_k$ for STS downlinks, we aim to determine an

allocation to maximize the sum throughput in all small cells during an allocation, subject to the power and delay constraints in small cells. Moreover, considering the coupled satellite backhaul capacity and the downlink capacity of each small cell as in (5), the optimization problem can be formulated as

$$\max \sum_{k=1}^{K}\sum_{m=1}^{M} v_{k,m}^{T}(\mathcal{S}_k, \mathcal{P}_k) \tag{7}$$

$$s.t. \sum_{m=1}^{M}\sum_{j=1}^{N} P_{ijm}^{k} s_{ijm}^{k} \le P_{\text{cell}}, \quad \forall i, k, \tag{8}$$

$$\sum_{m=1}^{M} s_{ijm}^{k} \le 1, \quad \forall i, j, k, s_{ijm}^{k} \in \{0, 1\}, \tag{9}$$

$$v_{k,m}^{T}(\mathcal{S}_k, \mathcal{P}_k) \le Q_m^k, \quad \forall m, k, \tag{10}$$

$$v_{k,m}^{t}(\mathcal{S}_k, \mathcal{P}_k) \ge \sum_{i=1}^{t} q_{W_m^k-1}^{i}, \forall t, k, m \in \mathcal{M}_1, \tag{11}$$

$$\sum_{k=1}^{K} P_s^k \le P_{\text{sat}}, \tag{12}$$

$$\sum_{m=1}^{M}\sum_{j=1}^{N} r_{ijm}^{k} s_{ijm}^{k} \le C_k, \quad \forall m, i, k, \tag{13}$$

where Q_m^k is the queue length at the satellite for user terminal m in cell k at the beginning of an allocation. (8) and (12) ensure the downlink power of each STS and the satellite power are upperbounded respectively. (9) states that one block (i, j) can only be assigned to at most one user. (10) shows that the overall bits one user can transmit during an allocation cannot be larger than its queue length, and (11) guarantees the delay requirements for relay-sensitive services. Finally, (13) shows the STS downlink traffic is restricted to be accommodated by the satellite backhaul capacity.

III. DELAY-AWARE RESOURCE ALLOCATION AND BACKHAULING

To solve the mixed combinatorial and non-convex optimization problem formulated above, we decomposed it into two types of subproblems by decoupling the optimization of satellite backhaul capacity and STS downlink capacity. Then, by taking advantage of a delay-violation parameter, the algorithm iteratively approaches the optimal power and subchannel solution, while satisfying the delay constraints.

A. DECOUPLING OF SATELLITE BACKHAULING AND STS DOWNLINKS

Inspired by [24], we first derive the Lagrangian function according to problem (7) and constraint (13), given as

$$\begin{aligned} &L(\mathcal{S}_k, \mathcal{P}_k, \mathcal{P}_S, \boldsymbol{\lambda}) \\ &= \sum_{k=1}^{K}\sum_{m=1}^{M} v_{k,m}^{T}(\mathcal{S}_k, \mathcal{P}_k) \\ &\quad + \sum_{t=1}^{T}\sum_{k=1}^{K} \lambda_k \left(C_k - \sum_{m=1}^{M}\sum_{j=1}^{N} r_{ijm}^{k} s_{ijm}^{k} \right) \\ &= \sum_{k=1}^{K} (1-\lambda_k) \sum_{m=1}^{M} v_{m,k}^{T}(\mathcal{S}_k, \mathcal{P}_k) + \sum_{k=1}^{K} \lambda_k C_k T, \end{aligned} \tag{14}$$

where λ_k is the Lagrangian multiplier with constraint (13). Accordingly, Lagrangian dual problem can be given by

$$\min_{\boldsymbol{\lambda}\ge 0} g(\boldsymbol{\lambda}) = \min_{\boldsymbol{\lambda}\ge 0} \max_{\mathcal{S}_k, \mathcal{P}_k, \mathcal{P}_S} L(\mathcal{S}_k, \mathcal{P}_k, \mathcal{P}_S, \boldsymbol{\lambda}) \tag{15}$$

For given $\boldsymbol{\lambda}$, the Lagrangian function (14) can be divided into two types of subproblems, including STS downlink allocation problem in each cell k as

$$\begin{aligned} \text{STSP}_k: \quad &\max \quad (1-\lambda_k) \sum_{m=1}^{M} v_{m,k}^{T}(\mathcal{S}_k, \mathcal{P}_k) \\ &s.t. \quad (8), (9), (10), (11), \end{aligned} \tag{16}$$

and the satellite backhauling problem as

$$\begin{aligned} \text{SatP}: \quad &\max \quad \sum_{k=1}^{K} \lambda_k C_k T \\ &s.t. \ (12), \end{aligned} \tag{17}$$

The optimization process consists of multiple iterations and in each iteration τ, with the fixed $\boldsymbol{\lambda}$, the STSP_k problem and SatP problem can be addressed independently. Then update λ_k according to $\lambda_k(\tau+1) = [\lambda_k(\tau) - \eta(\tau)\Delta(\tau)]^+$ where $\eta(\tau) > 0$ are proper step-sizes,

$$\Delta(\tau) = C_k - \max_{0<i<T} \left\{ \sum_{m=1}^{M}\sum_{j=1}^{N} r_{ijm}^{k} s_{ijm}^{k} \right\},$$

and $[x]^+$ means $\max(0, x)$. Compared to the STSP_k problem in each cell, the SatP problem is easier to deal with and thus, we concentrate on the solution of the STSP_k problem.

B. LAGRANGIAN DECOMPOSITION FOR STS DOWNLINKS

In order to make the STSP_k problem (16) more tractable, we first relax the integer constraint on $s_{ijm}^k \in \{0, 1\}$ to a time sharing factor $s_{ijm}^k \in [0, 1]$, which was proposed in [25] and indicates the portion of time the block (i, j) is allocated to user m. By introducing the factor, the mixed integer programming problem is converted into a continuous problem. Moreover, we define $p_{ijm}^k = P_{ijm}^k s_{ijm}^k$ for all i, j and m. With the help of the sharing factor, p_{ijm}^k becomes the actual power allocated to

user m. As a result, constraint (8), (9) and equation (1) become

$$\sum_{m=1}^{M}\sum_{j=1}^{N} p_{ijm}^{k} \le P_{\text{cell}}, \quad \forall i,k, \tag{18}$$

$$\sum_{m=1}^{M} s_{ijm}^{k} = 1, \quad \forall i,j,k, s_{ijm}^{k} \in [0,1], \tag{19}$$

$$r_{ijm}^{k} = B\log_2\left(1+\frac{p_{ijm}^{k}\gamma_{ijm}^{k}}{s_{ijm}^{k}}\right). \tag{20}$$

Based on (16), (18)-(20), the Lagrangian function of STTP$_k$ problem is derived as

$$\begin{aligned} &L\left(\mathcal{S}_k,\mathcal{P}_k,\boldsymbol{\alpha}_k,\boldsymbol{\beta}_k\right)\\ &=(1-\lambda_k)\sum_{m=1}^{M} v_{k,m}^{T}(\mathcal{S}_k,\mathcal{P}_k)+\sum_{m=1}^{M}\sum_{t=1}^{T}\alpha_{k,m}^{t}\left(v_{m,k}^{t}(\mathcal{S}_k,\mathcal{P}_k)\right.\\ &\left.-\sum_{i=1}^{t} q_{W_m^k-1}^{i}\right)+\sum_{i=1}^{T}\beta_k^i\left(P_{cell}-\sum_{m=1}^{M}\sum_{j=1}^{N} p_{ijm}^{k}\right), \end{aligned} \tag{21}$$

where β_k^i, and $\alpha_{k,m}^t$ are Lagrangian multipliers. The constraint (10) is considered in the algorithm proposed later, is omitted in the Lagrangian function. Then the Lagrangian dual function can be given by

$$D(\boldsymbol{\alpha}_k,\boldsymbol{\beta}_k)=\begin{cases}\max\limits_{\mathcal{S}_k,\mathcal{P}_k} & L\left(\mathcal{S}_k,\mathcal{P}_k,\boldsymbol{\alpha}_k,\boldsymbol{\beta}_k\right)\\ s.t. & \sum\limits_{m=1}^{M} s_{ijm}^{k}=1,\forall i,j,k,s_{ijm}^{k}\in[0,1]\end{cases} \tag{22}$$

And the corresponding dual optimization problem is

$$\min_{\boldsymbol{\alpha}_k,\boldsymbol{\beta}_k\ge 0} D(\boldsymbol{\alpha}_k,\boldsymbol{\beta}_k). \tag{23}$$

It can be easily proved that the objective function in (16) is concave and the constraint (11) is convex, which means the STSP$_k$ problem satisfies the time-sharing condition according to [25]. Therefore, it can be guaranteed that the STSP$_k$ problem and the dual problem (23) have the same solution and the duality gap is zero. Since the objective function in dual problem (23) is convex, there exists a globally optimal solution and the subgradient method can be utilized to minimize $D(\boldsymbol{\alpha}_k,\boldsymbol{\beta}_k)$.

From (21) we can observe that there are two subproblems needed to be solved: optimal power allocation and resource block allocation. Let p_{ijm}^{k*} and s_{ijm}^{k*} denote the expected optimal power allocation and block allocation solution in small cell k. Before applying the Karush-Kuhn-Tucker (KKT) optimality conditions, we first differentiate $v_{k,m}^t(\mathcal{S}_k,\mathcal{P}_k)$ with respect to p_{ijm}^k and can obtain

$$\frac{\partial v_{k,m}^{t}(\mathcal{S}_k,\mathcal{P}_k)}{\partial p_{ijm}^{k}}=\begin{cases}s_{ijm}^{k}\dfrac{\partial r_{ijm}^{k}}{\partial p_{ijm}^{k}}, & t\ge i\\ 0, & t<i\end{cases} \tag{24}$$

where $\frac{\partial r_{ijm}^{k}}{\partial p_{ijm}^{k}}=\frac{B\gamma_{ijm}^{k}}{\left(s_{ijm}^{k}+p_{ijm}^{k}\gamma_{ijm}^{k}\right)\ln 2}$. Thus, differentiating the Lagrangian in (21) with respect to p_{ijm}^k, we obtain

$$\begin{aligned}&\frac{\partial L\left(\mathcal{S}_k,\mathcal{P}_k,\boldsymbol{\alpha}_k,\boldsymbol{\beta}_k\right)}{\partial p_{ijm}}\\ &=\left(1-\lambda_k+\sum_{t=i}^{T}\alpha_{k,m}^{t}\right)s_{ijm}^{k}\frac{\partial r_{ijm}^{k}}{\partial p_{ijm}^{k}}-\beta_k^i.\end{aligned} \tag{25}$$

(25) clearly shows that given a resource block assignment, the optimal power computation for a fixed user requires not only the Lagrange multiplier $\alpha_{k,m}^t$ for the slot i, but also the ones for the following slots. This is because that the delay constraint (11) for terminal m at slot t lower bounds the sum amount of bits to be transmitted for the terminal from the beginning slot to current slot t, and accordingly, for slots $i\le t$ the optimal power solution is correlated with the delay constraints of slot t. As a consequence, to achieve the optimal power at a slot concerns both the Lagrange multiplier for the current slot and that for the following slots, which leads to an increase in complexity. Moreover, to update $M\times T$ Lagrange multipliers $\alpha_{k,m}^t$ brings in a great computational complexity. Hence, to reduce the computational complexity, we take advantage of a delay-violation parameter to optimize the problem.

C. DELAY-VIOLATION PARAMETER

To increase the efficiency of convergence, we exploit a delay-violation parameter ξ_m to derive a low-complexity algorithm of the problem. To explain the parameter, we first define $\delta_{k,m}^i=\sum_{t=i}^{T}\alpha_{k,m}^t$. Note that $\delta_{k,m}^i$ decreases with the increase of i, namely $\delta_{k,m}^1\ge\delta_{k,m}^2\ge\ldots\ge\delta_{k,m}^T$, and $\delta_{k,m}^i=\delta_{k,m}^{i+1}+\alpha_{k,m}^i$ for all slots. Before the delay constraint (11) for slot i is satisfied, $\alpha_{k,m}^i$ will increase at every iteration; hence, $\beta_{k,m}^i$ will not reach a stable value before $\beta_{k,m}^{i+1}$ reaches its optimum.

Let $x_{W_m^k-1}^i$ denotes, after an iteration, the number of bits left in $\mathcal{Q}_{W_m^k-1}^i$, which are supposed to be transmitted before slot $i+1$. When $\sum_{i=1}^{T} x_{W_m^k-1}^i=0$, the delay bounds of all packets are satisfied and $\boldsymbol{\alpha}_k$ reaches its optimum. At each iteration, $\alpha_{k,m}^t$ can be updated by the following subgradient method as

$$\Delta\alpha_{k,m}^{t}=v_{k,m}^{t}(s_{ijm}^{k*},p_{ijm}^{k*})-\sum_{i=1}^{t} q_{W_m^k-1}^{i}, \tag{26}$$

aiming at achieving more resources for the bits in $\mathcal{U}_{k,m}^t=\bigcup_{i=1}^{t}\mathcal{Q}_{W_m^k-1}^i$ and decreasing $\sum_{i=1}^{t} x_{W_m^k-1}^i$. The closer α_m^t approaches the optimum, the less bits $\mathcal{U}_{k,m}^t$ holds and the smaller $\sum_{i=1}^{t} x_{W_m^k-1}^i$ becomes.[1] Consequently, considering

[1]Note that $\Delta\alpha_{k,m}^i=0$ can not state that there are no bits left in $\mathcal{U}$. Only when all the $\{\alpha_{k,m}^t, t=1\ldots i\}$ reach their optimums, the bits in $\mathcal{U}$ will be transmitted before their delay bounds are expired.

$\delta^i_{k,m} = \sum_{t=i}^{T} \alpha^t_{k,m}$, the update of $\delta^i_{k,m}$ for any slot involves $\sum_{i=1}^{T} x^i_{W^k_m-1}$. Furthermore, the resources allocated to a user at a slot t can not contribute to transmit the bits in $\mathcal{U}^{t-1}_{k,m}$, where the delay bound is violated before slot t. Hence, when the bits in $\bigcup_{i=t}^{T} \mathcal{Q}^i_{W^k_m-1}$ are all scheduled while satisfying the delay constraints, which means that $\sum_{i=t}^{T} x^i_{W^k_m-1} = 0$, there is no need to update $\delta^i_{k,m}$ for slots $i = t \ldots T$.

Based on the analysis above, we substitute for $\delta^i_{k,m}$ by only one variable $\xi_{k,m}$, given as the delay-violation parameter, and its subgradient is

$$\Delta\xi_{k,m} = \sum_{i=1}^{T} x^i_{W^k_m-1}. \tag{27}$$

$\Delta\xi_{k,m}$ represents the amount of bits still waiting in queues while their delay bounds are expired. The concept of delay-violation parameter is introduced in our previous work [23], and in this paper we develop the concept to accommodate a more complicated scenario with further analysis. With the update of $\xi_{k,m}$ at each iteration, if the bits of $\bigcup_{i=t}^{T} \mathcal{Q}^i_{W^k_m-1}$ are all scheduled for transmission after the allocation, which means $\sum_{i=t}^{T} x^i_{W^k_m-1} = 0$, $\xi_{k,m}$ for the slots $i \geq t$ will not be updated in following iterations. Such updating algorithm coincides with two specialties of $\delta^i_{k,m}$: one is the decreasing specialty when i increases; the other is that $\delta^i_{k,m}$ reaches its optimum later than the following slots.

D. OPTIMAL POWER AND RESOURCE BLOCK ALLOCATION FOR STS DOWNLINK

Applying the KKT optimality conditions [26], we obtain the necessary and sufficient conditions for the power allocation p^{k*}_{ijm} as

$$\frac{\partial L\left(\mathcal{S}_k, \mathcal{P}_k, \boldsymbol{\alpha}_k, \boldsymbol{\beta}_k\right)}{\partial p^{k*}_{ijm}} = 0, \tag{28}$$

Based on the dual problem (23) and the optimality condition, we first study the optimal power allocation with a given resource block assignment. Let s^k_{ijm} be any given block assignment scheme. Substituting $\xi_{k,m}$ with $\delta^i_{k,m}$ in (25) and making it to zero according to the KKT condition (28), the optimal power allocation is given as

$$P^{k*}_{ijm} = \frac{p^{k*}_{ijm}}{s^k_{ijm}} = \left(\frac{1-\lambda_k+\xi_{k,m}}{\beta^i_k \ln 2} - \frac{1}{\gamma^k_{ijm}}\right)^+, \tag{29}$$

where $(x)^+$ means $\max(0, x)$. In our proposed policy the water level $\frac{1-\lambda_k+\xi_{k,m}}{\beta^i_k \ln 2}$ is sensitive to the delay constraints, which differs from the classical water-filling policy [27],

After achieving the optimal power, we derive the optimal resource block assignment. By substituting the optimal power allocation P^{k*}_{ijm} into (22), we can obtain

$$D(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k) = \begin{cases} \max\limits_{\mathcal{S}_k, \mathcal{P}_k} & \sum\limits_{m=1}^{M}\sum\limits_{i=1}^{T}\sum\limits_{j=1}^{N} G^k_{ijm}(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k) s^k_{ijm} - y_k \\ s.t. & \sum\limits_{m=1}^{M} s^k_{ijm} = 1, \forall i, j, k, s^k_{ijm} \in [0, 1] \end{cases} \tag{30}$$

where the function $G^k_{ijm}(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k)$ is given by

$$\begin{aligned} G^k_{ijm} &= (1 - \lambda_k + \sum_{t=i}^{T} \alpha^t_{k,m}) \log_2\left(1 + P^{k*}_{ijm}\gamma^k_{ijm}\right) - \beta^i_k P^{k*}_{ijm} \\ &= (1 - \lambda_k + \xi_{k,m}) \log_2\left(1 + P^{k*}_{ijm}\gamma^k_{ijm}\right) - \beta^i_k P^{k*}_{ijm}, \end{aligned} \tag{31}$$

and y_k is

$$y_k = \sum_{m=1}^{M}\sum_{t=1}^{T} \alpha^t_{k,m} \sum_{i=1}^{t} q^i_{W^k_m-1} + \sum_{i=1}^{T} \beta^i_k P_{cell}. \tag{32}$$

From (31) it can be observed that, with fixed resource block (i, j), each terminal m will obtain a corresponding rate given as the first term of G^k_{ijm}, and the power consumption given as the second term. Therefore, the terminal with maximum G^k_{ijm} is preferred to use the resource block and by choosing the terminal with the maximum G^k_{ijm} for each block, the objection function in (30) will be maximized. Accordingly, the optimal resource block assignment s^{k*}_{ijm} is given by

$$s^{k*}_{ijm} = \begin{cases} 1, & m^* = \arg\max\limits_{m} G^k_{ijm}, \\ 0, & \text{otherwise}, \end{cases} \tag{33}$$

Taking the partial derivative of G^k_{ijm} with respect to $\xi_{k,m}$ and γ^k_{ijm}, we find that G^k_{ijm} is a monotonous increasing function in $\xi_{k,m}$ and γ^k_{ijm}. Hence, for a resource block, the terminal with better channel condition is preferred to occupy the subchannel at this slot. And the increase of $\xi_{k,m}$ will bring more chances for the terminal to access more resources. Note that with the power and block allocation approaching the optimal solution iteratively, $x^i_{W^k_m-1}$ experiences a wide range of variation. As a result, we need to find a proper time to start updating $\xi_{k,m}$ and we hope that from that time on, $x^i_{W^k_m-1}$ can represent the actual delay requirements of delay-sensitive users.

Based on the above analysis, in the next subsection, we propose a two-step algorithm to find the optimal power and block allocation to maximize the STS downlink throughput under delay constraints. Meanwhile, according to III-A, the solution to satellite-terrestrial data offloading will be given by iteratively updating the λ_k.

E. DELAY-AWARE MECHANISM FOR SATELLITE-TERRESTRIAL DATA OFFLOADING

In this subsection, we first propose a two-step algorithm to achieve the maximum throughput for STS downlink while guaranteeing the delay requirements. Aiming at maximizing

the throughput, we first optimize the power and block allocation without considering delay constraints, which means $\xi_{k,m} = 0$. Once when the maximum throughput is achieved, we will initiate updating $\xi_{k,m}$ according to $\sum_{i=1}^{T} x^i_{W^k_m - 1}$ and increase $\xi_{k,m}$ with the decreasing $\sum_{i=1}^{T} x^i_{W^k_m - 1}$ in following iterations as described before. Meanwhile, the optimal block and power allocation will be determined based on relation (29) and (33).

The proposed algorithm is described in detail as follows.

1. **STS downlink throughput optimal solution without delay constraints.**

(a) Initialize the Lagrange multipliers $\boldsymbol{\beta}_k(0)$, $\xi_{k,m} = 0$.

(b) Given $\boldsymbol{\beta}_k$, $\forall m \in \mathcal{M}$, $Q'_{k,m} = Q_{k,m}$. For slot $i = 1 : T$, allocate subchannels one by one: for each subchannel $j = 1$ to N

1) Calculate P^k_{ijm} for each terminal according to (29). Solve optimal s^k_{ijm} via (33) and find m^* such that $s_{ijm^{k*}} = 1$.

2) Calculate $r_{ijm^{k*}}$. To satisfy the constraint (10) and avoid resource over-allocation, ensure $r_{ijm^{k*}} \le Q'_{k,m}$ and otherwise, $r_{ijm^{k*}} = Q'_{k,m}$. Update $Q'_{k,m^*} = Q'_{k,m^*} - r^k_{ijm^*}$.

(c) Check whether the power constraints (8) for all slots are satisfied or not. If satisfied, go to step 2; otherwise, update dual variables $\boldsymbol{\beta}_k$ according to the following relation.

$$\beta^i_k(t+1) = \left[\beta^i_k(t) + \Gamma(t)\left(\sum_{m=1}^{M}\sum_{j=1}^{N} P^k_{ijm}x^k_{ijm} - P_{\text{cell}}\right)\right]^+ \tag{34}$$

where t is the iteration index, $\Gamma(t) > 0$ are proper step-sizes, and $[x]^+$ means $\max(0, x)$.[2]

2. **Resource allocation while updating the delay-bound parameter.**

(a) For each delay-sensitive terminal, find the slot t after which the bits in $\mathcal{Q}^i_{W^k_m - 1}$ are all scheduled for transmission, which indicates after slot t the delay constraints of terminal m's packets are all satisfied. t can be determined by

$$t_{k,m} = \arg\max_{1 \le i \le T}\left\{\sum_{i=t}^{T} x^i_{W^k_m - 1} > 0\right\}. \tag{35}$$

At the slots $i > t_{k,m}$, $\xi_{k,m} = 0$ while in the slots $i \le t_{k,m}$, $\xi_{k,m}$ is updated the same as (34) with the subgradient given in (27).

(b) Given $\boldsymbol{\beta}_k$ and $\xi_{k,m}$, $\forall m \in \mathcal{M}$, $Q'_{k,m} = Q_{k,m}$. Achieve the resource allocation as step 1(b).

(c) If the power constraints (8) for all slots are satisfied and $\sum_{i=1}^{T} x^i_{W^k_m - 1} = 0$ for all terminals, the algorithm is done and the optimal solution is achieved; otherwise, update dual variables $\boldsymbol{\beta}_k$ according to (34) and return to step 2.

[2]The dual variables obtained by the subgradient method is guaranteed to converge to the optimal solution when the step sizes are chosen properly [25].

Note that step 1(b) guarantees the constraint (10), which is not under consideration in the Lagrangian function (21) for simplicity. Hence, the resource allocated to a terminal is no more than the amount of the bits in its queue.

Our optimization problem for STS downlink aims to maximize the throughput, however, to ensure the delay constraints of delay-sensitive terminals results in a sacrifice of throughput. When the throughput reaches the maximum without delay constraints, the amount of unsatisfied bits, $\sum_{i=1}^{T} x^i_{W^k_m - 1}$, indicates the resources needed to be reallocated to delay-sensitive terminals with a throughput loss. As a result, we pick the time when the throughput reaches the maximum to initialize the update of $\xi_{k,m}$ and the procedure to satisfy delay constraints.

Furthermore, following the power allocation procedure in STS downlink, the solution to SatP problem (12) can be easily obtained as

$$P^{k*}_s = \left(\frac{\lambda_k T}{\eta_k} - \frac{\sigma^2}{G_k}\right)^+, \tag{36}$$

where η_k is the Lagrange multiplier with constraint (12). Accordingly, the satellite-terrestrial data offloading allocation can be started with initiating λ_k; then, with fixed dual variable λ_k, the optimal solution to SatP problem and STSP$_k$ problem can be obtained respectively; finally, the optimal solution to the integrated resource allocation problem can be achieved by iteratively updating the dual variable λ_k till the backhaul capacity constraint (13) is satisfied. Although introducing $\xi_{k,m}$ leads to a decrease of complexity, there are still a large number of iterations. To reduce the computational complexity, a sub-optimal scheduling algorithm is proposed in the next section.

IV. LOW-COMPLEXITY SUB-OPTIMAL ALGORITHM FOR USER SCHEDULING

In the previous section, an optimal solution to the STSP$_k$ problem is derived by jointly allocating power and block with a computational burden. In this section, to reduce the computational complexity the iteration procedure brings about, a low-complexity greedy-based sub-optimal scheduling algorithm is proposed. In the algorithm, the delay constraints of delay-sensitive users are satisfied by scheduling users instead of jointly optimizing power and block allocation via iterations. In the first step, the power and block allocation solution is obtained without delay consideration, as in the previous section. In the second step, we fix the power allocated to blocks, and only user assignments will be changed in the scheduling process. Subject to the delay constraints of delay-sensitive users, we formulate a user scheduling problem aiming to minimize the throughput loss compared to the solution in the first step. To solve the problem, we propose a greedy scheduling algorithm where users search for favorable blocks by themselves.

Based on the analysis in section III, the optimal block allocation while ignoring the delay requirements can be

achieved by

$$m^* = \arg\max_m \left\{ \left(\log_2 \left(\frac{\gamma_{ijm}^k}{\beta_k^i \ln 2} \right) \right)^+ - \frac{\gamma_{ijm}^k - \beta_k^i \ln 2}{\gamma_{ijm}^k \ln 2} \right\}. \tag{37}$$

The corresponding optimal power allocation policy is given by

$$P_{ijm}^{k*} = \left(\frac{1}{\beta_k^i} - \frac{1}{\gamma_{ijm}^k} \right)^+. \tag{38}$$

Relation (38) shows that given a β_k^i, the optimal power allocation only depends on the channel condition, following the classical water-filling policy. The user with favorable channel condition is preferred to occupy the subchannel.

After an iterative search for the dual variable β_k^i, we can achieve power allocation P_{ijm}^{k*}, the user assignment s_{ijm}^{k*} with the optimal user m^* and the corresponding throughput $r_{ijm^*}^k$ with each block (i, j). Based on the solution, we propose a greedy scheduling algorithm to satisfy the delay constraints with the least sacrifice of throughput.

A. GREEDY USER SCHEDULING ALGORITHM FOR MINIMIZING THE THROUGHPUT LOSS UNDER DELAY CONSTRAINTS

In this subsection, based on the allocation policy obtained above, we propose a user scheduling algorithm for heterogeneous users in small cells. In the algorithm, delay-sensitive user terminals are scheduled to occupy some blocks already allocated to NRT terminals until the delay requirements of all RT terminals are satisfied. The principle of the scheduling algorithm is to sacrifice the minimum throughput in exchange for the satisfaction of the delay requirements.

It is to be noted here that the optimal power P_{ijm}^{k*} associated with a resource block (i, j) in small cell k remains unchanged while the user assignment with this block might be changed, and hence, we use P_{ij}^{k*} to denote the power allocated to the block instead of P_{ijm}^{k*}. Let U_{ijm}^k denote the throughput RT terminals can achieve with each block, and U_{ijm}^k can be given as $U_{ijm}^k = B\log_2\left(1 + P_{ij}^k \gamma_{ijm}^k\right)$. Define

$$\Delta U_{ijm}^k = r_{ijm^*}^k - U_{ijm}^k, \tag{39}$$

as the throughput gap with the block (i, j) between the throughput achieved by delay-sensitive terminal m and the throughput achieved by the optimal terminal m^* in the previous subsection. In addition, the resource blocks allocated to RT terminals in the previous subsection will not be reallocated in this scheduling algorithm. Hence, let $z_{ij}^k = 1$ denote that the block (i, j) is allocated to a delay-tolerant terminal and $z_{ij}^k = 0$ denote the opposite.

Based on the analysis above, we formulate the following optimization problem to find the block allocation.

$$\min \sum_{m \in M_1} \sum_{i=1}^{T} \sum_{j=1}^{N} \Delta U_{ijm}^k \delta_{ijm}^k z_{ij}^k \tag{40}$$

$$s.t. \sum_{m \in M_1} \delta_{ijm}^k \le 1, \quad \forall i, j, \delta_{ijm}^k \in \{0, 1\}, \tag{41}$$

$$\sum_{i=1}^{t} \sum_{j=1}^{N} \Delta U_{ijm}^k \delta_{ijm} z_{ij}^k \ge \sum_{i=1}^{t} x_{W_m^k - 1}^i, \quad \forall t, m, k \in \mathcal{M}_1, \tag{42}$$

where the binary assignment variable δ_{ijm} indicates whether block (i, j) is reallocated to delay-sensitive terminal m in the scheduling algorithm. (42) describes the delay requirements for delay-sensitive traffic with the help of $x_{W_m^k - 1}^i$, which is defined in section III and to be initialized after the allocation in the previous subsection. (41) implies that a given resource block (i, j) cannot be shared by more than one terminal.

Note that not all the blocks already allocated to delay-tolerant terminals will participate in the reallocation, since rescheduling some of them might be enough to satisfy the requirements of delay-sensitive terminals. It is hard to determine which of these blocks will be reallocated to delay-sensitive terminals. Hence, the approach adopted in section III, where an optimal user terminal is selected for a fixed block, is no longer applicable. To solve the problem (40)-(42), we propose a greedy scheduling algorithm as follows.

In order to minimize the throughput loss after user scheduling, we hope to achieve more throughput improvement for delay-sensitive terminals by reallocating the least number of blocks. As a result, for each delay-sensitive terminal m with the block (i, j) in small cell k, we define a weight as

$$w_{ijm}^k = \frac{\Delta U_{ijm}^k}{U_{ijm}^k}. \tag{43}$$

The smaller the weight w_{ijm}^k is, the block is the more favorable to the terminal. Define the weight matrix $\mathbf{W}_m^k = \left[w_{ijm}^k \right]$ for delay-sensitive terminal m where the (i, j) element is the weight with the block (i, j).

Based on the weight matrix $\mathbf{W}_m^k$, we propose a algorithm to optimize the problem (40)-(42). After the allocation in the previous subsection, if the deadlines of some packets in delay-sensitive terminal m's queue are expired at slot t, which is $x_{W_m^k - 1}^t > 0$, the user terminal will find its favorable blocks from the beginning slot to slot t to achieve more resources until the delay requirement is satisfied or there are no more available blocks. Accordingly, during the delay-sensitive terminals' self-search, different terminals might choose to occupy the same block. Hence, after all delay-sensitive terminals' searching, if there are such blocks that have been selected by more than one delay-sensitive terminal,

to minimize the throughput loss, the terminal with the least ΔU_{ijm}^k will be chosen to occupy the block and the other users will find alternative blocks.

The proposed algorithm is described in detail as follows.

1. Initialization.

(a) Based on the previous subsection, obtain z_{ij}^k with each block (i, j), and $\mathbf{W}_m^k$ for each delay-sensitive terminal.

(b) Let U_{im}^k denote the sum throughput terminal m obtains in this scheduling for slot i and set $U_{im}^k = 0$.

(c) Define a set A_{ij}^k such that if terminal m in cell k chooses to select the block (i, j) during the scheduling, add the terminal to the set A_{ij}^k as $A_{ij}^k = A_{ij}^k + \{m\}$.

2. Delay-sensitive terminals' greedy self-search for favorable blocks according to $\mathbf{W}_m^k$.

For time slot t, check whether the delay constraints of each delay-sensitive terminal in cell k is satisfied or not: for terminal $m = 1 : M_1$ in cell k, where M_1 is number of delay-sensitive terminals in cell k.

(a) If $x_{W_m^k-1}^t > 0$, the terminal will find the minimum weight of $\mathbf{W}_m^k$ from the first row to the tth row except the zero ones and determine its corresponding block (x, y) as follows.

$$(x, y) = \arg \min_{1 \le i \le t} \mathbf{W}_m^k.$$

If the minimum weight is zero, there are not available blocks for terminal m and return to step 2 for next terminal until $m = M_1$. If otherwise, go to next step.

(b) Update $A_{xy}^k = A_{xy}^k + \{m\}$, $\delta_{ijm}^k = 1$ and set the corresponding element of $\mathbf{W}_m^k$ to 0. Set $U_{tm}^k = U_{tm}^k + U_{xym}^k$. If $U_{tm}^k \ge x_{W_m^k-1}^t$, return to step 2 for next terminal until $m = M_1$ and otherwise, return to step 2(a).

3. Reschedule terminals when a block is occupied by more than one delay-sensitive terminals.

For each block $\{(i, j), 1 \le i \le t\}$ with $z_{ij}^k = 1$, find whether the block is allocated to more than one delay-sensitive terminal.

(a) If the number of elements in A_{ij}^k is more than one, go to step 3(b); otherwise, go to step 3(e).

(b) Compare the throughput loss ΔU_{ijm}^k of terminals in A_{ij}^k and find the terminal m^* with the least ΔU_{ijm} to occupy the subchannel. Thus, define $B_{ij}^k = A_{ij}^k - \{m^*\}$ to denote the terminals who need to be reallocated. For each terminal m in B_{ij}^k, $U_{tm}^k = U_{tm}^k - U_{ijm}^k$.

(c) For each terminal m in B_{ij}^k, find the minimum weight of $\mathbf{W}_m^k$ and determine (x, y) as in step 2(a). If the minimum weight is zero, return to step 3(c) for next terminal; otherwise, go to next step.

(d) If A_{xy}^k is not empty, set the corresponding element of $\mathbf{W}_m$ to 0 and repeat 3(c) for the terminal. If A_{xy}^k is empty, $U_{tm}^k = U_{tm}^k + U_{xym}^k$.

(e) Go to step 3(a) until all blocks have been checked by step 3.

If $t < T$, set $t = t + 1$ and repeat step 2-3; otherwise, the scheduling is done.

The operation of the proposed scheduling algorithm is carried out from the first to the last time slot in an allocation. For slot t, delay-sensitive terminals with delay-bound violations in each cell will find their favorable blocks from the beginning slot to slot t to achieve their target throughput until there are no more available blocks. Note that during the procedure, different terminals don't care whether the block has been selected by other delay-sensitive terminals during the scheduling. After all delay-sensitive terminals' searching, if a block is occupied by more than one delay-sensitive terminals, the terminal with the least throughout loss will be chosen to take the block, and each of the other terminals will find an alternative block.

V. COMPLEXITY ANALYSIS

The computational complexity of the proposed optimal algorithm is determined by the complexity of solving the dual problem (15) and the dual problems corresponding to SatP problem and STSP_k problem. Firstly, the complexity of the proposed STS downlink allocation algorithm for each cell consists of two parts. The first part lies in the number of iterations to obtain the throughput-optimal solution. MNT computations are needed to obtain the optimal user assignment for all blocks at each iteration. The subgradient method is exploited to update the Lagrange multipliers, and its computational complexity is polynomial in the number of dual variables [25]. Since updating $\boldsymbol{\beta}_k$ is executed for each slot independently, its complexity is $\mathcal{O}(1)$. Hence, the first part has a complexity of $\mathcal{O}(MNT)$. The other part is the complexity of iterations needed to satisfy the delay constraints. Since to guarantee the delay requirements of a delay-sensitive user only needs to update its delay-violation parameter, the complexity is $\mathcal{O}(M^2NT)$. Accordingly, the complexity of the proposed algorithm is $\mathcal{O}(M^2NT)$. Secondly, K computations are needed to obtain the optimal backhaul power allocation, and the complexity to satisfy the total power constraint by updating the dual variable η_k is $\mathcal{O}(1)$. Hence, the overall complexity to achieve the backhaul power allocation is $\mathcal{O}(K)$. Finally, the complexity to solve the dual problem (15) by updating λ_k is $\mathcal{O}(K)$. Therefore, the complexity of the whole integrated scheme is $\mathcal{O}(K^2M^2NT)$.

For the proposed sub-optimal algorithm, the complexity also consists of three parts. Differently, the complexity of the downlink allocation in small cells has changed. Besides the complexity of $\mathcal{O}(MNT)$ to achieve the maximum throughput without delay constraints, the additional complexity of the greedy-based algorithm is not significant compared to the optimal algorithm, because: (a) there are not iterations in the second step of the sub-optimal algorithm; (b) if $x_{W_m^k-1}^t > 0$, delay-sensitive terminal m only has to search for the blocks allocated to delay-tolerant terminals and the blocks already allocated to delay-sensitive terminals, are out of the terminal's searching list. Moreover, for a delay-sensitive terminal, the bits with delay bound violation exist at some slots rather than all slots, which further narrows the search. Therefore, the complexity of greedy-based searching

cannot be larger than $\mathcal{O}(MNT)$, and hence, the complexity of downlink resource allocation in small cells is $\mathcal{O}(MNT)$. Accordingly, the overall complexity of the sub-optimal algorithm has an upper bound $\mathcal{O}(K^2MNT)$.

VI. SIMULATION RESULTS

In this section, the performance of the proposed satellite-terrestrial backhauling algorithms are presented. The satellite works at 16 GHz with backhauling bandwidth 35 MHz, which is divided into 5 channels each allocated to one small cell and the total available satellite power for the 5 small cells is 40 W. The path loss of the satellite backhauling can be modeled as $20\log(4\pi df)$ where d is the distance between the satellite and the STS in each cell, and f is the backhaul frequency. The downlink in small cells operates at 6 GHz and each STS downlink bandwidth is 5 MHz with the number of physical resource blocks (PRB) equal to 25 and each PRB has 12 adjacent subchannels of bandwidth 15 KHz. A frame lasts 20 ms and is slotted into 40 slots. The total power at the STS is 20 W. We assume user terminals are uniformly distributed in a circular area of radius 1 Km and the STS is located at the center. There are $M = 8$ terminals in each small cell with $M_1 = 4$ delay sensitive terminals and $M_2 = 4$ delay-tolerant terminals. For the downlink in small cells, we assume each user experiences an independent rayleigh fading channel and a modified COST231-Hata propagation model is utilized with path loss $128.1 + 37.6\log(R)$.

The packets arrived following an independent Poisson process with a fixed 128 bytes size. If the delay bound of a packet is expired, the packet will be dropped from the queue. Each simulation is performed for at least 10000 time slots. We compare our proposed optimal algorithm, defined as algorithm 1, and the greedy-based sub-optimal algorithm, defined as algorithm 2, with (1) the pure maximum throughput algorithm (max-throughput algorithm) (2) the MLWDF algorithm [16] (there are minimum rate constraints for delay-sensitive terminals) (3) QoS scheduling scheme [12]. In addition, although delay constraints are not taken into account in the max-throughput scheme, compared to the maximum capacity achieved by the max-throughput scheme, the throughput loss while considering the delay constraints can be clearly observed with the increase of arrival rates.

A. CASE 1 HOMOGENEOUS DELAY BOUND

In this case, the delay bounds of all delay-sensitive terminals are set to 20 ms (the propagation delay is not included), and the loss probability requirement is 5%. Fig. 2 and Fig. 3 show the overall throughput in all small cells and the loss probability under different arrival rates. It can be observed that, except for the max-throughput scheme, our proposed algorithms achieve better overall throughput than other algorithms. Compared with these algorithms, the maximum throughput improvement occurs when the arrival rate is 6.5 Mbps, and the overall throughput of all cells achieved by our proposed algorithms is 12 Mbps more than the QoS scheduling scheme and 22 Mbps more than MLWDF. Although the max-throughput

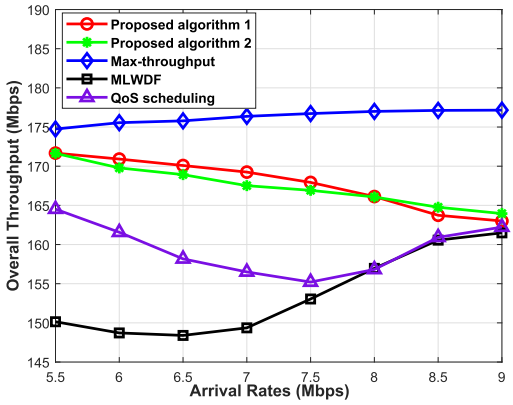

FIGURE 2. Overall throughput of different algorithms with homogeneous delay bound.

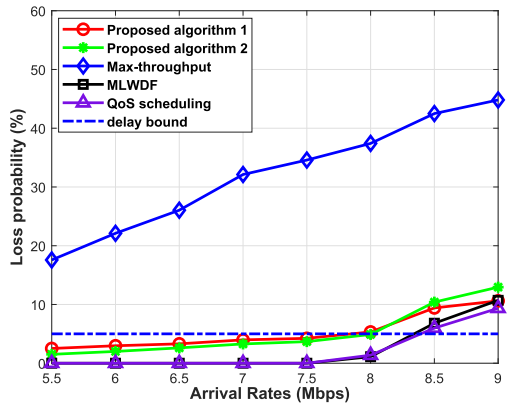

FIGURE 3. Loss probability of different algorithms with homogeneous delay bound.

algorithm achieves the maximum throughput, a large number of delay-sensitive traffic is dropped because their delay requirements are not taken into consideration, as shown in Fig. 3. It also shows that our algorithms obtain delay performance of our proposed algorithm is close to that of the MLWDF and the QoS scheduling scheme.

In Fig. 2, we can observe that the throughput of the algorithms considering the delay-sensitive traffic, including our algorithms, the MLWDF, and the QoS scheduling scheme, decreases with the increase of arrival rates. This is because, with more traffic arrived, more resources are scheduled to guarantee the delay requirements of delay-sensitive terminals, resulting in a throughput loss. Moreover, for the MLWDF and the QoS scheduling schemes, delay-sensitive terminals have the priority to occupy resources. As the arrival rate increases, all the resources will be scheduled to delay-sensitive terminals, and the ones with better channel conditions are preferred. Therefore, for these two algorithms, the throughput decreases at first and then increases with the increase of the arrival rate.

Moreover, as shown in Fig. 2 and Fig. 3, the throughput and loss probability performance of our proposed algorithms are very close with each other, and the reason is that in both

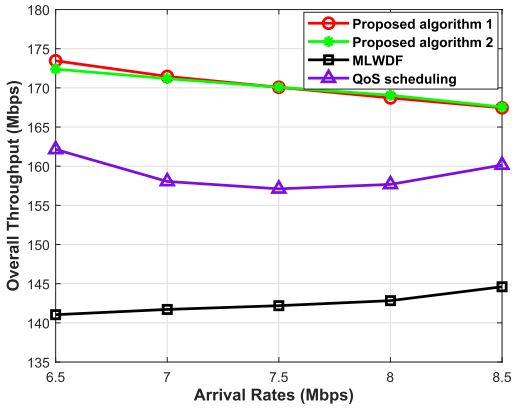

FIGURE 4. Overall throughput of different algorithms with heterogeneous delay bounds.

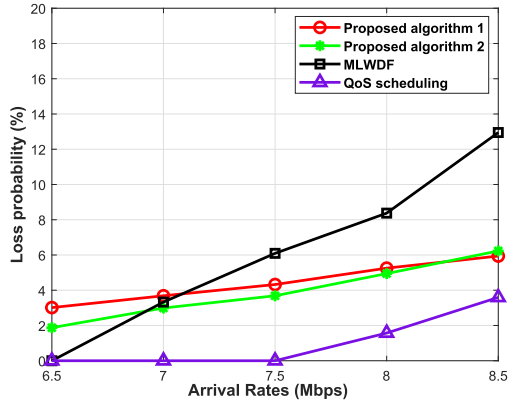

FIGURE 5. Loss probability of different algorithms with heterogeneous delay bounds.

algorithms, delay requirements are described by $x^i_{W^k_m-1}$. Note that when the arrival rate is larger than 8 Mbps, the loss probability requirement cannot be satisfied by the proposed algorithms, as shown in Fig. 3. In this situation, the sub-optimal algorithm prefers to allocate the resources to terminals which can achieve more throughput according to (43) while the optimal algorithm pursues to reduce the loss probability, which leads to a performance crossover at the arrival rate 8 Mbps as shown in Fig. 2 and Fig. 3. Accordingly, based on the close performance of the proposed algorithms, the sub-optimal algorithm would be preferred in practical cases due to the lower complexity. Fig. 2 and Fig. 3 only show the performance when packets arrive at the satellite with a relatively high rate, where the max-throughput algorithm cannot guarantee the delay requirements, and algorithms considering delay constraints need to schedule more resources to delay-sensitive terminals. And if the traffic load is low, as the arrival rate increases, the throughput achieved by all algorithms will increase, which is different from the performance achieved with a high arrival rate. This is because, with a low rate, it does not need a large number of resources to ensure the delay requirements of delay-sensitive traffic, and resources can be allocated to terminals under good channel conditions, while a high throughput is pursued.

B. CASE 2 HETEROGENEOUS DELAY BOUNDS

In this case, the delay bounds of different delay-sensitive terminals in each cell are set to 20 ms, 40 ms, 60 ms, and 80 ms. In Fig. 4 and Fig. 5, our proposed algorithms achieve better throughput performance than the other two algorithms, as in case 1. Compared to case 1 with the same delay bound, the delay requirements are relaxed, and both the proposed two algorithms pursue the high throughput, which leads to the close performance of the proposed algorithms, as shown in Fig. 4, including the crossover at 8.5 Mbps. When there are various delay bounds among terminals, the performance of the MLWDF is getting worse than that in case 1. This is because the scheduling principle of this algorithm is to allocate more resources to terminals that have waited a longer time than others without considering packets' deadlines.

VII. CONCLUSION

In this paper, we have presented a satellite-terrestrial backhauling mechanism where ground terminals in satellite-terrestrial small cells can access different services via the satellite-terrestrial station (STS) in each cell. Under the delay constraints of delay-sensitive services, we aim to maximize the sum throughput in all small cells while the downlink capacity in each small cell is bounded by the varying satellite backhaul capacity. We first decouple the satellite power allocation and STS downlink resource allocation into several respective problems. Then, a two-step allocation algorithm is proposed considering the deadlines of all delay-sensitive traffic packets by utilizing a delay-violation parameter, and the optimal power and subchannel allocation is approached. At last, a low-complexity greedy-based sub-optimal algorithm is proposed while delay requirements are guaranteed by terminals' self-search for favorable resources, aiming at sacrificing the minimum throughput in exchange for the delay performance. Simulation results show that our algorithms effectively strike a well-performed balance between throughput and delay performance.

REFERENCES

[1] X. Yan, K. An, T. Liang, G. Zheng, Z. Ding, S. Chatzinotas, and Y. Liu, "The application of power-domain non-orthogonal multiple access in satellite communication networks," *IEEE Access*, vol. 7, pp. 63531–63539, 2019.

[2] V. Bankey, P. K. Upadhyay, D. B. Da Costa, P. S. Bithas, A. G. Kanatas, and U. S. Dias, "Performance analysis of multi-antenna multiuser hybrid satellite-terrestrial relay systems for mobile services delivery," *IEEE Access*, vol. 6, pp. 24729–24745, 2018.

[3] G. Giambene, S. Kota, and P. Pillai, "Satellite-5G integration: A network perspective," *IEEE Netw.*, vol. 32, no. 5, pp. 25–31, Sep. 2018.

[4] U. Siddique, H. Tabassum, E. Hossain, and D. I. Kim, "Wireless backhauling of 5G small cells: Challenges and solution approaches," *IEEE Wireless Commun.*, vol. 22, no. 5, pp. 22–31, Oct. 2015.

[5] S. Cioni, R. De Gaudenzi, O. Del Rio Herrero, and N. Girault, "On the satellite role in the era of 5G massive machine type communications," *IEEE Netw.*, vol. 32, no. 5, pp. 54–61, Sep. 2018.

[6] M. Shaat, E. Lagunas, A. I. Perez-Neira, and S. Chatzinotas, "Integrated terrestrial-satellite wireless backhauling: Resource management and benefits for 5G," *IEEE Veh. Technol. Mag.*, vol. 13, no. 3, pp. 39–47, Sep. 2018.

[7] J. Du, C. Jiang, H. Zhang, Y. Ren, and M. Guizani, "Auction design and analysis for SDN-based traffic offloading in hybrid satellite-terrestrial networks," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 10, pp. 2202–2217, Oct. 2018.

[8] Y. Ruan, Y. Li, C.-X. Wang, R. Zhang, and H. Zhang, "Power allocation in cognitive satellite-vehicular networks from energy-spectral efficiency tradeoff perspective," *IEEE Trans. Cognit. Commun. Netw.*, vol. 5, no. 2, pp. 318–329, Jun. 2019.
[9] J. Lei and M. A. Vazquez-Castro, "Joint power and carrier allocation for the multibeam satellite downlink with individual SINR constraints," in *Proc. IEEE Int. Conf. Commun.*, Cape Town, South Africa, May 2010, pp. 1–5.
[10] A. Grundinger, M. Joham, and W. Utschick, "Bounds on optimal power minimization and rate balancing in the satellite downlink," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Ottawa, ON, Canada, Jun. 2012, pp. 3600–3605.
[11] D. Zhang, X. Tao, J. Lu, and M. Wang, "Dynamic resource allocation for real-time services in cooperative OFDMA systems," *IEEE Commun. Lett.*, vol. 15, no. 5, pp. 497–499, May 2011.
[12] Y. Kim, K. Son, and S. Chong, "QoS scheduling for heterogeneous traffic in OFDMA-based wireless systems," in *Proc. GLOBECOM - IEEE Global Telecommun. Conf.*, Honolulu, HI, USA, Nov. 2009, pp. 1–6.
[13] M. Tao, Y. Liang, and F. Zhang, "Resource allocation for delay differentiated traffic in multiuser OFDM systems," *IEEE Trans. Wireless Commun.*, vol. 7, no. 6, pp. 2190–2210, Jun. 2008.
[14] H. Zhang, Y. Ma, D. Yuan, and H.-H. Chen, "Quality-of-Service driven power and sub-carrier allocation policy for vehicular communication networks," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 1, pp. 197–206, Jan. 2011.
[15] T. Abrao, L. D. H. Sampaio, S. Yang, K. T. K. Cheung, P. J. E. Jeszensky, and L. Hanzo, "Energy efficient OFDMA networks maintaining statistical QoS guarantees for delay-sensitive traffic," *IEEE Access*, vol. 4, pp. 774–791, 2016.
[16] C. Mohanram and S. Bhashyam, "Joint subcarrier and power allocation in channel-aware queue-aware scheduling for multiuser OFDM," *IEEE Trans. Wireless Commun.*, vol. 6, no. 9, pp. 3208–3213, Sep. 2007.
[17] D. Wu and R. Negi, "Effective capacity: A wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, vol. 24, no. 5, pp. 630–643, May 2003.
[18] D. Wing Hui, V. Nang Lau, and W. Lam, "Cross-layer design for OFDMA wireless systems with heterogeneous delay requirements," *IEEE Trans. Wireless Commun.*, vol. 6, no. 8, pp. 2872–2880, Aug. 2007.
[19] F. Sun, V. O. K. Li, and Z. Diao, "Joint dynamic subcarrier allocation and flow control for real-time streaming over multiuser OFDM systems," in *Proc. IEEE Global Telecommun. Conf. (IEEE GLOBECOM)*, Washington, DC, USA, Nov. 2007, pp. 3719–3723.
[20] A. Mesodiakaki, F. Adelantado, L. Alonso, M. Di Renzo, and C. Verikoukis, "Energy- and spectrum-efficient user association in millimeter-wave backhaul small-cell networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 2, pp. 1810–1821, Feb. 2017.
[21] D. Tse and P. Vishwanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.
[22] S. A. Kanellopoulos, C. I. Kourogiorgas, A. D. Panagopoulos, S. N. Livieratos, and G. E. Chatzarakis, "Channel model for satellite communication links above 10 GHz based on Weibull distribution," *IEEE Commun. Lett.*, vol. 18, no. 4, pp. 568–571, Apr. 2014.
[23] Z. Ji, Y. Wang, W. Feng, and J. Lu, "Delay-aware power and bandwidth allocation for multiuser satellite downlinks," *IEEE Commun. Lett.*, vol. 18, no. 11, pp. 1951–1954, Nov. 2014.
[24] B. Di, H. Zhang, L. Song, Y. Li, and G. Y. Li, "Ultra-dense LEO: Integrating terrestrial-satellite networks into 5G and beyond for data offloading," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 47–62, Jan. 2019.
[25] W. Yu and R. Lui, "Dual methods for nonconvex spectrum optimization of multicarrier systems," *IEEE Trans. Commun.*, vol. 54, no. 7, pp. 1310–1322, Jul. 2006.
[26] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[27] A. Goldsmith, *Wireless Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

ZHE JI (Member, IEEE) received the Ph.D. degree in electronic engineering from Tsinghua University, Beijing, China, in 2015. She is currently a Postdoctoral Researcher with the Beijing University of Posts and Telecommunications, Beijing. Her research interests include networking and resource management in wireless and satellite networks.

SUZHI CAO (Member, IEEE) received the B.S. and M.S. degrees from Tianjin University, in 2004 and 2007, respectively, and the Ph.D. degree from the Academy of Opto-Electronics, Chinese Academy of Sciences, in 2010. She is currently an Associate Researcher with the Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences. Her research interests include satellite networks, edge computing, and distributed computing.

SHENG WU (Member, IEEE) received the B.S. and M.S. degrees from the Beijing University of Posts and Telecommunications, Beijing, China, in 2004 and 2007, respectively, and the Ph.D. degree in electronic engineering from Tsinghua University, Beijing, in 2014. He was a Postdoctoral Researcher with the Tsinghua Space Center, Tsinghua University. He is currently an Associate Professor with the Beijing University of Posts and Telecommunications. He has published over 60 journal and conference papers. His research interests include mainly in iterative signal processing, massive MIMO, and satellite communications.

WENBO WANG (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees from the Beijing University of Posts and Telecommunications (BUPT), in 1986, 1989, and 1992, respectively. He is currently a Professor and the Executive Vice Dean of the Graduate School, BUPT, where he is also the Assistant Director of the Key Laboratory of Universal Wireless Communication, Ministry of Education. He has authored over 200 journal and international conference papers and six books. His current research interests include radio transmission technology, wireless network theory, and software radio technology.