Predictive Two-Timescale Resource Allocation for VoD Services in Fast Moving Scenarios

Wanting Yang, *Student Member, IEEE*, Xuefen Chi, Linlin Zhao, *Member, IEEE*, and Ruizhe Qi

***Abstract*—With the burgeoning video applications in 5G/B5G, the requirements for video frame quality and playback smoothness are equally strict but challenging to satisfy simultaneously, especially in fast moving scenarios. To tackle this issue, in this paper, we propose a predictive two-timescale resource allocation scheme for video-on-demand (VoD) services by leveraging the long-term channel prediction. The scheme addresses two critical concerns. One is how to schedule packets over a large timescale to avoid transmission in poor channel states. The other is how to guarantee the delay-quality of service (QoS) over a small timescale to satisfy the high-quality requirements of video services. In the framework of network slicing, we split the VoD slice into two logical sub-slices to support the video contents that play immediately after downloading and the video contents to be pre-cached, respectively. To perform an efficient resource reservation, we propose a martingales-based resource estimation method for the video streams with statistical delay-QoS requirements. Based on this, our scheme is divided into two stages. First, on the time scale of prediction time slots, we propose a low-complexity heuristic algorithm to find a spectrum-efficient video delivery pattern. Then, in each transmission time interval, we develop a utility theory based resource allocation algorithm to balance the metrics of spectrum efficiency, fairness, and delay-QoS. The simulation results demonstrate the capability of the QoS guarantee and the promising gain of spectrum efficiency brought by our scheme.**

***Index Terms*—Predictive resource allocation, martingales theory, VoD, spectrum efficiency, delay-QoS guarantee.**

I. Introduction

IN THE five-generation (5G) and beyond 5G (B5G) eras, many new types of video on demand (VoD) services have emerged, such as 4 K/8 K ultra-high-definition videos, high-frame-rate videos and 360°/720° degree virtual reality videos [1]. To guarantee the impressive details of these videos, a stable communication link with a high data rate has to be maintained. As we all know, the achievable transmission rate is highly dependent on the distance between the user and the base station (BS). Hence, for the user in a running vehicle, its resource demand is bound to vary over time for satisfying the requirements on the video quality level and smooth playback. Moreover, since the initial position, the driving direction and the speed of each user are independent of each other, the variations in the total resource demands become more random in a multi-user scenario. Therefore, it is challenging to satisfy all users' resource demands under limited system bandwidth in the radio access network (RAN) slicing architecture.

Fortunately, thanks to the advanced techniques of big data analysis (BDA), the large-scale channel can be predicted according to the trajectories of the users in running vehicles. By leveraging the long-term channel prediction, the time-frequency resources can be foreseeingly and strategically allocated to individual users. Therefore, the predictive resource allocation (PRA) scheme has been recognized as a promising method to improve the quality of service (QoS) and enhance spectrum/energy efficiency (SE/EE) in VoD transmissions.

In [2] and [3], the PRA has been proved efficient in reducing resource consumption for VoD services. The basic idea of PRA is to take full advantage of the good channel states to deliver more VoD contents and avoid low-SE/EE transmission in the poor channel states. The idea is clear but there are still some tricky issues to be dealt with, as described below.

- Firstly, evaluated at any moment, the channel states of different users are independent and diversified. One user is suffering from poor channel states while another may be enjoying good channel states. Unlike real-time video services, the VoD contents can be transmitted in advance, which creates big space for the optimization of resource allocation. Considering the limited resources, how to jointly design the multi-user video delivery pattern to maximize the average EE/SE of the system within a long term is still an open issue.
- Secondly, the delay-QoS requirements of the emerging VoD services have become stricter. However, the existing works about the PRA only optimized resource allocation on a large time scale, and the resource consumption was estimated roughly according to historical average data rates. Due to the high speed of vehicles, the small-scale channel fading envelope fluctuates in the millisecond scale. Thus, the channel gain in each transmission time interval (TTI) may differ from the predicted average channel gain. How to deal with the small-scale channel fading to meet strict delay-QoS requirements is still a puzzle for the realization of the PRA.

Manuscript received July 20, 2020; revised January 2, 2021, April 17, 2021, and June 23, 2021; accepted July 4, 2021. Date of publication July 9, 2021; date of current version October 15, 2021. This work was supported in part by the Natural Science Foundation of the Department of Science and Technology of Jilin Province under Grant 20180101040JC, and in part by the National Natural Science Foundation of China under Grant 61801191. This article was presented in part at the 2020 IEEE International Conference on Communications. The review of this article was coordinated by Dr. Ngoc-Dung Dao. *(Corresponding author: Xuefen Chi.)*

The authors are with the Department of Communications Engineering, Jilin University, Changchun 130012, China (e-mail: yangwt18@mails.jlu.edu.cn; chixf@jlu.edu.cn; zhaoll13@mails.jlu.edu.cn; qirz18@mails.jlu.edu.cn).

Digital Object Identifier 10.1109/TVT.2021.3095917

- Thirdly, in the PRA, some VoD contents may be delivered well in advance of its playback time. As a result, the delay deadlines of the packets involved are considerably extended. In this sense, the delay-QoS requirements of these packets become looser, which means that these packets require relatively smaller bandwidth and could contribute to higher SE through well-designed resource allocation scheme. Nevertheless, the existing works ignored the difference in the delay-QoS requirements between these two types of packets (i.e., packets delivered on time and packets delivered in advance). How to allocate resources reasonably in each TTI to guarantee diverse delay-QoS in a spectrum-efficient way remains blank.

For tackling the above challenges, we propose a predictive two-timescale resource allocation scheme. Motivated by the network slicing architecture (where the services with different QoS requirements are supported by different slices), we split the VoD slice into two logical sub-slices supporting the video contents that play immediately after downloading and video contents to be pre-cached, respectively. On this basis, our scheme can be divided into two stages. In Stage 1, a statistical-QoS-aware delivery pattern reconfiguration (DPR) algorithm is proposed based on the long-term channel prediction. It operates in a mobile edge computing (MEC) server at the beginning of the prediction window (PW) and determines how many video packets to be transmitted in a sub-slice in a prediction time slot (PTS). In Stage 2, a delay-sensitive resource allocation algorithm is proposed to complete the physical resource block (PRB) allocation for the sub-slices in each TTI by jointly considering the historical transmission information and the instantaneous channel states. The contributions of this paper are summarized as follows:

- We propose a martingales-based resource estimation method to evaluate the minimum number of PRBs required in a PTS for transmitting a specific VoD flow. In this method, the features of time-varying wireless channels and the data flow are characterized in the perspective of martingales. By constructing a supermartingale for the delay, the minimum PRB consumption is studied with considering the impact of statistical delay-QoS requirement on the scheduling decisions.
- Relying on the martingales-based resource estimation method, we formulate the large-timescale DPR algorithm as a SE maximization problem. The constraints of this problem characterize the requirements of smooth playback, high quality of each video frame, and good isolation of the VoD slice. This is a high-dimensional non-convex optimization problem. To solve it, we propose a low-complexity heuristic algorithm to find a three-dimensional delivery pattern with high SE and the guarantee of statistical delay-QoS.
- We formulate the small-timescale PRB allocation algorithm as a binary integer programming problem based on utility theory. The utility function is related to the metrics of SE, delay-QoS, and fairness. Then, we provide an iterative algorithm to solve this problem and make a PRB allocation map. By jointly considering the queue length, the delay, and the instantaneous channel states, the algorithm achieves a satisfactory overall system performance.

The paper is organized as follows. Section II introduces the related works. Section III describes the system model and the overview of our scheme. The problem formulation and the solution algorithms of the two stages are detailed in Section IV and Section V, respectively. Section VI shows the performance of our proposed scheme. Section VII concludes the paper.

II. Related Works

A. Video Packet Scheduling and Resource Allocation

In recent years, with the development of multimedia technology, the major concerns of the research about wireless video transmission evolve gradually. The authors of [4] introduced a content-aware quality index to evaluate the quality contribution made by the successful transmission of a packet. Then they formulated an optimization problem aiming to maximize the perceived video quality by a synthetic consideration of video contents, multiuser packet scheduling, and wireless resource allocation. The authors of [5] considered the stochastic nature of wireless channels and expended the evaluation of video quality to a random field which is characterized by the cumulative distribution function (CDF) of video quality. Then, they studied the power allocation algorithm to satisfy the lower bound of target video quality.

Later, with the explosively growing traffic demands for multimedia applications, QoS guarantee in video communications becomes challenging due to the scarcity of radio resources. Hence, the scheduling scheme conceived for dynamic adaptive streaming over HTTP (DASH) has drawn more and more researchers' attention. In [6], the authors proposed a video packet scheduling scheme to minimize stalling probability based on Markov decision process for DASH, in which the effect of the current resource allocation decisions on the likelihood of future stalling was considered. In [7], based on the utility theory, the authors proposed a utility-based dynamic adaptive multimedia streaming scheme. In each scheduling interval, the scheduling policy was co-determined by quality utility, power consumption utility, packet error ratio utility, and remaining battery utility.

Meanwhile, Lei Xu *et al.* resorted to the heterogeneous cognitive networks to alleviate the pressure of resource scarcity. In [8], they formulated the joint video scheduling, subchannel assignment and power allocation problem in heterogeneous cognitive networks as a mixed-integer non-linear programming problem. Then, they proposed a packet scheduling scheme based on the auction theory to maximize the video quality for each secondary users. In [9], they further improved their work based on the non-orthogonal multiple access to enhance SE. They also proposed a video packet scheduling scheme with stochastic QoS for heterogeneous cognitive networks, in which the CDF of video quality was analyzed. However, none of the above works considered the long-term channel prediction technique. As a result, the existing resource allocation scheme did not take full advantage of the optimizable space for VoD transmission. They mainly guaranteed the smooth playback at the cost of sacrificing

video quality, which may be not suitable for the emerging VoD services in 5G/B5G eras.

B. Predictive Resource Allocation

Nowadays, the BDA-based channel prediction technique provides a large space for the PRA scheme to improve QoS and SE/EE of the system with non-realtime VoD services. In [10], Hossam S. Hassanein *et al.* first proposed the predictive green wireless access (PreGWA) scheme based on the ideal prediction of achievable data rates and then demonstrated effectiveness of the PreGWA scheme in improving energy efficiency through simulations. On the basis of this PreGWA scheme, they further put forward the robust PRA algorithms under the imperfect prediction of achievable data rates relying on the chance-constrained programming in [2], [11]. Besides, the particle filter and Kalman filter were adopted to track the channel states during the prediction processes, respectively. In recent years, they investigated the PRA under the two-dimensional imperfect prediction of data rates required by users and network resources provided by the system [3], [12]. Unfortunately, the above PRA schemes failed to guarantee critical delay-QoS requirements since the randomness in the resource allocation was neglected. Furthermore, all their proposed PRA algorithms only focused on the VoD delivery pattern on the time scale of seconds and did not study the specific PRB allocation scheme. In [13], Margolies $et\ al.$ proposed a predictive finite-horizon proportional fair framework in which the user was scheduled in the time slots with the good channel states. In [14], a new cross-layer transport protocol was put forward to minimize system utilization. Both of them can only avoid the video freezing, but cannot guarantee the specific delay violation probability or the specific video stop probability.

Besides, the long-term channel prediction can also be applied to DASH to assign the quality level of each video segment for saving energy [15], [16], and improving the quality of experience [17], [18]. In [19], the information of user mobility prediction was utilized to support seamless operation in Named Data Network by storing the data needed in the caches before the producer's handover. The authors in [20], [21] and [22] all focused on the small cells. In [20], Abou-Zeid $et\ al.$ used a utility function to improve user fairness over multiple cells through the long-term resource allocation. In [21], Guo $et\ al.$ proposed a two-threshold-based algorithm that employed the cell-level coarse-grained information to achieve a spectrum-efficient resource allocation. Furthermore, in [22], they further designed a deep neural network in an end-to-end manner to predict the knowledge of two thresholds of channel gains for making a decision, which could adapt to traffic variation and user mobility.

III. System Model and Scheme Overview

A. Scenario Description

As shown in Fig. 1, we focus on a single cell with a roadside BS and multiple users using online VoD services in the vehicles moving along the roads. The set of VoD users is defined as $\mathcal{M} = \{1, 2, \ldots, i, \ldots, M\}$, where i $(1 \leq i \leq M)$ denotes the user index. Assume that the historical information about the users' driving speeds and the road traffic is stored in the MEC server deployed at RANs. Furthermore, we assume that the users' trajectories within the PW can be predicted by the MEC server with BDA technology. The PW is divided into T PTSs. The set of the T PTSs is denoted by $\mathcal{T} = \{1, 2, \ldots, t, \ldots, T\}$, where t denotes the index of PTSs. The average channel gain experienced by each user in each PTS can be obtained by mapping the predicted trajectories to the Radio Environment Map [2] [23]. Accordingly, the average achievable data rate within each PTS for user i can be calculated and is denoted by $\bar{r}_{i,t}$ [bits/s/Hz]. Hence, the two-dimensional matrix $\bar{\mathbf{R}} = (\bar{r}_{i,t} : i \in \mathcal{M}, t \in \mathcal{T})$ can be known by the MEC server at the beginning of each PW.[1]

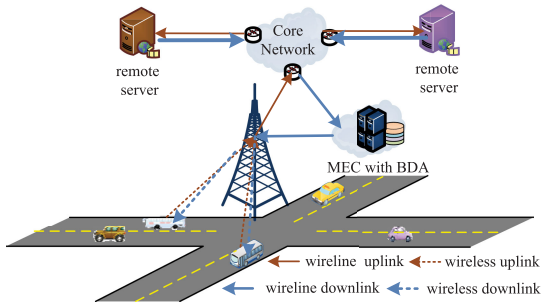


Fig. 1. Scenario description.

B. VoD Traffic Model

In the VoD communication scenario, the VoD contents stored in the remote servers is usually partitioned into small chunks containing the same number of frames [24]. Each frame can be encoded into several VoD packets with the length of L for the transmission in the network. When the user successfully receives all the packets belonging to a video frame, the video frame will be decoded and played. In this sense, the packets belonging to one frame have the same delay deadline; the delay deadline difference for any two consecutive frames equals the time interval between the two frame playbacks [9]. If the delay D of a packet exceeds the delay deadline, its corresponding video frame will be distorted or the playback of the video frame will be delayed. If there are many packets whose delay exceeds the delay deadline, the video will stutter. In this paper, the allowable delay bound of each packet in the same sub-slice is assumed to be same.[2] More details about the sub-slice will be introduced in the later sub-section. Obviously, the smaller start-up delay means the smaller allowable delay bound [25]; the precise quantitative relationship between the two is not analyzed in this paper. We assume that the wireless link is the bottleneck, and hence the waiting time to be scheduled in the BS is regarded as the

[1]The achievable data rate is related to both bandwidth and transmitting power. To facilitate the analysis, we assume that the transmitting power of BS is constant and aim to conceive a spectrum-efficient scheduling scheme.

[2]Actually, the allowable delay bounds of these packets are slightly different, as the packets corresponding to the same frame will arrive at the BS in different TTIs. However, because of the presence of start-up delay, the allowable delay bounds of packets can be approximated as the same.

TABLE I
THE MAIN SYMBOLS AND NOTATIONS

Symbols	Notations
$\bar{r}_{i,t}$	The average achievable data rate for user i in PTS t
L	The length of a data packet
$\hat{P}_{i,t}$	The number of packets involved by the video contents played in PTS t for user i
τ_{PTS}	The length of a PTS
$\Phi^{\max}$	The total number of the available PRBs in PTS t within the VoD slice
$P^q_{i,t}$	The number of packets to be transmitted to user i through sub-slice q in PTS t
$\hat{t}_i$	The PTS in which the transmission of VoD contents required by user i in the PW ends
Dm^q_i	The allowable delay bound of the packets transmitted to user i through sub-slice q
ε_i	The maximum acceptable DVP for user i
$\psi^q_{i,n}$	The indicator variable representing whether is assigned to sub-slice q of user i
$A^q_{i,t}(0,k)$	The number of cumulative arrival packets for sub-slice q of user i until TTI k in PTS t
$S^q_{i,t}(0,k)$	The number of cumulative served packets for sub-slice q of user i until TTI k in PTS t
$Q^q_{i,t}(k)$	The total number of the packets arriving and still waiting in the buffer for sub-slice q of user i until TTI k in PTS t
$U^q_{i,n}(k')$	The utility function for PRB n to be allocated to sub-slice q of user i in TTI k', where $k' = k + (t-1)\,\tau_{\mathrm{PTS}}$
$r_{i,n}(k')$	The instantaneous data rate for user i in TTI k' on PRB n
$Q^q_i(k')$	The queue length for sub-slice q of user i in TTI k'
$D^q_i(k')$	The delay of the first packet in queue of sub-slice q of user i in TTI k'

delay of the packet. Frequently used notations are summarized in Table I.

C. Scheme Overview

In our scheme, the remote servers receive the VoD user requests and deliver all of the video contents to be played in the PW to the MEC server. Assume that the MEC server knows the number of packets corresponding to the VoD contents to be played in each PTS for each user, which are constructed as $\hat{\mathbf{P}} = (\hat{P}_{i,t} : i \in \mathcal{M}, t \in \mathcal{T})$ [6]. Then, a multi-user strategic VoD delivery pattern $\mathbf{P}$ can be reconfigured from $\hat{\mathbf{P}}$ according to $\bar{\mathbf{R}}$ at the beginning of the PW. After that, the MEC server deliveries the video packets to the BS according to the delivery pattern $\mathbf{P}$. Then the BS transmits the packets received to the multiple users in the timescale of TTI according to the instantaneous channel states. In the rest of this subsection, we will introduce the motivations and analytical models of our two-timescale scheme in detail.

1) Predictive DPR: The DPR algorithm can be treated as an algorithm of resource pre-allocation in the large timescale. Its objective is to reduce the total transmission time of the VoD contents and to achieve higher SE. It is known that the bad channel state tends to cause low SE. Hence, the idea behind the DPR is to convert the transmission with strict delay-QoS requirement in the PTSs under bad channel states into the advance transmission with loose delay-QoS requirement in PTSs under good channel states as much as possible. Without loss of generality, we assume that the VoD slice is allocated a separate frequency band to ensure the isolation among the VoD and other services [26]. In addition, the total number of available PRBs within a PTS is assumed to be a constant, which is denoted by $\Phi^{\max}$. We take an example of two users to illustrate the method of DPR,[3] as shown in Fig. 3. Fig. 3(a) shows how the average achievable data rate within each PTS varies with the user's distance from the BS. Fig. 3(b) shows an example of the reconfigured delivery pattern of the DPR. Fig. 3(c) shows the illustration of the time-frequency resource consumption of the DPR. Take user 2 as an example. For taking full advantage of the better channel states, the packets that should be transmitted after PTS 34 are scheduled to be transmitted in advance in PTS 30 to PTS 33, as shown in Fig. 3(b). Consequently, the transmission can be completed before the end of the PW, as shown in Fig. 3(c).

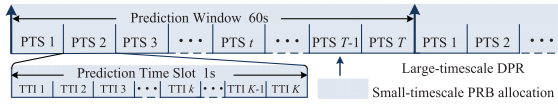


Fig. 2. Illustration of timescales.

In our problem, we denote the index of the PTS when the transmission for user i terminates as $\hat{t}_i$. The VoD packets are divided into two types: the packets related to video contents that play immediately after downloading, and the packets related to the pre-cached video contents to be played after PTS $\hat{t}_i$. In this way, the VoD slice can be naturally divided into two sub-slices which support these VoD packets with different delay requirements, and are denoted by $q = 1$ and $q = 2$ accordingly. Therefore the reconfigured delivery pattern can be constructed as a three-dimensional matrix $\mathbf{P} = (P^q_{i,t} : i \in \mathcal{M}, t \in \mathcal{T}, q \in \{1,2\})$, where $P^q_{i,t}$ represents the number of packets delivered in PTS t through sub-slice q to user i. And we denote the PRB consumption of sub-slice q of user i in PTS t by $\Phi^q_{i,t}$.

To guarantee the video frame quality, the packets in different types of frames are given with the same importance. Considering the stochastic nature of wireless channels, the delay violation probability (DVP) (i.e., $\Pr\{D^q_i > \mathit{Dm}^q_i\}$) is utilized to evaluate the video quality experienced by users qualitatively. During the large-timescale DPR, it is known that the video quality can be guaranteed by meeting statistical delay-QoS requirements, i.e., $\Pr\{D^q_i > \mathit{Dm}^q_i\} \le \varepsilon_i$, where ε_i represents the maximum user-acceptable DVP for user i.

2) Packet Scheduling and Resource Allocation: The small-timescale resource allocation can be treated as the implementation of the reconfigured delivery pattern in each TTI. In this work, we assume that each user has two logical links for the two sub-slices, each of which is modeled as a discrete-time queueing system with First-In-First-Out service discipline (as shown in Fig. 4(a)). SE, delay-QoS and fairness are the main factors in the design of an allocation policy [27]. To achieve a balance of the three metrics, the cross layer information including the instantaneous channel states in the physical layer and queue length information in the media access control (MAC) layer are all considered in the resource allocation in each TTI. Since one PW consists of T PTSs and one PTS consists of K TTIs, as shown in Fig. 2, the kth TTI within PTS t is the k'th TTI within the PW, and

[3]For ease of exposition, it is assumed that the video contents played in each PTS have the same amount of data in Fig. 3.

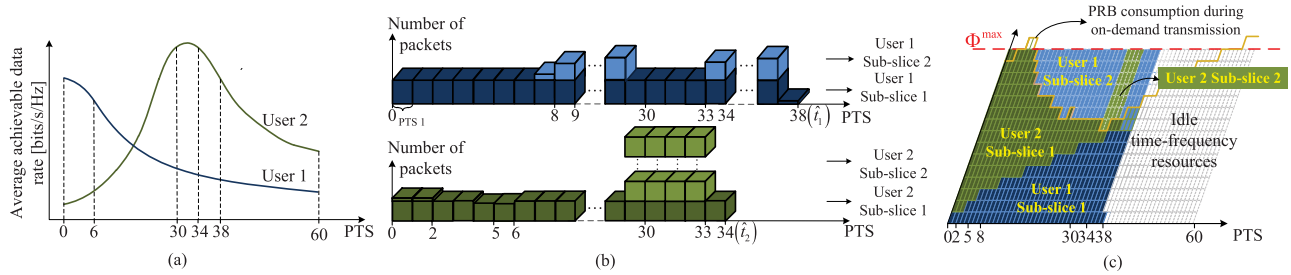


Fig. 3. The illustration of DPR with two users as examples. (a) variation trends of data rate. (b) reconfigured delivery pattern. (c) time-frequency resource consumption.

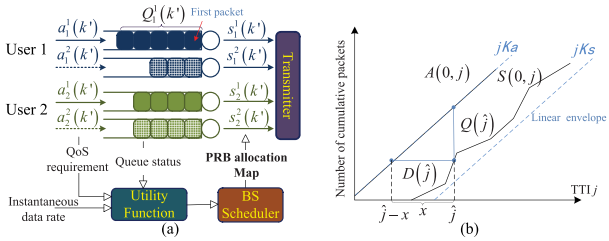


Fig. 4. Illustration of small-timescale resource allocation model and statistical delay analysis model.

$k' = k + (t-1)K$. Let $a_i^q(k')$ and $s_i^q(k')$ denote the number of packets arriving and being served in the queue of sub-slice q in TTI k', respectively. Then, the number of cumulative arrival packets is denoted $A_i^q(0,k')$, i.e., $A_i^q(0,k') = \sum_{x=0}^{k'} a_i^q(x)$. The number of cumulative served packets is denoted by $S_i^q(0,k')$, i.e., $S_i^q(0,k') = \sum_{x=0}^{k'} s_i^q(x)$. Thus, the queue length can be expressed by $Q_i^q(k') = \sup_{k'\geq 0}\{A_i^q(0,k') - S_i^q(0,k')\}$, where $\sup\{\cdot\}$ represents the supremum operator. The value of $s_i^q(k')$ is determined by the PRB allocation map $\mathbf{\Psi} = (\psi_{i,n}^q : i \in \mathcal{M}, q \in \{1,2\}, n \in \mathcal{N})$, where $\mathcal{N} = \{1,2,\ldots,N\}$ denotes the set of these available PRBs, and N represents the total number of the available PRBs in each TTI. If PRB n is assigned to sub-slice q of user i, $\psi_{i,n}^q = 1$; otherwise, $\psi_{i,n}^q = 0$. Assume that the frequency division duplex is supported and a PRB occupies 180 kHz in the frequency domain and lasts for 1 ms in the time domain.

IV. Large-Timescale Delivery Pattern Reconfiguration Algorithm

In this section, we will focus on the design of the DPR algorithm. The primary step of the DPR design is to uncover the relationship between the number of PRBs $\Phi_{i,t}^q$ and the number of the packets $P_{i,t}^q$. However, due to the random nature of the wireless channel, this relationship is non-linear, complex and difficult to characterize. In this paper, we utilize the martingales-based method to handle this issue. Before deriving the relationship between $P_{i,t}^q$ and $\Phi_{i,t}^q$, we first introduce the statistical delay analysis model based on martingales theory. To avoid notational confusion with the small-timescale resource allocation, the TTI is indexed by j in this section, where $j \in [0,\infty)$.

A. Preliminary

First, we revisit some important martingale-related definitions in the following.

- *Definition 1 (Supermartingales):* A discrete-time supermartingale process is stochastic process $X(j), j = 1,2,\ldots,$ that satisfies
  - $E[X(j)] < \infty$,
  - $E[X(j+1)|X(1),X(2),\ldots,X(j)] \leq X(j)$ for any $j \geq 1$.
- *Definition 2 (Arrival-Martingales) [28]:* The arrival process of a data flow admits arrival martingales if there is a $\boldsymbol{Ka} > 0$ and a function $\boldsymbol{ha} : \mathrm{rng}(a(k)) \to \mathrm{R}^+$ for $\theta > 0$ such that the process
$$\boldsymbol{M}a(j) = \boldsymbol{ha}(a(k))e^{\theta(A(0,j)-j\boldsymbol{Ka})}, j \geq 0 \quad (1)$$
is a supermartingale.
- *Definition 3 (Service-Martingales) [28]:* The service process admits service-martingales if there is a $\boldsymbol{Ks} > 0$ and a function $\boldsymbol{hs} : \mathrm{rng}(s(j)) \to \mathrm{R}^+$ for $\theta > 0$ such that the process
$$\boldsymbol{M}s(j) = \boldsymbol{hs}(s(j))e^{\theta(j\boldsymbol{Ks}-S(0,j))}, j \geq 0 \quad (2)$$
is a supermartingale.

In the definitions, the notation $\mathrm{rng}(\cdot)$ stands for the range operator. The parameters $\boldsymbol{Ka}$ and $\boldsymbol{ha}$ ($\boldsymbol{Ks}$ and $\boldsymbol{hs}$) implicitly depend on θ. For brevity the augmented notation $\boldsymbol{Ka}(\theta)$ and $\boldsymbol{ha}(\theta)$ ($\boldsymbol{Ks}(\theta)$ and $\boldsymbol{hs}(\theta)$) are omitted in this paper.

The parameter $\boldsymbol{ha}(a(j))$ ($\boldsymbol{hs}(s(j))$) reflects the correlation of the stochastic variables $a(j)$ ($s(j)$) in arrival (service) process. The parameter $\boldsymbol{Ka}$ ($\boldsymbol{Ks}$) determines a linear envelope for the process of $A(0,j)$ ($S(0,j)$), as shown by the blue dashed line in Fig. 4(b). The exponential transform determines the decay rate of queuing metrics, which assigns more weight to larger arrivals (smaller services).

According to the stochastic delay analysis model in *stochastic network calculus* [29], the queue delay can be defined by
$$D(j) := \min\{x \geq 0 \,|\, A(0,j-x) \leq S(0,j)\}. \quad (3)$$
As shown in Fig. 4(b), the queue delay is exactly equal to the delay of the first packet in the queue. According to (3), $(D(j) > x)$ and $(A(0,j-x) > S(0,j))$ are equivalent. Let $\boldsymbol{Dm}$ denote the allowable delay bound. Thus, the DVP can be expressed by
$$\Pr\{D(j) \geq \boldsymbol{Dm}\} = \Pr\{A(0,j-\boldsymbol{Dm}) > S(0,j)\}. \quad (4)$$

Relying on arrival(service)-martingales and the analysis of (4), we can construct a supermartingale for the delay of a data flow, and then yield the DVP via the optional stopping theorem for supermartingales. During the analysis, the parameter θ is related to both arrival-martingale and service-martingale. The value of θ is affected by the maximum user-acceptable DVP ε and it can be treated as a QoS exponent. The DVP is given by Theorem 1 [28].

Theorem 1: For one data flow, assume that its arrival process and service process are statistically independent. Additionally, as the queue stability condition, assume that

$$\hat{\theta} := \sup\{\theta > 0 : \mathit{Ka} \leq \mathit{Ks}\}. \tag{5}$$

If the arrival process and service process admit the arrival-martingale and service-martingale respectively, then the DVP holds for

$$\Pr\{D(k) \geq \mathit{Dm}\} \leq \frac{\mathrm{E}[\mathit{ha}(a(0))]\mathrm{E}[\mathit{hs}(s(0))]}{H} e^{-\hat{\theta}\mathit{KsDm}}, \tag{6}$$

where

$$H = \min\{\mathit{ha}(a(j))\mathit{hs}(s(j)) : a(j) - s(j) > 0, j \geq 0\}.$$

$Proof$: Please see Appendix A.

B. Martingale-Based Resource Estimation

In this subsection, we study the relationship between $P_{i,t}^q$ and $\Phi_{i,t}^q$. First, based on the martingales theory, we derive the minimum average bandwidth that can support the statistical-delay-QoS constrained transmission for a given arrival process in a PTS .[4] Then, the PRB consumption is calculated according to the minimum average bandwidth and the predicted average achievable data rate.

As the VoD flow is from wireline network, the arrival process of a sub-slice is treated as a determinate process. In this case, the number of the arrival packets in the queue of sub-slice q of user i in TTI k in PTS t (i.e., $a_{i,t}^q(k)$) equals the value of packet arrival rate which is calculated as $P_{i,t}^q/\tau_{\mathrm{PTS}}$, where τ_{PTS} is the length of a PTS in units of TTIs. Thus, $\mathit{ha}(a_{i,t}^q(k))$ should be a constant, and, by the definition of supermartingale, $\mathit{Ka}_{i,t}^q \geq (P_{i,t}^q/\tau_{\mathrm{PTS}})$ must be satisfied to ensure that the expression in (1) is supermartingale. Therefore, without loss of generality we take $\mathit{ha}(a_{i,t}^q(k)) = 1$ and $\mathit{Ka}_{i,t}^q = P_{i,t}^q/\tau_{\mathrm{PTS}}$.

For the service process, considering the time-varying channel states and the downlink scheduling policy, we assume that the service rate $s_{i,t}^q(k)$ follows the independent identical distributed (IID) Poisson distribution with the parameter $\lambda_{i,t}^q$ across TTIs within PTS t. Under this assumption, $\lambda_{i,t}^q$ also represents the average service rate for user i in PTS t through sub-slice q. In order to support a video flow with a specific statistical delay-QoS requirement, we utilize the service-martingale framework to derive the value of $\lambda_{i,t}^q$. From [28], for the IID service processes, when $\mathit{hs}_{i,t}^q(s_{i,t}^q(k))$ is a constant and $\mathrm{E}[e^{-\theta_{i,t}^q s_{i,t}^q(k)}] = e^{-\theta_{i,t}^q \mathit{Ks}_{i,t}^q}$, the service-martingale process admits supermartingales. Thus, we take $\mathit{hs}_{i,t}^q(s_{i,t}^q(k)) = 1$ and $\mathit{Ks}_{i,t}^q$ can be expresses by

$$\mathit{Ks}_{i,t}^q = -\frac{\log \mathrm{E}[e^{-\theta_{i,t}^q - 1}]}{\theta_{i,t}^q} = -\frac{\lambda_{i,t}^q(e^{-\theta_{i,t}^q} - 1)}{\theta_{i,t}^q}. \tag{7}$$

To ensure the stability of queue, according to (5), we have $\mathit{Ks}_{i,t}^q = \mathit{Ka}_{i,t}^q$, which is given as

$$\frac{P_{i,t}^q}{\tau_{\mathrm{PTS}}} = -\frac{\lambda_{i,t}^q\left(e^{-\hat{\theta}_{i,t}^q} - 1\right)}{\hat{\theta}_{i,t}^q}. \tag{8}$$

It can be seen from (8) that $\hat{\theta}_{i,t}^q$ has to be calculated first for obtaining the relationship between $P_{i,t}^q$ and $\lambda_{i,t}^q$. According to (6), the constraint of the DVP for one sub-slice can be expressed by

$$e^{-\hat{\theta}_{i,t}^q P_{i,t}^q \mathit{Dm}_i^q/\tau_{\mathrm{PTS}}} \leq \varepsilon_i. \tag{9}$$

In order to save resources, we focus on the upper bound of the inequality (9). The QoS exponent $\hat{\theta}_{i,t}^q$ defined in (5) is taken as

$$\hat{\theta}_{i,t}^q = \frac{-\tau_{\mathrm{PTS}} \ln \varepsilon_i}{P_{i,t}^q \mathit{Dm}_i{}^q} \tag{10}$$

Then (8) can be rewritten as

$$\lambda_{i,t}^q(k) = \frac{\ln \varepsilon_i}{\left(e^{\frac{\tau_{\mathrm{PTS}} \ln \varepsilon_i}{P_{i,t}^q \mathit{Dm}_i^q}} - 1\right)\mathit{Dm}_i^q} \tag{11}$$

Now, the equality (11) characterizes the relationship between the number of packets $P_{i,t}^q$ and the average service rate $\lambda_{i,t}^q$. Next, we turn to studying the relationship between $P_{i,t}^q$ and $\Phi_{i,t}^q$. Considering the predicted average achievable data rate $\bar{r}_{i,t}$, the total number of the PRBs consumed (i.e.,$\Phi_{i,t}^q$) in each TTI is given as

$$\Phi_{i,t}^q = \tau_{\mathrm{PTS}} \frac{\lambda_{i,t}^q L}{\bar{r}_{i,t} B_{\mathrm{PRB}}}, \tag{12}$$

where B_{PRB} denotes the size of a PRB and is equal to 180 [Hz · s]. According to (8) and (10), (12) can be rewritten as

$$\Phi_{i,t}^q = \frac{\tau_{\mathrm{PTS}} L \ln \varepsilon_i}{\left(e^{\frac{\tau_{\mathrm{PTS}} \ln \varepsilon_i}{P_{i,t}^q \mathit{Dm}_i^q}} - 1\right)\bar{r}_{i,t} B_{\mathrm{PRB}} \mathit{Dm}_i{}^q}, \tag{13}$$

which is written as $\Phi_{i,t}^q = f(P_{i,t}^q)$ for brevity. Fig. 5 shows the comparison between the analytical and simulation results in terms of the DVP in different system loads (i.e., $P_{i,t}^q(\lambda_{i,t}^q \cdot \tau_{\mathrm{PTS}})$) for the video flows with different frame rates. The allowable delay bounds of the video flows with the frame rates of 240 fps, 120 fps, 60 fps, and 25 fps are approximated as 4 ms, 8 ms, 16 ms, and 40 ms, respectively. The analytical results in Fig. 5 are obtained by the martingales-based analysis (i.e., the left hand of inequality (9)). For the simulation results, we simulate the transmission process for dozens of PWs, during which the DVP is recorded according to different allowable delay bounds. From Fig. 5, the analytical results and simulation results match well,

[4]In our analysis, we assume that the queue length within each PTS achieves steady states. The initial queue length (i.e., the impact of delay in the preceding PTS) does not affect the distribution of the queue length and delay within the next PTS.

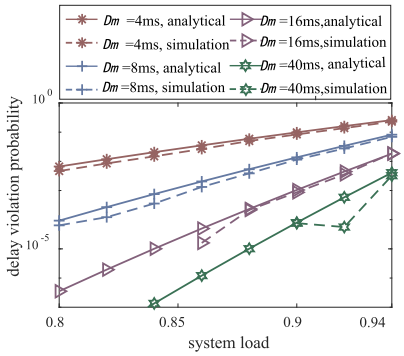


Fig. 5. Comparison between the analytical and simulation results in terms of the DVP.

which verifies the accuracy of the martingales-based delay analysis and demonstrates the feasibility of the martingales-based resource estimation method we proposed. From Fig. 5, it also can be seen that the VoD flows with stricter delay requirements require higher service rates.

C. Problem Formulation

In order to achieve high SE, we take the user demands $\hat{\mathbf{P}}$ and the predicted average achievable data rates $\bar{\mathbf{R}}$ as known parameters and formulate the DPR into an optimization problem to find the optimal delivery pattern $\mathbf{P}$, which is shown as below.

$$\max_{\mathbf{P},\hat{\mathbf{t}}} \frac{\sum_{t=1}^{T}\sum_{i=1}^{M}\hat{P}_{i,t}}{\sum_{t=1}^{T}\sum_{i=1}^{M}\sum_{q=1}^{2}\Phi_{i,t}^{q}} \tag{14}$$

$$\text{s.t.} \sum_{t'=1}^{t} P_{i,t'}^{1} \geq \sum_{t'=1}^{t}\hat{P}_{i,t}, \forall i \in \mathcal{M}, \forall t \in \left\{1,2,\ldots \hat{t}_i\right\} \tag{14a}$$

$$\sum_{t=1}^{\hat{t}_i}\sum_{q=1}^{2} P_{i,t}^{q} = \sum_{t=1}^{T}\hat{P}_{i,t}, \qquad \forall i \in \mathcal{M} \tag{14b}$$

$$\Phi_{i,t}^{q} = f\left(P_{i,t}^{q}\right), \quad \forall i \in \mathcal{M}, \forall t \in \mathcal{T}, \forall q \in \{1,2\} \tag{14c}$$

$$\sum_{i=1}^{M}\sum_{q=1}^{2}\Phi_{i,t}^{q} \leq \Phi^{\max}, \qquad \forall t \in \mathcal{T} \tag{14d}$$

$$P_{i,t}^{q} \geq 0, \qquad \forall i \in \mathcal{M}, \forall t \in \mathcal{T}, \forall q \in \{1,2\} \tag{14e}$$

$$\hat{t}_i = \max\left\{t_i \,\middle|\, t_i \in \mathcal{T}, P_{i,t_i}^{1} > 0\right\} \qquad \forall i \in \mathcal{M} \tag{14f}$$

The objective function evaluates the average SE within the PW. The numerator of the objective function represents the total data amount of all the VoD packets required by the users in the whole PW. The denominator of the objective function represents the total PRB consumption. The constraints in (14a) ensure the smooth video playback for user i, which requires that the cumulative number of the packets delivered in sub-slice 1 is more than or equal to the number of packets corresponding to the total VoD contents to be played until PTS t. The constraints in (14b) ensure that the cumulative number of VoD packets transmitted to user i until PTS $\hat{t}_i$ is equal to the total demands in the PW of the user. Since the video contents transmitted in sub-slice 2 will be played after PTS $\hat{t}_i$, the transmission in sub-slice 2 does not affect the smoothness of video playback. The constraints in (14c) represent the functional relationship of $\Phi_{i,t}^{q}$ and $P_{i,t}^{q}$ for given Dm_i^q and ε_i, which is derived in Subsection A. The constraints in (14d) represent the limitation of total PRB consumption of all users in one PTS. Therefore, if the constraints in (14c) and (14b) are satisfied, the DVP constraint $\Pr\{D_{i,t}^{q}(k) > Dm_i^q\} \leq \varepsilon_i$ can be satisfied as well. The constraints in (14e) ensure the nonnegative feature of $P_{i,t}^{q}$. The constraints in (14f) represent that PTS $\hat{t}_i$ is the PTS in which the transmission of user i in sub-slice 1 terminates.

D. Heuristic Solution Method

In this section, a low-complexity heuristic algorithm is proposed to find a real-time feasible solution to the high-dimensional optimization problem in (14). The heuristic algorithm consists of two phases. The first phase focuses on satisfying the QoS constraints, and it is so called *QoS provisioning phase*. The second phase focuses on optimizing the SE, and it is so called *SE optimization phase*. The flowchart of the two-phase algorithm is depicted in Fig. 6. The details about the algorithm are described as follow.

1) QoS Provisioning Phase: Due to the fast movement of users, the total resource consumption may vary significantly over time as shown by the yellow curve in Fig. 3(c). In this sense, the constraints in (14d) may be violated in some PTSs. To handle this issue, some video contents in these PTSs must be delivered in advance. Based on this idea, we design the QoS provisioning phase, the main steps in which are shown as below.

Step 1 (Initialization): The delivery pattern $\mathbf{P}$ is initialized according to the video contents played in each PTS, i.e., $P_{i,t}^{1} = \hat{P}_{i,t}$ and $P_{i,t}^{2} = 0$.

Step 2 (Checking): The goal of Step 2 is to check if there are any excessive packets in PTS t that need to be transmitted earlier to a preceding PTS t'[5] for satisfying the constraints in (14d). According to (14c), estimate $\Phi_{i,t}^{1}$. Define Φ_t as the total PRB consumption in PTS t, and calculate $\Phi_t = \sum_{i=1}^{M}\Phi_{i,t}^{1}$. The inequality $\Phi_t > \Phi^{\max}$ means that the constraints in (14d) will be violated. If it happens, go straight to Step 3. Otherwise, check PTS $t-1$. The reverse order of checking PTSs (i.e., from PTS T to PTS 1) is adopted to prevent the re-check, because moving excessive packets to a preceding PTS t' may cause a new QoS constraint violation in PTS t'. When $t = 1$, if Φ_1 is still larger than $\Phi^{\max}$, it means that there is no chance to satisfy the QoS constraint, then the whole QoS provisioning phase terminates.

Step 3 (QoS provisioning): The goal of Step 3 is to ensure that the constraints in (14d) is satisfied when $\Phi_t > \Phi^{\max}$. In this step, the QoS provisioning algorithm (i.e., Algorithm 1) is performed to determine which users' video contents to be moved to the preceding PTS (i.e., PTS $t-1$ in our algorithm) and how much to be moved. Accordingly, update the delivery pattern in PTS t and PTS $t-1$. Then go back to Step 2 to continue to check PTS $t-1$.

[5]Here, "move excessive packets in PTS t to a preceding PTS t'" means that the packets that should be transmitted in PTS t are scheduled to be transmitted in PTS t'.

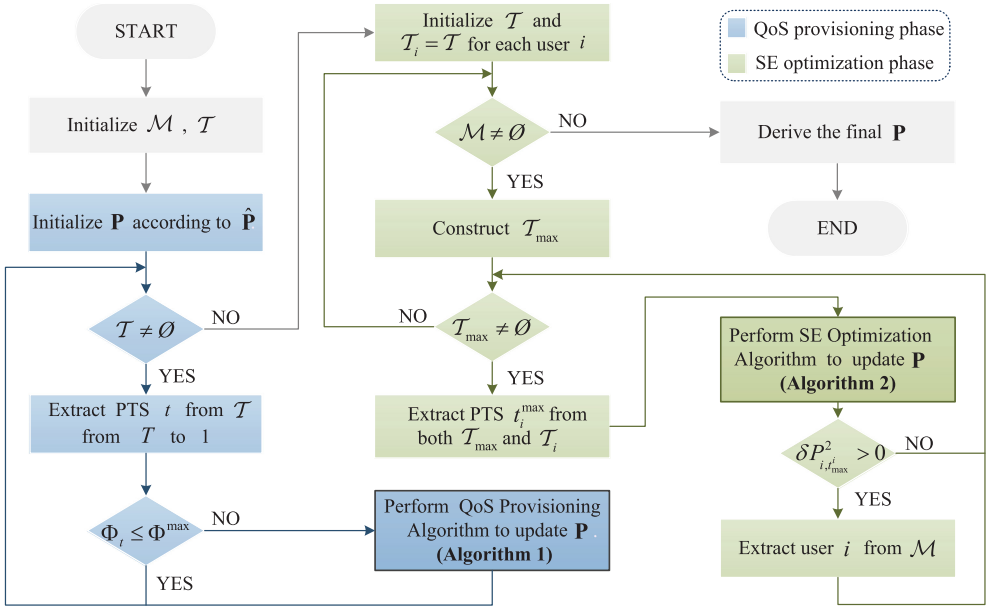


Fig. 6. The flow diagram of the heuristic algorithm.

Next we go into details about Algorithm 1 in Step 3. The main concern of this algorithm is how to satisfy the constraints in (14d) with as few iterations as possible but without compromising on SE. Naturally, the average data rates in PTS $t-1$ may be smaller than those in PTS t for some users. To avoid degrading the SE, we first divide the users into two sets according to the variation tends of data rates. If $\bar{r}_{i,t-1}/\bar{r}_{i,t} \geq 1$, put user i into set $\mathcal{M}^0$, otherwise put user i into set $\mathcal{M}^1$. Here, the users in $\mathcal{M}^0$ are to be selected first, and their packets are moved to the preceding PTS.

Let $\Delta\Phi_t = \Phi_t - \Phi^{\max}$. For a selected user i^*, if $\Delta\Phi_t < \Phi^1_{i^*,t}$, some of its packets (that should be delivered in PTS t) are moved to PTS $t-1$ so that the updated Φ_t equals $\Phi^{\max}$. As a result, all the constraints in (14d) are satisfied, and Algorithm 1 terminates. If $\Delta\Phi_t > \Phi^1_{i^*,t}$, all of its packets are moved to PTS $t-1$.[6] Then, the next user is selected, and its packets are moved. The above process is repeated until $\Phi_t = \Phi^{\max}$. In addition, in the set $\mathcal{M}^0$ or $\mathcal{M}^1$, the user calling for more PRBs is to be selected first for reducing iterations. The user selection rule is shown in Lines 4–8. The calculation procedure for the number of these moved packets is outlined in Lines 9–18.

2) SE Optimization Phase: After the QoS provisioning phase, if there exist surplus PRBs, the SE optimization phase is implemented. Its key idea is to transmit the partial video contents through sub-slice 2 in advance in each user's good-channel-gain PTSs as much as possible. The main steps in this phase are shown as below.

Step 1 (Initialization): Build an individual PTS set $\mathcal{T}_i = \mathcal{T}$ for each user $i \in \mathcal{M}$, which contains all the PTSs waiting for optimization.

Step 2 (Selection): Considering the optimal delivery pattern can be obtained by fully utilizing the good channels of each user,[7] a set $\mathcal{T}_{\max} = \{t^i_{\max} | t^i_{\max} = \arg\max_{t_i \in \mathcal{T}_i}\{\bar{r}_{i,t}\}, \forall i \in \mathcal{M}\}$ is built based on $\bar{\mathbf{R}}$ in each iteration, where $t^i_{\max}$ is the index of the PTS with the highest predicted achievable data rate for user i within set $\mathcal{T}_i$. The delivery pattern in all the PTSs in set $\mathcal{T}_{\max}$ will be optimized next.

Step 3 (Ordering): It is known that optimizing the delivery pattern in the preceding PTSs might provide more surplus PRBs. This facilitates the earlier completion of VoD transmission in the PW, especially for the users whose peak-data-rate PTSs are near the end of the PW. Hence, in this step, the PTSs in set $\mathcal{T}_{\max}$ are arranged in the ascending order of PTS index. In other words, go to Step 4 starting by the PTS with the smallest index in set $\mathcal{T}_{\max}$.

Step 4 (SE Optimization): Perform the SE optimization algorithm (i.e., Algorithm 2) for each PTS $t^i_{\max}$ and the corresponding user i in turn. Then, PTS $t^i_{\max}$ is removed from the set $\mathcal{T}_i$. If more than one user has the same $t^i_{\max}$, Algorithm 2 is performed for the user with the smallest index in this paper. If the delivery pattern of user i has been optimal,[8] user i is removed from user set $\mathcal{M}$. After handling all the PTSs in set $\mathcal{T}_{\max}$, if $\mathcal{M} = \emptyset$, the heuristic algorithm terminates, otherwise, go back to Step 2.

[6] Actually, the delay requirements also become looser for the packets moved to preceding PTSs in QoS provisioning phase. However, this phase is only operated for a few special cases. Even with the occurrence of $\Phi_t > \Phi^{\max}$ in some PTSs, the number of packets to be transmitted in advance is relatively small. Hence, in order to simplify the operation of MEC server and user device, this partial video contents are not allowed to be transmitted as a separate video stream in sub-slice 2, so that the continuity of serial numbers can be maintained during the reading and storing of video contents. Although there will be some wastage of resources caused by the over-guaranteed delay-QoS for sub-slice 1, we believe that these small amounts of wasted resources are acceptable.

[7] Since the heuristic algorithm is developed based on intuition and experience, the delivery pattern found in the our heuristic algorithm may be not the optimal one.

[8] The optimal delivery pattern for one user in our heuristic algorithm is that all subsequent video contents are transmitted within the user's peak-channel-gain PTS.

Algorithm 1: QoS Provisioning Algorithm.

Input:
delivery pattern $\mathbf{P}_{M\times T\times 2}$, predicted average achievable data rates $\bar{\mathbf{R}}_{M\times T}$, PRB consumption limitation $\Phi^{\max}$, PRB consumption Φ_t, PTS t

Output:
delivery pattern $\mathbf{P}_{M\times T\times 2}$,

1: Divide users into two categories:
$\mathcal{M}^0=\{i|i\in\mathcal{M},\bar{r}_{i,t-1}/\bar{r}_{i,t}\geq 1\}$
$\mathcal{M}^1=\{i|i\in\mathcal{M},\bar{r}_{i,t-1}/\bar{r}_{i,t}< 1\}$
2: $\Delta\Phi_t=\Phi_t-\Phi^{\max}$
3: **while** $\Delta\Phi_t>0$ **do**
4: **if** $\mathcal{M}^0\neq\emptyset$ & $\max\{\Phi^1_{i,t}:i\in\mathcal{M}^0\}\neq 0$ **then**
5: $i^*=\arg\max_{i\in\mathcal{M}^0}\{\Phi^1_{i,t}\}$
6: **else**
7: $i^*=\arg\max_{i\in\mathcal{M}^1}\{\Phi^1_{i,t}\}$
8: **end if**
9: $\Phi_0\leftarrow\Phi^1_{i^*,t}$
10: $\Phi^1_{i^*,t}\leftarrow\max\{\Phi^1_{i^*,t}-\Delta\Phi_t,0\}$
11: $\Delta\Phi_t\leftarrow\max\{\Delta\Phi_t-\Phi_0,0\}$
12: $P_0\leftarrow P^1_{i^*,t}$
13: Update the value of $P^1_{i^*,t}$ based on $P^1_{i^*,t}=f^{-1}(\Phi^1_{i^*,t})$
14: **if** $t>1$ **then**
15: $P^1_{i^*,t-1}\leftarrow P^1_{i^*,t-1}+(P_0-P^1_{i^*,t})$
16: **else**
17: $\Delta\Phi_t\leftarrow 0$
18: **end if**
19: **end while**

Algorithm 2: SE Optimization Algorithm.

Input:
delivery pattern $\mathbf{P}_{M\times T\times 2}$, predicted average achievable data rates $\bar{\mathbf{R}}_{M\times T}$, PRB consumption Φ_t, PRB consumption limitation $\Phi^{\max}$, PTS $t^i_{\max}$, PTS set $\mathcal{T}_i$

Output:
delivery pattern $\mathbf{P}_{M\times T\times 2}$

1: **if** $\Phi_{t^i_{\max}}<\Phi^{\max}$ **then**
2: $\Delta\Phi_{t^i_{\max}}=\Phi^{\max}-\Phi_{t^i_{\max}}$
3: Calculate $\Delta P^2_{i,t^i_{\max}}$ according to $\Delta\Phi_{t^i_{\max}}$
4: $\Delta P^2_{i,t^i_{\max}}=\min\{\Delta P^2_{i,t^i_{\max}},\sum_{t\in\mathcal{T}_i^C}P^1_{i,t}\}$,
5: $P^2_{i,t^i_{\max}}\leftarrow\Delta P^2_{i,t^i_{\max}}$
6: **while** $P^2_{i,t^i_{\max}}>0$ **do**
7: $\hat{t}_i=\max\{t|t\in\mathcal{T}_i,P^1_{i,t}>0\}$
8: $P_0\leftarrow P^1_{i,\hat{t}_i}$
9: $P^1_{i,\hat{t}_i}\leftarrow\max\{P^1_{i,\hat{t}_i}-\Delta P^2_{i,t^i_{\max}},0\}$
10: $\Delta P^2_{i,t^i_{\max}}\leftarrow\max\{\Delta P^2_{i,t^i_{\max}}-P_0,0\}$
11: **end while**
12: **end if**

Now, we elaborate on Algorithm 2 and the criterion for determining whether the delivery pattern of a user achieves optimality in Step 4. Algorithm 2 consists of two parts. One is to determine the number of packets delivered to user i in PTS $t^i_{\max}$ via sub-slice 2, (i.e., $P^2_{i,t^i_{\max}}$). Let $\Delta P^2_{i,t^i_{\max}}$ denote the maximum number of the packets supported by the surplus PRBs in PTS $t^i_{\max}$, which is calculated as $\Delta P^2_{i,t^i_{\max}}=f^{-1}(\Delta\Phi_{\max})$. If $\Delta P^2_{i,t^i_{\max}}$ is less than the total number of the packets intended to deliver in the PTSs after PTS $t^i_{\max}$ within $\mathcal{T}_i$, (i.e., $\Delta P^2_{i,t^i_{\max}}<\sum_{h\in\mathcal{T}_i^C}P^1_{i,h}$ where $\mathcal{T}_i^C=\{t\in\mathcal{T}_i|t>t^i_{\max},P^1_{i,t}>0\}$,) we have $P^2_{i,t^i_{\max}}=\Delta P^2_{i,t^i_{\max}}$. Otherwise, we have $P^2_{i,t^i_{\max}}=\sum_{h\in\mathcal{T}_i^C}P^1_{i,h}$. The specific procedure is outlined in Lines 2–5. The second part in Algorithm 2 is to update PTS $\hat{t}_i$ according to the value of $P^2_{i,t^i_{\max}}$, which is outlined in Lines 6–10.

In addition, let $\delta P^2_{i,t^i_{\max}}=\Delta P^2_{i,t^i_{\max}}-\sum_{h\in\mathcal{T}_i^C}P^1_{i,h}$. If $\delta P^2_{i,t^i_{\max}}>0$, it means that the delivery patten of user i is optimal, because there have been no packets to be transmitted in the PTSs in set $\mathcal{T}_i^C$. And the residual PRBs in PTS $t^i_{\max}$ are to be handled in the subsequent process for the remaining users.

V. Small-Timescale Resource Allocation Algorithm

In this section, we study how to instantiate the delivery pattern in each TTI. This instantiation problem is formulated as a PRB allocation problem to achieve a favorable system performance. To handle this high dimensional optimization problem, we propose a utility theory based resource allocation algorithm, which can decouple the interdependency among the allocation of each PRB. Before detailing the algorithm, we first introduce the design of utility function.

A. Utility Function Construction

Considering the metrics of SE, delay-QoS and fairness, the per-PRB utility function $U^q_{i,n}(k')$ for PRB n and sub-slice q of user i is constructed as

$$U^q_{i,n}(k')=a(e_{i,n}(k'))^{\alpha}+b(d^q_i(k'))^{\beta}+c(f^q_i(k'))^{\gamma}. \quad (15)$$

The parameters a,b,c and α,β,γ denote the weight coefficients of the three sub-functions. The three sub-functions are discussed in detail as below.

1) $e_{i,n}(k')$*:* This sub-function aims for high SE while ensuring the scheduling fairness among the users at different locations in each TTI. The achievable data rate over one PRB is affected by the distances between the BS and the users, as well as the multipath effect. Similar to [27], in order to provide the same priority to the users in different locations in the cell, the normalized achievable data rate $e_{i,n}(k')$, is defined as

$$e_{i,n}(k')=\frac{r_{i,n}(k')}{r_i(k')}, \quad (16)$$

where $r_{i,n}(k')$ denotes the achievable data rate over PRB n for user i in TTI k', and $r_i(k')$ denotes the the achievable data rate over all PRBs for user i, i.e., $r_i(k')=\sum_{n=1}^{N}r_{i,n}(k')$. It can be seen that the impact of the path loss is mitigated by the normalized achievable data rate $e_{i,n}(k')$, hence the fairness among users with different distances to the BS in a TTI can be ensured. A higher $e_{i,n}(k')$ indicates that PRB n is more likely to be assigned to user i.

2) $d_i^q(k')$: This sub-function is designed to guarantee delay-QoS in each TTI. As the delay-QoS requirement of the packets transmitted in sub-slice 1 is stricter than that of the packets transmitted in sub-slice 2, the delay of sub-slice 1 has a greater impact on making a scheduling decision. Here, the form of Sigmoid function is utilized to characterize the impact of the delay and the sub-slice type on the delay-QoS metric. The sub-function $d_i^q(k')$ is given as

$$d_i^q(k') = \frac{1}{1 + e^{-\kappa^q\left(D_i^q(k') - \mathit{Dm}_i^q\right)}}, \tag{17}$$

where Dm_i^q is the allowable delay bound of sub-slice q for user i and $D_i^q(k')$ denotes the delay of sub-slice q for user i in TTI k', which is expressed by

$$D_i^q(k') = \frac{Q_{i,t}^q(k)\tau_{\mathrm{PTS}}}{P_{i,t}^q} + \bar{D}_{i,t}^q(K)\mathrm{I}_{S_{i,t}^q(0,k)}. \tag{18}$$

$Q_{i,t}^q(k)$ is the total number of the packets arriving and still waiting in the buffer until TTI k in PTS t. It can be calculated as $Q_{i,t}^q(k) = A_{i,t}^q(0,k) - \max\{S_{i,t}^q(0,k) - Q_{i,t-1}^q(K), 0\}$ and $Q_{i,t-1}^q(K)$ is the queue length at the end of PTS $t-1$. $\bar{D}_{i,t}^q(K)$ denotes the delay at the end of PTS $t-1$, and $\mathrm{I}_{S_{i,t}^q(0,k)}$ is an indicator of the impact of the packets arriving in PTS $t-1$ on the delay, i.e.,

$$\mathrm{I}_{S_{i,t}^q(0,k)} = \begin{cases} 0, Q_{i,t-1}^q(K) - S_{i,t}^q(0,k) \le 0 \\ 1, Q_{i,t-1}^q(K) - S_{i,t}^q(0,k) > 0. \end{cases}$$

Additionally, $(D_i^q(k') - \mathit{Dm}_i^q)$ characterizes the degree of urgency to be scheduled for sub-slice q of user i. The coefficient $\kappa^q \in (0,1)$ is utilized to adjust the sensitivity of the function value to the delay $D_i^q(k')$. In our work, we take κ^1 as 0.2, and take κ^2 as 0.1.

3) $f_i^q(k')$: This sub-function defines a long-term fairness from the view of approximated weighted load by jointly considering the numbers of cumulative arrival packets, cumulative packets served, and the total number of packets to be transmitted in PTS t. We formulate this sub-function as below,

$$f_i^q(k') = \frac{\frac{1}{k}(1 - \frac{S_{i,t}^q(0,k)}{P_{i,t}^q})S_{i,t}^q(0,k) + (1 - \frac{1}{k})\frac{Q_{i,t}^q(k)}{P_{i,t}^q}Q_{i,t}^q(k)}{A_{i,t}^q(0,k)}, \tag{19}$$

where the value of $P_{i,t}^q$ is taken from the reconfigured delivery pattern. Since the sub-slice with a higher arrival rate requires higher service capability, the user with more packets served is given higher priority at the beginning period of the PW. However, as time goes on, the proportion of the queue length to the number of cumulative arrival packets becomes an increasingly important role in the metric of fairness. What is more, for one considered sub-slice with a specific queue length, the value of $f_i^q(k')$ becomes higher with the increase of k within a PTS. In one TTI, for the sub-slices with the same queue length, the sub-slice with the lower value of $P_{i,t}^q$ is given a higher priority.

Algorithm 3: Small-Timescale Resource Allocation Algorithm.

Input:
QoS for all sub-slices; queue states and historical transmission information, delivery pattern $\mathbf{P}_{M\times N\times 2}$
Output:
PRB allocation map $\mathbf{\Psi}_{M\times N\times 2}$

```
Initialize utility function U_{i,n}^q(k')
Set ψ_{i,n}^q to 0
repeat
for all n ∈ 𝒩 do
    (i*, q*) = arg max_{i∈ℳ, q∈{1,2}} {U_{i,n}^q(k')}
    ψ_{i*,n}^{q*} ← 1
end for
for all i ∈ ℳ do
    for all q ∈ {1, 2} do
        if Σ_{n=1}^{N} ψ_{i,n}^q r̃_{i,n}(k') > Q_i^q(k') L then
            Set ψ_{i,n}^q and U_{i,n}^q(k') on the excessive PRBs
            with relatively low achievable data rates to 0
        end if
    end for
end for
until The constraints in (20c) are all satisfied
```

B. Utility Theory Based Algorithm

Based on the utility function, the priority of the sub-slice to be scheduled over the PRB can be obtained. Hence, a favorable system performance can be achieved by allocating each PRB to the sub-slice with the highest utility function value. To avoid wasting resources, the optimization problem of PRB allocation map $\mathbf{\Psi}$ can be formulated as below.

$$\max_{\mathbf{\Psi}} \sum_{i=1}^{M}\sum_{n=1}^{N}\sum_{q=1}^{2} U_{i,n}^q(k')\psi_{i,n}^q \tag{20}$$

$$\text{s.t. } \psi_{i,n}^q \in \{0,1\}, \quad \forall i \in \mathcal{M}, \forall n \in \mathcal{N}, \forall q \in \{1,2\} \tag{20a}$$

$$\sum_{i=1}^{M}\sum_{q=1}^{2} \psi_{i,n}^q \le 1, \quad \forall n \in \mathcal{N} \tag{20b}$$

$$\sum_{n=1}^{N} \psi_{i,n}^q r_{i,n}(k') \le Q_i^q(k')L, \forall i \in \mathcal{M}, \forall q \in \{1,2\} \tag{20c}$$

The objective of the optimization problem is to find the optimal PRB allocation map. The indicator variable $\psi_{i,n}^q$ in (20a) represents whether PRB n is assigned to sub-slice q of user i in TTI k'. The constraints in (20b) ensure that each PRB is constrained to be assigned to one sub-slice of one user in a given TTI. The constraints in (20c) ensure that the number of packets that can be transmitted does not exceed the number of packets in the queues.

Let $\hat{\mathbf{\Psi}} = \arg\max_{\mathbf{\Psi}}\{\sum_{i=1}^{M}\sum_{n=1}^{N}\sum_{q=1}^{2} U_{i,n}^q(k')\psi_{i,n}^q | (20a), (20b)\}$. Let $\mathbf{\Psi}^*$ represent the optimal solution of the optimization problem (20). The difference between $\hat{\mathbf{\Psi}}$ and $\mathbf{\Psi}^*$ lies in whether the constraints in (20c) are considered. From the definition of $\hat{\mathbf{\Psi}}$, it is obvious that less PRBs may be allocated to the sub-slice

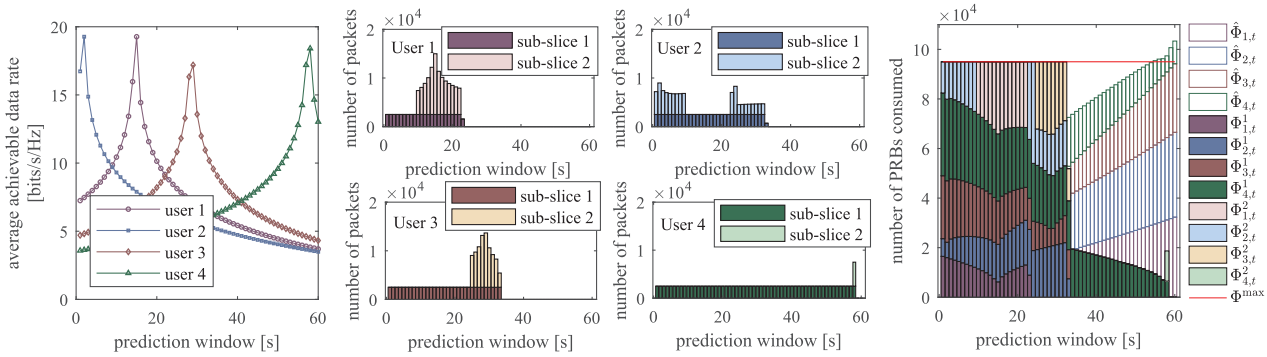


Fig. 7. Illustration of DPR in case 1 (with user 1, user 2, user 3, and user 4 in Table III). In Fig. 7(c), $\hat{\Phi}_{i,t}$ represents the estimated PRB consumption of user i in PTS t when VoD contents are delivered on demand and $\Phi^q_{i,t}$ represents the estimated PRB consumption through sub-slice q of user i in PTS t. $\Phi^{\max}$ is the number of total available PRB in one PTS. (a) Long-term channel state variation. (b) Reconfigured delivery patterns for four users. (c) Estimated PRB consumption for four users.

TABLE II
SIMULATION PARAMETER

Parameters	Value
BS transmit power	43 dBm
noise power	-55 dBm
cell radius	600 m
video flow rate (from MEC server to BS)	20 Mbps
packet size	1000 bytes
allowable delay bound (sub-slice 1)	40 ms
allowable delay bound (sub-slice 2)	1 s
maximum user-acceptable DVP	10^{-4}
PW duration	60 s
PTS duration	1 s
TTI duration	1 ms
a;b;c $(q=1)$	4;6;5
$\alpha;\beta;\gamma$ $(q=1)$	0.8;0.1;0.1
a;b;c $(q=2)$	3;3;3
$\alpha;\beta;\gamma$ $(q=2)$	0.5;0.5;0.5

TABLE III
PARAMETERS FOR USER GENERATION

User index	Initial distance	Initial direction[1]	Drive speed	User index	Initial distance	Initial direction[1]	Drive speed
1	177m	1	43km/h	7	256m	1	36km/h
2	17m	1	36km/h	8	133m	1	50km/h
3	400m	1	50km/h	9	510m	1	43km/h
4	576m	1	36km/h	10	410m	1	25km/h
5	19m	0	36km/h	11	445m	1	29km/h
6	290m	1	47km/h	12	450m	1	25km/h

[81] The number "1" indicates that the user is driving towards the BS; the number "0" indicates that the user is driving away from the BS.

with shorter queue length. Naturally, $\hat{\mathbf{\Psi}}$ may only make the constraints in (20c) unsatisfied for a few users' sub-slices with much longer queue length. Hence, we propose an iterative algorithm which starts from $\hat{\mathbf{\Psi}}$ and modifies itself according to the constraints in (20c) in each iteration, until all the constraints are satisfied. The specific procedure is outlined in Algorithm 3. In the initialization of utility function matrix $\mathbf{U}_{M\times N\times 2}$, the utility function values for sub-slices with empty queues are set directly to 0. The process to find $\hat{\mathbf{\Psi}}$ is outlined in Lines 4–7. In case of over allocation, the utility function values on the excessive PRBs with relatively poor channel states are set to 0, which is outlined in Lines 8–15. Then repeat the above two processes until there is no over allocation or there are no packets in the queues of all the sub-slices.

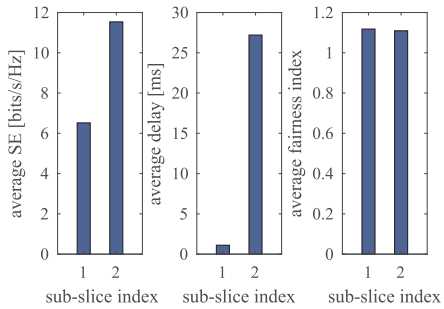


Fig. 8. Performance of SE, delay and fairness.

VI. PERFORMANCE EVALUATIONS

In this section, we evaluate the performance of the proposed scheme with simulations by comparing with the blind deadline-based resource allocation (BDRA) scheme proposed in [6].[9] All results are obtained via the MATLAB simulations.

A. Simulation Setup

We consider the cell served by a roadside BS which is connected to the MEC server. In our simulation, the large-scale channel gain within each PTS remains constant, and it may change from one PTS to another due to the users' fast mobility. Additionally, the trend of the large-scale channel depends on the users' moving trajectories. Based on Shannon's formula, the average achievable transmission rate is obtained according to the large-scale channel gain. In this paper, we ignore the impact of shadow fading on large-scale channel gain. The large-scale

[9]In order to ensure the identical scenario for both schemes in the comparison, we assume that the instantaneous data rate in the current TTI is known by BS scheduler in the BDRA scheme. In addition, we extend the service mechanism of serving one packet per TTI for one user to allocate the full bandwidth to a user at a TTI. To make delay requirements in both algorithms the same, we assume that the packet deadline is the playback time of the frame involving this packet and the duration of the video segment we focus on to minimize the delay violation probability is set 1 s, i.e., a PTS.

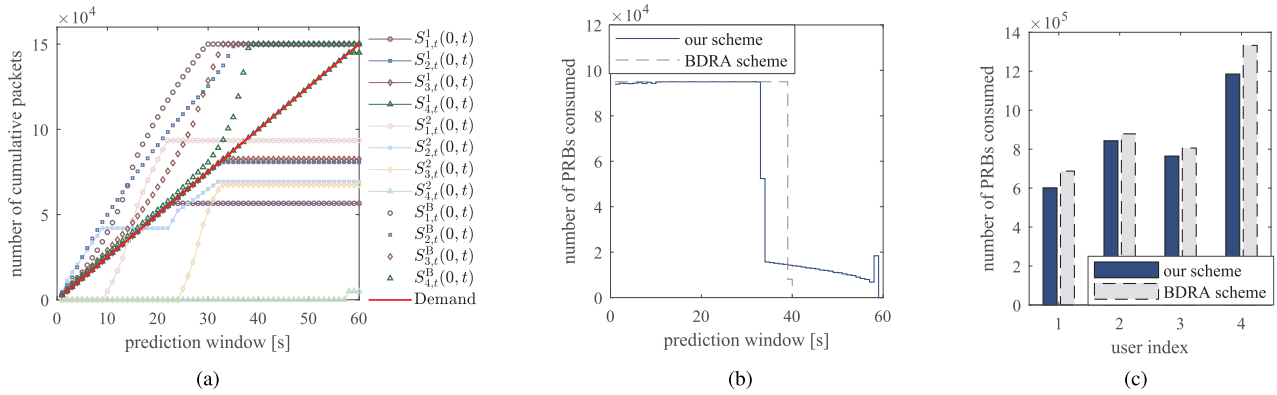


Fig. 9. The comparison of the two schemes in case 1. In Fig. 9(a), $S_{i,t}^{q}(0,k)$ represents the number of the cumulative packets of user i through sub-slice q until PTS t in the predictive two-timescale resource allocation scheme, and $S_{i,t}^{\mathrm{B}}(0,k)$ represents the number of the cumulative packets of user i until PTS t in the BDRA scheme. (a) Number of cumulative packets served. (b) PRB consumption in each PTS. (c) PRB consumption per user.

channel gain of user i is given as $g_i(d) = d^{-l}$, where d denotes the distance between the BS and user i and l denotes the path-loss exponent, which is taken as 3. In the small scale, we model the instantaneous date rate over a PRB as a Gaussian distribution with a variance of 80 [30]. Other simulation parameters are listed in Table II. The generating parameters of the twelve users involved in the simulation process are listed in Table III.

B. Simulation Results

Firstly, we take a four-user case (case 1) to demonstrate the reconfigured delivery pattern found in the DPR algorithm with the allowable bandwidth bound of 17.1 MHz. The total number $\Phi^{\max}$ of the available PRBs in one PTS is 95 000. Fig. 7(a) shows the variation of the average achievable data rates of the four users in the PW, where the peak channel gain for each user occurs in different PTSs. Fig. 7(b) shows the reconfigured delivery pattern, and Fig. 7(c) shows the estimated PRB consumption per PTS for both the on-demand delivery pattern and reconfigured delivery pattern. It can be seen that owing to the DPR, the surplus PRBs have been utilized at the time when the user is experiencing good channel states. Take user 1 as an example. When user 1 experiences its good channel gains during the 10th–22nd PTSs, the BS can transmit the video contents played after the 23 rd PTS to it in advance through sub-slice 2. In this case, the transmission of all the VoD contents played in the PW terminates in the 23 rd PTS. Whereas, for user 4, its peak channel gain occurs at the end of the PW, as shown in Fig. 7(a). In this case, as most of VoD contents cannot wait to be transmitted until the end of the PW, little space is left for DPR to improve SE. As a result, the transmission of user 4 terminates till the 58th PTS, as shown in Fig. 7(b). These results indicate that the DPR can improve the SE more significantly for the users that experience good channel states earlier within the PW.

Then the delivery pattern is instantiated based on the small-timescale resource allocation algorithm. Fig. 8 shows the performance of the three metrics considered in the utility function and the comparison between the two kinds of sub-slices. As shown in Fig. 8(a), the average SE for sub-slice 2 is almost

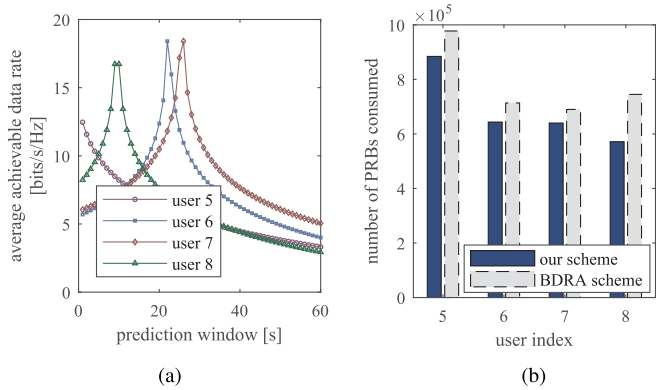


Fig. 10. Simulation results in case 2 (with user 5, user 6, user 7, and user 8 in Table III). (a) Long-term channel state variation. (b) PRB consumption per user.

double that for sub-slice 1. The reason lies in that a heavier weight is given to SE metric in the utility function for sub-slice 2 than that for sub-slice 1. In addition, as shown in Fig. 8(b), the average delay for sub-slice 1 is significantly less than that for sub-slice 2. This result is reasonable as the delay metric is given a heavier weight for sub-slice 1. However, it also can be seen that the delay-QoS requirements of both kinds of sub-slices are somewhat over-guaranteed. The reasons are described as follows. In this paper, the average achievable data rate is obtained according to the average channel gain, without considering the benefit of scheduling scheme. For the frequency-selective wireless channel, a higher average achievable data rate can be achieved through a well-designed scheduling scheme. Moreover, we utilize the reciprocal of cumulative variance of fairness sub-function values in each TTI within a PTS as the fairness index. From Fig. 8(c), we can see that the performance of fairness is similar despite the large difference in average delay.

In Fig. 9 we compare our resource allocation scheme with the BDRA scheme from three perspectives. Fig. 9(a) presents the curves of the cumulative number of packets served for each user. In the BDRA scheme, the BS will transmit subsequent video contents to the users with equal probability whenever there

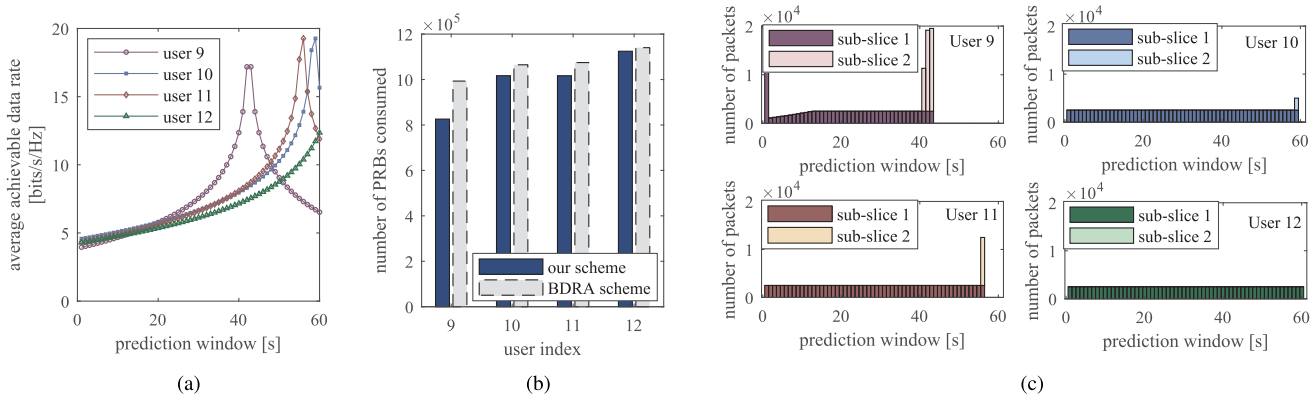


Fig. 11. Simulation results in case 3 (with user 9, user 10, user 11, and user 12 in Table III). (a) Long-term channel state variation. (b) PRB consumption per user. (c) Reconfigured delivery patterns for four users.

remain surplus PRBs after ensuring smooth video playback. Whereas, in our scheme, the BS will allocate the surplus PRBs to a well-chosen user with good channel gain experienced within the PW. Fig. 9(b) shows the PRB consumption per PTS. As seen in Fig. 9(b), the transmission process in the BDRA scheme terminates earlier than that in our scheme. However, the massive blind advance transmissions make the total PRB consumption in the BDRA scheme is much more than that in our scheme. Fig. 9(c) shows the total PRB consumption per user. From Fig. 9(c), all the users in our scheme consume fewer resources than those in the BDRA scheme. In particular, the benefit of our scheme is greatest for user 4. This is because our scheme can take full advantage of the good channel states of user 4 for transmissions, while the BDRA scheme can not.

To analyze the effect of user trajectory on the performance, we consider two other four-user cases shown in Fig. 10(a) and Fig. 11(a), which are labeled as case 2 and case 3, respectively. As can be seen in Fig. 10(b) and Fig. 11(b), for each user in both cases, the total PRB consumption in our scheme is less than that in the BDRA scheme. Furthermore, the superiority of our scheme is most significant for user 8 in case 2. This is because that the time of the best channel gain occurring of the user 8 is extremely different from those of other users. This means that few users contend for the surplus PRBs with user 8 for the transmission in advance. Consequently, more surplus PRBs can be allocated to user 8, and more video contents of user 8 can be transmitted with high SE. For the same reason, user 9 in case 3 makes the greatest contribution to system SE in our scheme. Additionally, it can be seen in Fig. 11(c) that no modification is made to the initial delivery pattern of user 12 in our scheme. However, even in this case, our scheme is still superior to the BDRA scheme in terms of saving PRBs for user 12, as shown in Fig. 11(b).

To analyze the effect of the number of users on the performance, we consider two other cases with eight and twelve users, labeled as case 4 and case 5 respectively. The allowable bandwidth bounds in the two cases are set to 34.2 MHz and 51.3 MHz. Fig. 12 shows the total PRB consumption and the average SE in the whole PW in case 1, case 4 and case 5. The PRB consumption of each user in case 5 is shown in Fig. 13.

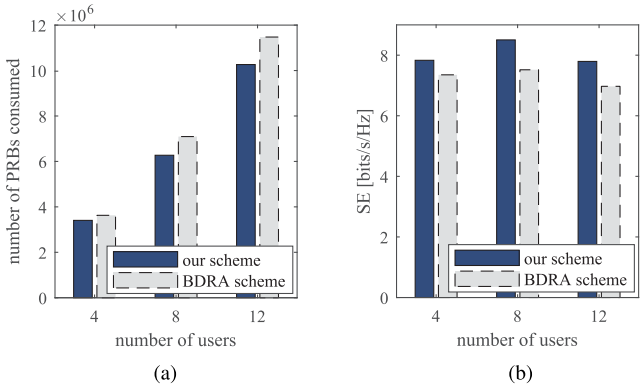


Fig. 12. Comparisons of PRB consumption and average SE in case 1, case 4 (with user 1, user 2, user 3, user 4, user 5, user 6, user 9, and user 12 in Table III), and case 5 (with twelve users in the Table III). (a) Total PRB consumption. (b) Average SE.

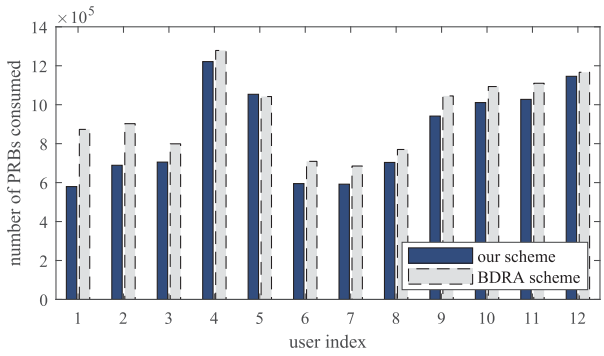


Fig. 13. PRB consumption per user in case 5.

Comparing Fig. 9(c) and Fig. 13, we can see that for the first three users, the gap of PRB consumption between the two schemes becomes larger as the number of users increases. This demonstrates that larger system bandwidth provides more potential for our scheme to improve SE. Besides, among all the users, user 5 is an exception, who consumes more PRBs in our scheme than in the BDRA scheme. This results from the fact that user 5's channel gain peaks at the beginning of the PW. In this case,

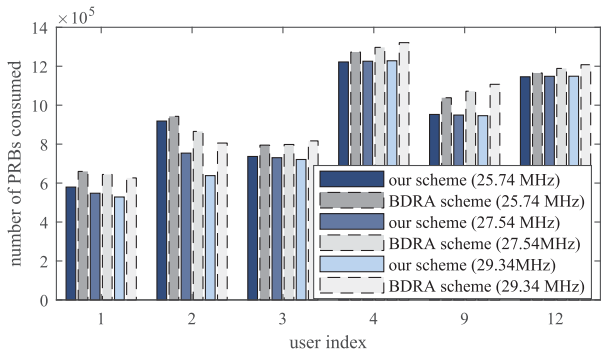


Fig. 14. PRB consumption per user in case 6 (with user 1, user 2, user 3, user 4, user 9, and user 12 in Table III).

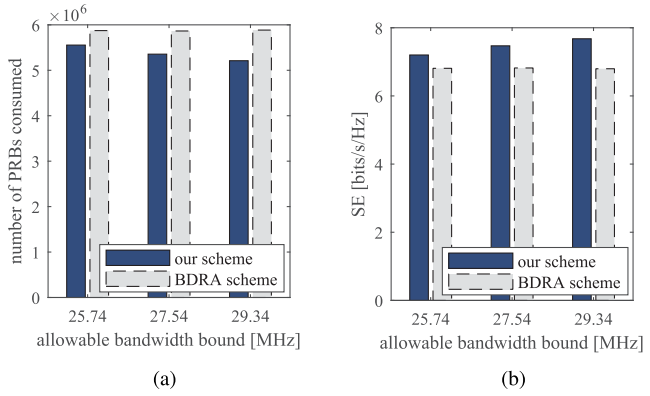


Fig. 15. Comparisons of PRB consumption and average SE in cases with different allowable bandwidth bounds. (a) Total PRB consumption. (b) Average SE.

the more video contents of user 5 are transmitted in advance, the more resources are saved. But the massive advance transmission for user 5 leaves other users with less resource share in their PTSs with good channel gain. In addition, in view of the overall system via Fig. 12, the difference in total PRB consumption between the two schemes becomes increasingly significant with the increase of the number of the users. Whereas, the gain of the average SE is largest when $M = 8$, which indicates that there may exist the optimum number of users to maximize the SE, which we will study in the future.

Fig. 14 shows the PRB consumption per user in a six-user case (case 6) within different allowable bandwidth bounds. As shown in Fig. 14, the PRB consumption of the first three users decreases gradually with the relaxation of the allowable bandwidth bound. This is because that the increased number of available PRBs allows more video contents of the three users to be transmitted in advance in good channel states. Whereas, for user 4, user 9 and user 12, the PRB consumption of these users seems to be the same for different bandwidth bounds in our scheme, and increases slightly along with the relaxation of bandwidth bound in the BDRA scheme. This is also reasonable. For user 4, user 9 and user 12, their peak channel gains all occur near the end of the PW, which forces most of the video contents to be transmitted in the poorer channel states. In this sense, their optimum delivery pattern (i.e., all subsequent video contents are transmitted within the user's peak-channel-gain PTS) can be achieved under a relative strict allowable bandwidth bound. Hence, the PRB consumption of the last three users in our scheme shows no apparent changes. Whereas, in the BDRA scheme, the BS does not know the long-term channel states. Naturally, more packets can be transmitted in advance when the bandwidth bound is looser.As a result, more PRBs are consumed by the last three users in the BDRA scheme. Furthermore, from the view of total PRB consumption of the users as shown in Fig. 15, we can see that our scheme will present better performance in terms of both the PRB consumption and the average SE. In other words, our scheme is more effective in improving SE in moderate and low load system.

VII. Conclusion

In this paper, we proposed a two-timescale resource allocation scheme for VoD services in fast moving scenarios. In the large timescale, we put forward a BDA-aided DPR algorithm, which leverages the non-realtime nature of VoD and provides a new way to enhance the smoothness of video playback without sacrificing video frame quality. Meanwhile, we proposed a martingales-based resource estimation method on the time scale of PTSs for the VoD flows with different delay-QoS requirements. Furthermore, the VoD slice was divided into two logical sub-slices to serve the two kinds of packets delivered on time and in advance. In the small timescale, the PRB allocation problem for the sub-slices in each TTI was formulated as a binary integer programming problem based on the utility theory, and solved by an iterative algorithm. In the future, we will further investigate how to combine our scheme with the DASH architecture to avoid frequent video-quality level switching when system load is relatively high.

Appendix A
Theorem 1

$\hat{\theta}$ is defined as (5). From (2), we have

$$\begin{aligned}
&\Pr\{D(j) \geq \mathit{Dm}\} \\
&\leq \Pr\left\{\max_{j \geq \mathit{Dm}} \{A(\mathit{Dm}, j) - S(0, j)\} \geq 0\right\} \\
&\leq \Pr\left\{\max_{j \geq \mathit{Dm}} \{A(\mathit{Dm}, j) - S(0, j) \right. \\
&\quad \left. - (j - \mathit{Dm})(\mathit{Ks} - \mathit{Ks})\} \geq 0\right\} \\
&\leq \Pr\left\{\max_{j \geq \mathit{Dm}} \{A(\mathit{Dm}, j) - (j - D^{\max})\mathit{Ka}\right. \qquad (21) \\
&\quad \left. + j\mathit{Ks} - S(0, j)\} \geq \mathit{DmKs}\right\}. \qquad (22)
\end{aligned}$$

For one data stream, construct a process as follows:

$$M(k) = \mathit{ha}(a(j)) \times \mathit{hs}(s(j)) e^{\{A(\mathit{Dm}, j) - (j - \mathit{Dm})\mathit{Ka} + j\mathit{Ks} - S(0, j)\}}. \qquad (23)$$

If the arrival process and service process of the data stream at the queue maintained in the BS admit the arrival-martingale and service-martingale, then we have

$$M(j) = \boldsymbol{Ma}(j)\boldsymbol{Ms}(j). \tag{24}$$

If the arrival process and service process are statistically independent, then we have

$$\begin{aligned}
&\mathrm{E}\left[M(j+1)\,|M(1),\ldots,M(j)\right] \\
&= \mathrm{E}\left[\boldsymbol{Ma}(j+1)\boldsymbol{Ms}(j+1)\,|M(1),\ldots,M(j)\right] \\
&= \boldsymbol{Ka}\mathrm{E}\left[\boldsymbol{Ma}(j+1)\,|\boldsymbol{Ma}(1),\ldots,\boldsymbol{Ma}(j)\right] \\
&\quad\times \mathrm{E}\left[\boldsymbol{Ms}(j+1)\,|\boldsymbol{Ms}(1),\ldots,\boldsymbol{Ms}(j)\right] \\
&\le \boldsymbol{Ma}(j)\boldsymbol{Ms}(j) \\
&= M(j).
\end{aligned} \tag{25}$$

So the process $M(j)$ is a supermartingales. Define the stopping time j^0 as the first time when $A(\boldsymbol{Dm},j) - (j-\boldsymbol{Dm})\boldsymbol{Ka} + j\boldsymbol{Ks} - S(0,j)$ exceeds $\boldsymbol{DmKs}$. Then we have

$$\begin{aligned}
j^0 = \min\{j : A(\boldsymbol{Dm},j) - (j-\boldsymbol{Dm})\boldsymbol{Ka} \\
+ j\boldsymbol{Ks} - S(0,j) \ge \boldsymbol{DmKs}\}.
\end{aligned} \tag{26}$$

Note that it is possible that $j^0 = \infty$ and $\Pr\{\max_{j\ge \boldsymbol{Dm}} \{A(\boldsymbol{Dm},j) - (j-\boldsymbol{Dm})\boldsymbol{Ks} + j\boldsymbol{Ks} - S(0,j)\} \ge \boldsymbol{DmKs}\}$. Define $j^0 \wedge j = \min\{j^0, j\}$ for $j \ge 0$. According to the optional stopping theorem, we have

$$\begin{aligned}
\mathrm{E}\left[M(0)\right] &\ge \mathrm{E}\left[M(j^0)\mathrm{l}_{j^0<j}\right] \\
&= \mathrm{E}\left[\boldsymbol{ha}\left(a\left(j^0\right)\right)\boldsymbol{hs}\left(s\left(k^0\right)\right)e^{\boldsymbol{DmKs}}\mathrm{l}_{j^0<j}\right] \\
&\ge He^{\boldsymbol{DmKs}}\Pr\left\{j^0 < j\right\} \\
&\ge He^{\boldsymbol{DmKs}}\Pr\left\{j^0 < \infty\right\},
\end{aligned} \tag{27}$$

where l_y denotes the indicator function. If the event y is true, $\mathrm{l}_y = 1$; otherwise, $\mathrm{l}_y = 0$.

Because $\mathrm{E}[M(0)] = E[\boldsymbol{ha}(a(0))]\mathrm{E}[\boldsymbol{hs}(s(0))]$, from (22), we have

$$\begin{aligned}
\Pr\left\{D(j) \ge \boldsymbol{Dm}\right\} &\le \Pr\left\{j^0 < \infty\right\} \\
&\le \frac{\mathrm{E}\left[\boldsymbol{ha}\left(a(0)\right)\right]\mathrm{E}\left[\boldsymbol{hs}\left(s(0)\right)\right]}{H}e^{-\hat{\theta}\boldsymbol{KsDm}}.
\end{aligned} \tag{28}$$

REFERENCES

[1] M. Zink, R. Sitaraman, and K. Nahrstedt, "Scalable 360 video stream delivery: Challenges, solutions, and opportunities," *Proc. IEEE*, vol. 107, no. 4, pp. 639–650, Apr. 2019.

[2] R. Atawia, H. Abou-zeid, H. S. Hassanein, and A. Noureldin, "Joint chance-constrained predictive resource allocation for energy-efficient video streaming," *IEEE J. Select. Areas Commun.*, vol. 34, no. 5, pp. 1389–1404, May 2016.

[3] R. Atawia, H. S. Hassanein, and A. Noureldin, "Energy-efficient predictive video streaming under demand uncertainties," in *Proc. IEEE Int. Conf. Commun.*, 2017, pp. 1–6.

[4] F. Li, P. Ren, and Q. Du, "Joint packet scheduling and subcarrier assignment for video communications over downlink OFDMA systems," *IEEE Trans. Veh. Technol.*, vol. 61, no. 6, pp. 2753–2767, Jul. 2012.

[5] M. Ismail, and W. Zhuang, "Mobile terminal energy management for sustainable multi-homing video transmission," *IEEE Trans. Wireless Commun.*, vol. 13, no. 8, pp. 4616–4627, Aug. 2014.

[6] E. Ozfatura, O. Ercetin, and H. Inaltekin, "Optimal network-assisted multiuser DASH video streaming," *IEEE Trans. Broadcast.*, vol. 64, no. 2, pp. 247–265, Jun. 2018.

[7] P. K. Barik, C. Singhal, and R. Datta, "Energy-efficient user-centric dynamic adaptive multimedia streaming in 5G cellular networks," in *Proc. Nat. Conf. Commun.*, 2020, pp. 1–6.

[8] L. Xu, A. Nallanathan, and X. Song, "Joint video packet scheduling, subchannel assignment and power allocation for cognitive heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1703–1712, Mar. 2017.

[9] L. Xu, Y. Zhou, P. Wang, and W. Liu, "Max-min resource allocation for video transmission in NOMA-based cognitive wireless networks," *IEEE Trans. Commun.*, vol. 66, no. 11, pp. 5804–5813, Nov. 2018.

[10] H. Abou-Zeid, and H. S. Hassanein, "Predictive green wireless access: Exploiting mobility and application information," *IEEE Wireless Commun.*, vol. 20, no. 5, pp. 92–99, Oct. 2013.

[11] R. Atawia, H. S. Hassanein, H. Abou-zeid, and A. Noureldin, "Robust content delivery and uncertainty tracking in predictive wireless networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 4, pp. 2327–2339, Apr. 2017.

[12] R. Atawia, H. S. Hassanein, N. Abu Ali, and A. Noureldin, "Utilization of stochastic modeling for green predictive video delivery under network uncertainties," *IEEE Trans. Green Commun. Netw.*, vol. 2, no. 2, pp. 556–569, Jun. 2018.

[13] R. Margolies *et al.*, "Exploiting mobility in proportional fair cellular scheduling: Measurements and algorithms," *IEEE/ACM Trans. Netw.*, vol. 24, no. 1, pp. 355–367, Feb. 2016.

[14] Z. Lu, and G. de Veciana, "Optimizing stored video delivery for wireless networks: The value of knowing the future," *IEEE Trans. Multimedia*, vol. 21, no. 1, pp. 197–210, Jan. 2019.

[15] R. Atawia, H. S. Hassanein, and A. Noureldin, "Robust long-term predictive adaptive video streaming under wireless network uncer- tainties," *IEEE Trans. Wireless Commun.*, vol. 17, no. 2, pp. 1374–1388, Feb. 2018.

[16] H. Abou-Zeid, H. S. Hassanein,, and S. Valentin, "Energy-efficient adaptive video transmission: Exploiting rate predictions in wireless networks," *IEEE Trans. Veh. Technol.*, vol. 63, no. 5, pp. 2013–2026, Jun. 2014.

[17] I. Triki, R. El-Azouzi, and M. Haddad, "NEWCAST: Anticipating resource management and QoE provisioning for mobile video streaming," in *Proc. IEEE World Wireless, Mobile Multimedia Netw. (WoWMoM)*, 2016, pp. 1–9.

[18] X. K. Zou *et al.*, "Can accurate predictions improve video streaming in cellular networks" in *Proc. Assoc. Comput. Machinery (ACM) HotMobile*, pp. 57–62, 2015.

[19] H. Farahat, R. Atawia, and H. S. Hassanein, "Robust proactive mobility management in named data networking under erroneous content prediction," in *Proc. IEEE Global Commun. Conf.*, 2017, pp. 1–6.

[20] H. Abou-Zeid, H. S. Hassanein,, and N. Zorba, "Long-term fairness in multi-cell networks using rate predictions," in *Proc. 7th IEEE Gulf Cooperation Council (GCC) Conf. Exhib.*, 2013, pp. 131–135.

[21] J. Guo, C. She, and C. Yang, "Predictive resource allocation with coarse-grained mobility pattern and traffic load information," in *Proc. IEEE Int. Conf. Commun.*, 2018, pp. 1–6.

[22] J. Guo, C. Yang and I. Chih-Lin, "Exploiting future radio resources with end-to-end prediction by deep learning," *IEEE Access*, vol. 6, pp. 75729–75747, 2018.

[23] H. Abou-Zeid, H. S. Hassanein, Z. Tanveer,, and N. AbuAli, "Evaluating mobile signal and location predictability along public transportation routes," in *Proc. IEEE Wireless Commun. Netw. Conf.*, 2015, pp. 1195–1200.

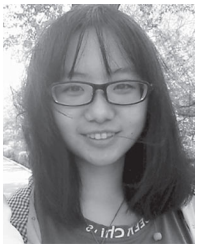
[24] E. Baccour, A. Erbad, A. Mohamed, K. Bilal, and M. Guizani, "Proactive video chunks caching and processing for latency and cost minimization in edge networks," in *Proc. IEEE Wireless Commun. Netw. Conf.*, 2019, pp. 1–7.

[25] T. Huang, R. Johari, and N. McKeown, "A buffer-based approach to rate adaptation: Evidence from a large video streaming service," in *Proc. ACM Special Interest Group Data Commun. (SIGCOMM)*, 2014, vol. 44, no. 4, pp. 187–198.

[26] H. Yu, F. Musumeci, J. Zhang, M. Tornatore, and Y. Ji, "Isolation-aware 5G RAN slice mapping over WDM metro-aggregation networks," *J. Lightw. Technol.*, vol. 38, no. 6, pp. 1125–1137, Mar. 2020.

[27] F. Capozzi, G. PSPro, L. A. Grieco, G. Boggia, and P. Camarda, "Downlink packet scheduling in LTE cellular networks: Key design issues and a survey," *IEEE Commun. Surv. Tuts.*, vol. 15, no. 2, pp. 678–700, Apr.–Jun. 2013.

[28] F. Poloczek, and F. Ciucu, "Service-martingales: Theory and applications to the delay analysis of random access protocols," in *Proc. IEEE Conf. Comput. Commun.*, 2015, pp. 945–953.
[29] Y. Jiang and Y. Liu, *Stochastic Network Calculus*. London, U.K.: Springer, 2008.
[30] X. Lu, Q. Ni, D. Zhao, W. Cheng, and H. Zhang, "Resource virtualization for customized delay- bounded QoS provisioning in uplink VMIMO- SC-FDMA systems," *IEEE Trans. Commun.*, vol. 67, no. 4, pp. 2951–2967, Apr. 2019.



Wanting Yang (Student Member, IEEE) received the B.S. degree in 2018 from the Department of Communications Engineering, Jilin University, Changchun, China, where she is currently working toward the Ph.D. degree with the College of Communication Engineering. Her research interests include wireless video transmission, learning, ultra-reliable, and low-latency communications.

Xuefen Chi received the B.Eng. degree in applied physics from the Beijing University of Posts and Telecommunications, Beijing, China, in 1984, and the M.S. and Ph.D. degrees from the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China, in 1990 and 2003, respectively. She was a Visiting Scholar with the Department of Computer Science, Loughborough University, Loughborough, U.K., in 2007 and the School of Electronics and Computer Science, University of Southampton, Southampton, U.K., in 2015. She is currently a Professor with the Department of Communications Engineering, Jilin University, Changchun, China. Her current research interests include machine type communications, indoor visible light communications, random access algorithms, delay-QoS guarantees, and queuing theory and its applications.

Linlin Zhao (Member, IEEE) received the B.Eng., M.S., and Ph.D. degrees from the Department of Communications Engineering, Jilin University, Changchun, China, in 2009, 2012, and 2017, respectively. From 2017 to 2019, she was a Postdoctor with the Department of Communications Engineering, Jilin University, and in 2019, she joined Jilin University. She is currently a Postdoctoral Research Fellow with the State Key Laboratory of Internet of Things for Smart City, University of Macau, China. Her current research interests include throughput optimal random access algorithms, resource allocation schemes, and delay and reliability analysis and optimization, especially for reliability analysis of ultra-reliable low-latency communications. Dr. Zhao was the recipient of the Best Ph.D. Thesis Award of Jilin University in 2017, and acquired the Macau Young Scholars Program in 2019.

Ruizhe Qi received the B.S. degree in communication engineering from the Xi'an University of Posts and Telecommunications, Xi'an, China, in 2018. She is currently working toward the M.S. degree with the College of Communication Engineering, Jilin University, Changchun, China. Her research interests include random access algorithms, ultra-reliable and low-latency communications, and delay-QoS guarantees.