Proactive VoD delivery pattern reconfiguration based on temporal-spatial channel prediction

1st Wanting Yang, 2st Xuefen Chi, 3rd Linlin Zhao
Department of Communications Engineering
Jilin University
Changchun, China
yangwt18@mails.jlu.edu.cn, chixf@jlu.edu.cn, zhaoll13@mails.jlu.edu.cn

***Abstract*—With the help of big data analytics, predictive resource allocation (PRA) techniques for video on demand (VoD) have been recognized as promising methods to save time-frequency resources, for a number of VoD packets can be transmitted in good channels in advance to avoid the predicted transmissions in bad channel conditions. With the increasing demands on a fantastic user quality of experience, a smooth playback and a low start-up delay are of equal importance to the emerging VoD with high fidelity, which inevitably leads to a critical delay requirement of VoD packets. However, the issue of resource estimation with quality of service (QoS) requirements is still an unsolved puzzle in PRA. In this paper, we propose a martingales-based physical resource block (PRB) abstraction method, where the random characteristics of the service process are embedded in the minimum PRB consumption. Based on the method, a proactive QoS-guaranteed reconfiguration algorithm is developed to optimize the multi-user delivery pattern applied in the prediction window, aiming to maximize spectrum efficiency. In this algorithm, since the delay sensitivity of VoD content transmitted in advance is dulled compared with the original VoD stream, we divide the original VoD slice into two sub-slices and derive a three-dimensional delivery pattern. The gain of resource saving and the capability of QoS guarantee brought by the reconfiguration have been demonstrated by the simulation results.**

***Index Terms*—channel state prediction, reconfiguration, delivery pattern, martingales, VoD, spectrum efficiency, delay-QoS**

I. Introduction

In the 5G/B5G era, new video on demand (VoD) services, such as 4K/8K ultra high definition videos and three dimensional holographic videos, bring us a more captivating quality of experience at the cost of much more bandwidth consumption [1]. A smooth video playback and high fidelity of each video frame are equally important to the new VoD. Further, the delay requirements of online VoD packet transmission become more stringent. Different from live videos, the VoD content can be pre-buffered in user equipments while users are watching videos online. This distinctive feature of VoD provides more space on the improvement of quality of service (QoS) and spectrum efficiency.

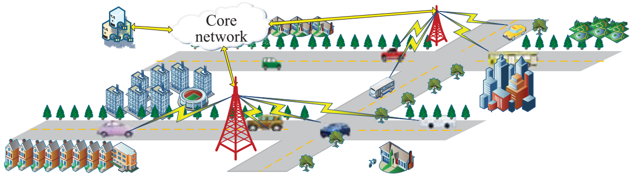

Fig. 1. Scenario diagram.

Recently, utilizing the big data analytics (BDA) to improve the performance of communication system is a new trend. In most cases, the user behaviors and channel states are predictable [2]. The prediction-based resource allocation has attracted growing attention in the VoD research field. In 2013, Hossam S. Hassanein team first proposed the predictive green wireless access (PreGWA) scheme based on the ideal data rate prediction in [3]. The main idea is to transmit as many VoD packets as possible to the user who is experiencing the better channel states and its effectiveness on resource saving was proved. In recent years, Hossam's team further put forward robust predictive resource allocation (PRA) algorithms under the imperfect prediction of link data rates, data rates required by users and network resources in [4] [5]. Some other teams applied the channel state prediction to the research on a finite-horizon proportional fair scheduling framework [6] or a cross-layer transport protocol to minimize the system utilization [7]. All of these methods could only avoid the video freezing if at all possible. The resource estimation method involved in the above work was roughly based on the predicted average data rate and could only ensure the data rate in a large timescale.

In practice, if a packet arrives later than its decoding time it has to be discarded [8], which results in a distorted video frame. And excessive packet delay violations leads to video freezings. Since the rates of packet departure are random in the timescale of transmission time interval (TTI), the existing PRA algorithms fail to guarantee the packet-level delay QoS and to provide high-fidelity VoD services. It is also worth noting that the VoD content transmitted in advance changes to be "dull" content with a loose delay requirement, which means that a packet can acquire the flexibility to determine its departure time and then contribute to higher spectrum efficiency. Yet the existing work adopted the same strategy to transmit these two kinds of packets, i.e., departure-on-time

This work was supported in part by the Natural Science Foundation of the Department of Science and Technology of Jilin Province named by *Research and Development of Medical Information Releasing System Based on MIMO OCC*, and in part by the National Natural Science Foundation of China under Grant 61801191.

978-1-7281-5089-5/20/$31.00 ©2020 IEEE

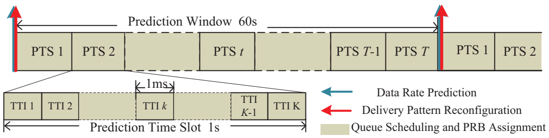

Fig. 2. The timescales in our work. The reconfiguration is performed at the beginning of the PW in a short time. The computing time of reconfiguration algorithm is ignored in the process of delay analysis.

packets and departure-in-advance packets.

For tackling such challenges, we propose a proactive reconfiguration algorithm for VoD delivery pattern to save more physical resource blocks (PRBs). The delivery pattern reconfiguration can be considered as the process of adjusting the numbers of the packets transmitted in each prediction time slot (PTS) within the prediction window (PW) for avoiding the hard transmissions. In this reconfiguration, one VoD slice is divided into two sub-slices. One is the delay-sensitive sub-slice, which carries the arriving-on-time packets. The other is the delay-insensitive sub-slice, which carries the arriving-in-advance packets. The original two-dimensional delivery pattern is converted into a three-dimensional delivery pattern, which contains the information about the number of packets that need to be transmitted to each user in each sub-slice in each PTS. The main contributions of this paper are summarized as follows:

1) We propose a martingales-based method to abstract the PRB consumption for a given data stream. Relying on the powerful martingale methodology, we establish the relationship between the packet arrival rate and the average PRB consumption for a certain data stream, where channel states and QoS requirements are considered.
2) We formulate the problem of delivery pattern reconfiguration as a matrix optimization, aiming to maximize spectrum efficiency with the provisions of 1) smooth playbacks of VoD, 2) high fidelity of each video frame, and 3) isolation of VoD slice.
3) We propose a heuristic algorithm to resolve the reconfiguration problem. Firstly, a QoS-guaranteed delivery pattern is derived by pre-checking and modifying the initial pattern. Then the pattern is optimized through a series of complementary optimizations, which are performed successively in a new PTS set built in each iteration.

The paper is organized as follows. Section II describes the system model. The problem formulation and the solution algorithm are presented in Section III. Section IV simulates and evaluates the performances of our proposed algorithm. Section V concludes the paper.

II. System Architecture

The system model is shown in Fig. 1. One BS provides VoD services for users in the vehicles moving along roads. We assume that user traces within a PW are predictable. One PW $\mathcal{T}$ contains T PTSs, which contains K TTIs. $\mathcal{M}$ is defined as the set of active users who are served by the BS with VoD services. In order to guarantee the isolation between the VoD slice and other slices, the number of PRBs allocated to the VoD users in one PTS is constrained to be the maximum value $\Phi^{\max}$. Predicted average reachable data rates of all users in each PTS are constructed as a matrix $\bar{R} = (\bar{r}_{i,t} : i \in \mathcal{M}, t \in \mathcal{T})$. $i \in \mathcal{M}$ denotes the index of user and $t \in \mathcal{T}$ denotes the index of PTS.

In our work, we assume that the VoD content transmitted from remote servers to the BS has been segmented into the proper packet size L to facilitate the next transmission from the BS to users. Each chunk with one second duration of the VoD content requested by user i is divided into $V_{i,t}$ packets in the transmission process. The packet rates required by users within a PW are constructed as $V = (V_{i,t} : i \in \mathcal{M}, t \in \mathcal{T})$, which is determined by the VoD content per second and the quality requested by users.

III. Problem Formulation and Solution Algorithm

A. Problem Formulation

During the delivery pattern reconfiguration, a resource estimation method is the basis and core of QoS guarantee. Before formulating the reconfiguration problem, the martingales-based PRB abstraction method is introduced in detail.

The PRB abstraction method proposed in this paper is to establish a relationship between the arrival process and the average PRB consumption (i.e., Φ) for a certain data stream, where QoS requirements and channel states are both considered.

For one data stream, the numbers of arrival packets and transmitted packets in its queue maintained in the BS in TTI k are defined as $a(k)$ and $s(k)$, respectively. The waiting delay of a data stream is defined as the staying time of the head packet in its queue, which is denoted by $D(k)$. We have $A(0,k) = \sum_{x=0}^{x=k} a(x)$ and $S(0,k) = \sum_{x=0}^{x=k} s(x)$. Then $D(k)$ can be expressed as

$$D(k) := \min \left\{ x \geq 0 \,|\, A(0, k-x) \leq S(0,k) \right\}. \tag{1}$$

For brevity, we write $X(n) := X(0,n)$ for both arrival and service processes. Considering the random characteristics in the network, the statistical delay-QoS adopted in our works is given as

$$\begin{aligned} \Pr\left\{D(k) \geq \mathit{Dm}\right\} &= \Pr\left\{A(k-\mathit{Dm}) \geq S(k)\right\} \\ &= \Pr\left\{ \max_{k \geq \mathit{Dm}} \left\{A(\mathit{Dm},k) - S(k)\right\} \geq 0 \right\}, \end{aligned} \tag{2}$$

where Dm denotes the delay bound of the packet and $A(\mathit{Dm},k)$ denotes the number of accumulated arrival packets from the TTI Dm to the TTI k. According to the definition of supermartingales in mathematics, we revisit the definitions of the arrival-martingale and the service-martingale for the arrival process and the service process of one data stream.

$Definition\ 1\ (Arrival-Martingales)$ [9]: For one data stream, its arrival process $a(k)$ admits arrival martingales if for every $\theta > 0$ there is a $\mathit{Ka}(\theta) > 0$ and a function ha :

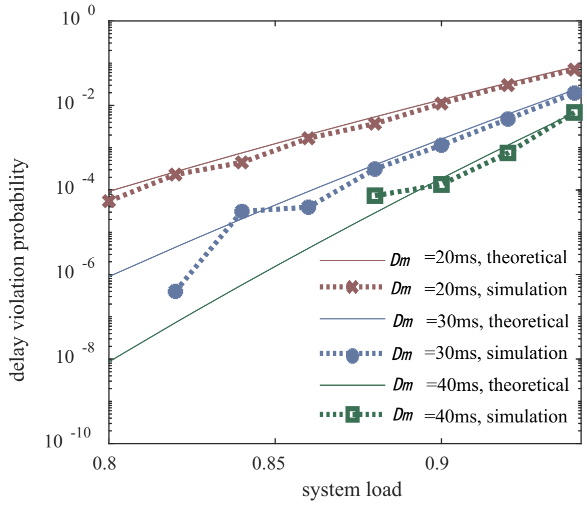

Fig. 3. Comparison between the theoretical and simulation results in terms of the delay-bound violation probability.

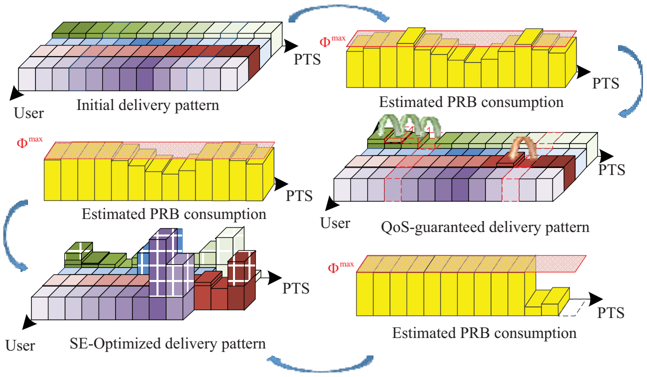
User
Initial delivery pattern
PTS
Φmax
Estimated PRB consumption
PTS
Φmax
Estimated PRB consumption
PTS
User
QoS-guaranteed delivery pattern
PTS
User
SE-Optimized delivery pattern
PTS
Φmax
Estimated PRB consumption
PTS

Fig. 4. Reconfiguration diagram. The color depth of these blocks represents the value of the predicted average reachable data rate. The height of these blocks represents the number of the VoD packets transmitted in each PTS. The height of the yellow blocks represents the number of the consumed PRB. The red plane represents the PRB consumption limitation. The blocks with the pure color represent the delay-sensitive sub-slice and the gridding blocks represent the delay-insensitive sub-slice. The minimum request unit is not considered in our paper.

$\mathrm{rng}(a(k)) \to \mathrm{R}^{+}$ such that the process

$$\mathit{Ma}(k) = \mathit{ha}(a(k))e^{\theta(A(k)-k\mathit{Ka}(\theta))}, k \geq 0 \tag{3}$$

is a supermartingale. $\mathrm{rng}(\cdot)$ stands for the range operator.

Definition 2 (Service − Martingales) [9]: For one data stream, its service process admits service-martingales if for every $\theta > 0$ there is a $\mathit{Ks}(\theta) > 0$ and a function $\mathit{hs} : \mathrm{rng}(s^{q}_{i,t}(k)) \to \mathrm{R}^{+}$ such that the process

$$\mathit{Ms}(k) = \mathit{hs}(s(k))e^{\theta(k\mathit{Ks}(\theta)-S(k))}, k \geq 0 \tag{4}$$

is a supermartingale.

For one data stream, $\mathit{Ma}(k)$ and $\mathit{Ms}(k)$ characterize the stochastic features of its arrival (service) process from the perspective of supermartingales with parameters $\mathit{ha}(a(k))$ and $\mathit{Ka}(\theta)$ [$\mathit{hs}(s(k))$ and $\mathit{Ks}(\theta)$]. Relaying on arrival (service)-martingales and the analysis of (2), we can construct a supermartingales for the delay of the data stream and then yield the delay-bound violation probability via the optional stopping theorem for supermartingales. The delay-bound violation probability is given by Theorem 1 [9].

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

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

If the arrival process and service process respectively admit the arrival-martingale and service-martingale for this data stream, then its delay-bound violation probability holds for

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

where

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

Proof: Please see Appendix A.

In this paper, we sever one VoD stream into two individual data streams. Accordingly the VoD slice is divided into two sub-slices. The type of the sub-slices is denoted by $q \in \mathcal{Q}$. $q = 1$ represents the delay-sensitive sub-slice, which carries the arriving-on-time packets and $q = 2$ represents the delay-insensitive sub-slice, which carries the arriving-in-advance packets. The arrival process of the VoD stream in sub-slice q of user i in one PTS is treated as a constant process. The number of the arrival packets in the queue of the sub-slice in TTI k is calculated as $a^{q}_{i,t}(k) = P^{q}_{i,t}/\tau_{PTS}$, where $P^{q}_{i,t}$ denotes the number of VoD packets to be transmitted through sub-slice q in PTS t. For brevity, $a^{q}_{i,t}(k)$ is denoted as $C^{q}_{i,t}$. And the accumulated number of the arrival packets can be expressed as

$$A^{q}_{i,t}(0,k) = kC^{q}_{i,t}. \tag{7}$$

So we take $\mathit{Ka}^{q}_{i,t}(\theta) = C^{q}_{i,t}$, and $\mathit{ha}(a^{q}_{i,t}(k)) = 1$, and the arrival process of one VoD stream in sub-slice q admits the arrival martingale $\mathit{Ma}^{q}_{i,t}(k)$ denoted as a constant. For the service process, considering the impacts of the time-varying channel states and the downlink scheduling policy, we model the packet departure process $s^{q}_{i,t}(k)$ across TTIs within one PTS t as an independent identical distributed (IID) Poisson distribution with the parameter $\lambda^{q}_{i,t}$. The parameters of the service-martingale for IID service processes have been investigated in [9]. The results we utilize in this paper are given as follow:

1) $\mathit{hs}(s(k)$ is a constant.
2) $\mathrm{E}[e^{-\theta s(k)}] = e^{-\theta \mathit{Ks}(\theta)}$.

So the $H^{q}_{i,t} = 1$ and the $\mathit{Ks}^{q}_{i,t}(\theta^{q}_{i,t})$ in our service-martingales can be expressed as

$$\mathit{Ks}^{q}_{i,t}(\theta^{q}_{i,t}) = -\frac{\log \mathrm{E}[e^{-\theta^{q}_{i,t}-1}]}{\theta^{q}_{i,t}} = -\frac{\lambda^{q}_{i,t}(e^{-\theta^{q}_{i,t}} - 1)}{\theta^{q}_{i,t}}. \tag{8}$$

According to the definition (5), $\hat{\theta}$ can be calculated as the solution of $\mathit{Ks}^q_{i,t}(\theta^q_{i,t}) = \mathit{Ka}^q_{i,t}(\theta^q_{i,t})$, which can be rewritten as

$$C^q_{i,t} = -\frac{\lambda^q_{i,t}\left(e^{-\hat{\theta}^q_{i,t}} - 1\right)}{\hat{\theta}^q_{i,t}}. \tag{9}$$

According to (6), the delay-bound constraint for one sub-slice can be expressed as below

$$e^{-\hat{\theta}^q_{i,t}\mathit{Ks}^q_{i,t}(\hat{\theta}^q_{i,t})\mathit{Dm}^q_i} \le \varepsilon_i, \tag{10}$$

where ε_i represents the tolerance of user i to the video distortion. In order to maximize the spectrum efficiency, we focus on the upper bound of the inequality (10). Combining (9) and (10), we can acquire the relationship between $C^q_{i,t}$ and $\lambda^q_{i,t}$. We define the ratio of $C^q_{i,t}$ and $\lambda^q_{i,t}$ as the system load. We simulate the VoD delivery process for ten PWs via MATLAB, during which the delay values are recorded to calculate the delay violation probability according to different delay bound as simulation value. The comparison of the theoretical results and simulation results is shown in Fig. 3. It demonstrates that the data streaming with stricter delay bound need stronger service capability. Then we translate the packet-level average departure rate $\lambda^q_{i,t}$ into the bit-level rate $\hat{\lambda}_{i,t} = \lambda_{i,t}L$. Assume that frequency division duplex is supported by the radio interface. According to predicted average reachable data rate $\bar{r}_{i,t}$, the average number of the consumed PRBs $\phi^q_{i,t}$ at each TTI can be expressed as

$$\phi^q_{i,t} = \frac{\hat{\lambda}^q_{i,t}}{\bar{r}_{i,t}B_{\mathrm{PRB}}}, \tag{11}$$

where B_{PRB} denotes the size of one PRB, which spans over 180 kHz in frequency domain and over 0.5ms in time domain. Further, we estimate the total PRB consumption $\Phi^q_{i,t} = \phi^q_{i,t}\tau_{PTS}$. So far the relationship between $P^q_{i,t}$ and $\Phi^q_{i,t}$ is established by the martingales-based PRB consumption abstraction method. Based on this method, aiming to maximizing spectrum effectiveness, we formulate the reconfiguration problem as follows

$$\max_{P} \frac{\sum_{t=1}^{|\mathcal{T}|}\sum_{i=1}^{|\mathcal{M}|} V_{i,t}}{\sum_{t=1}^{|\mathcal{T}|}\sum_{i=1}^{|\mathcal{M}|}\sum_{q=1}^{2}\Phi^q_{i,t}} \tag{12}$$

subject to

$$\Pr\left\{D^q_{i,t}(k) \ge \mathit{Dm}^q_i\right\} \le \varepsilon_i, \forall i \in \mathcal{M}, \forall t \in \mathcal{T}, \forall q \in \mathcal{Q} \tag{12a}$$

$$\sum_{i=1}^{|\mathcal{M}|}\sum_{q=1}^{|\mathcal{Q}|}\Phi^q_{i,t} \le \Phi^{\max}, \forall t \in \mathcal{T} \tag{12b}$$

$$P^q_{i,t} \ge 0, \forall i \in \mathcal{M}, \forall t \in \mathcal{T}, \forall q \in \mathcal{Q} \tag{12c}$$

$$\sum_{t'=1}^{t}\sum_{q=1}^{|\mathcal{Q}|}P^q_{i,t'} \ge \sum_{t'=1}^{t} V_{i,t'}, \forall i \in \mathcal{M}, \forall t \in \mathcal{T} \tag{12d}$$

$$\sum_{t=1}^{|\mathcal{T}|}\sum_{q=1}^{|\mathcal{Q}|}P^q_{i,t} = \sum_{t=1}^{|\mathcal{T}|} V_{i,t}, \forall i \in \mathcal{M} \tag{12e}$$

The constraint (12a) ensures the statistical delay-QoS of each sub-slice during the spectrum efficiency optimization process. The constraint (12b) represents the limitation of the total PRB consumption of all users. Constraint (12c) ensures nonnegative $P^q_{i,t}$. Constraint (12d) guarantees that the accumulated online VoD content up to PTS t can support a smooth playback for user i. Constraint (12e) implies that the total VoD content transmitted to each user remains unchanged in the reconfiguration.

Algorithm 1 QoS provisioning algorithm

```
Input:
    delivery pattern $P_{|\mathcal{M}|\times|\mathcal{T}|\times|\mathcal{Q}|}$, predicted average reachable data rates $\bar{R}_{|\mathcal{M}|\times|\mathcal{T}|}$, PRB consumption limitation $\Phi^{\max}$, PRB consumption $\Phi_t$, PTS $t$
Output:
    delivery pattern $P_{|\mathcal{M}|\times|\mathcal{T}|\times|\mathcal{Q}|}$
while $\Phi_t > \Phi^{\max}$ do
    Select the user set $\mathcal{M}'$ in which the users are driving away from the BS.
    if $\mathcal{M}' \neq \emptyset$ then
        $i^* := \arg\max_{i\in\mathcal{M}'}\{\bar{r}_{i,t}\}$ //select the user with the highest average data rate
    else
        $i^* := \arg\max_{i\in\mathcal{M}}\{\bar{r}_{i,t}\}$
    end if
    $\delta\Phi_t \leftarrow \Phi_t - \Phi^{\max}$
    Calculate $\delta P^1_{i^*,t}$ according to $\delta\Phi_t$ // calculate the number of the packets of user $i^*$ which need to be transmitted in PTS $t-1$
    if $\delta P^1_{i^*,t} \le P^1_{i^*,t}$ then
        $P^1_{i^*,t} \leftarrow P^1_{i^*,t} - \delta P^1_{i^*,t}$; $\Phi^1_{i^*,t} \leftarrow \Phi^1_{i^*,t} - \delta\Phi_t$
    else
        $\delta P^1_{i^*,t} \leftarrow P^1_{i^*,t}$; $P^1_{i^*,t} \leftarrow 0$; $\Phi^1_{i^*,t} \leftarrow 0$
        $\delta\Phi \leftarrow \Phi^1_{i^*,t}$
    end if
    if $t > 1$ then
        $P^1_{i^*,t-1} \leftarrow P^1_{i^*,t-1} + \delta P^1_{i^*,t}$
        $\Phi_t \leftarrow \Phi_t - \delta\Phi_t$
    else
        $\Phi_t \leftarrow \Phi^{\max}$
    end if
end while
return $P_{|\mathcal{M}|\times|\mathcal{T}|\times|\mathcal{Q}|}$
```

B. Solution Algorithm

The reconfiguration problem can be regarded as a high-dimension matrix problem. Since the delay-QoS constraint is related to both (9) and (10), constraint (12a) is non-linear. To resolve the high-dimension non-convex optimization problem,

Algorithm 2 SE optimization algorithm

```
Input:
    delivery pattern $P_{|\mathcal{M}|\times|\mathcal{T}|\times|\mathcal{Q}|}$, predicted average reachable data rates $\bar{R}_{|\mathcal{M}|\times|\mathcal{T}|}$, PRB consumption $\Phi_t$, PRB consumption limitation $\Phi^{\max}$, PTS $t^i_{\max}$, PTS set $\mathcal{T}_i$
Output:
    delivery pattern $P_{|\mathcal{M}|\times|\mathcal{T}|\times|\mathcal{Q}|}$
t ← $t^i_{\max}$
if $\Phi_t < \Phi^{\max}$ then
    $\Delta\Phi_t = \Phi^{\max} - \Phi_t$
    Calculate $\Delta P^2_{i,t}$ according to $\Delta\Phi_t$
    Construct $\mathcal{T}_i^C = \left\{h \in \mathcal{T}_i \,\middle|\, h > t, P^1_{i,t} > 0\right\}$ //the set of the candidate PTSs in which the number of VoD packets to be transmitted can be reduced
    $\Delta P^2_{i,t} = \min\left\{\Delta P^2_{i,t}, \sum_{h\in\mathcal{T}_i^C} P^1_{i,t}\right\}$
    $P^2_{i,t} \leftarrow P^2_{i,t} + \Delta P^2_{i,t}$
    while $\Delta P^2_{i,t} > 0$ do
        $h^l = \arg\min_{h\in\mathcal{T}_i^C} \{\bar{r}_{i,h}\}$ // choose one PTS from $\mathcal{T}_i^C$ with the worst the channel state
        $P_0 \leftarrow P^1_{i,h^l}$
        $P^1_{i,h^l} \leftarrow \max\left\{P^1_{i,h^l} - \Delta P^2_{i,t}, 0\right\}$; Update $\mathcal{T}_i^C$
        $\Delta P^2_{i,t} \leftarrow \max\left\{\Delta P^2_{i,t} - P_0, 0\right\}$
    end while
end if
return $P_{|\mathcal{M}|\times|\mathcal{T}|\times|\mathcal{Q}|}$
```

we propose a low complexity heuristic algorithm consisting of two phases: QoS provisioning and SE optimization, as shown in Fig. 4.

First, the delivery pattern $P = \left(P^q_{i,t} : i \in \mathcal{M}, t \in \mathcal{T}, q \in \mathcal{Q}\right)$ is initialized according to the number $V_{i,t}$ of packets the VoD content contains per PTS, i.e., $P^1_{i,t} = V_{i,t}$. Then the PRB consumption $\Phi^1_{i,t}$ of the user i in each PTS t is estimated by the proposed PRB consumption abstraction method according to $P^1_{i,t}$, $\bar{r}_{i,t}$ and the delay-QoS requirement. In the QoS provisioning phase, pre-check the total PRB consumption $\Phi_t = \sum_{i=1}^{|\mathcal{M}|} \Phi^1_{i,t}$ in each PTS from the end to the beginning of the PW. If a PTS t, in which $\Phi_t > \Phi^{\max}$, is found out, the QoS provisioning algorithm is performed, which is shown in Algorithm 1. The main idea of this algorithm is to pick one or more users and transmit all or part of $P^q_{i,t}$ packets in the previous PTS $t-1$ until the constraint ($12b$) is satisfied. In order to save PRBs, the picking rule for the users whose delivery pattern need to be modified in PTS t is summarized as follows:

- Among the users in $\mathcal{M}$, the user driving away from the BS is picked first.
- Among the users driving in the same direction, the user experiencing the highest average reachable data rate is picked first.

At the beginning of the SE optimization phase, each user maintains an own PTS set $\mathcal{T}_{\max} = \left\{t^i_{\max} \,\middle|\, t^i_{\max} = \arg\max_{t\in\mathcal{T}_i} \{\bar{r}_{i,t}\}, \forall i \in \mathcal{M}\right\}$ of PTSs is built. For each PTS $t^i_{\max}$ in $\mathcal{T}_{\max}$, the delivery pattern of the user i is chosen to perform the optimization algorithm, and the PTS $t^i_{\max}$ is removed from $\mathcal{T}_i$ at the same time. The SE optimization algorithm concerns two aspects: 1) to determine how many non-required VoD packets can be transmitted in the current PTS t, and 2) to update the delivery pattern in the following PTSs, determining in which PTSs the number of the transmitted packets can be reduced. The specific process is shown in Algorithm. 2. After one complementary optimization process, there is no VoD packets to be transmitted after the picked PTS within its own $\mathcal{T}_i$, the overall SE optimization process of user i ends and user i is extracted from the user set $\mathcal{M}$. After each iteration, the PTS group $\mathcal{T}_{\max}$ is updated. When all the users' optimizing processes end, our final VoD delivery pattern will have been developed.

IV. Performance Evaluations

We focus on a single cell in which there are four vehicles with the users watching online VoD moving along a road that crosses the cell. The radius of the cell is set to be 600m. Here, we treat the Shannon capacity calculated based on the average channel gain as the ideal predicted average reachable data rate. The average channel gain at the area with the distance d to the BS is denoted as $g(d) = d^{-l}$, where l denotes the path-loss exponent and is set to be 2. At the beginning of the prediction window, the four vehicles are 100m, 200m, 300m, 350m away from the BS at the speed of 44km/h, 60km/h, 36km/h, 28km/h heading to the BS, respectively. Suppose the users in different vehicles request different quality video with the streaming rate 40Mbps, 50Mbps, 30Mbps, 20Mbps, respectively. The other simulation parameters are listed in Table. I and all results are obtained via the MATLAB simulations.

The reconfiguration process consists two phases: QoS provisioning and optimization. The results of the reconfiguration with the bandwidth limitation 24.3MHz for the four users are shown in Fig. 5(b)-Fig. 5(e). As shown in the green stem diagram in Fig. 6, the total PRB consumption exceeds

TABLE I
Simulation Parameter

Parameters	Value
BS transmit power	43 dBm
noise power	-55 dBm
packet size	1000 bytes
delay bound (sub-slice1)	40ms [8]
toleration to video distortion	10^{-4}
prediction window	60s
prediction time slot	1s
TTI duration	1ms

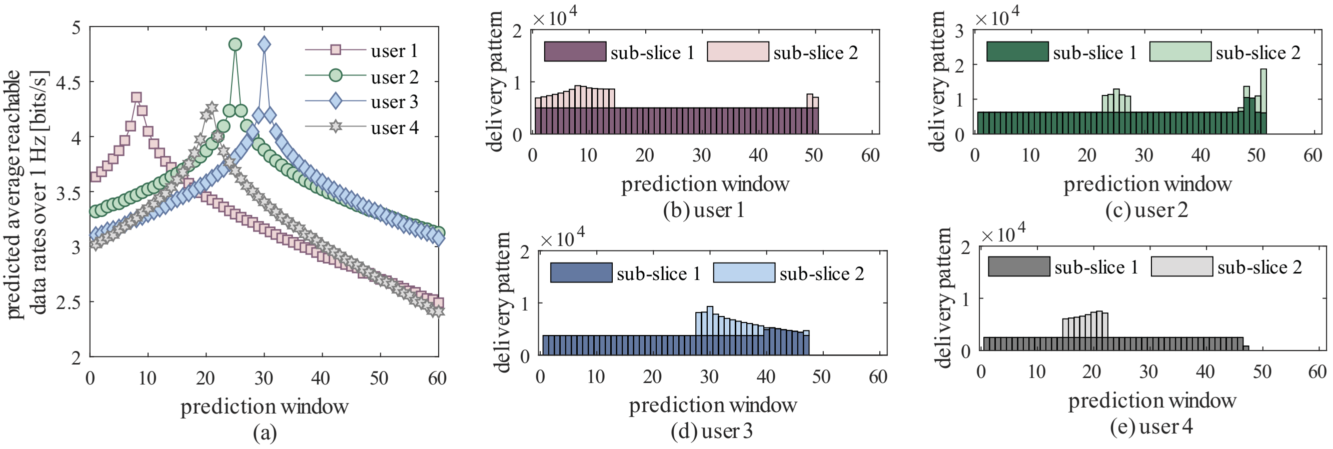

Fig. 5. Predicted reachable data rates over the bandwidth with 1 Hz and VoD delivery pattern after reconfiguration within 24.3MHz bandwidth.

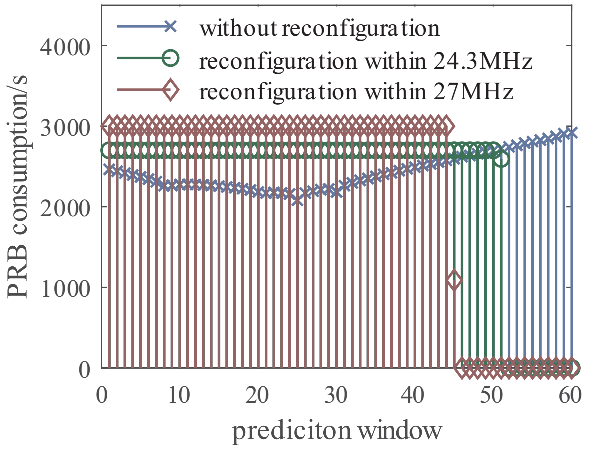

Fig. 6. Comparison of the number of the consumed PRBs in each PTS in the three cases.

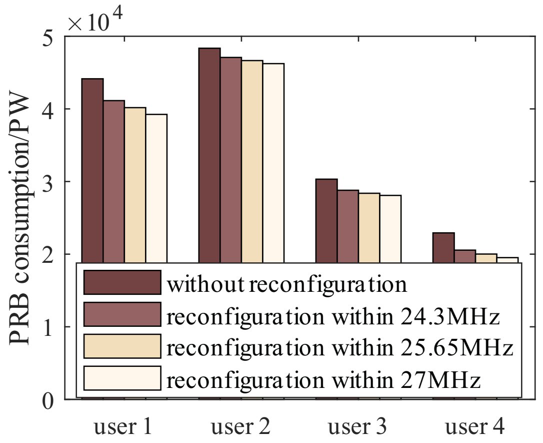

Fig. 7. Comparison of the number of the consumed PRBs for users in different cases.

the consumption limitation within 24.3MHz in the 51th-60th PTSs. From Fig. 5(a) we can see that, at end of the PW, all the users drive away from the BS and user 2 experiences the highest average data rate. Hence, at the beginning of the QoS provisioning phase, the VoD content of user 2 in the 51th-60th PTSs is transmitted in advance sequentially.

Whereas in 28th-50th PTSs the average data rate of user 3 is higher than user 2, BS turns to transmit the VoD content of user 3 in advance to meet PRB consumption limitation. And the QoS provisioning process ends in the 40th PTS. Compared the Fig. 5(a) and the four figures on the right, we can see that when the four users experience good channel states, the BS transmits the VoD content which should be transmitted in the end of the PW with low data rate in advance. The transmission for all the users ends in the 51th PTS. In the following PTSs, the full bandwidth is available. And the BS can provide the bandwidth for other services.

The comparison of the total PRB consumption in each PTS after the reconfiguration is shown in Fig. 6. It is easy to see that the delivery pattern reconfiguration with a larger bandwidth limitation ends earlier. And if without a reconfiguration, the BS cannot support all the users when the bandwidth limitation is less than 29MHz. The comparison of the total PRB consumption of each user in different cases is shown in Fig. 7. Compared with the PRB consumption without a reconfiguration, the PRB consumption of each user is reduced in different degrees after the reconfiguration. The reduction depends on both the user traces and the value of bandwidth limitation. As can be seen in Fig. 7, with the increase of bandwidth limitation, for most users the PRB consumption is reduced obviously, whereas for others, the PRB consumption shows no obvious changes.

V. Conclusion

In this paper, with the help of BDA, we presented a proactive reconfiguration algorithm for VoD services. The reconfiguration algorithm is performed at the beginning of the PW according to the predicted average reachable data rates. In order to guarantee the QoS of users, we put forward a martingales-based PRB consumption abstraction method, where the impacts of the downlink scheduling policy and channel states on the packet departure rate are considered. To derive a spectrum-efficient delivery pattern in a short time, we proposed a heuristic algorithm consisting two phases: modification and optimization. In the optimization phase, we divided the original VoD slice into two sub-slices, carrying arrival-on-time packets and arrival-in-advance packets, respectively. The PRB saving gain and the capability of QoS guarantee of resource allocation scheme have been demonstrated in the simulations.

Appendix A
Theorem 1

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

$$\begin{aligned}
&\Pr\{D(k) \geq \mathsf{Dm}\} \\
&\quad \leq \Pr\left\{ \max_{k \geq \mathsf{Dm}} \{A(\mathsf{Dm},k) - S(k)\} \geq 0 \right\} \\
&\quad \leq \Pr\left\{ \max_{k \geq \mathsf{Dm}} \{A(\mathsf{Dm},k) - S(k) \right. \\
&\qquad\qquad \left. - (k - \mathsf{Dm})\left(\mathsf{Ks}\left(\hat{\theta}\right) - \mathsf{Ks}\left(\hat{\theta}\right)\right)\right\} \geq 0 \Big\} \\
&\leq \Pr\left\{ \max_{k \geq \mathsf{Dm}} \left\{ A(\mathsf{Dm},k) - (k - D^{\max})\,\mathsf{Ka}\left(\hat{\theta}\right) \right.\right. && (13) \\
&\qquad\qquad \left.\left. + k\mathsf{Ks}\left(\hat{\theta}\right) - S(k) \right\} \geq \mathsf{DmKs}\left(\hat{\theta}\right) \right\}. && (14)
\end{aligned}$$

For one data stream, construct a process as follows:

$$\begin{aligned}
M(k) &= \mathsf{ha}(a(k)) \\
&\times \mathsf{hs}(s(k))e^{\left\{A(\mathsf{Dm},k)-(k-\mathsf{Dm})\mathsf{Ka}\left(\hat{\theta}\right)+k\mathsf{Ks}\left(\hat{\theta}\right)-S(k)\right\}}.
\end{aligned} \tag{15}$$

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

$$M(k) = \mathsf{Ma}(k)\mathsf{Ms}(k)e^{\mathsf{DmKa}\left(\hat{\theta}\right)}. \tag{16}$$

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

$$\begin{aligned}
&\mathrm{E}\left[M(k+1)\,|M(1),\cdots,M(k)\right] \\
&= \mathrm{E}\left[\mathsf{Ma}(k+1)\mathsf{Ms}(k+1)e^{\mathsf{DmKa}\left(\hat{\theta}\right)}\,|M(1),\cdots,M(k)\right] \\
&= e^{\mathsf{Dm}}\mathsf{Ka}\left(\hat{\theta}\right)\mathrm{E}\left[\mathsf{Ma}(k+1)\,|\mathsf{Ma}(1),\cdots,\mathsf{Ma}(k)\right] \\
&\qquad\qquad \times \mathrm{E}\left[\mathsf{Ms}(k+1)\,|\mathsf{Ms}(1),\cdots,\mathsf{Ms}(k)\right] \\
&\leq \mathsf{Ma}(k)\mathsf{Ms}(k)e^{\mathsf{DmKa}\left(\hat{\theta}\right)} \\
&= M(k).
\end{aligned} \tag{17}$$

So the process $M(k)$ is a supermartingales. Define the stopping time k^0 as the first time when $A(\mathsf{Dm},k) - (k - \mathsf{Dm})\mathsf{Ka}\left(\hat{\theta}\right) + k\mathsf{Ks}\left(\hat{\theta}\right) - S(k)$ exceed $\mathsf{DmKs}\left(\hat{\theta}\right)$. Then we have

$$\begin{aligned}
k^0 = \min\Big\{k : A(\mathsf{Dm},k) - (k - \mathsf{Dm})\mathsf{Ka}\left(\hat{\theta}\right) \\
+ k\mathsf{Ks}\left(\hat{\theta}\right) - S(k) \geq \mathsf{DmKs}\left(\hat{\theta}\right)\Big\}.
\end{aligned} \tag{18}$$

Note that it is possible that $k^0 = \infty$ and $\Pr\left\{k^0 < \infty\right\} = \Pr\left\{ \max_{k \geq \mathsf{Dm}} \left\{ A(\mathsf{Dm},k) - (k - \mathsf{Dm})\,\mathsf{Ka}\left(\hat{\theta}\right) + k\mathsf{Ks}\left(\hat{\theta}\right) - S(k)\right\} \geq \mathsf{DmKs}\left(\hat{\theta}\right)\right\}$ Define $k^0 \wedge k = \min\left\{k^0, k\right\}$ for $k \geq 0$. According to the optional stopping theorem, we have

$$\begin{aligned}
\mathrm{E}\left[M(0)\right] &\geq \mathrm{E}\left[M(k^0)\mathrm{l}_{k^0<k}\right] \\
&= \mathrm{E}\left[\mathsf{ha}\left(a\left(k^0\right)\right)\mathsf{hs}\left(s\left(k^0\right)\right)e^{\mathsf{DmKs}(\hat{\theta})}\mathrm{l}_{k^0<k}\right] \\
&\geq He^{\mathsf{DmKs}(\hat{\theta})}\Pr\left\{k^0 < k\right\} \\
&\geq He^{\mathsf{DmKs}(\hat{\theta})}\Pr\left\{k^0 < \infty\right\},
\end{aligned} \tag{19}$$

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

Because $\mathrm{E}\left[M(0)\right] = \mathrm{E}\left[\mathsf{ha}\left(a(0)\right)\right]\mathrm{E}\left[\mathsf{hs}\left(s(0)\right)\right]$, from (14), we have

$$\begin{aligned}
\Pr\{D(k) \geq \mathsf{Dm}\} &\leq \Pr\left\{k^0 < \infty\right\} \\
&\leq \frac{\mathrm{E}\left[\mathsf{ha}\left(a(0)\right)\right]\mathrm{E}\left[\mathsf{hs}\left(s(0)\right)\right]}{H}e^{-\hat{\theta}\mathsf{Ks}(\hat{\theta})\mathsf{Dm}}.
\end{aligned} \tag{20}$$

References

[1] M. Zink, R. Sitaraman and K. Nahrstedt, ”Scalable 360° Video Stream Delivery: Challenges, Solutions, and Opportunities,” in Proceedings of the IEEE, vol. 107, no. 4, pp. 639-650, April 2019.

[2] M. Elazab, A. Noureldin and H. S. Hassanein, ”Integrated Cooperative Localization for Connected Vehicles in Urban Canyons,” 2015 IEEE Global Communications Conference (GLOBECOM), San Diego, CA, 2015, pp. 1-6.

[3] 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.

[4] R. Atawia, H. S. Hassanein, H. Abou-zeid and A. Noureldin, ”Robust Content Delivery and Uncertainty Tracking in Predictive Wireless Networks,” in IEEE Transactions on Wireless Communications, vol. 16, no. 4, pp. 2327-2339, April 2017.

[5] R. Atawia, H. S. Hassanein, N. Abu Ali and A. Noureldin, ”Utilization of Stochastic Modeling for Green Predictive Video Delivery Under Network Uncertainties,” in IEEE Transactions on Green Communications and Networking, vol. 2, no. 2, pp. 556-569, June 2018.

[6] R. Margolies et al., ”Exploiting Mobility in Proportional Fair Cellular Scheduling: Measurements and Algorithms,” in IEEE/ACM Transactions on Networking, vol. 24, no. 1, pp. 355-367, Feb. 2016.

[7] Z. Lu and G. de Veciana, ”Optimizing Stored Video Delivery for Wireless Networks: The Value of Knowing the Future,” in IEEE Transactions on Multimedia, vol. 21, no. 1, pp. 197-210, Jan. 2019.

[8] F. Li, P. Ren and Q. Du, ”Joint Packet Scheduling and PRB Assignment for Video Communications Over Downlink OFDMA Systems,” in IEEE Transactions on Vehicular Technology, vol. 61, no. 6, pp. 2753-2767, July 2012.

[9] F. Poloczek and F. Ciucu, ”Service-martingales: Theory and applications to the delay analysis of random access protocols,” 2015 IEEE Conference on Computer Communications (INFOCOM), Kowloon, 2015, pp. 945-953.