Energy Efficient Predictive Resource Allocation for VoD and Real-time Services

Changyang She and Chenyang Yang

Abstract

This paper studies how to exploit the predicted information to maximize energy efficiency (EE) of a system supporting hybrid services. To obtain an EE upper bound of predictive resource allocation, we jointly optimize resource allocation for video on-demand (VoD) and real-time (RT) services to maximize EE by exploiting perfect future large-scale channel gains. We find that the EE-optimal predictive resource allocation is a two-timescale policy, which makes a resources usage plan at the beginning of prediction window and allocates resources in each time slot. Analysis shows that if there is only VoD service, predicting large-scale channel gains and distribution of small-scale channel gains are necessary to achieve the EE upper bound. If there is only RT service, future large-scale channel gains cannot help improve EE. However, if there are both VoD and RT services, predicting large-scale channel gains of both kinds of users are helpful. A low-complexity is proposed, which is robust to prediction errors. Simulation results show that the optimal policy is superior to the relevant counterparts, and the heuristic policy can achieve higher EE than the optimal policy when the large-scale channel gains are inaccurate.

Index Terms

Energy efficiency, predictive resource allocation, VoD services, real-time services

I. INTRODUCTION

Energy efficiency (EE) is a key performance metric for the fifth generation (5G) cellular networks [2,3]. Inspired by the finding in [4] that user mobility is highly predictable, improving EE by exploiting predicted information has drawn significant attention as the smart phone popularizes and big data analytics flourishes [5,6]. With predicted trajectory of a mobile user [7], EE can be boosted by sending more data to the user when it is close to a base station (BS).

A part of this work was presented in IEEE/CIC ICCC 2015 [1].

Changyang She and Chenyang Yang are with the School of Electronics and Information Engineering, Beihang University, Beijing 100191, China (email:{cyshe,cyyang}@buaa.edu.cn).
5G networks are expected to support diverse services with different quality-of-service (QoS) provision [8]. As shown in [9], more than half of the overall data traffic is mobile video in 2014, and the percentage is anticipated to become 72% by 2019. In order to satisfy the users experience of video on-demand (VoD) services, the video quality and playback interruption are two important metrics [10]. In future cellular networks, there still exist many real-time (RT) services such as video conference and voice over IP that require stringent QoS [8], which is characterized by a delay bound and a small delay bound violation probability for packets [11].

Three kinds of services have been considered in existing predictive resource allocations.

The first kind is RT services (see [12, 13] and references therein). To improve the admission level QoS, the cell-level mobility prediction is exploited, which has been widely studied in existing literature (e.g., [14] and the references in [6]). By predicting the future handoff time and the cell that a RT user will access to, the bandwidth at the next BS was reserved for the user [12], and a call admission control scheme was proposed in [13]. These works aim to tradeoff the handoff call dropping rate and the new call blocking rate, and implicitly assume that fixed bandwidth is reserved for each user to ensure the QoS.

The second kind is VoD services [15–21]. To improve the packet level QoS, the trajectory or rate prediction is exploited, which has been investigated in [7, 18] and the references in [6]. With the trajectory prediction and the help of a radio map, the future large scale channel gains can be predicted. Based on the predicted large scale channel gain or data rate, resource allocation among future time slots was studied in [15, 16], either to minimize video degradation or to maximize EE. In [18], the average rates at different locations measured in the past days were used as the average rate prediction with the help of user trajectory, with which the QoS improvement of video streaming was demonstrated. In [17], a practical two-timescale resource allocation was proposed. In the first timescale, time resource allocation is optimized based on the rate prediction, while in the second timescale, subcarriers are allocated based on the small-scale channel gains. In [19], future data rate was allocated to minimize the usage of resources, where an iterative allocation algorithm was proposed to account for the uncertainty on predicting user locations and number of users in a cell. Considering that future data rate cannot be predicted without error, a robust predictive resource allocation was proposed in [20], where the prediction errors on future rate is modelled as Gaussian distribution. A closed-form relation between prediction errors on data rate and probabilistic QoS guarantees was obtained in [21]. These studies assume that the future
data rate is predictable, but the small-scale channel fading is not considered. However, the data rate of wireless link highly depends on small-scale channel fading, simply ignoring such fading in predictive resource allocation may lead to the degradation of QoS.

The third kind is the delay tolerant services such as file downloading [15, 22, 23]. By exploiting perfect instantaneous data rate prediction, time resource allocation among multiple users was optimized in [15], either to maximize the total throughput over a prediction window or to maximize the minimal throughput. By exploiting future large-scale channel gains, a proportional fair scheduling policy was optimized in [22]. With both future large-scale channel gains and average arrival rate of RT traffic, a predictive resource allocation policy was proposed in [23] to save energy at the BSs, where RT services are regarded as a background traffic with higher priority, and resources are reserved for RT services to ensure the QoS.

In all previous studies, the predictive resource allocation is only designed for a single kind of services. Yet a real-world cellular network needs to support different kinds of services. It is no doubt that jointly allocating resources to different services can improve EE, and the policy that reserves resources for RT services is inevitably conservative.

Besides, all existing predictive resource allocation policies except those proposed in [17, 23] are in one timescale, which were designed at the beginning of a prediction window. In practice, the large-scale channel gains are predictable in the timescale of seconds, but the small-scale channel gains are hard to predict beyond the channel coherence time, which is in the timescale of milliseconds. As a consequence, the one timescale policies can neither fully use the radio resources nor guarantee the QoS in time-varying fading channels. While the policies in [17, 23] are implemented in two-timescale, the policies in different time scales are designed separately.

In this paper, we optimize predictive resource allocation jointly for hybrid services and for the policies in two timescales. We consider an orthogonal frequency division multiple access (OFDMA) network serving two types of users, one type of users request VoD services (delay tolerant service can be regarded as a special case of VoD services), and the other users request RT services. We study EE-optimal resource allocation exploiting both large-scale channel gains predictable within window and small-scale channel gains estimatable in each time slot. We further find which type channel information needs to predict, which is of practical interest since channel prediction inevitably consumes computing (e.g., trajectory predicting) and storage (e.g., radio map) resources. The major contributions of this work are summarized as follows,
• To obtain the EE upper bound achieved by predictive resource allocation and show that predicting which kinds of channel information are necessary to achieve the upper bound, resource allocation is jointly optimized for both services and for two timescale. At the beginning of a prediction window, the average transmit power and bandwidth are assigned to each user based on future large-scale channel gains and small-scale channel distribution. At the start of each time slot, instantaneous transmit power is allocated to the subcarriers of each user according to the assigned resources with the available small-scale channel gains.

• Our analysis shows that predicting small-scale channel gains for VoD users cannot improve EE. When there are only RT services, predicting large-scale channel gains cannot help improve EE. When there are both VoD and RT services, predicting large-scale channel gains for both type of users are necessary to achieve the EE upper bound. This is because by optimizing the resource allocation plan for RT users, we can predict how much resources they will occupy, which is useful for making the resource allocation plan for users with VoD services. Simulation results show that joint resource allocation for the two kinds of services can improve EE significantly, and decoupling the resource allocation in two time scales leads to considerable EE loss.

• To provide a viable scheme for practice use, a heuristic policy is proposed, which is with low complexity and robust to prediction errors. Simulation results show that the heuristic policy performs closely to the optimal policy if the prediction of large-scale channel gain is error-free and outperforms the optimal policy when the prediction is with large errors.

The rest of the paper is organized as follows. In section II, we introduce system model and the QoS requirements of both services. In section III, we optimize predictive resource allocation with perfect large-scale channel gains, first for a single cell scenario and then extended to multi-cell scenario. In section IV, we show which kind of channel information needs to be predicted for each type of users. In section V, we propose a heuristic policy robust to prediction uncertainty. In section VI, we provide simulation results, and in section VII, we conclude the paper.

II. SYSTEM MODEL AND QoS REQUIREMENTS

Consider the scenario that multiple mobile users travel through an OFDMA network, which request VoD and RT services, respectively. For notational simplicity, we first consider a single cell scenario, and then extend to the multi-cell scenario in the end of next section.
A. Transmission and Channel Models

Consider frequency-selective block fading channel. Time is discretized to frames each with duration $\Delta T$ and time slots each with duration $\tau$. The durations are defined according to the channel variation, i.e., the variation of large scale channel gain (caused by path-loss and shadowing) and small scale channel gain (caused by fast fading) due to user mobility. Assume that: (1) the large scale channel gain remains constant within each frame and may vary among frames, and (2) the small scale channel gain remains constant within each time slot and is independent and identically distributed (i.i.d.) among different time slots and subcarriers in each frame. In typical scenarios, large-scale channel gain varies in the order of seconds. $\tau$ is the channel coherence time, which is in the order of milliseconds [24]. With the predicted user location along the trajectory [4] and measured radio map, the large scale channel gain is predictable [5], but the small scale channel gain is hard to predict beyond the channel coherence time. For notational simplicity, we assume $\Delta T = N_S \tau$. In practice, $\Delta T \gg \tau$ [24].

| Symbol | Description |
|--------|-------------|
| $\tau$ | duration of each time slot |
| $N_S$ | number of time slots in each frame |
| $M_D$ | number of VoD users |
| $g_{mijk}$ | small-scale channel gain for the $m$th user in the $j$th time slot of the $i$th frame on the $k$th subcarrier |
| $\alpha^m_i$ | large-scale channel gain for the $m$th user in the $i$th frame |
| $P_{\text{max}}$ | maximal transmit power |
| $p_{mijk}$ | transmit power allocated to the $m$th user in the $j$th time slot of the $i$th frame on the $k$th subcarrier |
| $K_m^i$ | number of subcarriers allocated to the $m$th user in the $i$th frame |
| $B$ | subcarrier spacing |
| $\rho$ | power amplifier efficiency |
| $s_{mij}$ | instantaneous channel capacity for the $m$th user in the $j$th time slot of the $i$th frame |
| $S_m^i$ | amount of data that can be transmitted to the $m$th user during the $i$th frame |
| $R_m^i$ | amount of data played for the $m$th VoD user in the $i$th frame |
| $\bar{s}_m^i$ | average service rate for the $m$th user in the $i$th frame |
| $b_{mij}$ | departure rate for the $m$th RT user in the $j$th time slot of the $i$th frame |
| $D_m^{\text{max}}$ | delay bound of the $m$th RT user |
| $\theta^m$ | QoS exponent of the $m$th user |
| $E_m^i(\theta^m)$ | effective capacity of the $m$th user in the $i$th frame |
| $P_c$ | circuit power consumption on each subcarrier |
| $P_0$ | the fixed circuit power consumption |

Denote the number of users that have accessed to the BS at the beginning of a prediction

---

July 7, 2017
as $M_D + M_R$, where $M_D$ and $M_R$ are the numbers of VoD and RT users, respectively.

The prediction window includes $N_L$ successive frames. For the $m$th user, $\alpha^m_i$ is the large-scale channel gains in the $i$th frame, and $g^m_{ijk}$ is the small-scale channel gain on the $k$th subcarrier in the $j$th time slot of the $i$th frame. At the beginning of the prediction window, we assume that $\alpha^m_i, i = 1, ..., N_L$ are perfectly predicted by the BS, but $g^m_{ijk}, k = 1, ..., K_{\text{max}}, j = 1, ..., N_S, i = 1, ..., N_L$ are unknown for $m = 1, ..., M_D + M_R$, where $K_{\text{max}}$ is the total number of subcarriers. During the transmission procedure, $g^m_{ijk}$ is available at both the $m$th user and the BS after channel estimation in the $j$th time slot of the $i$th frame. A list of symbols is given in Table I.

The achievable instantaneous data rate for the $m$th user can be expressed as follows [25],

$$s^m_{ij} = B \sum_{k=1}^{K^m_i} \log_2 \left( 1 + \frac{\alpha^m_i \phi \sigma^2_0 p^m_{ijk} g^m_{ijk}}{\phi \sigma^2_0} \right) \text{ bits/s},$$

where $B$ is the subcarrier spacing, $p^m_{ijk}$ is the transmit power allocated to the $m$th user on the $k$th subcarrier in the $j$th time slot of the $i$th frame, $\phi > 1$ captures the gap between capacity and achievable rate with practical modulation and coding schemes, $\sigma^2_0$ is the variance of the additive Gaussian noise, and $K^m_i$ is the number of subcarriers assigned to the $m$th user in the $i$th frame.

### B. QoS Requirement for VoD Services

Since the key factor that determines the experience of a user requesting VoD service is playback interruption, we consider the queue in the buffer at each user. We assume that the video segments to be played within the prediction window are available at the BS [16, 26, 27]. The queueing model for VoD services is shown in Fig. 1. $R^m_i$ is the amount of data played at the $m$th user in the $i$th frame, which is given when a certain quality level of the video is chosen by the user (e.g., high definition video). The amount of data that can be transmitted to the $m$th user during the $i$th frame is given by $S^m_i = \tau \sum_{j=1}^{N_S} s^m_{ij}$.

Denote the duration of each video segment as $T_{\text{seg}}$, which is determined by the video sources and does not depend on the mobility of users. For notational simplicity and without loss of generality, we set $T_{\text{seg}} = \Delta T$. Then, there are $N_L$ video segments in a prediction window. Assume that the buffer size is larger than the size of $N_L$ video segments, which is reasonable.

---

1The users arriving at the cell during a prediction window will wait to be served in the next prediction window, where some of newly arrived users will not be admitted by the BS if their QoS cannot be ensured.
Fig. 1. Queueing model for VoD services.

for smart phones since storage devices are cheap nowadays. This assumption will be removed in Section V, where we design a policy that is aware of limited buffer size.

To guarantee the requested video quality, each video segment should be delivered to the user before it is played. Then, the QoS requirement of the VoD services can be reflected by the following constraint [16],

\[
Q_0^m + \sum_{i=1}^{l} S_i^m \geq \sum_{i=1}^{l+1} R_i^m, l = 1, ..., N_L, m = 1, ..., M_D,
\]

(2)

where \(Q_0^m = R_1^m\) is the initial queue length and \(R_{N_L+1}^m\) is the number of bits in the first video segment to be played in the next prediction window. In other words, the first video segment to be played in a prediction window has been conveyed to the user in the previous prediction window. Hence, no interruption occurs between the adjacent prediction windows.

Since the number of time slots in each frame is large in practice, by channel coding among time slots, the data rate in a frame can approach the average data rate [28]. From (1), the average data rate for the \(m\)th user in the \(i\)th frame can be expressed as follows,

\[
\bar{s}_i^m = B \sum_{k=1}^{K_i^m} \mathbb{E}_h \left[ \log_2 \left( 1 + \frac{\alpha_i^m \phi \sigma_i^m}{2 P_{ijk} g_{ijk}} \right) \right] \text{ bits/s},
\]

(3)

where the average is taken over small-scale channel fading. Then, we have \(S_i^m = \Delta T \bar{s}_i^m\), and the constraint in (2) can be equivalently written as

\[
\sum_{i=1}^{l} \bar{s}_i^m \geq \frac{1}{\Delta T} \sum_{i=2}^{l+1} R_i^m, l = 1, ..., N_L, m = 1, ..., M_D.
\]

(4)

Remark 1. For delay tolerant service such as file downloading, the user demand can be characterized as to transmit a file with size \(\tilde{R}^m\) in \(N_L\) frames. Then, the required data rate can also be formulated as \(\sum_{i=1}^{N_L} \bar{s}_i^m \geq \tilde{R}^m\), which is similar to (4). Therefore, the delay tolerant service can also be included in our framework.
C. QoS Requirement for Real-time Services

Different from VoD services, the data of RT services are randomly generated by users rather than stored in the server, hence the data from RT users cannot be transmitted in advance. After the data from users randomly arrive at the BS, they are waiting in the queue at the BS for transmission but cannot wait too long in the buffer to satisfy the QoS. The queueing model for the $m$th user requesting a RT service is shown in Fig. 2 where $a_{ij}^m$ represents the data arrival rate in the $j$th time slot of the $i$th frame. The queueing delay in the buffer of the BS should satisfy the statistical QoS requirement, characterized by a delay bound $D_{\text{max}}^m$ and a delay violation probability $\epsilon_D^m$. If the queueing delay in the $m$th queue exceeds $D_{\text{max}}^m$ with probability less than $\epsilon_D^m$, then the QoS requirement of the $m$th user can be satisfied. For example, the upper bound on $\epsilon_D^m$ for VoIP is 2% while $D_{\text{max}}^m$ is 50 ms for radio access network.

![Queueing model for the $m$th user requesting a RT service. There are $M_R$ queues for the $M_R$ RT users.](image)

Effective bandwidth and effective capacity are widely applied tools in designing resource allocation with statistical QoS requirement. For uncorrelated random arrival process $\{a_{ij}^m, i = 1, ..., N_L, j = 1, ..., N_S\}$, the effective bandwidth can be expressed as [29]

$$E_B^m (\theta^m) = \frac{1}{\theta^m \tau} \ln \mathbb{E} \left[ \exp \left( \theta^m \tau a_{ij}^m \right) \right] \text{ (bits/s)},$$

(5)

where $\theta^m$ is the QoS exponent.

For the RT services with short delay requirement, the duration of each frame is much longer than the delay bound, i.e., $\Delta T \gg D_{\text{max}}^m$. The small-scale channel gains are i.i.d. in different time slots, and the power allocated in the $j$th time slot only depends on $g_{ijk}^m$. Consequently, $s_{ij}^m, j = 1, ..., N_S$ are also i.i.d.. Then, the effective capacity in the $i$th frame for the $m$th user can be expressed as [30]

$$E_C^m (\theta^m) = -\frac{1}{\theta^m \tau} \ln \mathbb{E} \left[ \exp \left( -\theta^m \tau s_{ij}^m \right) \right] \text{ (bits/s)}.$$ 

(6)

Denote the steady state delay for the $m$th user as $D_{\infty}^m$. Then, the required QoS exponent...
θ^m to guarantee \((D^m_{\text{max}}, \varepsilon^m_D)\) can be obtained from [31] as 
\[
\Pr\{D^m_\infty > D^m_{\text{max}}\} \approx \Pr\{D^m_\infty > 0\} \exp\left[-\theta^m E^m_B(\theta^m) D^m_{\text{max}}\right] \leq \exp\left[-\theta^m E^m_B(\theta^m) D^m_{\text{max}}\right] = \varepsilon^m_D, \quad i = 1, ..., N_L, 
\]
where the approximation is accurate when the delay bound is much longer than the duration of each time slot, which is true for mobile users requesting typical RT services like video conference and VoIP [11]. To guarantee the required \(\theta^m\), the following constraint should satisfy [32]
\[
E^m_{C_i}(\theta^m) \geq E^m_{B}(\theta^m), \quad m = M_D + 1, ..., M_D + M_R, \quad i = 1, ..., N_L.
\]

### D. Power Consumption Model and EE Definition

The total energy consumed at the BS serving \(M_D + M_R\) users in the prediction window (i.e., in \(N_L\) frames) can be modeled as [33–35]
\[
\sum_{i=1}^{N_L} E_i = \sum_{i=1}^{N_L} \left( \frac{1}{\rho} \sum_{m=1}^{M_D+M_R} \sum_{j=1}^{N_S} \sum_{k=1}^{K^m_{ij}} \tau p^m_{ijk} + \Delta T P_c \sum_{m=1}^{M_D+M_R} K^m_i + \Delta T P_0 \right),
\]
(8)
where \(E_i\) is the energy consumption in the \(i\)th frame, \(\rho \in (0, 1]\) is the power amplifier efficiency, \(P_c\) is the circuit power consumed for baseband processing such as channel estimation on each subcarrier, and \(P_0\) is the fixed circuit power consumption for the BS.

According to the bits per Joule metric in [36], EE of a system is the ratio of the amount of data transmitted to the energy consumed during a certain period. For predictive resource allocation, the period is the prediction window. However, since only the large-scale channel gains are available at the beginning of the prediction window, both the amount of data to be transmitted and the energy to be consumed in the upcoming \(N_L\) frames are random variables, which depend on the small-scale channel gains. As a result, we cannot optimize predictive resource allocation to maximize the EE metric in [36]. Since the number of time slots in each frame is large, i.e., \(N_S\) is large, maximizing the above EE metric is equivalent to maximizing the ratio of the average amount of transmitted data to the average energy consumption, where the average is taken over the small-scale channel gains. Hence, we define the EE as follows,
\[
\eta \triangleq \left[ \mathbb{E}_h \left( \sum_{m=1}^{M_D} \sum_{i=1}^{N_L} \sum_{j=1}^{N_S} s^m_{ij} \right) \right] + \mathbb{E}_h \left( \sum_{m=M_D+1}^{M_D+M_R} \sum_{i=1}^{N_L} \sum_{j=1}^{N_S} b^m_{ij} \right) \bigg/ \mathbb{E}_h \left( \sum_{i=1}^{N_L} E_i \right).
\]
(9)
For VoD services, the amount of data transmitted equals to the amount of data that needs to
transmit. Thus, 
\[ \mathbb{E}_h \left( \sum_{m=1}^{M_D} \sum_{i=1}^{N_L} \sum_{j=1}^{N_S} s_{ij}^m \right) = \sum_{m=1}^{M_D} N_L \sum_{i=1}^{N_L} \Delta T s_i^m = \sum_{m=2}^{M_D} N_L \sum_{i=1}^{N_L} R_i^m, \]
which is determined at the beginning of the prediction window by the requested video level and network status. For RT services, when the queues are in steady states, the average departure rates equal to the average arrival rates \[29\]. Thus, 
\[ \mathbb{E}_h \left( \sum_{m=M_D+1}^{M_D+M_R} N_L \sum_{i=1}^{N_L} b_{ij}^m \right) = \mathbb{E}_h \left( \sum_{m=M_D+1}^{M_D+M_R} N_L \sum_{i=1}^{N_L} \sum_{j=1}^{N_S} a_{ij}^m \right), \]
which is determined by the arrival processes. Therefore, the numerator of \(9\) does not depend on the resource allocation policy, and maximizing the EE in \(9\) is equivalent to minimizing the average energy consumption. Substituting \(8\) without the last term into the denominator of \(9\), maximizing the EE is equivalent to minimizing the following expression, 
\[ \frac{1}{\rho} \mathbb{E}_h \left( \sum_{m=1}^{M_D+M_R} \sum_{i=1}^{N_L} \sum_{j=1}^{N_S} \sum_{k=1}^{K_i^m} T P_{ijk}^m \right) + \Delta T P_c \sum_{m=1}^{M_D+M_R} \sum_{i=1}^{N_L} K_i^m. \quad (10) \]

### III. Energy Efficient Predictive Resource Allocation

In this section, we optimize predictive resource allocation for OFDMA systems supporting both VoD and RT services to show the potential of predictive policy in improving EE. To exploit future large-scale channel gains and current small-scale channel gains in a joint manner, we formulate a functional extreme problem and obtain the global optimal solution. We first consider single cell scenario, and then extend to multi-cell scenario.

#### A. Problem Formulation

At the beginning of the prediction window (i.e., the 1st time slot of the 1st frame), we cannot optimize \(p_{ijk}^m\) to minimize \(10\) since future small-scale channel gains are unknown. Yet we can optimize the average transmit power \(\bar{P}_i^m \triangleq \mathbb{E}_h \left( \sum_{k=1}^{K_i^m} P_{ijk}^m \right)\) and the number of subcarriers (i.e., bandwidth) \(K_i^m\) assigned to the \(m\)th user in the \(i\)th frame, when future large-scale channel gains and the distribution of the small-scale channel are known. We refer to \(\{\bar{P}_i^m, K_i^m, m = 1, ..., M_D + M_R, i = 1, ..., N_L\}\) as the resource allocation plan.

At the beginning of each time slot, we can optimize \(p_{ijk}^m\) based on \(g_{ijk}^m, k = 1, ..., K_i^m\) and \(\{\bar{P}_i^m, K_i^m\}\), since the small-scale channel gains are available at the BS after channel estimation. We denote the power allocation policies for the VoD services and the RT services as \(p_{ijk}^m = f_D(\bar{P}_i^m, K_i^m, g_{ijk}^m), m = 1, ..., M_D\) and \(p_{ijk}^m = f_R(\bar{P}_i^m, K_i^m, g_{ijk}^m), m = M_D + 1, ..., M_D + M_R\).
respectively, where \( i = 1, \ldots, N_L, j = 1, \ldots, N_S \) and \( k = 1, \ldots, K_{i}^{m} \). The forms of the functions \( f_{D}(\cdot) \) and \( f_{R}(\cdot) \) differ for different power allocation policies.

The optimization of resource allocation plan and power allocation policies are closely coupled. In what follows, we formulate the joint optimization problem for the two-timescale policy.

Substituting the power allocation policy for VoD service \( p_{ijk}^{m} = f_{D}(\bar{P}_{i}^m, K_{i}^m, g_{ijk}^m) \) into (3), the average service rate in the \( i \)th frame can be expressed as follows,

\[
\bar{s}_{i}^{m} = K_{i}^{m} \int_{0}^{\infty} B \log_{2} \left[ 1 + \frac{\alpha_{i}^{m}}{\phi \sigma_{0}^{2}} f_{D}(\bar{P}_{i}^{m}, K_{i}^{m}, g) \right] e^{-g} dg,
\]

where \( m = 1, \ldots, M_D \).

Substituting the power allocation policy for the RT service \( p_{ijk}^{m} = f_{R}(\bar{P}_{i}^m, K_{i}^m, g_{ijk}^m) \) into (1) and then into (6), the effective capacity in the \( i \)th frame can be obtained as

\[
E_{C_{i}}^{m}(\theta^{m}) = -\frac{K_{i}^{m}}{\theta_{m}^{m} \tau} \ln \left\{ \int_{0}^{\infty} \left[ 1 + \frac{\alpha_{i}^{m}}{\phi \sigma_{0}^{2}} f_{R}(\bar{P}_{i}^{m}, K_{i}^{m}, g) \right]^{-\beta_{m}} e^{-g} dg \right\} \text{ (bits/s)},
\]

where \( m = M_D + 1, \ldots, M_D + M_R, \beta_{m} \Delta \equiv \frac{\theta_{m}^{m} \tau B}{\ln 2} \). (11) and (12) are obtained with Rayleigh fading where \( g_{ijk}^m \) are exponentially distributed with mean of 1.

Further considering (4) and (7), the optimal resource allocation plan and power allocation policies that minimize the average energy consumption under the QoS constraints for both VoD and RT services can be obtained by solving the following problem,

\[
\min_{f_{D}(\cdot), f_{R}(\cdot), P_{i}^{m}, K_{i}^{m}} \frac{1}{\Delta T} \sum_{m=1}^{M_D+M_R} \sum_{i=1}^{N_L} \left( \frac{1}{\rho} \bar{P}_{i}^{m} + P_{i} K_{i}^{m} \right),
\]

s.t.

\[
\sum_{i=1}^{l} K_{i}^{m} \int_{0}^{\infty} B \log_{2} \left[ 1 + \frac{\alpha_{i}^{m}}{\phi \sigma_{0}^{2}} f_{D}(\bar{P}_{i}^{m}, K_{i}^{m}, g) \right] e^{-g} dg \geq \frac{1}{\Delta T} \sum_{i=2}^{l+1} R_{i}^{m},
\]

\( m = 1, \ldots, M_D, l = 1, \ldots, N_L \),

\[
-\frac{K_{i}^{m}}{\theta_{m}^{m} \tau} \ln \left\{ \int_{0}^{\infty} \left[ 1 + \frac{\alpha_{i}^{m}}{\phi \sigma_{0}^{2}} f_{R}(\bar{P}_{i}^{m}, K_{i}^{m}, g) \right]^{-\beta_{m}} e^{-g} dg \right\} \geq E_{B}^{m}(\theta_{m}^{m}),
\]

\( m = M_D + 1, \ldots, M_D + M_R, i = 1, \ldots, N_L \),

\[2\text{We take Rayleigh fading as an example in this work, but the methodology can be extended to the other channels.}\]
where the objective function in (13) is obtained by substituting \( \bar{P}^m_i = \mathbb{E}_h \left( \sum_{k=1}^{K^m_i} p^m_{ijk} \right) \) into (10) and ignoring a constant \( \Delta T = N_S \tau \), constraints in (13a) and (13b) are obtained by substituting (11) and (12) into (4) and (7), respectively, and (13c) and (13d) are the constraints on the average transmit power and the total number of subcarriers. Since only channel statistics are known at the start of the window, this problem only optimize average power and bandwidth in each frames for each user. With constraint (13d), we can always allocate each subcarrier only to one user.

The constraints in (13a) and (13b) depend on the forms of the functions \( f_D(\cdot) \) and \( f_R(\cdot) \). This indicates that the resource allocation planning depends on the power allocation policies. In other words, the optimal value of the objective function in (13) is a function of \( f_D(\cdot) \) and \( f_R(\cdot) \). We denote it as \( E^*_{\text{ave}}(f_D, f_R) \). The optimal power allocation policies can be obtained by minimizing \( E^*_{\text{ave}}(f_D, f_R) \), and are denoted as \( f^*_D(\cdot) \) and \( f^*_R(\cdot) \). It is worth noting that finding the optimal form of function is a functional extreme problem, and cannot be solved by standard convex optimization tools. In next subsection, we first find the forms of the functions of \( f^*_D(\cdot) \) and \( f^*_R(\cdot) \) that minimizes \( E^*_{\text{ave}}(f_D, f_R) \). Then, the optimal resource allocation planning, \( \{ \bar{P}^m_i, K^m_i, m = 1, \ldots, M_D + M_R, i = 1, \ldots, N_L \} \), can be found from problem (13).

**Remark 2.** The terms inside the sum of left hand side of (13a) (i.e., \( \bar{s}^m_i \) in (11)) is the average rate in each frame. In many existing works [15–21], this average rate is assumed known by prediction. As a natural result, the predictive resource allocation is either only in one timescale (i.e., only make the plan) [15, 16, 18–21], or decoupled into independently designed policies in the two timescales [17]. However, it is clear from problem (13) that the future average rate depends on \( \{ \bar{P}^m_i, K^m_i \} \) and \( f_D(\cdot) \) even when the system only supports VoD services.

### B. Optimal Power Allocation Policies

A policy that maximizes the average service \( \bar{s}^m_i \) (or effective capacity \( E^m_{Ci}(\theta^m) \)) with given average transmit power \( \bar{P}^m_i \) and number of subcarriers \( K^m_i \) (i.e., bandwidth) can minimize \( P^m_i \)
with given \( \bar{s}_i^m \) (or \( E_C^m(\theta^m) \)) and \( K_i^m \) \cite{24,37}. Inspired by such a fact, we first find the power allocation policy that maximizes \( \bar{s}_i^m \) (or \( E_C^m(\theta^m) \)) with given \( P_i^m \) and \( K_i^m \), and then prove that the policy is optimal to minimize the average energy consumption, i.e., minimize (13).

1) Power allocation policy for VoD services: As shown in \cite{24}, the policy that maximizes \( \bar{s}_i^m \) with given \( \bar{P}_i^m \) and \( K_i^m \) is the water-filling policy, which is

\[
\mathcal{F}^w_D \left( \frac{\bar{P}_i^m}{K_i^m}, g \right) = \left\{ \begin{array}{ll}
\frac{\phi \sigma_0^2}{\alpha_i^m} \left( \frac{1}{\nu_i^m} - \frac{1}{g} \right), & g \geq \nu_i^m, \\
0, & g < \nu_i^m,
\end{array} \right.
\]

(14)

where \( m = 1, ..., M_D \), \( i = 1, ..., N_L \), and the water level \( \nu_i^m \) can be obtained from

\[
\int_{\nu_i^m}^{\infty} \frac{\sigma_0^2}{\alpha_i^m} \left( \frac{1}{\nu_i^m} - \frac{1}{g} \right) e^{-g} dg = \frac{\bar{P}_i^m}{K_i^m}.
\]

(15)

Note that the form of the function in (14) does not depend on the value of \( g \).

2) Power allocation policy for RT services: As shown in \cite{37}, the optimal power allocation that maximizes \( E_C^m(\theta^m) \) with given \( \bar{P}_i^m \) and \( K_i^m \) also follows a water-filling structure, but the water-level is time-varying and the instantaneous power allocated to each subcarrier depends on the small-scale channel gains on all the subcarriers assigned to the user.

For mathematical tractability, we consider independent power allocation policy\(^3\) that maximizes \( E_C^m(\theta^m) \) with given \( P_i^m \) and \( K_i^m \), which can be expressed as follows \cite{37},

\[
\mathcal{F}^w_R \left( \frac{\bar{P}_i^m}{K_i^m}, g \right) = \left\{ \begin{array}{ll}
\frac{\phi \sigma_0^2}{\alpha_i^m} \left[ \frac{1}{(\nu_i^m)^{\beta m}} g^{\beta m} - \frac{1}{g} \right], & g \geq \nu_i^m, \\
0, & g < \nu_i^m,
\end{array} \right.
\]

(16)

where \( m = M_D + 1, ..., M_D + M_R, i = 1, ..., N_L, \beta_m = \frac{\theta_m + R}{\ln 2} \), and the water level \( \nu_i^m \) over Rayleigh fading channel can be obtained from

\[
\int_{\nu_i^m}^{\infty} \frac{\phi \sigma_0^2}{\alpha_i^m} \left[ \frac{1}{(\nu_i^m)^{\beta m}} g^{\beta m} - \frac{1}{g} \right] e^{-g} dg = \frac{\bar{P}_i^m}{K_i^m}.
\]

(17)

\(^3\)“Independent power allocation policy” means that the instantaneous transmit power on a certain subcarrier only depends on the small-scale channel gain on this subcarrier and is independent of the small-scale channel gains on the other subcarriers. This policy is near optimal when \( \theta_m \) is small \cite{37}. For VoIP service, the delay requirement is not very stringent. For video conference service, the average arrival rate is high. For both RT services, \( \theta_m \) is small, and hence the policy is near optimal. In the sequel, we refer to the optimal “independent power allocation policy” as the optimal power allocation policy for simplicity.
3) Optimality of the power allocation policies: The following proposition indicates that (14) is the optimal power allocation policy for VoD services and (16) is the optimal independent power allocation policy for RT services that maximizes the EE.

**Proposition 1.** For ANY power allocation policies $f'_D\left(\bar{P}_i^m, K_i^m, g\right)$ and $f'_R\left(\bar{P}_i^m, K_i^m, g\right)$,

$$E_{ave}^{*}\left(f'_D, f'_R\right) \leq E_{ave}^{*}\left(f''_D, f''_R\right). \quad (18)$$

The proposition is derived based on the result in [24]: the water-filling policy $f''_D\left(\bar{P}_i^m, K_i^m, g\right)$ can minimize $\bar{P}_i^m$ with given $K_i^m$ and average service rate $\bar{s}_i^m$. See proof in Appendix A.

Proposition 1 indicates that $f''_D\left(\bar{P}_i^m, K_i^m, g\right) = f''_D\left(\bar{P}_i^m, K_i^m, g\right)$ and $f''_R\left(\bar{P}_i^m, K_i^m, g\right) = f''_R\left(\bar{P}_i^m, K_i^m, g\right)$.

**C. Optimal Resource Allocation Planning**

Substituting the optimal power allocation policies in (14) and (16) into (13a) and (13b), the optimal resource allocation plan can be obtained from the following problem,

$$\min_{P_i^m, K_i^m} \sum_{m=1}^{M_D+M_R} \sum_{i=1}^{N_L} \left(\frac{1}{\rho} \bar{P}_i^m + P_c K_i^m\right), \quad (19)$$

s.t.

$$\sum_{i=1}^{l} K_i^m F_D\left(\bar{P}_i^m / K_i^m\right) \geq \frac{1}{\Delta_T} \sum_{i=2}^{l+1} R_i^m, \quad m = 1, ..., M_D, \quad l = 1, ..., N_L, \quad (19a)$$

$$- \frac{K_i^m}{\theta_i^m \tau} \ln \left[F_R\left(\bar{P}_i^m / K_i^m\right)\right] \geq E_B^m(\theta^m), \quad m = M_D + 1, ..., M_D + M_R, \quad i = 1, ..., N_L, \quad (19b)$$

(13c), (13d) and (13e),

where

$$F_D\left(\frac{P_i^m}{K_i^m}\right) = \int_0^{\infty} B \log_2 \left[1 + \frac{\alpha_i^m}{\phi \sigma_0^2} f_D\left(\frac{P_i^m}{K_i^m}, g\right)\right] e^{-g} dg, \quad (20)$$

$$F_R\left(\frac{P_i^m}{K_i^m}\right) = \int_0^{\infty} \left[1 + \frac{\alpha_i^m}{\phi \sigma_0^2} f_R\left(\frac{P_i^m}{K_i^m}, g\right)\right]^{-\beta_i^m} e^{-g} dg. \quad (21)$$

The following two properties indicate that the feasible region of problem (19) is a convex set.

**Property 1.** The left hand side of (19a) is jointly concave in $\bar{P}_i^m$ and $K_i^m$, $i = 1, ..., l$.

The proof of Property 1 is shown in [1].

**Property 2.** The left hand side of (19b) is jointly concave in $\bar{P}_i^m$ and $K_i^m$. 

July 7, 2017
Proof: See Appendix B.

Since the objective function in (19) is linear, problem (19) is a convex programming, whose global optimal solution can be solved numerically by interior-point method if it is feasible [38]. Problem (19) could be infeasible. When it is infeasible, the video quality of VoD services has to be reduced, i.e., the values of \( \{R_{m}^{i}, m = 1, ..., M_{D}\} \) in constraints (19a) need to be reduced to make this problem feasible. To minimize the quality deterioration, the video quality of VoD services should be optimized. Such problem has been studied in existing literatures, e.g., [39, 40]. In this work, we only consider a simple method: if problem (19) is infeasible, then the system will reduce the video quality of all VoD users to a lower level to make it feasible.

The optimal predictive resource allocation operates in two timescales:

- At the beginning of each prediction window, the BS makes the resource allocation plan for both VoD and RT users (i.e., assigns average transmit power and bandwidth for the forthcoming frames), with the knowledge of large-scale channel gains, small-scale channel distribution and QoS requirements of all users, and the statistics of the RT arrival processes.
- At the beginning of each time slot during transmission procedure, the BS allocates transmit power to different subcarriers respectively for the VoD users and RT users according to the resource allocation plan, and with the knowledge of small-scale channel gains.

D. Extension to Multicell Scenario

Now we consider a scenario where the \( M_{D} + M_{R} \) users are served by \( N_{B} \) BSs. We assume that the BSs can share the future large-scale channel gains of all the users in a prediction window among each other, which does not need high capacity and low latency backhaul links. We assume that the inter-cell interference can be treated as noise.

It is not hard to show that Proposition 1 can be extended into the multi-cell scenario, and hence the power allocation policies in (14) and (16) are optimal for VoD services and RT services, respectively. Denote \( \bar{P}_{n}^{m} \) and \( K_{n}^{m} \) as the average transmit power and the number of subcarriers assigned to the \( m \)th user in the \( i \)th frame by its accessed BS (i.e., the \( n \)th BS). Denote \( M_{n}^{i} \) as the set of indices of the users that are served by the \( n \)th BS in the \( i \)th frame. The difference between single cell scenario and multi-cell scenario lies in the constraints on average transmit power and total number of subcarriers. Specifically, the power and bandwidth assigned to the users that
access to the same BS in the multi-cell scenario should satisfy the following constraints

\[ \sum_{m \in \mathcal{M}_i^n} \bar{P}^m_{n i} \leq P_{\text{ave}}, n = 1, \ldots, N_B, i = 1, \ldots, N_L, \quad (22) \]

\[ \sum_{m \in \mathcal{M}_i^n} K^m_{n i} \leq K_{\text{max}}, n = 1, \ldots, N_B, i = 1, \ldots, N_L. \quad (23) \]

The user association \( \{\mathcal{M}_i^n, n = 1, \ldots, N_B, i = 1, \ldots, N_L\} \) and resource allocation plan can be jointly optimized, but the resulting problem is a mixed integer optimization problem, which is much more challenging than problem (19). To save energy, it is reasonable to assume that each user is accessed to the BS with the highest large-scale channel gain. Then, \( \{\mathcal{M}_i^n, n = 1, \ldots, N_B, i = 1, \ldots, N_L\} \) are known by the BSs in the beginning of each prediction window since the trajectories of users are predictable. Similar to problem (19), the optimal resource allocation plan in multi-cell scenario is also convex programming, and can be solved numerically by the interior-point method.

IV. IMPACTS OF PREDICTED INFORMATION OF DIFFERENT KINDS OF SERVICES ON EE

With the prediction of user trajectory and the assistance of radio map, the large-scale channel gains are predictable. On the other hand, the small-scale channel gains (i.e., channel state information (CSI)) are hard to predict beyond the horizon of channel coherence time [41]. This fact naturally leads to the following questions: (1) To maximize the EE of a system, do we really need to know the future CSI? (2) To maximize EE, for which kind of services the future large-scale channel gains are beneficial?

In this section, we strive to answer the questions by separately considering VoD users and RT users. For notational simplicity, we consider the single cell scenario.

A. Predicted Information of VoD Users

To study whether or not the future CSI of VoD users are necessary for improving EE, we assume that there is no RT user, i.e., \( M_R = 0 \).

If the future CSI is available at the BS at the beginning of each prediction window and \( M_R = 0 \), then minimizing (10) is equivalent to minimizing the following objective function,

\[
\frac{1}{\rho} \left( \sum_{m=1}^{M_D} \sum_{i=1}^{N_L} \sum_{j=1}^{N_S} \sum_{k=1}^{K^m_{i j k}} \tau_{p_{i j k}^m} \right) + \Delta T \rho_p \sum_{m=1}^{M_D} \sum_{i=1}^{N_L} K^m_{i}, \quad (24)
\]
where the transmit powers on different subcarriers in the $i$th frame $\{p_{ijk}^m, k = 1, ..., K_i^m\}$ depend on the small-scale channel gains $\{g_{ijk}^m, k = 1, ..., K_i^m\}$. Denote the total transmit power for the $m$th user in the $j$th time slot of the $i$th frame as $P_{ij}^m = \sum_{k=1}^{K_i^m} p_{ijk}^m$. Since the fast fading in different time slots in each frame are i.i.d., if the number of time slots in each frame is large, then the time average transmit power converges to the ensemble average transmit power, i.e.,

$$\frac{1}{N_S} \sum_{j=1}^{N_S} P_{ij}^m \to P_i^m \text{ when } N_S \to \infty.$$ Furt her considering that $\Delta T = N_S T$, minimizing (24) is equivalent to minimizing the following expression,

$$\frac{1}{\rho} \left( \sum_{m=1}^{M_D} \sum_{i=1}^{N_L} \hat{P}_i^m \right) + P_c \sum_{m=1}^{M_D} \sum_{i=1}^{N_L} K_i^m,$$

which is the same as the objective function in (13) when $M_R = 0$.

The optimal policy that minimizes the energy consumption of VoD services with perfect future CSI can be obtained by minimizing (25) under constraints (13a), (13c), (13d) and (13e). Since the optimization problem is the same as problem (13), the optimal power allocation policy, the optimal average transmit power and number of subcarriers, and the minimal total energy consumption are the same for the two problems. This suggests the following observation.

**Observation 1**: To maximize the EE of a system, predicting CSI of VoD users is not beneficial, but the prediction of their large-scale channel gains are necessary.

Essentially, this is because the optimal power allocation policy only depends on the distribution of small scale channels (e.g., Rayleigh Fading as we considered) [24].

**B. Predicted Information of Real-Time Users**

To study whether or not the future large-scale and small-scale channel gains of RT users are necessary for improving EE, we assume that there is no VoD user, i.e., $M_D = 0$.

For RT services, $\tau \ll D_{\max}^m \ll \Delta T$. If the future large-scale channel gains are not available at the BS at the beginning of the prediction window but only available at the beginning of each frame, then the BS can assign the average transmit power and number of subcarriers to each RT user at the beginning of each frame. The resource allocation assigned to each RT user in the $i$th frame can be obtained from the following problem,

$$\min_{P_i^m, K_i^m} \sum_{m=1}^{M_R} \left( \frac{1}{\rho} \hat{P}_i^m + P_c K_i^m \right)$$

(26)
According to the expressions in (13b), (13c), (13d) and (13e), we can see that the resource allocation in the $i$th frame does not depend on the resource allocation in the other frames, and hence problem (13) can be decomposed into $N_L$ independent problems as problem (26). Knowing the large-scale channel gains in the future frames cannot help improve the QoS (i.e., the delay bound $D_{max}^m$ and delay bound violation probability $\varepsilon^m_\tau$) or the EE of a system only with RT services, since most of the data should be transmitted within one frame (except the data arrive the BS in the end of a frame). This gives rise to another observation as follows.

Observation 2: To maximize the EE of a system only with RT services, the future large-scale channel gains of RT users is no need to know at the beginning of the prediction window.

Remark 3. For VoD services, the requested data can be pre-buffered at the user terminal before playback. Since the duration of a prediction window exceeds the duration of each frame $\Delta T$, the BS can choose the frames with high large-scale channel gains to transmit data in advance to save energy. By contrast, for RT services, $\tau \ll D_{max}^m \ll \Delta T$. As a result, the EE can only be improved by adjusting resources among the time slots within $D_{max}^m$, and the future large-scale channel gains of RT users cannot help improve the EE of a system only serving RT services. Nevertheless, with the prediction of large-scale channel gains of RT users and the proposed joint optimization, the network resource usage status available for serving VoD users becomes predictable, which is beneficial in improving the EE of a network with both VoD and RT services.

V. A LOW COMPLEXITY POLICY ROBUST TO PREDICTION ERRORS

The solution of problem (13) is with high computational complexity, which consumes extra energy that may counteract the EE gain from the joint optimization. Besides, large-scale channel gains can never be predicted error-free. To provide a viable scheme for practice use, we propose a heuristic policy in this section, which is with low complexity and robust to prediction errors.

Recall that the basic idea of improving EE with predictive resource allocation is to transmit more data to a delay tolerant user under good channel condition, and transmit less or even no data to the user under bad channel condition [5]. In order to develop a low complexity policy, we can decouple the design of improving EE and user experience. To increase EE, we find a “ruler” to judge whether the large-scale channel gain in a frame is high or low. To improve user
experience, we decide how many video segments should be transmitted in a frame considering the queueing status at the VoD user. For RT service, the resource allocation is non-predictive.

To find a “ruler” (i.e., a threshold) robust to prediction errors, we resort to classical statistical theory. As shown in [42], the median of a set of samples is insensitive to outliers, which is defined as the 50th percentile that separating the first half of the data samples with large values from the second half with small values. Compared to mean value, another widely used statistic, median is less sensitive to outliers. For the problem at-hand, outliers are the large-scale channel gains with large prediction errors. Hence, we adopt the median of the predicted large-scale channel gains, denoted as $\alpha_m^{\text{med}}$, as the threshold. Then, at the beginning of the prediction window, the BSs only need to predict the median $\alpha_m^{\text{med}}$.

To avoid stalling and buffer overflow for the VoD users with very limited buffer sizes, the number of segments transmitted in each frame needs to be controlled. The number depends on the traffic load of the network, the buffer size and channel condition of each VoD user. Since we use the median as the threshold, in average the BS transmits data to a VoD user in 50% time during streaming. Then, it is reasonable to transmit two segments to a user with good channel, if there is still room in the buffer. Denote the maximal buffer size as $Q_{\text{max}}$, and the queue length of the $m$th user at the beginning of the $i$th frame as $Q_m^i$.

At the beginning of the $i$th frame, the large-scale channel gain of the $m$th user, $\alpha_i^m$, can be estimated at its associated BS. Denote $\tilde{i}$ as the index of last video segment that has been transmitted before the $i$th frame. Then, the indices of segments to be transmitted are $\{\tilde{i} + 1, \ldots\}$.

If $\alpha_i^m \geq \alpha_m^{\text{med}}$, then the $m$th user is in good channel condition. Two segments will be transmitted in the $i$th frame if the buffer has enough residual space, i.e., $Q_i^m + R_{i+1}^m + R_{i+2}^m - R_i^m \leq Q_{\text{max}}$. Then, the required average service rate with the heuristic policy is $\bar{s}_{\text{heu}}^i = \frac{1}{\Delta T} (R_{i+1}^m + R_{i+2}^m)$. One segment will be transmitted if $Q_i^m + R_{i+1}^m + R_{i+2}^m - R_i^m > Q_{\text{max}}$ but $Q_i^m + R_{i+1}^m - R_i^m \leq Q_{\text{max}}$, then $\bar{s}_{\text{heu}}^i = \frac{1}{\Delta T} R_{i+1}^m$. If $Q_i^m + R_{i+1}^m - R_i^m > Q_{\text{max}}$, then no video segment will be transmitted, and hence $\bar{s}_{\text{heu}}^i = 0$.

If $\alpha_i^m < \alpha_m^{\text{med}}$, then the $m$th user is in bad channel condition. No data will be transmitted in the $i$th frame if $\tilde{i} \geq i + 1$ (i.e., the video segment that to be played in the $i + 1$th frame has been transmitted). If $\tilde{i} = i$, we set $\bar{s}_{\text{heu}}^i = \frac{1}{\Delta T} R_{i+1}^m$ to avoid playback interruption, which means that the video segment to be played in the next frame will be transmitted in the $i$th frame.

Given the required average service rate $\bar{s}_{\text{heu}}^i$ of VoD users, the resource allocation plan in the
ith frame can be optimized from the following problem,

\[
\min_{P^m_i, K^m_i, m=1,...,M_D+M_R} \sum_{m=1}^{M_D+M_R} \left( \frac{1}{\rho} P^m_i + P_c K^m_i \right),
\]

\[\text{s.t.} \quad K^m_i F_D \left( \frac{P^m_i}{K^m_i} \right) \geq \bar{s}_i \text{heu}, m = 1, ..., M_D, \]

\[\quad - \frac{K^m_i}{\theta^m i} \ln \left[ F_R \left( \frac{P^m_i}{K^m_i} \right) \right] \geq E^m_B (\theta^m), m = M_D + 1, ..., M_D + M_R, \]

(13c), (13d) and (13e).

Except that the value of \( \bar{s}_i \text{heu} \) depends on \( \alpha_i^m \text{med} \), problem (27) does not depend on future information. Because the average service rate constraint in (27a) is a special case of the effective capacity constraint in (27b) with \( \theta^m \to 0 \) [30], many existing low-complexity algorithms in [43] and [44] can be applied to find the solution of this problem. The complexity of the heuristic policy is almost the same as the non-predictive joint resource allocation policy for VoD and RT users. This is because the only difference between the two policies lies in the service rate requirement in constraint (27a). Without predicted information, the video segment to be played in the \( i + 1 \)th frame should be transmitted in the \( i \)th frame, and hence the required average service rate in the \( i \)th frame is \( \frac{1}{\Delta T} R^m_i \theta^m \) rather than \( \bar{s}_i \text{heu} \).

The heuristic predictive resource allocation policy can be implemented in three timescales:

- At the beginning of prediction window, the median of large scale channel gains is predicted.
- At the beginning of each frame, the BS assigns average transmit power and bandwidth for the frame with estimated large-scale channel gains, QoS requirements of all users, and the statistics of RT traffic arrival processes.
- At the beginning of each time slot, the BS allocates transmit power to different subcarriers respectively for the VoD users and RT users according to the plan with the estimated small-scale channel gains.

A possible way to predict the median of large-scale channel gains in the prediction window is as follows. In each BS, we can pre-store the median of the large-scale channel gains of all possible locations in the cell, say by drive test or crowd-sourcing. For each VoD user, we only need to predict the cells that it will access in the prediction window. Then, from the median of large-scale channel gains in each cell the user accessed, we can predict the median in the
window. Applying the heuristic policy does not need to construct and store fine-grained radio map and to predict accurate user trajectory. As a result, the storage and computing resources can be reduced significantly.

VI. SIMULATION RESULTS

In this section, we evaluate the EE of the proposed optimal policy and heuristic policy. We consider both scenarios with perfect and imperfect prediction of large-scale channel gains.

A. Simulation Setup

For VoD service, we use scalable video coding in [45] (each segment includes one base layer and five enhance layers) to evaluate the performance of different policies. The bit rate of each layer can be found in [46]. The average streaming rate of each VoD service is around 2 Mbits/s. For RT service, the packets of each user arrive at the buffer of BS according to a Poisson process with average rate $\lambda_a = 500$ packets/s. The size of each packet follows exponential distribution with average $1/\lambda_u = 4$ kbits/packet. Hence, the average data arrival rate of RT service is 2 Mbits/s. All the users move along a road from point A at $(0, 0)$ m to point B, as shown in Fig. 3. To save transmit power, each user is accessed to its nearest BS. The distances between BSs are 500 m, and the minimal distance between the BSs to the road are 100 m. The path loss model is $35.3 + 37.6 \log_{10} D_m$ dB, where $D_m$ is the distance in meters between the $m$th user and its accessed BS in the $j$th time slot. The circuit powers of different components in [33] are measured in the year of 2012. The scaling law in [47] is further applied to predict $P_c$ and $P_0$ in 2020, which are used in our simulation. The EE is the ratio of the amount of transmitted data to the amount of energy consumed by the BSs to serve the VoD and RT services in the prediction window. The prediction window is with duration $N_L \Delta T = 60$ s. The total simulation
time is 6000 s. All the simulation parameters are listed in Table III. This setup will be used in the following unless otherwise specified.

| Maximal transmit power $P_{\text{max}}$ | 40.0 W |
|----------------------------------------|--------|
| Number of available subcarriers $K_{\text{max}}$ | 512 |
| Bandwidth of each subcarrier $B$ | 15 kHz |
| Power amplifier efficiency $\rho$ | 38.8 \% |
| Circuit power consumption for one subcarrier $P_c$ | 72 mW/MHz |
| Fixed circuit power consumption $P_0$ | 136 mW/MHz |
| Single-sided noise spectral density $N_0$ | -173 dBm/Hz |
| Duration of each frame $\Delta T$ and each time slot | 1 s and 5 ms |

We compare the optimal predictive resource allocation policy with three baseline policies:

- Non-predictive resource allocation (legend “Baseline 1”): This baseline is a simple extension of the policy in [43], where only RT services are considered. The video segments to be played in the $i$th frame are transmitted in the $(i-1)$th frame (i.e., $s_{i-1}^{mn} = \frac{1}{\Delta T} R_i^{mn}$). The resource allocation is obtained by solving problem (27) in multi-cell scenarios, where $K_i^{mn}$ and $P_i^{mn}$ are replaced by $K_i^{mn}$ and $P_i^{mn}$, respectively, and constraints in (13c) and (13d) are replaced by those in (22) and (23). The gain of the optimal policy over Baseline 1 comes from predicting large scale channel gains for both VoD and RT users.

- Predictive resource allocation only with future large-scale channel gains for VoD users (legend “Baseline 2”): This is a simple extension of the policy in [5], where only VoD service is considered. The unknown distances between BS and RT users are set as the radius of the cell in all the frames, and then the resource allocation for VoD and RT users are jointly optimized. By considering the worst case, the QoS of the RT users can be guaranteed no matter where they are located. The gain of the optimal policy over Baseline 2 comes from predicting large scale channel gains for RT users.

- Decoupled resource allocation in the two timescales (legend “Baseline 3”): This is extended from the two-timescale policy in [17], where only VoD services are considered. The extended

---

4 We do not compare with the robust policies in [20][21] due to two reasons. First, there is no simple extension of these policies to the system with both VoD and RT services. Second, in this paper we consider scalable video coding with multiple data layers, but in [20][21] a single data layer video coding is considered.
policy optimizes bandwidth allocation at the beginning of the prediction window (equivalent to allocating transmission time in [17]), where the total transmit power is equally allocated to all subcarriers in order to predict future average rate. In each time slot, the instantaneous transmit power is allocated to subcarriers with (14) and (16), i.e., allocating transmit power to subcarriers with good channels (similar to subcarrier selection in [17]). The gain of the optimal policy over Baseline 3 comes from joint resource allocation in two timescales.

B. Perfect Prediction of Large-scale Channel Gains

The users move at the same constant velocity of 20 m/s, and the future large-scale channel gains are available at the beginning or prediction window.

![Figure 4](image)

(a) EE v.s. numbers of VoD users $M_D$, $M_D + M_R = 5$. (b) EE v.s. streaming rate of VoD user, where $M_D = M_R = 1$.

Fig. 4. EE achieved by different policies.

EE achieved by different policies are illustrated in Fig. 4. In Fig. 4(a), the total number of users is fixed as $M_D + M_R = 5$, and the numbers of different kinds of users vary. In Fig. 4(b), the total data rate required by all users are fixed as $E(R^{m}_i/\Delta T) + \lambda_a/\lambda_u = 10$ Mbits, where the arrival data rate of RT user (or the streaming rate of VoD user) varies. Simulation results show that when there is no VoD user, the achieved EE of the optimal policy and Baseline 1 are the same. The results are consistent with the analysis in Section IV-B, i.e., the predicted information can not help improve the EE of the system if there are only RT users. When there are both VoD and RT users, the achieved EE of the optimal policy could be $50 \sim 100\%$ higher than the EE achieved by the baselines. The achieved EE of Baseline 2 is lower than Baseline 1.
when the number (or arrival data rate) of RT users is large, because the preserved resources for the RT users is conservative. The EE achieved by the heuristic policy is closed to that of the optimal policy. The EE achieved by Baseline 3 is low, which suggests the necessity of the two timescale joint resource allocation. Since Baseline 3 performs the worst in most cases, we no longer provide its performance.

### Table III

| $M_D = M_R$ | ≤ 8 | 9   | 10  | 11  | ≥ 12 |
|--------------|-----|-----|-----|-----|------|
| Optimal      | 5   | 4.9667 | 4.8833 | 4.6500 | NA   |
| Heuristic     | 5   | 4.9167 | 4.7833 | 4.6167 | NA   |
| Baseline 1    | 5   | 4.8333 | 4.7500 | 4.5800 | NA   |
| Baseline 2    | 5   | 4.4833 | 4.0800 | NA     | NA   |

Video quality with different number of users is shown in Table III which is the average number of enhance layers transmitted to VoD users. “NA” means that at least in one frame, the QoS of RT users cannot be satisfied or the data in base layer cannot be transmitted to VoD users (i.e., playback interruption occurs for VoD users). As shown in the table, the maximal total number of users that the system can support with ensured QoS is 20, i.e., $M_D = M_R = 10$. Again, the results indicate that the video quality of heuristic policy is near optimal.

![EE vs. total number of users](image-url)

Fig. 5. EE v.s. total number of users, where $M_D = M_R$.

The relation between EE and the total number of users are illustrated in Fig. 5. The results show that EE achieved by the optimal and heuristic policies are much higher than that achieved...
by the baselines when the number of users is small (i.e., the traffic load is light). When the number of users approaches to the maximal number of users that the system can support, the EE achieved by different policies are almost identical. This is because when the traffic load is heavy, the BSs need to serve the users with all the resources, and then there is no chance to save energy. By reserving resources for RT users, Baseline 2 is even inferior to Baseline 1. This means that joint predictive resource allocation for RT and VoD users are critical.

C. Imperfect Prediction of Large-scale Channel Gains

The prediction errors may come from many sources such as erroneous mobility route prediction \[48\], inaccurate velocity prediction \[49\], reported or estimated user location with errors \[50\], and inaccurate radio map \[51\]. Here we take the velocity prediction error as an example to illustrate the impact of imperfect prediction, since it leads to large accumulative error and hence causes more severe performance degradation than other type of prediction errors.

Markov chain is widely used in modeling the mobility of vehicles (see \[41\] and references therein). We use a discrete time Markov chain to characterize the velocity of each user. Specifically, the velocity of each user lies in \( V = \{ v_1, v_2, ..., v_U \} \), where \( \Delta v \triangleq v_{u+1} - v_u = 1 \text{ m/s} \), \( v_1 = 0 \text{ m/s} \), and \( v_U = 30 \text{ m/s} \), and the velocities are constant within each frame of duration \( \Delta T \).

With this velocity model, the velocity of a user may varies from \( 0 \sim 30 \text{ m/s} \) in the prediction window. Denote the velocity of the \( m \)th user in the \( i \)th frame as \( V^m_i \). We set \( \Delta v/\Delta T \) equal to the maximal acceleration of vehicles (e.g., \( 1 \text{ m/s}^2 \) \[52\]). The velocity can only transit between adjacent states (i.e., it can change \( \Delta v \) after \( \Delta T \)). The transition matrix of the Markov chain is

\[
P = \begin{bmatrix}
1 - q & q & 0 & \cdots \\
q & 1 - 2q & q & \cdots \\
0 & q & 1 - 2q & \cdots \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\end{bmatrix}, \tag{28}
\]

where the element at the \( i \)th row \( j \)th column represents the probability that the velocity transits from \( v_i \) to \( v_j \). For example, from the first column of \( P \), we know that if \( V^m_i = v_1 \), then \( \Pr\{V^m_{i+1} = v_1|V^m_i = v_1\} = 1 - q \), \( \Pr\{V^m_{i+1} = v_2|V^m_i = v_1\} = q \) and \( \Pr\{V^m_{i+1} = v_l|V^m_i = v_1\} = 0, l > 2 \). For
\( q = 0 \), the velocity is constant, and always equals to the initial value. By increasing the value of \( q \), the prediction uncertainty of velocities in the upcoming \( N_L \) frames increases.

Since EE can be improved evidently when the traffic load is light, we set \( M_D = M_R = 5 \), which is half of the maximal number of users that can be supported by the BSs. The initial positions of the users are uniformly distributed in the first cell. The initial velocity is set to be 20 m/s. We do not study how to predict the trajectory of each user, and apply a simple way to illustrate the performance of different policies. Specifically, the predicted locations of the user is obtained by assuming that each user travels along the predicted route with the initial velocity.

When the predicted large-scale channel gains, denoted as \( \{ \hat{\alpha}_m^i, i = 1, ..., N_L \} \), are inaccurate, the optimal policy needs to adjust resources to ensure QoS. With \( \{ \hat{\alpha}_m^i, i = 1, ..., N_L \} \), resource allocation plan \( \{ \hat{P}_m^i, \hat{K}_m^i, i = 1, ..., N_L \} \) can be obtained by solving problem (19). At the beginning of the \( i \)th frame, if the large-scale channel gain estimated at the BS \( \alpha_m^i \neq \hat{\alpha}_m^i \), then we apply a simple adjustment that does not need to solve another optimization problem: the BS adjusts average transmit power according to \( \alpha_m^i \hat{P}_m^i = \hat{\alpha}_m^i \hat{P}_m^i \), and \( \hat{K}_m^i \) does not change. With the adjustment, the constraints of problem (19) can be satisfied.

![Fig. 6. EE v.s. uncertainty of velocity, the adjusted optimal policy is with legend “Optimal-A”.

The EE achieved by different policies is shown in Fig. 6 where the qualities of all the VoD users are the same (all six data layers are transmitted before playback). The results show that the

\(^5\)According to simulations, the uncertainty of velocity modelled in the sequel will lead to 200 ∼ 300 % prediction errors on large-scale channel gains at the end of a prediction window with 60 seconds duration.
heuristic policy is robust to the prediction uncertainty. Even when $q = 0.5$, which leads to over 300% prediction errors on the large scale channel gains at the end of the prediction window, the EE loss of heuristic policy is negligible.

VII. CONCLUSION

In this paper, we studied how to optimize predictive resource allocation to maximize EE of a system with both VoD and RT services. The resource allocation policy was jointly optimized in two timescales for OFDMA system by harnessing the prediction of large-scale channel gains. At the beginning of prediction window, a resource allocation plan is made to assign the average transmit power and bandwidth for future frames, with the predicted large-scale channel gains. At the beginning of each time slot during the procedure of transmission, the transmit power is allocated to subcarriers according to the plan, with the estimated small-scale channel gains. We showed that predicting small-scale channel gains of VoD users are not beneficial to improve EE. Predicting the large-scale channel gains of RT users does not help improve EE and QoS if there are only RT users, but can improve the EE if there are both VoD and RT users. A heuristic policy was proposed, which is of low complexity and robust to the prediction errors on large-scale channel gains. Simulation results showed that with joint resource allocation for the two kinds of services in the two timescales, EE can be improved significantly. The heuristic policy performs closely to the optimal policy if the prediction on large-scale channel gains is accurate, and outperforms the optimal policy if the prediction is with large uncertainty.

APPENDIX A

PROOF OF PROPOSITION 1

Proof: To prove Proposition 1, we first prove that $f^w_D \left( \frac{\bar{P}_m}{K_i^m}, g \right)$ is the optimal power allocation policy for VoD services. For arbitrary power allocation policy for RT services $f'_R \left( \bar{P}_i^m, K_i^m, g \right)$, the optimal solutions of problem (13) with policies $f^w_D \left( \frac{\bar{P}_m}{K_i^m}, g \right)$ and $f'_D \left( \bar{P}_i^m, K_i^m, g \right)$ are denoted as $\{ \bar{P}_i^m, K_i^m, m = 1, ..., M_D + M_R, i = 1, ..., N_L \}$ and $\{ P_i^{m'}, K_i^{m'}, m = 1, ..., M_D + M_R, i = 1, ..., N_L \}$, respectively. Then, we need to prove

$$E^*_{ave} (f^w_D, f'_R) \leq E^*_{ave} (f'_D, f'_R), \quad (A.1)$$

where $E^*_{ave} (f^w_D, f'_R) = \sum_{m=1}^{M_D+M_R} \sum_{i=1}^{N_L} \left( \frac{1}{\rho} \bar{P}_i^m + P_c K_i^m \right)$ and $E^*_{ave} (f'_D, f'_R) = \sum_{m=1}^{M_D+M_R} \sum_{i=1}^{N_L} \left( \frac{1}{\rho} P_i^{m'} + P_c K_i^{m'} \right)$. 

July 7, 2017 DRAFT
Denote the average service rates achieved by the power allocation policies $f_D^w (\bar{P}_i^m, K_i^m, g)$ with resource allocation planning $\{P_i^{m'}, K_i^{m'}\}$ as $s_i^{m'}, m = 1, ..., M_D + M_R, i = 1, ..., N_L.$

To prove (A.1), we need the following result: the water-filling policy $f_D^w (\frac{\bar{P}_i^m}{K_i^m}, g)$ can minimize $\bar{P}_i^m$ with given $K_i^m$ and average service rate $s_i^{m'}$ [24]. According to this result, given $K_i^{m'}$ and $s_i^{m'}, i = 1, ..., N_L$, the average transmit power is minimized with $f_D^w (\frac{\bar{P}_i^m}{K_i^m}, g)$. Denote the related minimal average transmit power for the $m$th user in the $i$th frame as $\min(\bar{P}_i^m).$ Then, $\min(\bar{P}_i^m) \leq P_i^{m'}, m = 1, ..., M_D, i = 1, ..., N_L.$ Hence

$$\sum_{m=1}^{M_D} \sum_{i=1}^{N_L} \left[ \frac{1}{\rho} \min(\bar{P}_i^m) + P_i K_i^{m'} \right] + \sum_{m=M_D}^{M_D+M_R} \sum_{i=1}^{N_L} \left( \frac{1}{\rho} P_i^{m'} + P_i K_i^{m'} \right) \leq \sum_{m=1}^{M_D} \sum_{i=1}^{N_L} \left( \frac{1}{\rho} P_i^{m'} + P_i K_i^{m'} \right).$$

(A.2)

Moreover, with $f_D^w (\frac{\bar{P}_i^m}{K_i^m}, g)$, the optimal resource allocation plan is $\{\bar{P}_i^m, \tilde{K}_i^m, m = 1, ..., M_D + M_R, i = 1, ..., N_L\}.$ Thus,

$$\sum_{m=1}^{M_D+M_R} \sum_{i=1}^{N_L} \left( \frac{1}{\rho} \bar{P}_i^m + P_i \tilde{K}_i^m \right) \leq \sum_{m=1}^{M_D} \sum_{i=1}^{N_L} \left[ \frac{1}{\rho} \min(\bar{P}_i^m) + P_i K_i^{m'} \right] + \sum_{m=M_D}^{M_D+M_R} \sum_{i=1}^{N_L} \left( \frac{1}{\rho} P_i^{m'} + P_i K_i^{m'} \right).$$

(A.3)

From (A.2) and (A.3), we have (A.1).

Similarly, given power allocation policy for VoD services $f_D^w (\frac{p_i^m}{K_i^m}, g)$, we can prove

$$E_\text{ave}^* (f_D^w, f_R^w) \leq E_\text{ave}^* (f_D^w, f_R^w).$$

(A.4)

The proof is omitted for conciseness. From (A.1) and (A.4) we have

$$E_\text{ave}^* (f_D^w, f_R^w) \leq E_\text{ave}^* (f_D^w, f_R^w) \leq E_\text{ave}^* (f_D^w, f_R^w).$$

(A.5)

This completes the proof.  

APPENDIX B

PROOF OF PROPERTY 2

Proof: The left hand side of (19b) is the perspective of $-\frac{1}{\theta m} \ln \left[ F_R (\bar{P}_i^m) \right]$, where $\bar{P}_i^m = \frac{\bar{P}_i^m}{K_i^m}$. To prove that the left hand side of (19b) is jointly concave in $\bar{P}_i^m$ and $K_i^m$, we only need to prove that $-\frac{1}{\theta m} \ln \left[ F_R (\bar{P}_i^m) \right]$ is concave in $\bar{P}_i^m$. For notational simplicity, we omit
Substituting (B.4), (B.5) and (B.6) into (B.3), we can derive that

\[
\frac{d^2 \ln F_R(\bar{P}_S)}{d^2 P_S} = \frac{F_R(\bar{P}_S) \frac{d^2 F_R(\bar{P}_S)}{d\bar{P}_S^2} - \left[ \frac{dF_R(\bar{P}_S)}{d\bar{P}_S} \right]^2}{\left[ \frac{dF_R(\bar{P}_S)}{d\bar{P}_S} \right]^2} > 0. \tag{B.1}
\]

Substituting (16) into (21), we can obtain that

\[
F_R(\bar{P}_S) = 1 - e^{-\nu} + \nu^{\frac{\beta}{\alpha}} \int_{\nu}^{\infty} g^{-\frac{\beta}{\alpha + 1}} e^{-g} dg \tag{B.2}
\]

where the relation between \( \nu \) and \( \bar{P}_S \) can be obtained from (17). Then, \( F_R(\bar{P}_S) \) can be regarded as a composition function \( F_R[\nu(\bar{P}_S)] \), and thus

\[
\frac{dF_R[\nu(\bar{P}_S)]}{dP_S} = \frac{dF_R}{d\nu} \frac{d\nu}{dP_S}, \quad \frac{d^2 F_R[\nu(\bar{P}_S)]}{dP_S^2} = \frac{d^2 F_R}{d\nu^2} \left( \frac{d\nu}{dP_S} \right)^2 + \frac{d^2 F_R}{d\nu} \frac{d^2 \nu}{dP_S^2}. \tag{B.3}
\]

From (17), we can derive the relation between \( \bar{P}_S \) and \( \nu \), i.e.,

\[
\frac{d\bar{P}_S}{d\nu} = -\frac{\phi \sigma_0^2}{\alpha(\beta + 1)} \nu^{-\frac{\beta + 2}{\beta + 1}} \int_{\nu}^{\infty} g^{-\frac{\beta}{\beta + 1}} e^{-g} dg.
\]

According to the characteristic of inverse function (i.e., \( \frac{d\bar{P}_S}{d\nu} \frac{d\nu}{dP_S} = 1 \) at any point \((\nu, \bar{P}_S)\)), we can derive \( \frac{d\nu}{dP_S} \) from (17), i.e.,

\[
\frac{d\nu}{dP_S} = -\frac{\alpha (\beta + 1)}{\phi \sigma_0^2} \frac{\nu^{\frac{\beta + 2}{\beta + 1}}}{\nu^\beta e^{-\nu}} - \frac{1}{\int_{\nu}^{\infty} g^{-\frac{\beta}{\beta + 1}} e^{-g} dg}. \tag{B.4}
\]

From \( \frac{d^2 \nu}{dP_S^2} = \frac{d^2 \nu}{d\nu} \frac{d\nu}{dP_S} \), we can derive that

\[
\frac{d^2 \nu}{dP_S^2} = \left( \frac{\alpha}{\phi \sigma_0^2} \right)^2 \left[ (\beta + 2) \nu^{\frac{1}{\beta + 1}} \frac{1}{\varphi} + (\beta + 1) \nu^{-\frac{\beta}{\beta + 1}} e^{-\nu} \frac{1}{\varphi^2} \right] \left[ (\beta + 1) \nu^{\frac{\beta + 2}{\beta + 1}} \frac{1}{\varphi} \right], \tag{B.5}
\]

where \( \varphi = \int_{\nu}^{\infty} g^{-\frac{\beta}{\beta + 1}} e^{-g} dg \).

From (B.2), we have

\[
\frac{dF_R}{d\nu} = \frac{\beta}{\beta + 1} \nu^{-\frac{\beta}{\beta + 1}} \varphi, \quad \frac{d^2 F_R}{d\nu^2} = -\frac{\beta}{(\beta + 1)^2} \nu^{-\frac{\beta + 2}{\beta + 1}} \varphi - \frac{\beta}{\beta + 1} \nu^{-1} e^{-\nu}. \tag{B.6}
\]

Substituting (B.4), (B.5) and (B.6) into (B.3), we can derive that

\[
\frac{dF_R[\nu(\bar{P}_S)]}{dP_S} = -\frac{\alpha}{\phi \sigma_0^2} \beta \nu, \quad \frac{d^2 F_R[\nu(\bar{P}_S)]}{dP_S^2} = \left( \frac{\alpha}{\phi \sigma_0^2} \right)^2 (\beta + 1) \nu^{\frac{\beta + 2}{\beta + 1}} \frac{1}{\varphi}. \tag{B.7}
\]
Upon substituting (B.7), the numerator of (B.1) can be derived as follows,

\[(1 - e^{-\nu}) \left( \frac{\alpha}{\phi \sigma^2_0} \right)^{\beta} (\beta + 1) \nu^{\frac{\beta + 2}{\phi}} + \left( \frac{\alpha}{\phi \sigma^2_0} \right)^{\beta} \nu^2. \] (B.8)

Since \(\varphi = \int_0^\infty g^{-\frac{\beta}{\beta+1}} e^{-g} dg > 0\), \(\beta = \frac{\theta \tau B}{\ln 2} > 0\), (B.8) is positive. Therefore, we have (B.1).

The proof follows.

REFERENCES

[1] C. She and C. Yang., “Context aware energy efficient optimization for video on-demand service over wireless networks,” in IEEE ICC, 2015.
[2] S. Zhang, Q. Wu, S. Xu, and G. Y. Li, “Fundamental green tradeoffs: Progresses, challenges, and impacts on 5G networks,” IEEE Commun. Surveys Tuts., vol. 19, no. 1, pp. 33–56, 2017.
[3] Q. Wu, G. Y. Li, W. Chen, D. W. K. Ng, and R. Schober, “An overview of sustainable green 5G networks,” IEEE Wireless Commun. [Online]. Available: https://arxiv.org/abs/1609.09773
[4] C. Song, Z. Qu, N. Blumm, and A.-L. Barabasi, “Limits of predictability in human mobility,” Science, vol. 327, no. 5968, pp. 1018 – 1021, Feb. 2010.
[5] 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.
[6] N. Bui, M. Cesana, S. A. Hosseini, Q. Liao, I. Malanchini, and J. Widmer, “A survey of anticipatory mobile networking: Context-based classification, prediction methodologies, and optimization techniques,” IEEE Commun. Surveys Tuts., early access, 2017.
[7] A. Nadembega, A. S. Hafid, and T. Taleb, “A destination and mobility path prediction scheme for mobile networks,” IEEE Trans. Veh. Technol., vol. 64, no. 6, pp. 2577–2590, Jun. 2015.
[8] A. Osseiran, F. Boccardi and V. Braun, et al., “Scenarios for 5G mobile and wireless communications: The vision of the METIS project,” IEEE Commun. Mag, vol. 52, no. 5, pp. 26 – 35, May. 2014.
[9] Cisco, “Cisco visual networking index: Global mobile data traffic forecast update, 2014-2019,” Cisco white paper, 2015.
[10] P. Juluri, V. Tamarapalli, and D. Medhi, “Measurement of quality of experience of video-on-demand services: A survey,” IEEE Commun. Surveys Tuts., vol. 18, no. 1, pp. 401–418, Firstquarter, 2016.
[11] 3GPP, Further Advancements for E-UTRA Physical Layer Aspects. TSG RAN TR 36.814 v9.0.0, Mar. 2010.
[12] S. Wee-Seng and S. K. Hyong, “A predictive bandwidth reservation scheme using mobile positioning and road topology information,” IEEE/ACM Trans. Netw., vol. 14, no. 5, pp. 1078–1091, Oct. 2006.
[13] A. Nadembega, A. Hafid, and T. Taleb, “Mobility-prediction-aware bandwidth reservation scheme for mobile networks,” IEEE Trans. Veh. Technol., vol. 64, no. 6, pp. 2561–2576, Jun. 2015.
[14] T. Anagnostopoulos, C. Anagnostopoulos, and S. Hadjispyridis, “Efficient location prediction in mobile cellular networks,” Springer Int. J. Wireless Inf. Netw., vol. 19, no. 2, pp. 97–111, Jun. 2012.
[15] H. Abou-zeid, H. S. Hassanein, and S. Valentin, “Optimal predictive resource allocation: Exploiting mobility patterns and radio maps,” in Proc. IEEE Globecom, Dec. 2013.
[16] ———, “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] H. Abou-zeid, H. S. Hassanein, and R. Atawia, “Towards mobility-aware predictive radio access: Modeling, simulation, and evaluation in LTE networks,” in ACM MSWiM, 2014.
[18] J. Yao, S. S. Kanhere, and M. Hassan, “Improving QoS in high-speed mobility using bandwidth maps,” IEEE Trans. Mobile Comput., vol. 11, no. 4, pp. 603–617, Apr. 2012.
[19] N. Bui and J. Widmer, “Mobile network resource optimization under imperfect prediction,” in Symp. IEEE WoWMoM, Jun. 2015.
[20] R. Atawia, H. Abou-zeid, H. S. Hassanein, and A. Noureldin, “Joint chance-constrained predictive resource allocation for energy-efficient video streaming,” IEEE J. Sel. Areas Commun., vol. 34, no. 5, pp. 1389–1404, May 2016.
[21] 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.
[22] R. Margolies, A. Sridharan, V. Aggarwal, R. Jana, N. K. Shankaranarayanan, V. A. Vaishampayan, and G. Zussman, “Exploiting mobility in proportional fair cellular scheduling: Measurements and algorithms,” in Proc. IEEE INFOCOM, Apr. 2014.
[23] C. Yao, C. Yang, and Z. Xiong, “Energy-saving predictive resource planning and allocation,” IEEE Trans. Commun., vol. 64, no. 12, pp. 5078–5095, Dec. 2016.
[24] A. Goldsmith, Wireless Communications. Cambridge University Press, 2005.
[25] L. Liu, Y. Yi, C. J.-F., and J. Zhang, “Energy-efficient power allocation for delay-sensitive multimedia traffic over wireless systems,” IEEE Trans. Veh. Technol., vol. 63, no. 5, pp. 2038 – 2047, Mar. 2014.
[26] T. Taleb and A. Ksentini, “Follow me cloud: Interworking federated clouds and distributed mobile networks,” IEEE Network, vol. 27, no. 5, pp. 12–19, Oct. 2013.
[27] A. Nadembega, A. S. Hafid, and R. Brisebois, “Mobility prediction model-based service migration procedure for follow me cloud to support QoS and QoE,” in Proc. IEEE ICC, May 2016.
[28] D. Bethanabhotla, G. Caire, and M. J. Neely, “Adaptive video streaming for wireless networks with multiple users and helpers,” IEEE Trans. Commun., vol. 63, no. 1, pp. 268–285, Jan. 2015.
[29] C.-S. Chang, “Stability, queue length, and delay of deterministic and stochastic queueing networks,” IEEE Trans. Automatic Control, vol. 39, no. 5, pp. 913–931, May 1994.
[30] D. Wu and R. Negi, “Effective capacity: A wireless link model for support of quality of service,” IEEE Trans. Wireless Commun., vol. 2, no. 4, pp. 630–643, Jul. 2003.
[31] A. Helmy, L. Musavian, and T. Le-Ngoc, “Energy-efficient power adaptation over a frequency-selective fading channel with delay and power constraints,” IEEE Trans. Wireless Commun., vol. 12, no. 9, pp. 4529 – 4541, Sep. 2013.
[32] L. Liu, P. Parag, J. Tang, W.-Y. Chen, and J.-F. Chamberland, “Resource allocation and quality of service evaluation for wireless communication systems using fluid models,” IEEE Trans. Inf. Theory, vol. 53, no. 5, pp. 1767–1777, May 2007.
[33] C. Desset, and B. Debaillie, V. Giannini and et al., “Flexible power modeling of LTE base stations,” in Proc. IEEE WCNC, Apr. 2012.
[34] Q. Wu, W. Chen, M. Tao, J. Li, H. Tang, and J. Wu, “Resource allocation for joint transmitter and receiver energy efficiency maximization in downlink OFDMA systems,” IEEE Trans. Commun., vol. 63, no. 2, pp. 416–430, Dec. 2015.
[35] S. Zhang, Y. Chen, and S. Xu, “Joint bandwidth-power allocation for energy efficient transmission in multi-user systems,” in IEEE Globecom Workshops, Dec. 2010.
[36] G. Auer, O. Blume, V. Giannini, I. Góbor, et al., “D 2.3: Energy efficiency analysis of the
reference systems, areas of improvements and target breakdown,” *EARTH*, Jan. 2012. [Online]. Available: [https://www.ict-earth.eu/publications/deliverables/deliverables.html](https://www.ict-earth.eu/publications/deliverables/deliverables.html)

[37] J. Tang and X. Zhang, “Quality-of-service driven power and rate adaptation for multichannel communications over wireless links,” *IEEE Trans. Wireless Commun.*, vol. 6, no. 12, pp. 4349–4360, Dec. 2007.

[38] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge Univ. Press, 2004.

[39] H. Riiser, T. Endestad, P. Vigmostad, C. Griwodz, and P. Halvorsen, “Video streaming using a location-based bandwidth-lookup service for bitrate planning,” *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 8, no. 3, pp. 24:1–24:19, Jul. 2012.

[40] X. K. Zou, J. Erman, V. Gopalakrishnan, E. Halepovic, R. Jana, X. Jin, J. Rexford, and R. K. Sinha, “Can accurate predictions improve video streaming in cellular networks?” in *ACM HotMobile*, 2015.

[41] N. Bui, F. Michelinakis, and J. Widmer, “A model for throughput prediction for mobile users,” in *Proc. of European Wireless*, 2014.

[42] J. Devore, *Probability and Statistics for Engineering and the Sciences*. Cengage Learning, 2015.

[43] C. Xiong, G. Y. Li, Y. Liu, Y. Chen, and S. Xu, “Energy-efficient design for downlink OFDMA with delay-sensitive traffic,” *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 3085–3095, Jun. 2013.

[44] A. Aijaz, M. Tshangini, M. R. Nakhai, X. Chu, and A.-H. Aghvami, “Energy-efficient uplink resource allocation in LTE networks with M2M/H2H co-existence under statistical QoS guarantees,” *IEEE Trans. Wireless Commun.*, vol. 62, no. 7, pp. 2353–2365, Jul. 2014.

[45] P. Seeling and M. Reisslein, “Video transport evaluation with H.264 video traces,” *IEEE Commun. Surveys Tuts.*, vol. 14, no. 4, pp. 1142 – 1165, 2012.

[46] Video Trace Library. [Online]. Available: [http://trace.eas.asu.edu/video traces2/cgs/cif/Sony_G16B15_CIF_DQP6_5EL_48_42_36_30_24_18/](http://trace.eas.asu.edu/video traces2/cgs/cif/Sony_G16B15_CIF_DQP6_5EL_48_42_36_30_24_18/)

[47] C. Desset, B. Debaillie, and F. Louagie, “Modeling the hardware power consumption of large scale antenna systems,” in *Proc. IEEE OnlineGreenComm*, Nov. 2014.

[48] J. Froehlich and J. Krumm, “Route prediction from trip observations,” Tech. Rep., Apr. 2008.

[49] H. Abou-zeid, H. Hassanein, Z. Tanveer, and N. AbuAli, “Evaluating mobile signal and location predictability along public transportation routes,” in *IEEE WCNC*, 2015.

[50] P. A. Zandbergen, “Accuracy of iPhone locations: A comparison of assisted GPS, WiFi and cellular positioning,” *Trans. GIS*, vol. 13, no. s1, pp. 5–25, June 2009.

[51] M. Kasparick, R. Cavalcante, S. Valentim, S. Stanczak, and M. Yukawa, “Kernel-based adaptive online reconstruction of coverage maps with side information,” *IEEE Trans. Veh. Technol., early access*, July 2015.

[52] H. A. Omar, W. Zhang, A. Abd Rabou, and L. Li, “Performance evaluation of VeMAC supporting safety applications in vehicular networks,” *IEEE Trans. Emerging Topics in Computing*, vol. 1, no. 1, pp. 69 – 83, Jun. 2013.