Energy Efficient Resource Allocation for Hybrid Services with Future Channel Gains

Changyang She and Chenyang Yang

Abstract

In this paper, we propose a framework to maximize energy efficiency (EE) of a system supporting real-time (RT) and non-real-time services by exploiting future average channel gains of mobile users, which change in the timescale of seconds and are reported predictable within a minute-long time window. To demonstrate the potential of improving EE by jointly optimizing resource allocation for both services by harnessing both future average channel gains and current instantaneous channel gains, we optimize a two-timescale policy with perfect prediction, by taking orthogonal frequency division multiple access system serving RT and video-on-demand (VoD) users as an example. Considering that fine-grained prediction for every user is with high cost, we propose a heuristic policy that only needs to predict the median of average channel gains of VoD users. Simulation results show that the optimal policy outperforms relevant counterparts, indicating the necessity of the joint optimization for both services and for two timescales. Besides, the heuristic policy performs closely to the optimal policy with perfect prediction while becomes superior with large prediction errors. This suggests that the EE gain over non-predictive policies can be captured with coarse-grained prediction.

Index Terms

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

I. Introduction

To support the ever-growing traffic demands, the main trend techniques of future mobile communications are exploring wider spectrum and deploying more antennas or base stations (BSs). Yet due to temporal-spatial traffic fluctuations, existing cellular networks, usually optimized for fully loaded scenarios, are often observed not-busy in many places. According to the recent data analysis in [2], only a small portion of radio resources in Long Term Evolution (LTE) networks are truly used in average. To avoid wasting resources when the system is not fully loaded, energy efficiency (EE) becomes a key performance metric for cellular networks [3]. Recently, energy-efficient resource allocation has been investigated extensively in literature [4–8].

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

The dynamic nature of traffic load comes from user behaviors such as mobility and activities, which change in a much longer timescale than channel state information (CSI) and have long been regarded as random in wireless system design. However, the research efforts in other domains demonstrate that some user behaviors, say mobility pattern, are highly predictable [9–12]. With the predicted trajectory, radio resources can be allocated adaptive to network dynamics caused by user mobility. This provides a promising way to circumvent the resource under-utilization. By harnessing future information in a minute-level time horizon, predictive resource allocation (PRA) has been shown to provide remarkable gain in terms of improving network EE, throughput, and user experience than the non-predictive counterparts [12–21]. The gain of PRA has been validated by recent data-driven analysis [2, 16, 22].

Prevalent resource allocation policies are non-predictive, which are optimized with instantaneous or average channel gain in the current time slot or frame varying in the timescales of milliseconds or seconds [4–8]. Different from these policies, PRA leverages future information in a minutes-long window [13, 18, 19]. By predicting trajectory [10, 23] and constructing radio map [24, 25], the future average channel gain (also called large-scale channel gain interchangeably in the sequel) in each frame can be predicted [23]. By using the historical record of the modulation and coding scheme for a mobile user, its average data rate in each frame of a 30-seconds time window is predicted in [2]. Though predicting information in such horizon is possible with machine learning, the prediction itself incurs extra costs for training and data gathering [10,23–25]. To achieve the gain of PRA at affordable costs, it is critical to study what information needs to be predicted and how to exploit different information effectively.

Maximizing the EE of a network should not compromise the quality-of-service (QoS) of users. Future cellular networks need to support diverse services with different QoS provision [26]. One type is real-time (RT) services such as video conference and voice over IP that require stringent QoS [26]. For this type of services, a data packet becomes useless once its required delay is violated. Hence, the QoS is characterized by the *statistical QoS requirement*, defined as a delay bound and a delay bound violation probability, whose values depend on specific service [27]. The other type is non-real-time (NRT) services such as file downloading and video-on-demand (VoD). For VoD services, the video quality and playback interruption are key metrics for user experience [28]. To meet the demands of different services efficiently, softwarization techniques such as network function virtualization (NFV) and software-defined networking (SDN) are proposed for the fifth generation (5G) networks [29, 30]. SDN manages radio resources and traffic flows in a centralized manner with a global view of the network state [30]. NFV is a viable way to provide a network slice tailored to each service [31]. In fact, with the global view of future average channel gains, the network performance can be improved by jointly optimizing predictive resource allocation for different types of services subject to the QoS of each user. However, existing works in the area of NFV/SDN focus on how to meet the demands of each kind of services rather than ensure the QoS of each user, and do not investigate how to harness the predictable trajectories of mobile users.

A. Related Works

Predictive resource allocation has been optimized separately for RT and NRT services.

For users requesting RT services, PRA is usually designed for improving admission level QoS via mobility management with cell-level mobility prediction [11, 12, 14, 32]. By predicting the future handoff time and the BS that a RT user will access to, the bandwidth at the next BS was reserved for the user [32], and a call admission control scheme was proposed in [14]. By predicting the next several cells a RT user will enter, the delay caused by handoff and signaling is reduced significantly [11].

For users requesting NRT services, PRA is usually designed for boosting network performance such as EE or QoS of mobile NRT users with fine-grained prediction [13, 15–19, 22, 33]. Most existing works of PRA consider VoD service. With known future instantaneous channel gains, the trade-off between the required resources and the stalling time was investigated in [15]. With known future average data rates in the frames of a prediction window, the number of time slots in each frame was optimized in [13] to save energy for ensuring the QoS of each VoD user. In [33], a practical two-timescale PRA was proposed for LTE systems. In the first timescale, the number of time slots is optimized based on the rate prediction at the start of the prediction window, while in the second timescale the subcarriers are allocated in each time slot based on the instantaneous channel gains. Considering that future data rates cannot be predicted without errors, a robust PRA policy was optimized in [18] by assuming bounded prediction errors. Further considering that the time resource occupied by RT services is uncertain due to the random request arrival, a robust PRA for VoD service was optimized in [19]. A common assumption in [13, 18, 19, 33] is that the future data rate of each user is predictable. However, the data rate of a wireless link depends on the resource allocation among users, i.e., the rate prediction is coupled with predictive resource allocation. This suggests that PRA with rate prediction is non-optimal. There also exist a few works of PRA considering the service of file downloading [16, 17]. By using future average channel gains, a proportional fair scheduling policy was proposed in [16]. With both future average channel gains of NRT users and average arrival rate of RT traffic, an energy-saving PRA policy was proposed in [17], where radio resources are reserved for RT services.

B. Motivations and Contributions

All previous PRA policies are only optimized for a single kind of services. All policies are either optimized in one timescale or separately designed in two timescales. All existing policies are designed based on the fine-grained information (say trajectory or rate in each second) of every mobile user. While technically viable, predicting fine-grained information incurs high costs. For example, to predict fine-grained average channel gains, one needs to predict a fine-grained trajectory for every user. This requires a large number of training samples and high computational complexity for the off-line training [23]. Besides, one needs to establish a fine-grained radio map for the network, where the average channel gains between each location and surrounding BSs need to be measured and stored, say by expensive drive tests [25].

PRA policies can be optimized toward different objectives, which need very different techniques to find the optimal solutions. In this paper, we consider a not-fully-loaded network. While throughput-maximal PRA can boost the maximal number of users/requests that the network is able to support, EE-maximal PRA can save resources when the traffic load is not heavy, which is often the case in real-world cellular networks [2].

Despite that prior works have demonstrated the potential of PRA, the following questions, which are important before PRA is put into practice use, remain open: 1) To maximize the EE of a network, do we need to jointly optimize PRA for different types of services over multiple timescales? 2) Which kinds of future channel information are needed to maximize EE? 3) Is it possible to approach the maximal EE with coarse-grained future information? We strive to answer these questions in this paper. Since the majority of data traffic is from mobile videos, we take VoD as an example of NRT services. Our major contributions are summarized as follows:

- To show the potential of the joint optimization, we propose a framework to joint optimize PRA that maximizes the EE of network subject to the QoS requirements for both VoD and RT users by using two-timescale channel information. Finding the optimal solution is challenging, because optimizing the policies in two timescales turns out a functional extreme problem, which can not be solved by directly using convex optimization tools. To provide a baseline of comparison for the heuristic policy with coarse-grained prediction, we assume that the fine-grained future average channel gains for both types of users are perfectly known as in the existing works. Simulation results show that jointly optimizing PRA for both types of services and for two timescales can improve EE significantly.

- To show which kind of future information is necessary to maximize EE, we analyze the degenerated optimization problem for the system serving only RT or VoD users. We find that predicting the average channel gains in the prediction window is helpful, but further predicting instantaneous channel gains in future time slots cannot improve EE.
- To illustrate that PRA can achieve high EE even with coarse-grained prediction, we propose a heuristic policy, inspired by the structure of the optimization problem. This policy only needs the median of future average channel gains of VoD users. Surprisingly, the heuristic policy outperforms the optimal policy when the prediction errors are large, thanks to the fact that a median is insensitive to errors.

II. System Model and QoS requirements

Consider the scenario that multiple mobile users travel across the cells of an orthogonal frequency division multiple access (OFDMA) network. A user either requests for VoD or requests for RT service. For notational simplicity, we first consider a single cell scenario in this section and then extend to the multi-cell scenario at the end of the next section.

A. Transmission and Channel Models

Consider frequency-selective block fading channel. Time is discretized to frames each with duration ΔT and time slots each with duration τ. The durations are defined according to the variation of large-scale channel gain and small-scale channel gain due to user mobility, respectively. The large-scale channel gains are predictable within a prediction window, with the predicted trajectories and a measured radio map [13]. The small-scale channel gains (called instantaneous channel gains interchangeably in this work, also called CSI in literature) are predictable [34] within the channel coherence time (i.e., within τ). For simplicity, we 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 time slots in each frame and subcarriers.

Each frame includes $N_{\rm S}$ time slots, i.e., $\Delta T = N_S \tau$. A prediction window includes N_L successive frames. At the beginning of a prediction window, the average channel gains in future frames within the window of both types of users are assumed known at the BS. However, CSI is only known at the BS and the user at the beginning of each time slot.

There are $M_D + M_R$ users that access to the BS at the beginning of a prediction window, where M_D and M_R are the numbers of users requesting VoD and RT services, respectively. For the mth user, α_i^m is the average channel gain in the ith frame, and g_{ijk}^m is the CSI on the kth subcarrier in the jth time slot of the ith frame.

The achievable instantaneous data rate for the mth user can be expressed as follows,

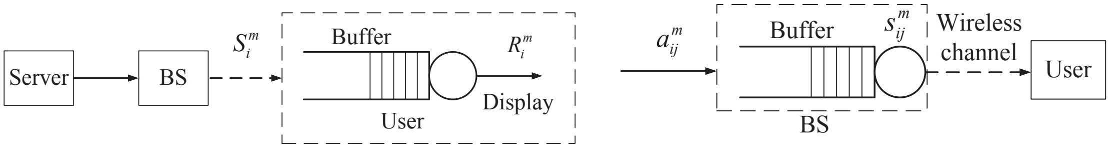
$$s_{ij}^m = B \sum_{k=1}^{K_i^m} \log_2 \left(1 + \frac{\alpha_i^m}{\phi \sigma_0^2} p_{ijk}^m g_{ijk}^m \right) \quad \text{bits/s}, \tag{1}$$

where B is the subcarrier spacing, p_{ijk}^m is the transmit power allocated to the mth user on the kth subcarrier in the jth time slot of the ith frame, $\phi > 1$ captures the gap between capacity and achievable rate with practical modulation and coding schemes, σ_0^2 is the variance of the additive Gaussian noise, and K_i^m is the number of subcarriers assigned to the mth user in the ith frame.

B. QoS Requirement for VoD Service

Since the key factor that determines the experience of a user requesting a 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 as in [13, 18, 35]. The queueing model for VoD service is shown in Fig. 1(a), where R_i^m is the amount of data played at the mth user in the ith frame. The value of R_i^m 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 mth user during the ith frame is given by $S_i^m = \tau \sum_{j=1}^{N_S} s_{ij}^m$.

Denote the duration of each video segment as T_{seg}. Without loss of generality, we set $T_{seg} = \Delta T$ for notational simplicity. 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. This is reasonable for smartphones since storage devices are cheap nowadays. The assumption will be removed in Section IV, where we design a heuristic policy that is aware of limited buffer size.



(a) Queueing model for the mth VoD user.

(b) Queueing model for the mth RT user.

Fig. 1. Queueing models for VoD and RT services.

To avoid stalling during playback, each video segment should be delivered to a VoD user before the segment is played. Then, the QoS required by the VoD user can be reflected by the following constraint [13],

$$Q_0^m + \sum_{i=1}^{l} S_i^m \geq \sum_{i=1}^{l+1} R_i^m, l = 1, ..., N_L, m = 1, ..., M_D, \tag{2}$$

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. Hence, no interruption occurs between the adjacent prediction windows.[1] Scalable video coding (SVC) is used to encode videos, i.e., each video segment is encoded into one base layer and multiple enhancement layers [36]. When the channel quality is not good such that the data rate cannot satisfy the requirement in (2), we can reduce the value of R_i^m by not transmitting some enhancement layers. In this way, we can reduce the stalling probability at the cost of sacrificing the definition of the video.

Since the number of time slots in each frame is large in practice, by channel coding among time slots, the time-average data rate in a frame can approach the ensemble-average data rate [37]. From (1), the average data rate for the mth user in the ith frame can be expressed as,

$$\bar{s}_i^m = B \sum_{k=1}^{K_i^m} \mathbb{E}_h \left[\log_2 \left(1 + \frac{\alpha_i^m}{\phi \sigma_0^2} p_{ijk}^m g_{ijk}^m \right) \right] \quad \text{bits/s}, \tag{3}$$

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

$$\sum_{i=1}^{l} \bar{s}_i^m \geq \frac{1}{\Delta T} \sum_{i=2}^{l+1} R_i^m, l = 1, ..., N_L, m = 1, ..., M_D. \tag{4}$$

[1] At the beginning of the first prediction window, the user only needs to download the video segment played in the first frame. This will not lead to long waiting time at the beginning of the service.

Remark 1. Other NRT services such as file downloading, whose user demand can be characterized as to transmit a file with size $\tilde{R}^m$ in N_L frames, can also be included in our framework. Its QoS requirement can be expressed as $\sum\limits_{i=1}^{N_L} \bar{s}_i^m \geq \tilde{R}^m$, which is similar to (4).

C. QoS Requirement for Real-time Service

The queueing model for the mth user requesting RT service is shown in Fig. 1(b), where a_{ij}^m represents the data arrival rate in the jth time slot of the ith frame. If the queueing delay in the mth queue exceeds a delay bound $D_{\max}^m$ with a delay violation probability less than ε_D^m, then the QoS requirement of the mth RT user can be satisfied [4, 5, 38, 39].

Effective bandwidth and effective capacity are widely applied tools in designing resource allocation with such statistical QoS requirement [40,41].[2] For uncorrelated random arrival process and service process, $\{a_{ij}^m\}$ and $\{s_{ij}^m\}$, the effective bandwidth and effective capacity can be expressed as $E_B^m\left(\theta^m\right) = \frac{1}{\theta^m \tau} \ln \mathbb{E}\left[\exp\left(\theta^m \tau a_{ij}^m\right)\right]$ (bits/s) and

$$E_{C_i}^m\left(\theta^m\right) = -\frac{1}{\theta^m \tau} \ln \mathbb{E}_{g_{ijk}^m}\left[\exp\left(-\theta^m \tau s_{ij}^m\right) | \alpha_i^m\right] \quad \text{(bits/s)}, \tag{5}$$

respectively [40, 41], where where θ^m is the *QoS exponent*. The required QoS exponent θ^m to guarantee $(D_{\max}^m, \varepsilon_D^m)$ can be obtained from [5], i.e.,

$$\Pr\{D_\infty^m > D_{\max}^m\} \approx \exp\left[-\theta^m E_B^m(\theta^m) D_{\max}^m\right] = \varepsilon_D^m, \tag{6}$$

where D_∞^m is the steady state delay for the mth user. To ensure the QoS of the mth RT user over wireless channels, the following constraint should be satisfied [39]

$$E_{C_i}^m\left(\theta^m\right) \geq E_B^m\left(\theta^m\right), m = M_D + 1, ..., M_D + M_R, i = 1, ..., N_L. \tag{7}$$

D. Power Consumption Model and EE Definition

The total energy consumed by transmit power and circuit power at the BS for serving $M_D + M_R$ users in the prediction window (i.e., in N_L frames) can be modeled as [42]

$$\sum_{i=1}^{N_L} E_i = \sum_{i=1}^{N_L} \left(\frac{1}{\rho} \sum_{m=1}^{M_D+M_R} \sum_{j=1}^{N_S} \sum_{k=1}^{K_i^m} \tau p_{ijk}^m + \Delta T P_c \sum_{m=1}^{M_D+M_R} K_i^m + \Delta T P_0 \right), \tag{8}$$

[2] The term "effective bandwidth" is not the spectrum resource in radio access networks. According to the definition in [40], it is the minimal constant service rate that is required to ensure the QoS of a RT user with random arrived packets.

where E_i is the energy consumption in the ith 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 [43], EE of a system is the ratio of the amount of data transmitted to the energy consumed during a certain period. For PRA, the period is the duration of the prediction window. However, since only the average 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 instantaneous channel gains. As a result, we cannot optimize PRA to maximize the EE metric in [43]. 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,

$$\eta \triangleq \left[\mathbb{E}_h\left(\sum_{m=1}^{M_D}\sum_{i=1}^{N_L}\tau\sum_{j=1}^{N_S} s_{ij}^m\right) + \mathbb{E}_h\left(\sum_{m=M_D+1}^{M_D+M_R}\sum_{i=1}^{N_L}\tau\sum_{j=1}^{N_S} b_{ij}^m\right)\right] \Bigg/ \left[\mathbb{E}_h\left(\sum_{i=1}^{N_L} E_i\right)\right]. \tag{9}$$

For VoD services, the amount of data transmitted equals the amount of data that needs to transmit. Thus, $\mathbb{E}_h\left(\sum\limits_{m=1}^{M_D}\sum\limits_{i=1}^{N_L}\tau\sum\limits_{j=1}^{N_S} s_{ij}^m\right) = \sum\limits_{m=1}^{M_D}\sum\limits_{i=1}^{N_L}\Delta T \bar{s}_i^m = \sum\limits_{m=1}^{M_D}\sum\limits_{i=2}^{N_L+1} R_i^m$, 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 [44]. Thus, $\mathbb{E}_h\left(\sum\limits_{m=M_D+1}^{M_D+M_R}\sum\limits_{i=1}^{N_L}\tau\sum\limits_{j=1}^{N_S} b_{ij}^m\right) = \mathbb{E}_h\left(\sum\limits_{m=M_D+1}^{M_D+M_R}\sum\limits_{i=1}^{N_L}\tau\sum\limits_{j=1}^{N_S} a_{ij}^m\right)$, which is determined by the arrival processes. Therefore, the numerator of (9) does not depend on the resource allocation. Further considering that the last term in (8) is a constant, maximizing EE is equivalent to minimizing the following average energy consumption,

$$\frac{1}{\rho}\mathbb{E}_h\left(\sum_{m=1}^{M_D+M_R}\sum_{i=1}^{N_L}\sum_{j=1}^{N_S}\sum_{k=1}^{K_i^m}\tau p_{ijk}^m\right) + \Delta T P_c \sum_{m=1}^{M_D+M_R}\sum_{i=1}^{N_L} K_i^m. \tag{10}$$

III. Energy Efficient Predictive Resource Allocation

In this section, we optimize predictive resource allocation under the assumption that the average channel gains in future frames of the window are perfectly known. We formulate a functional extreme problem and obtain the global optimal solution, referred to as *ideal policy* for short. We first consider the single cell scenario and then extend to the multi-cell scenario.

A. Problem Formulation

At the beginning of a prediction window, we cannot optimize p_{ijk}^m to minimize (10) since the instantaneous channel gains in future time slots 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 mth user in the ith frame, since the future average channel gains 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*. It determines the amount of resources assigned to the users in all frames of the prediction window.

At the beginning of each time slot, we can optimize p_{ijk}^m according to the assigned resources in the corresponding frame $\{\bar{P}_i^m, K_i^m\}$, since the instantaneous channel gains g_{ijk}^m, $k = 1, ..., K_i^m$ are available at the BS. To gain useful insight, here we only consider power allocation, but not subcarrier allocation. We denote the *power allocation policies* for the VoD users and the RT users 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 $i = 1, ..., N_L$, $j = 1, ..., N_S$ and $k = 1, ..., K_i^m$. 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 coupled. In what follows, we formulate the joint optimization problem for the two-timescale policy. We take Rayleigh fading as an example, but the methodology can be extended to the other channels.

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 ith frame for Rayleigh fading can be expressed as follows,

$$\bar{s}_i^m = K_i^m \int_0^\infty B\log_2\left[1 + \frac{\alpha_i^m}{\phi\sigma_0^2} f_D\left(\bar{P}_i^m, K_i^m, g\right) g\right] e^{-g} \mathrm{d}g, \tag{11}$$

where $m = 1, ..., M_D$, and g is exponentially distributed with the mean of 1.

Substituting the power allocation policy for RT service $p_{ijk}^m = f_R(\bar{P}_i^m, K_i^m, g_{ijk}^m)$ into (1) and then into (5), the effective capacity in the ith frame for Rayleigh fading can be obtained as

$$E_{C_i}^m(\theta^m) = -\frac{K_i^m}{\theta^m \tau} \ln\left\{\int_0^\infty \left[1 + \frac{\alpha_i^m}{\phi\sigma_0^2} f_R\left(\bar{P}_i^m, K_i^m, g\right) g\right]^{-\beta^m} e^{-g} \mathrm{d}g\right\} \quad \text{(bits/s)}, \tag{12}$$

where $m = M_D + 1, ..., M_D + M_R$, and $\beta^m \triangleq \frac{\theta^m \tau B}{\ln 2}$.

Then, the optimal two-timescale policy that maximizes the EE with satisfied QoS requirement of each RT user and each VoD user can be obtained by solving the following problem,

$$\min_{\substack{f_D(\cdot),f_R(\cdot),\bar{P}_i^m,K_i^m,\\ i=1,...,N_L,\\ m=1,...,M_D+M_R}} E_{\text{ave}} \triangleq \sum_{m=1}^{M_D+M_R} \sum_{i=1}^{N_L} \left(\frac{1}{\rho} \bar{P}_i^m + P_c K_i^m \right), \tag{13}$$

$$\text{s.t.} \quad \sum_{i=1}^{l} K_i^m \int_0^\infty B \log_2 \left[1 + \frac{\alpha_i^m}{\phi \sigma_0^2} f_D \left(\bar{P}_i^m, K_i^m, g \right) g \right] e^{-g} \mathrm{d}g \geq \frac{1}{\Delta T} \sum_{i=2}^{l+1} R_i^m,$$

$$m = 1, ..., M_D, l = 1, ..., N_L, \tag{13a}$$

$$- \frac{K_i^m}{\theta^m \tau} \ln \left\{ \int_0^\infty \left[1 + \frac{\alpha_i^m}{\phi \sigma_0^2} f_R \left(\bar{P}_i^m, K_i^m, g \right) g \right]^{-\beta^m} e^{-g} \mathrm{d}g \right\} \geq E_B^m \left(\theta^m \right),$$

$$m = M_D + 1, ..., M_D + M_R, i = 1, ..., N_L, \tag{13b}$$

$$\sum_{m=1}^{M_D+M_R} \bar{P}_i^m \leq P_{\text{ave}}^{\max}, i = 1, ..., N_L, \tag{13c}$$

$$\sum_{m=1}^{M_D+M_R} K_i^m \leq K_{\max}, i = 1, ..., N_L, \tag{13d}$$

$$\bar{P}_i^m \geq 0, K_i^m \geq 0, K_i^m \in \mathbb{Z}, m = 1, ..., M_D + M_R, i = 1, ..., N_L, \tag{13e}$$

where the objective function in (13) is obtained by substituting $\bar{P}_i^m = \mathbb{E}_h \left(\sum_{k=1}^{K_i^m} p_{ijk}^m \right)$ 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 constraints in (13c) and (13d) ensure that the average transmit power and the number of subcarriers allocated to all the users do not exceed the maximal average transmit power $P_{\text{ave}}^{\max}$ and the total number of subcarriers $K_{\max}$. With constraint (13d), we can always allocate each subcarrier only to one user at each time slot. Due to the bandwidth and power constraints, the problem could be infeasible when the system is heavy loaded and the channels are not good. In this case, the system can drop some enhancement layers of the SVC for VoD service, and then the value of R_i^m is reduced. To minimize the quality deterioration, R_i^m should be optimized when the system is fully loaded as in the literature, e.g., [45, 46]. In this work, we study how to improve EE when the system is not fully loaded. In this case, the problem is feasible.

Finding the optimal solution of problem (13) is non-trivial. On the one hand, the constraints in (13a) and (13b) depend on the forms of the functions $f_D(\cdot)$ and $f_R(\cdot)$. As a result, the optimal values of E_{ave} in (13) is a function of power allocation policies. We denote the minimal energy consumption with given power allocation policies 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 $f^*_R(\cdot)$. Finding the optimal form of functions results in an optimization problem in the calculus of variations [47], where the optimization variables are functions that can be regarded as vectors with infinite dimensions. Since convex optimization tools can only be used to solve finite dimensional optimization problems, they are not applicable here. On the other hand, there are no closed-form expressions of constraints (13a) and (13b). Although the achievable rate is concave in transmit power and bandwidth with equal power allocation [4], whether or not the convexity still holds with optimal power allocation policies is unknown.

To address the challenge of deriving the optimal two-timescale policy, we first find the functions of $f^*_D(\cdot)$ and $f^*_R(\cdot)$ that minimizes $E^*_{\text{ave}}(f_D, f_R)$, by proving that two spectral-efficient power allocation policies are fortunately also able to maximize EE. Then, we find the optimal resource allocation planning from problem (13) upon substituting to $f^*_D(\cdot)$ and $f^*_R(\cdot)$.

Remark 2. The terms inside the sum of the left-hand side of (13a) are the average rates in different frames (i.e., $\bar{s}^m_i$ in (11)). In many existing works [13, 18, 19, 33], this average rate is assumed known by prediction. 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 if the system only supports VoD services. This suggests that making the resource allocation plan based on average rate prediction is non-optimal.

B. Optimal Power Allocation Policies

A policy that maximizes the average service rate $\bar{s}^m_i$ (or effective capacity $E^m_{C_i}(\theta^m)$) with given average transmit power $\bar{P}^m_i$ and number of subcarriers K^m_i (i.e., bandwidth) can also minimize $\bar{P}^m_i$ with given $\bar{s}^m_i$ (or $E^m_{C_i}(\theta^m)$) and K^m_i [38, 48]. Yet this does not mean that the policy is optimal to minimize the average energy consumption in (13), i.e., maximize the EE.

1) Power allocation policy for VoD service: As shown in [48], the policy that maximizes $\bar{s}^m_i$ with given $\bar{P}^m_i$ and K^m_i is water-filling, which is

$$f^w_D\left(\frac{\bar{P}^m_i}{K^m_i}, g\right) = \begin{cases} \frac{\phi\sigma_0^2}{\alpha^m_i}\left(\frac{1}{\nu^m_i} - \frac{1}{g}\right), g \geq \nu^m_i, \\ \quad 0, \qquad\quad g < \nu^m_i, \end{cases} \tag{14}$$

where the water level ν_i^m can be obtained from $\int_{\nu_i^m}^{\infty} \frac{\sigma_0^2}{\alpha_i^m} \left(\frac{1}{\nu_i^m} - \frac{1}{g} \right) e^{-g} \mathrm{d}g = \frac{\bar{P}_i^m}{K_i^m}$.

2) Power allocation policy for RT service: As shown in [38], the policy that maximizes $E_{C_i}^m(\theta^m)$ with given $\bar{P}_i^m$ and K_i^m also follows a water-filling structure, which is

$$f_R^w\left(\frac{\bar{P}_i^m}{K_i^m}, g\right) = \begin{cases} \frac{\phi\sigma_0^2}{\alpha_i^m}\left[\frac{1}{\left(\nu_i^m\right)^{\frac{1}{\beta^m+1}} g^{\frac{\beta^m}{\beta^m+1}}} - \frac{1}{g}\right], g \geq \nu_i^m, \\ 0, \qquad g < \nu_i^m, \end{cases} \tag{15}$$

where $m = M_D + 1, ..., M_D + M_R, i = 1, ..., N_L$, $\beta^m = \frac{\theta^m \tau B}{\ln 2}$, and the water level ν_i^m over Rayleigh fading channel can be obtained from

$$\int_{\nu_i^m}^{\infty} \frac{\phi\sigma_0^2}{\alpha_i^m}\left[\frac{1}{\left(\nu_i^m\right)^{\frac{1}{\beta^m+1}} g^{\frac{\beta^m}{\beta^m+1}}} - \frac{1}{g}\right] e^{-g}\mathrm{d}g = \frac{\bar{P}_i^m}{K_i^m}. \tag{16}$$

The water-level is time-varying and the instantaneous power allocated to each subcarrier depends on the instantaneous channel gains on all the subcarriers assigned to the user.

3) Optimality of the power allocation policies: The following proposition (see proof in Appendix A.) indicates that (14) is the optimal power allocation policy for VoD service and (15) is the optimal power allocation policy for RT service in terms of maximizing the EE. In other words, $f_D^*(\bar{P}_i^m, K_i^m, g) = f_D^w\left(\frac{\bar{P}_i^m}{K_i^m}, g\right)$ and $f_R^*(\bar{P}_i^m, K_i^m, g) = f_R^w\left(\frac{\bar{P}_i^m}{K_i^m}, g\right)$.

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

$$E_{\text{ave}}^*\left(f_D^w, f_R^w\right) \leq E_{\text{ave}}^*\left(f_D', f_R'\right). \tag{17}$$

C. Optimal Resource Allocation Planning

Substituting the optimal power allocation policies in (14) and (15) into (13a) and (13b), the optimal resource allocation plan can be obtained from the following problem,

$$\min_{\substack{\bar{P}_i^m, K_i^m, \\ m=1,...,M_D+M_R, i=1,...,N_L}} \sum_{m=1}^{M_D+M_R} \sum_{i=1}^{N_L} \left(\frac{1}{\rho}\bar{P}_i^m + P_c K_i^m\right), \tag{18}$$

$$\text{s.t.} \quad \sum_{i=1}^{l} K_i^m F_D\left(\frac{\bar{P}_i^m}{K_i^m}\right) \geq \frac{1}{\Delta T}\sum_{i=2}^{l+1} R_i^m, m = 1, ..., M_D, l = 1, ..., N_L, \tag{18a}$$

$$-\frac{K_i^m}{\theta^m \tau}\ln\left[F_R\left(\frac{\bar{P}_i^m}{K_i^m}\right)\right] \geq E_B^m\left(\theta^m\right), m = M_D + 1, ..., M_D + M_R, i = 1, ..., N_L, \tag{18b}$$

$$(13c), (13d) \text{ and } (13e),$$

where

$$F_D\left(\frac{\bar{P}_i^m}{K_i^m}\right) = \int_0^\infty B\log_2\left[1 + \frac{\alpha_i^m}{\phi\sigma_0^2} f_D^w\left(\frac{\bar{P}_i^m}{K_i^m}, g\right) g\right] e^{-g}\mathrm{d}g, \tag{19}$$

$$F_R\left(\frac{\bar{P}_i^m}{K_i^m}\right) = \int_0^\infty \left[1 + \frac{\alpha_i^m}{\phi\sigma_0^2} f_R^w\left(\frac{\bar{P}_i^m}{K_i^m}, g\right) g\right]^{-\beta^m} e^{-g}\mathrm{d}g. \tag{20}$$

By relaxing the numbers of subcarriers as continuous variables, we can obtain the following property (See proof in Appendix B).

Property 1. The left-hand sides of (18a) and (18b) are jointly concave in $\bar{P}_i^m$ and K_i^m.

The above property indicates that the feasible region of problem (18) is a convex set. Since the objective function in (18) is linear, problem (18) is convex programming, whose global optimal solution can be solved numerically by the interior-point method if it is feasible [49].

The *ideal policy*, i.e., the optimal solution of problem (13), consists of making a plan and allocating power that operate in two timescales. The resource allocation plan for a user is made at the start of the prediction window with predicted average channel gains, which is optimized from problem (18). The transmit power is allocated at the start of each time slot with estimated CSI, which is optimized from (14) and (15).

D. Impacts of Predicted Information on EE

Predicting the instantaneous channel gains (i.e., CSI) and average channel gains of every user are possible, but inevitably incurs cost for training [23, 34]. In the sequel, we discuss which kinds of future information are necessary to maximize the EE.

1) Predicted Information of VoD Users: Here we consider a system without RT users, 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,

$$\frac{1}{\rho}\left(\sum_{m=1}^{M_D}\sum_{i=1}^{N_L}\sum_{j=1}^{N_S}\sum_{k=1}^{K_i^m} \tau p_{ijk}^m\right) + \Delta T P_c \sum_{m=1}^{M_D}\sum_{i=1}^{N_L} K_i^m, \tag{21}$$

where the transmit powers on different subcarriers in the ith frame $\{p_{ijk}^m, k = 1, ..., K_i^m\}$ depend on the instantaneous channel gains $\{g_{ijk}^m, k = 1, ..., K_i^m\}$. Denote the total transmit power for the

mth user in the jth time slot of the ith frame as $P_{ij}^m = \sum_{k=1}^{K_i^m} p_{ijk}^m$. Since the fast fading channels among the 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., $\frac{1}{N_S}\sum_{j=1}^{N_S} P_{ij}^m \to \bar{P}_i^m$ when $N_S \to \infty$. Further considering that $\Delta T = N_S\tau$, minimizing (21) is equivalent to minimizing the following expression,

$$\frac{1}{\rho}\left(\sum_{m=1}^{M_D}\sum_{i=1}^{N_L}\bar{P}_i^m\right) + P_c\sum_{m=1}^{M_D}\sum_{i=1}^{N_L}K_i^m, \tag{22}$$

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

The ideal policy that minimizes the energy consumed for VoD users with future CSI can be obtained by minimizing (22) under constraints (13a), (13c), (13d) and (13e). Since the optimization problem is the same as problem (13), the optimal power allocation policies, the optimal average transmit power and numbers of subcarriers, and the minimal total energy consumptions obtained from the two problems are equal. This leads to the following observation.

Observation 1: Predicting the CSI of each VoD user in future time slots does not help improve the system EE, but predicting the average channel gains of the VoD user can improve EE.

2) Predicted Information of RT Users: Similarly, here we consider a system without VoD users, i.e., $M_D = 0$. For RT service, $\tau \ll D_{\max}^m \ll \Delta T$, where τ and ΔT are the time slot and frame durations. At the beginning of each frame, when the average channel gain is available by estimation, the BS can assign the average transmit power and number of subcarriers to each RT user. The amount of resources assigned to the mth RT user in the ith frame can be obtained from the following problem,

$$\min_{\bar{P}_i^m, K_i^m,} \sum_{m=1}^{M_R}\left(\frac{1}{\rho}\bar{P}_i^m + P_c K_i^m\right) \tag{23}$$

$$\text{s.t. } (13\text{b}), (13\text{c}), (13\text{d}) \text{ and } (13\text{e}).$$

From the expressions of the constraints, we can see that the amount of resources assigned in the ith frame does not depend on the amount of resources assigned in other frames. Therefore, problem (13) can be decomposed into N_L independent problems as problem (23). Knowing the average channel gains (and hence CSI) in future frames cannot help improve the QoS (i.e., $D_{\max}^m$ and ε_D^m) or the EE *of a system only with RT services*. This implies that making the resource allocation plan for RT users is unnecessary, and gives rise to another observation as follows.

Observation 2: Predicting the average channel gains and CSI of each RT user in the prediction window does not help improve the EE of a system only with RT service.

Remark 3. VoD service is delay-tolerant. The QoS of a VoD user can be satisfied if the requested segment can be downloaded to the user before playback. Hence, the BS can choose some frames in the prediction window with high average channel gains to transmit data in advance to save energy. By contrast, RT service is delay-sensitive. The QoS of a RT user in terms of delay (i.e., $D^m_{\max}$) is less than the frame duration ΔT. Hence, to improve the EE under the QoS constraint, the BS can only adjust resources among the time slots within $D^m_{\max}$. This explains why the future average channel gains of each RT user cannot help improve the EE of a system only with RT users. Nevertheless, predicting the average channel gains of RT users helps improve the EE of a network with both VoD and RT services.

E. Extension to Multicell Scenario

Now we consider a scenario where the $M_D + M_R$ users are served by N_B BSs. The BSs are connected with a central processor (CP) and send the future average channel gains of all the users in a prediction window to the CP. To focus on the EE-maximal optimization for both services and in two timescales, we assume that the inter-cell interference can be treated as noise. A simple way to avoid strong interference is using orthogonal resources in adjacent cells, say by soft frequency reuse. This assumption is reasonable for the problem at hand, because we consider a non-fully-loaded network. The problem to optimize PRA with strong interference is nontrivial, as demonstrated in [21], where a PRA is designed for file downloading in heterogeneous networks.

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 (15) are optimal for VoD service and RT service, respectively. Denote $\bar{P}^{mn}_i$ and K^{mn}_i as the average transmit power and the number of subcarriers assigned to the mth user in the ith frame by its accessed BS (i.e., the nth BS). Denote $\mathcal{M}^n_i$ as the set of indices of the users that are served by the nth BS in the ith frame. The difference between single-cell and multi-cell scenarios lies in the constraints on the average transmit power and the total number of subcarriers. Specifically, the amount of resources assigned to the users that access to the same BS in the multi-cell scenario should satisfy the following constraints

$$\sum_{m\in\mathcal{M}_i^n} \bar{P}_i^{mn} \le P_{\rm ave}^{\max}, \sum_{m\in\mathcal{M}_i^n} K_i^{mn} \le K_{\max}, n=1,...,N_B, i=1,...,N_L. \tag{24}$$

The user association $\{\mathcal{M}_i^n, n = 1, ..., N_B, i = 1, ..., N_L\}$ and resource allocation plan can be jointly optimized, but the resulting problem is a mixed integer optimization problem, which is much more challenging than the problem for the single-cell scenario. To save energy, it is reasonable to assume that each user is accessed to the BS with the highest large-scale channel gain. Then, $\{\mathcal{M}_i^n, n = 1, ..., N_B, i = 1, ..., N_L\}$ are known by the CP at the beginning of each prediction window with the predicted trajectories of mobile users. Similar to problem (18), the optimal resource allocation plan is also convex programming.

IV. A Heuristic Joint Resource Allocation Policy with Low Costs

In this section, we propose a heuristic policy that can perform close to the optimal solution with coarsely predicted knowledge. The queue states in the buffers of VoD users are also taken into account. To be more specific, each VoD user sends two bits information to the BS to indicate whether the buffer will overflow and whether playback interruption will occur.[3]

This policy is inspired by the structure of problem (18), which suggests that the amount of resources assigned for a RT user in each frame is independent of the resources assigned in other frames if only RT users exist. This implies that predictive resource allocation for RT users cannot improve EE. In other words, we only need to design the resource allocation plan for VoD users. If we can also decompose the resource allocation planning problem for VoD users into N_L independent problems, then many existing low-complexity algorithms can be applied to assign resources in each frame for both types of users.

To decouple the resource allocation planning problem, we come back to the basic idea of PRA that only serves VoD users: transmit more data to a user when the user undergoes good average channels [13]. Such an idea can be translated as: which frame is with good channel to boost the EE and how much data should be transmitted to satisfy the QoS.

To increase EE, we find a "ruler" to judge whether the average channel gain in a frame is high or low. According to the results in [1] obtained from optimizing for a single VoD user,

[3] In practice, the user can send a request for stopping transmission if the buffer will overflow. If the last video segment in the buffer is played in the current frame, the user can send a transmission request to avoid interruption in the next frame.

the ideal policy transmits data only when the average channel gains exceed a certain threshold, which occurs over around half of the frames in the prediction window. This inspires us to use the median of the average channel gains in the window, denoted as α^m_{med}, as the threshold. Then, at the beginning of the window, the CP only needs to predict the median for each VoD user.

To avoid stalling and buffer overflow for the VoD users with limited buffer sizes, we should consider the queueing status of each VoD user and control the number of segments transmitted in each frame. 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 slots during streaming. Then, it is reasonable to transmit several segments (we consider two segments for illustration in the sequel, but more segments can be transmitted) to a user with good channel in a frame, if there is still room in the buffer. With given average power and bandwidth in each frame, the transmission procedure for each VoD user (say the mth user) of the heuristic policy is as follows.

At the beginning of the ith frame, the average channel gain of the mth user, α^m_i, can be estimated at its associated BS. Denote $\tilde{i}$ as the index of last video segment that has been transmitted before the ith frame. Then, the indices of segments to be transmitted are $\{\tilde{i}+1, ...\}$.

If $\alpha^m_i < \alpha^m_{\text{med}}$, then the mth user is in bad channel condition. No data will be transmitted in the ith frame if the video segment to be played in the $i+1$th frame has been transmitted before the ith frame. Otherwise, one segment will be transmitted in the ith frame. Thus, the required average service rate can be expressed as follows,

$$\hat{s}^m_i = \begin{cases} 0, & \text{if } \tilde{i} \geq i+1, \\ \frac{1}{\Delta T} R^m_{i+1}, & \text{otherwise.} \end{cases} \tag{25}$$

If $\alpha^m_i \geq \alpha^m_{\text{med}}$, then the mth user is in good channel condition. Two segments will be transmitted in the ith frame if the buffer has enough space for two segments. One segment will be transmitted if the buffer only has space for one more segment. If there is no enough space, no data will be transmitted. The required average service rate is given by

$$\hat{s}^m_i = \begin{cases} \frac{1}{\Delta T}(R^m_{\tilde{i}+1} + R^m_{\tilde{i}+2}), & \text{if } Q^m_i + R^m_{\tilde{i}+1} + R^m_{\tilde{i}+2} - R^m_i \leq Q_{\max}, \\ 0, & \text{if } Q^m_i + R^m_{\tilde{i}+1} - R^m_i > Q_{\max}, \\ \frac{1}{\Delta T} R^m_{\tilde{i}+1}, & \text{otherwise,} \end{cases} \tag{26}$$

where Q_i^m and $Q_{\max}$ are the queue length of the mth user at the beginning of the ith frame and the maximal buffer size, respectively.

The average power and bandwidth assigned to each VoD and RT user in the ith frame can be jointly optimized from the following problem,

$$\min_{\substack{\bar{P}_i^m, K_i^m, \\ m=1,...,M_D+M_R}} \sum_{m=1}^{M_D+M_R} \left(\frac{1}{\rho} \bar{P}_i^m + P_c K_i^m \right), \tag{27}$$

$$\text{s.t.} \quad K_i^m F_D \left(\frac{\bar{P}_i^m}{K_i^m} \right) \geq \hat{s}_i^m, m = 1, ..., M_D, \tag{27a}$$

$$-\frac{K_i^m}{\theta^m \tau} \ln \left[F_R \left(\frac{\bar{P}_i^m}{K_i^m} \right) \right] \geq E_B^m(\theta^m), m = M_D + 1, ..., M_D + M_R, \tag{27b}$$

$$\text{(13c), (13d) and (13e),}$$

where the optimal power allocation policies in (14) and (15) are adopted.

Compared with problem (18), the only difference lies in the QoS constraints of VoD users on the average service rate in (18a) and (27a). In problem (18), $\bar{s}_i^m$ is optimized implicitly through optimizing K_i^m and $\bar{P}_i^m$ according to all average channel gains in the prediction window. In problem (27) the value of $\hat{s}_i^m$ is only determined by the average channel gain in the ith frame and the threshold. According to the way we obtain $\hat{s}_i^m$ (i.e., (25) and (26)) in the heuristic policy, the average rate $\bar{s}_i^m$ will satisfy constraint (18a) if constraint (27a) is satisfied. Thus, a feasible solution of problem (18) can be obtained with the heuristic policy.

Because the average service rate constraint in (27a) is a special case of the effective capacity constraint in (27b) with $\theta^m \to 0$ [41], many existing algorithms can be applied to find the solution of problem (27) [4]. This indicates that except the cost in predicting the "ruler" for each VoD user (i.e., α_{med}^m) at the CP, the heuristic policy needs the same complexity as existing non-predictive counterparts. After obtaining average power and bandwidth assigned to each VoD and RT user in each frame, the optimal power allocation policies in (14) and (15) can be used at each time slot. The heuristic policy is summarized in Table I.

Remark 4. With the heuristic policy, future information is required only when we determine $\hat{s}_i^m$ by the "ruler", which is the median of the future average channel gains in the prediction window of the mth VoD user, α_{med}^m. To predict α_{med}^m, we can design a fully connected neural network

(NN), which input is the historical average channel gains of VoD users denoted as $\boldsymbol{x}$ and the output is the channel median denoted as y. The NN is trained with the training set $\{\boldsymbol{x}^{(n)}, y^{(n)}\}_{n=1}^{N}$ to minimize the cost function $J(\boldsymbol{W}, \boldsymbol{b}) = \frac{1}{N}\sum_{n=1}^{N} |\hat{y}^{(n)} - y^{(n)}|^2 + \frac{\nu}{2}\|\boldsymbol{W}\|_{\mathrm{F}}^2$, where $\boldsymbol{W}$ and $\boldsymbol{b}$ are the weight matrix between layers of the NN and the bias of neurons, $y^{(n)}$ is the label, $\hat{y}^{(n)}$ is the output with input $\boldsymbol{x}^{(n)}$, and the regularization term is added to reduce overfitting. By contrast, to predict the fine-grained average channel gain in each frame of the prediction window (say α_i^m), recurrent neural networks need to be used for predicting fine-grained trajectory [23]. For training and testing the prediction, we consider mobile users with trajectories in [23], the only difference is that the length of each road segment is 1 km and the random stoping time at the red lights is $1 \sim 30$ s here to reduce training time. For a one minute-long prediction window, after tuning the hyper-parameters of our NN and the long short term memory model proposed in [23], the results are as follows. When the users move along a road with minimal distance from BSs as 200 m, the average relative error (i.e., prediction errors normalized by the true values in the test set divided by the number of test samples) of $\hat{\alpha}_{\text{med}}^m$ is 42% and that of $\hat{\alpha}_i^m$ is 170%. The EE loss caused by prediction errors for the heuristic and optimal policies are 5.4% and 65%, respectively. 2000 and 56000 training samples are required for predicting α_{med}^m and α_i^m, and the training time for predicting α_{med}^m is about 8‰ of that for predicting α_i^m.

TABLE I
THE HEURISTIC POLICY

```
Input: $R_i^m$, $i = 1, ..., N_L$, $m = 1, ..., M_D$, $Q_{\max}$, $\alpha_{\text{med}}^m$, $m = 1, ..., M_D$.
Output: $P_i^m$ and $K_i^m$, $i = 1, ..., N_L$, $m = 1, ..., M_D$.
$i := 1$, $\tilde{i} := 1$, and $Q_1^m := R_1^m$, $m = 1, ..., M_D$.
while $i \le N_L$ do
    if $\tilde{i} < N_L + 1$. then
        while $m \le M_D$ do
            if $\alpha_i^m \ge \alpha_{\text{med}}^m$ then
                $\hat{s}_i^m := \frac{1}{\Delta T}(R_{\tilde{i}+1}^m + R_{\tilde{i}+2}^m)$, $\tilde{i} := \tilde{i} + 2$, if $Q_i^m + R_{\tilde{i}+1}^m + R_{\tilde{i}+2}^m - R_i^m \le Q_{\max}$.
                $\hat{s}_i^m := \frac{1}{\Delta T} R_{\tilde{i}+1}^m$, $\tilde{i} := \tilde{i} + 1$, if $Q_i^m + R_{\tilde{i}+1}^m + R_{\tilde{i}+2}^m - R_i^m > Q_{\max}$ and $Q_i^m + R_{\tilde{i}+1}^m - R_i^m \le Q_{\max}$.
                $\hat{s}_i^m := 0$, if $Q_i^m + R_{\tilde{i}+1}^m - R_i^m > Q_{\max}$.
            else
                $\hat{s}_i^m := \frac{1}{\Delta T} R_{\tilde{i}+1}^m$, $\tilde{i} := \tilde{i} + 1$, if $i = \tilde{i}$.
                $\hat{s}_i^m := 0$, if $\tilde{i} > i$.
            end if
        end while
        $\hat{s}_i^m := 0$.
    else
        Solve problem (27), and obtain $P_i^m$ and $K_i^m$, $m = 1, ..., M_D$.
    end if
end while
return $P_i^m$ and $K_i^m$, $i = 1, ..., N_L$, $m = 1, ..., M_D$.
```

V. Simulation Results

In this section, we evaluate the EE of the proposed ideal policy and heuristic policy. We consider both scenarios with perfect and imperfect predictions of average channel gains.

We compare the proposed policies (with legends "Ideal" and "Heuristic") with four baselines.

- Non-predictive joint resource allocation (legend "Baseline 1"): This is extended from a policy only considering RT users in [4] by jointly optimizing average transmit power and bandwidth in each frame for VoD and RT users. For VoD users, the segments to be played in the ith frame are transmitted in the $i-1$th frame (i.e., $\bar{s}_{i-1}^{mn} = \frac{1}{\Delta T} R_i^m$).
- PRA only with future average channel gains for VoD users (legend "Baseline 2"): This is extended from a policy only considering VoD users in [13]. The unknown distances between BS and RT users are set as cell radius in all frames, and then the resource allocation for VoD and RT users are jointly optimized.
- Decoupled PRA in two timescales (legend "Baseline 3"): This is extended from a two-timescale policy only considering VoD users in [33]. The extended policy optimizes bandwidth allocation at the beginning of prediction window (equivalent to allocating the number of time slots in [33]), where transmit power is equally allocated to all subcarriers (i.e., $P_{\max}/K_{\max}$) in order to predict average rate. In each time slot, the transmit power is allocated to subcarriers with (14) and (15) (similar to subcarrier allocation in [33]).
- Decoupled PRA for two services (legend "Baseline 4"): This is extended from a two-timescale policy that optimizes resource allocation for file-downloading users with the residual bandwidth and power after serving RT users [17]. To obtain the residual resources, we first assign bandwidth to RT users with fixed transmit power on each subcarrier as $P_{\max}/K_{\max}$. In this way, the residual bandwidth is proportional to the residual power, as assumed in [17]. By solving problem (18) with given residual resources, the resource allocation for VoD services in two timescales is jointly optimized.

By comparing with Baseline 1, we can illustrate the gain of improving EE by harnessing future average channel gains. By comparing with Baseline 2, we can show when the future average channel gains of RT users is helpful. By comparing with Baseline 3, we can show the EE loss from decoupling the optimization in the two timescales. By comparing with Baseline 4, we can show the EE loss due to reserving resources for RT service.

Since the predicted average channel gains, denoted as $\{\hat{\alpha}_i^m,\ i = 1, ..., N_L\}$, are not error-free, the ideal policy needs to adjust resources to ensure QoS. With $\{\hat{\alpha}_i^m,\ i = 1, ..., N_L\}$, resource allocation plan $\{\hat{P}_i^m, \hat{K}_i^m, i = 1, ..., N_L\}$ can be obtained by solving problem (18). At the beginning of the ith frame, if the average channel gain estimated at the BS $\alpha_i^m \neq \hat{\alpha}_i^m$, then we apply a simple adjustment that does not need to solve another optimization problem: the BS adjusts average transmit power according to $\alpha_i^m \tilde{P}_i^m = \hat{\alpha}_i^m \hat{P}_i^m$, and $\hat{K}_i^m$ does not change. Such a modified ideal policy to address prediction errors is referred to as *extended ideal policy*.

A. Simulation Setup

For VoD services, we consider SVC in [36] (each segment includes one base layer and five enhance layers). The bit rate of each layer can be found in [50]. The average streaming rate of each VoD service is around 2 Mbits/s. For RT services, 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 an exponential distribution with average $1/\lambda_u = 4$ kbits/packet. Hence, the average data arrival rate of each RT service is 2 Mbits/s.

All the users move along a road with minimal distance from BSs as 100 m, where the distance between two adjacent BSs is 500 m. In the scenario with perfect prediction, the velocities of users are constant, i.e., 20 m/s. In the scenario with imperfect prediction, the velocities are random variables, as to be detailed later. To save transmit power, each user is accessed to its nearest BS. The path loss model is $35.3 + 37.6 \log_{10} D$ dB, where D is the distance in meters between a user and its accessed BS in a frame. The circuit powers of different components are obtained from those measured in the year of 2012 in [42], where the scaling law in [51] is further applied to predict P_c and P_0 in 2020. The prediction window is with duration $N_L \Delta T = 60$ s. The results for the scenarios with perfect and imperfect prediction are respectively obtained from 100 and 1000 simulation trails. In each trail, the user trajectory in the prediction window, the packet arrival and packet size of RT services, and Rayleigh fading channels are randomly generated. Matlab is used for the simulation. The simulation parameters are listed in Table II. This setup will be used in the sequel unless otherwise specified.

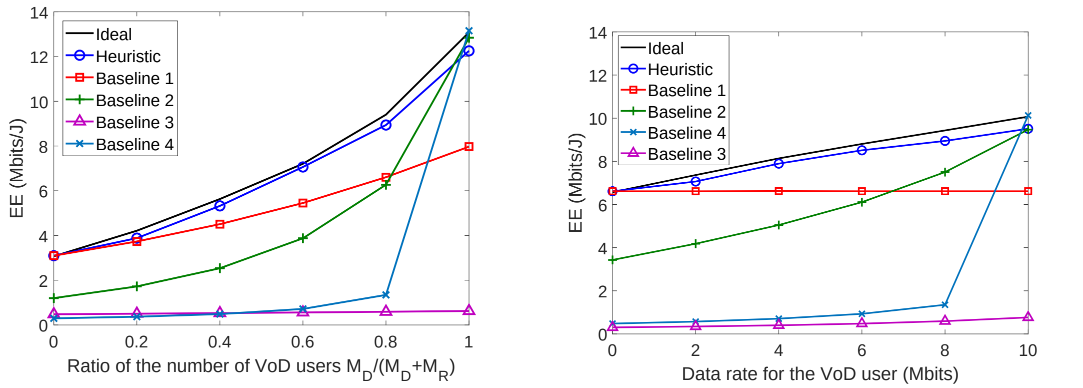
B. Perfect Prediction of Average Channel Gains

The EE achieved by different policies is shown in Fig. 2. In Fig. 2(a), the total number of users is fixed as $M_D + M_R = 5$. In Fig. 2(b), the sum data rate required by all users are fixed as

TABLE II
LIST OF SIMULATION PARAMETERS [42, 51]

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

$\mathbb{E}(R_i^m/\Delta T) + \lambda_a/\lambda_u = 10$ Mbits, where the arrival data rate of RT user (or the streaming rate of VoD user) varies. We can see that when there is no VoD user, the achieved EE of the ideal policy and Baseline 1 are identical. The results are consistent with Observation 2, i.e., predicting average channel gains cannot help improve the EE of the system only with RT users. When there are both VoD and RT users, the achieved EE of the ideal 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 resources reserved for the RT users are too conservative. The achieved EEs of Baseline 3 and Baseline 4 are much lower than Baseline 1, which is a non-predictive policy that jointly optimizes resource allocation for the two types of services. This result suggests the necessity of the joint optimization for two types of services in the two timescales. Since Baselines 3 and 4 perform the worst almost in all cases, we no longer provide their performance in the sequel unless necessary. The heuristic policy performs closely to the ideal policy.



(a) EE v.s. ratios of the number of VoD users.

(b) EE v.s. streaming rate of VoD user, $M_D = M_R = 1$.

Fig. 2. EE achieved by different policies.

To show the throughput of the considered network in terms of maximal total number of RT and VoD users without stalling, we provide the video quality with different numbers of users in Table

TABLE III
NUMBERS OF LAYERS OF SCV WITH DIFFERENT NUMBER OF USERS

$M_D = M_R$	≤ 8	9	10	11	≥ 12
Ideal	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

III, which provides the average number of enhanced layers transmitted to VoD users. To obtain the results, we first set the video quality at level 5 (all the 5 enhanced layers are transmitted) and find the solutions with different policies. If the problem is infeasible, we reduce the video quality of VoD users to a lower level until the problem is feasible. Finally, we calculate the average video quality of VoD users in 100 minutes. "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). Due to the lack of space, we do not show the performance when stalling occurs, which is acceptable in practice. We can see that the maximal total number of users that the system can support with ensured QoS is 20 (i.e., $M_D = M_R = 10$) if using Baseline 2, and is 22 if using other three policies. Again, the heuristic policy performs closely to the ideal policy in term of the video quality. Since the EE can be improved only when the traffic load is not high (say $M_D, M_R < 10$ for the considered setup), in the sequel we only consider the scenario where the QoS of all VoD users can be ensured.

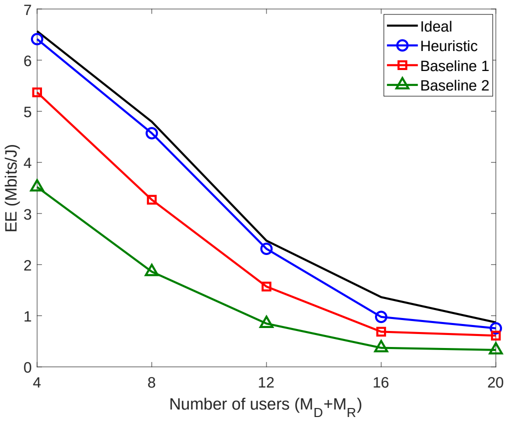


Fig. 3. EE v.s. total number of users, where $M_D = M_R$.

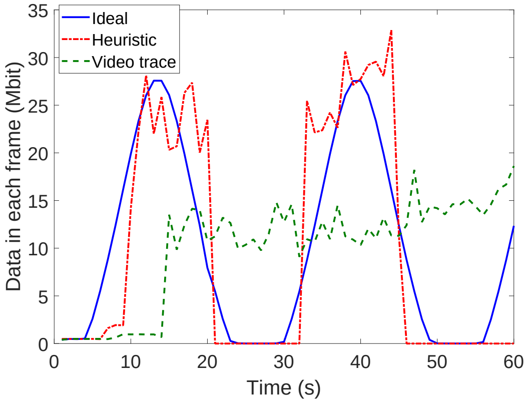


Fig. 4. Data transmitted in different frames, where the average streaming rate of the VoD service and the average data arrival rate of the RT service are set as 10 Mbps.

The EE of the system supporting different traffic loads is shown in Fig. 3. The results show that EE achieved by the ideal and heuristic policies are much higher than that achieved by the baselines when the number of users is small. 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 network is fully loaded, the BSs need to serve the users with maximal transmit power and bandwidth, and hence there is no chance to save energy.

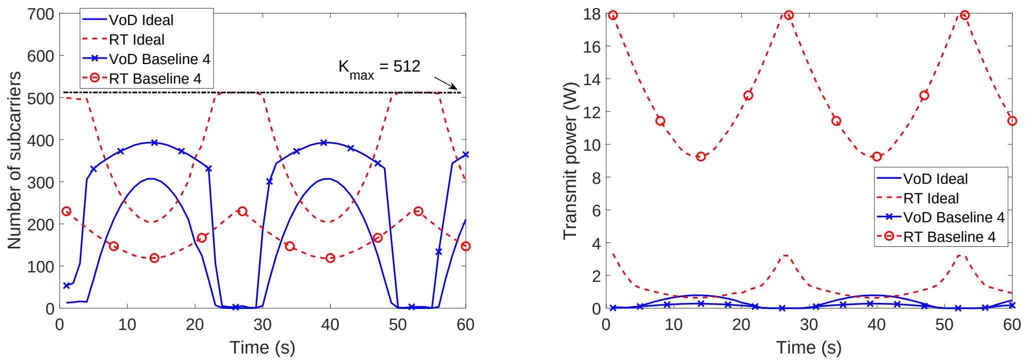
To understand the behavior of the heuristic policy, we consider a simplified scenario with one VoD user and one RT user. The amount of data transmitted in different frames and the amount of video data played in each frame (with legend "video trace") are shown in Fig. 4. The results show that at the first several frames in the prediction window (i.e., the beginning of the service), the data amount transmitted in each frame equals to the data amount to be played in the next frame. After the first several seconds, both ideal and heuristic policies transmit data when the large-scale channel gains are good. There is no stalling since the video segments are transmitted before playback.

To understand why the resource allocation between VoD and RT services should be jointly optimized, we again consider the simple scenario with one VoD user and one RT user. The average transmit power and bandwidth assigned by the ideal policy and Baseline 4 in each frame to each service are shown in Fig. 5. The results show that if the resource allocation is jointly optimized, more subcarriers will be allocated to the RT users if the average data rate of a VoD user (say the mth user), $\bar{s}_i^m$, is zero in a frame (e.g., the user is located in cell-edge). However, with Baseline 4, the BS first assigns subcarriers (i.e., reserves resources) to RT service without considering the bandwidth required by VoD service. To leave some bandwidth for VoD service, the BS will not assign all the subcarriers to the RT service. With less bandwidth, more transmit power is consumed by the RT service, and hence the EE of the system is low.

C. Imperfect Prediction of Large-scale Channel Gains

The prediction errors may come from many sources such as erroneous mobility route prediction, inaccurate velocity prediction, user location estimation and constructed radio map. Here we take the velocity prediction error as an example to illustrate the impact of imperfect prediction, since it leads to the errors on average channel gains accumulative with frames and hence causes more severe performance degradation than other types of prediction errors.

Markov chain is widely used to model the mobility of vehicles [12]. We use a discrete time Markov chain to characterize the velocity of each user. Specifically, the velocity of each user



(a) Bandwidth allocation.

(b) Average power allocation.

Fig. 5. Coupled resource allocation between VoD and RT services, where the average streaming rate of the VoD service and the average data arrival rate of the RT service are set as 10 Mbps. We use solid and dash lines and further add cross and dot marks in these lines to distinguish different results.

lies in $\mathcal{V} = \{v_1, v_2, ..., v_U\}$, where $\Delta v \triangleq v_{u+1} - v_u = 1$ m/s, $v_1 = 0$ m/s, and $v_U = 30$ m/s, and the velocities are constant within each frame of duration ΔT. With this model, the velocity of a user may vary from $0 \sim 30$ m/s in the prediction window. Denote the velocity of the mth user in the ith frame as V_i^m. We set $\Delta v / \Delta T$ equal to the maximal acceleration of vehicles (e.g., 1 m/s^2 [52]). The velocity can only transit between adjacent states (i.e., it can change Δv after ΔT). The $U \times U$ transition matrix of the Markov chain is denoted as $\mathbf{U}$, where $u_{i,j}$ is the probability that the velocity transits from v_i to v_j. For the considered scenario, $u_{11} = u_{U,U} = 1-q$, $u_{i,i} = 1-2q$, $i \neq 1, U$, and $u_{i,i+1} = u_{i+1,i} = q, i = 1, ..., U-1$. When $q = 0$, the velocity is constant and always equals to the initial value. By increasing the value of q, the prediction errors of velocities in the upcoming N_L frames increase. 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 are obtained by assuming that each user travels along the predicted route with the initial velocity, setting as 20 m/s.[4] The initial positions of the users are uniformly distributed in the first cell. The median of each VoD user is computed from the predicted average channel gains of the user in the window.

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 EE achieved by different policies[5] is shown in Fig. 6, where the qualities of all the VoD

[4] According to simulations, the uncertainty of velocity modeled in the sequel will lead to $200 \sim 300$ % prediction errors on average channel gains at the end of a prediction window with 60 seconds duration.

[5] We do not compare with the robust policies in [18,19]. This is because existing methods can only reformulate non-deterministic linear constraints as deterministic constraints [53], but the constraint in (13a) is non-linear.

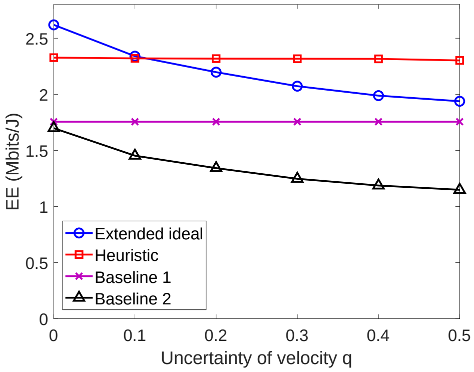


Fig. 6. EE v.s. uncertainty of velocity.

users are the same (all six layers are transmitted before playback). Since the ideal policy is optimized under the assumption of perfect prediction of average channel gains, it is no longer optimal when the prediction is non-perfect. When q is large, the heuristic policy outperforms the ideal policy. Even when $q = 0.5$, which leads to over 300% prediction errors on the average channel gains at the end of the prediction window, the EE loss of heuristic policy is negligible.

VI. Conclusion

In this paper, we demonstrated the gain in maximizing EE of an OFDMA system with both VoD and RT services by harnessing the prediction of average channel gains in a time window, and investigated which kinds of future information are needed to improve EE. To this end, an optimal two-timescale PRA policy was obtained, which needs fine-grained prediction as in existing works. By analyzing the optimal policy, we found that using instantaneous channel gains in future time slots can not improve EE. To reduce the training cost incurred by the fine-grained prediction, a heuristic policy was proposed. This policy only needs the median of future average channel gains of VoD users, and jointly optimizes the average power and bandwidth for RT and VoD users in each frame. Simulation results showed that EE can be improved remarkably by jointly optimizing resource allocation for the two types of services in the two timescales, which suggests the necessity of sharing resources among different services. The heuristic policy performs closely to the optimal policy if the prediction is error-free, and outperforms the optimal policy if the prediction is with large uncertainty. This demonstrates that the EE gain over non-predictive policies can be achieved by non-perfect coarse-grained prediction.

Appendix A

Proof of Proposition 1

Proof: To prove Proposition 1, we first prove that $f_D^w\left(\frac{\bar{P}_i^m}{K_i^m}, g\right)$ 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 $f_D^w\left(\frac{\bar{P}_i^m}{K_i^m}, g\right)$ and $f_D'\left(\bar{P}_i^m, K_i^m, g\right)$ are denoted as $\{\tilde{P}_i^m, \tilde{K}_i^m, m = 1, ..., M_D + M_R, i = 1, ..., N_L\}$ and $\{P_i^{m'}, K_i^{m'}, m = 1, ..., M_D + M_R, i = 1, ..., N_L\}$, respectively. Then, we need to prove $E_{\text{ave}}^*\left(f_D^w, f_R'\right) \leq E_{\text{ave}}^*\left(f_D', f_R'\right)$, where

$$E_{\text{ave}}^*\left(f_D^w, f_R'\right) = \sum_{m=1}^{M_D+M_R} \sum_{i=1}^{N_L} \left(\frac{1}{\rho}\tilde{P}_i^m + P_c\tilde{K}_i^m\right); E_{\text{ave}}^*\left(f_D', f_R'\right) = \sum_{m=1}^{M_D+M_R} \sum_{i=1}^{N_L} \left(\frac{1}{\rho}P_i^{m'} + P_c K_i^{m'}\right).$$

Denote the average service rates achieved by the power allocation policies $f_D'\left(\bar{P}_i^m, K_i^m, g\right)$ 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 $E_{\text{ave}}^*\left(f_D^w, f_R'\right) \leq E_{\text{ave}}^*\left(f_D', f_R'\right)$, we need the following result: the water-filling policy $f_D^w\left(\frac{\bar{P}_i^m}{K_i^m}, g\right)$ can minimize $\bar{P}_i^m$ with given K_i^m and average service rate $\bar{s}_i^m$ [48]. According to this result, given $K_i^{m'}$ and $s_i^{m'}$, $i = 1, ..., N_L$, the average transmit power is minimized with $f_D^w\left(\frac{\bar{P}_i^m}{K_i^m}, g\right)$. Denote the related minimal average transmit power for the mth user in the ith 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

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

Moreover, with $f_D^w\left(\frac{\bar{P}_i^m}{K_i^m}, g\right)$, the optimal resource allocation plan is $\{\tilde{P}_i^m, \tilde{K}_i^m$, $m = 1,...,M_D+$ M_R, $i = 1,...,N_L\}$. Thus, $\sum\limits_{m=1}^{M_D+M_R} \sum\limits_{i=1}^{N_L} \left(\frac{1}{\rho}\tilde{P}_i^m + P_c\tilde{K}_i^m\right) \leq \sum\limits_{m=1}^{M_D} \sum\limits_{i=1}^{N_L} \left[\frac{1}{\rho}\min(\bar{P}_i^m) + P_c K_i^{m'}\right] + \sum\limits_{m=M_D}^{M_D+M_R} \sum\limits_{i=1}^{N_L} (\frac{1}{\rho}P_i^{m'} + P_c K_i^{m'})$. Further considering (A.1), we have $E_{\text{ave}}^*\left(f_D^w, f_R'\right) \leq E_{\text{ave}}^*\left(f_D', f_R'\right)$.

Similarly, given power allocation policy for VoD services $f_D^w\left(\frac{\bar{P}_i^m}{K_i^m}, g\right)$, we can prove that $E_{\text{ave}}^*\left(f_D^w, f_R^w\right) \leq E_{\text{ave}}^*\left(f_D^w, f_R'\right)$. The proof is omitted for conciseness. Therefore, we can obtain that $E_{\text{ave}}^*\left(f_D^w, f_R^w\right) \leq E_{\text{ave}}^*\left(f_D^w, f_R'\right) \leq E_{\text{ave}}^*\left(f_D', f_R'\right)$. □

Appendix B
Proof of Property 1

Proof: The proof of the convexity of (18a) is shown in the conference version [1].

We only prove the convexity of (18b). The left-hand side of (18b) is the perspective of $-\frac{1}{\theta^m\tau}\ln\left[F_R\left(\bar{P}_{S_i}^m\right)\right]$, where $\bar{P}_{S_i}^m = \frac{\bar{P}_i^m}{K_i^m}$. To prove that the left-hand side of (18b) is jointly concave in $\bar{P}_i^m$ and K_i^m, we only need to prove that $-\frac{1}{\theta^m\tau}\ln\left[F_R\left(\bar{P}_{S_i}^m\right)\right]$ is concave in $\bar{P}_{S_i}^m$. For notational simplicity, we omit indices m and i of all the variables in this appendix. In the sequel, we will prove that

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

Substituting (15) into (20), we can obtain that

$$F_R\left(\bar{P}_S\right)=1-e^{-\nu}+\nu^{\frac{\beta}{\beta+1}}\int_{\nu}^{\infty}g^{-\frac{\beta}{\beta+1}}e^{-g}\mathrm{d}g \tag{B.2}$$

where the relation between ν and $\bar{P}_S$ can be obtained from (16). Then, $F_R\left(\bar{P}_S\right)$ can be regarded as a composition function $F_R\left[\nu\left(\bar{P}_S\right)\right]$, and thus

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

From (16), we can derive the relation between $\bar{P}_S$ and ν, i.e., $\frac{d\bar{P}_S}{d\nu}=-\frac{\phi\sigma_0^2}{\alpha(\beta+1)}\nu^{-\frac{\beta+2}{\beta+1}}\int_{\nu}^{\infty}g^{-\frac{\beta}{\beta+1}}e^{-g}\mathrm{d}g$. According to the characteristic of inverse function (i.e., $\frac{d\bar{P}_S}{d\nu}\frac{d\nu}{d\bar{P}_S}=1$ at any point $(\nu,\bar{P}_S)$), we can derive $\frac{d\nu}{d\bar{P}_S}$ from (16), i.e.,

$$\frac{d\nu}{d\bar{P}_S}=-\frac{\alpha\left(\beta+1\right)}{\phi\sigma_0^2}\nu^{\frac{\beta+2}{\beta+1}}\frac{1}{\int_{\nu}^{\infty}g^{-\frac{\beta}{\beta+1}}e^{-g}\mathrm{d}g}. \tag{B.4}$$

From $\frac{d^2\nu}{d\bar{P}_S{}^2}=\frac{d\frac{d\nu}{d\bar{P}_S}}{d\nu}\frac{d\nu}{d\bar{P}_S}$, we can derive that

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

where $\varphi = \int_{\nu}^{\infty} g^{-\frac{\beta}{\beta+1}} e^{-g} \mathrm{d}g$. From (B.2), we have

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

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

$$\frac{dF_R\left[\nu\left(\bar{P}_S\right)\right]}{d\bar{P}_S} = -\frac{\alpha}{\phi\sigma_0^2}\beta\nu, \quad \frac{d^2F_R\left[\nu\left(\bar{P}_S\right)\right]}{d\bar{P}_S^2} = \left(\frac{\alpha}{\phi\sigma_0^2}\right)^2\beta\left(\beta+1\right)\nu^{\frac{\beta+2}{\beta+1}}\frac{1}{\varphi}. \tag{B.7}$$

Upon substituting (B.7), the numerator of (B.1) can be derived as follows,

$$\left(1-e^{-\nu}\right)\left(\frac{\alpha}{\phi\sigma_0^2}\right)^2\beta\left(\beta+1\right)\nu^{\frac{\beta+2}{\beta+1}}\frac{1}{\varphi} + \left(\frac{\alpha}{\phi\sigma_0^2}\right)^2\beta\nu^2. \tag{B.8}$$

Since $\varphi = \int_{\nu}^{\infty} g^{-\frac{\beta}{\beta+1}} e^{-g} \mathrm{d}g > 0$, $\beta = \frac{\theta\tau B}{\ln 2} > 0$, (B.8) is positive, and hence we have (B.1). □

REFERENCES

[1] C. She and C. Yang., "Context aware energy efficient optimization for video on-demand service over wireless networks," in *Proc. IEEE ICCC*, 2015.

[2] N. Bui and J. Widmer, "Data-driven evaluation of anticipatory networking in LTE networks," *IEEE Trans. on Mobile Comput.*, vol. 17, no. 10, pp. 2252–2265, Oct. 2018.

[3] 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.

[4] 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.

[5] 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.

[6] C. She, C. Yang, and L. Liu, "Energy-efficient resource allocation for MIMO-OFDM systems serving random sources with statistical QoS requirement," *IEEE Trans. Commun.*, vol. 63, no. 11, pp. 4125–4141, Nov. 2015.

[7] Z. Yang, W. Xu, Y. Pan, C. Pan, and M. Chen, "Energy efficient resource allocation in machine-to-machine communications with multiple access and energy harvesting for IoT," *IEEE Internet of Things J.*, vol. 5, no. 1, pp. 229–245, Feb. 2018.

[8] L. Xu and W. Zhuang, "Energy-efficient cross-layer resource allocation for heterogeneous wireless access," *IEEE Trans on Wireless Commun.*, vol. 17, no. 7, pp. 4819–4829, Jul. 2018.

[9] 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.

[10] F. Altche and A. de La Fortelle, "An LSTM network for highway trajectory prediction," in *Proc. IEEE ITSC Workshop*, 2017.

[11] C.-Y. Lin, K.-C. Chen, D. Wickramasuriya, S.-Y. Lien, and R. D. Gitlin, "Anticipatory mobility management by big data analytics for ultra-low latency mobile networking," in *Proc. IEEE ICC*, 2018.

[12] N. Bui, M. Cesana, S. A. Hosseini *et al.*, “A survey of anticipatory mobile networking: Context-based classification, prediction methodologies, and optimization techniques,” *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1790–1821, Thirdquarter, 2017.

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

[14] 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.

[15] D. Tsilimantos, A. Nogales-Gomez, and S. Valentin, “Anticipatory radio resource management for mobile video streaming with linear programming,” in *Proc. IEEE ICC*, 2016.

[16] R. Margolies, A. Sridharan, V. Aggarwal *et al.*, “Exploiting mobility in proportional fair cellular scheduling: Measurements and algorithms,” *IEEE/ACM Trans. Networking*, vol. 24, no. 1, pp. 355–367, Feb. 2016.

[17] 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.

[18] 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.

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

[20] C. Yao, J. Guo, and C. Yang, “Achieving high throughput with predictive resource allocation,” in *IEEE GlobalSIP*, 2016.

[21] K. Guo, T. Liu, C. Yang, and Z. Xiong, “Interference coordination and resource allocation planning with predicted average channel gains for HetNets,” *IEEE Access*, vol. 6, pp. 60 137–60 151, 2018.

[22] S. Ahmad, R. Reinhagen, L. S. Muppirisetty, and H. Wymeersch, “Predictive resource allocation evaluation with real channel measurements,” in *Proc. IEEE ICC*, 2017.

[23] W. Zhang, Y. Liu, T. Liu, and C. Yang, “Trajectory prediction with recurrent neural networks for predictive resource allocation,” in *Proc. IEEE ICSP*, 2018.

[24] J. Chen, U. Yatnalli, and D. Gesbert, “Learning radio maps for UAV-aided wireless networks: A segmented regression approach,” in *Proc. IEEE ICC*, 2017.

[25] J. Thrane, M. Artuso, D. Zibar, and H. L. Christiansen, “Drive test minimization using deep learning with bayesian approximation,” in *Proc. IEEE VTC Fall*, 2018.

[26] 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.

[27] 3GPP, *Further Advancements for E-UTRA Physical Layer Aspects*. TSG RAN TR 36.814 v9.0.0, Mar. 2010.

[28] 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, First quarter, 2016.

[29] 5GPPP Architecture Working Group, “View on 5G architecture,” in *5G Architecture White Paper*, Dec. 2017.

[30] S. Bera, S. Misra, and A. V. Vasilakos, “Software-defined networking for internet of things: A survey,” *IEEE Internet of Things J.*, vol. 4, no. 6, pp. 1994–2008, Dec. 2017.

[31] A. Ksentini and N. Nikaein, “Toward enforcing network slicing on RAN: Flexibility and resources abstraction,” *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 102–108, Jun. 2017.

[32] 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.

[33] H. Abou-zeid, H. S. Hassanein, and R. Atawia, “Towards mobility-aware predictive radio access: Modeling, simulation, and evaluation in LTE networks,” in *Proc. ACM MSWiM*, 2014.

[34] W. Jiang and H. D. Schotten, “Neural network-based channel prediction and its performance in multi-antenna systems,” in *Proc. IEEE VTC Fall*, 2018.

[35] 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 *IEEE ICC*, 2016.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] C. Chang and J. A. Thomas, “Effective bandwidth in high-speed digital networks,” *IEEE J. Sel. Areas Commun.*, vol. 13, no. 6, pp. 1091–1100, Aug. 1995.

[41] 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.

[42] C. Desset *et al.*, “Flexible power modeling of LTE base stations,” in *Proc. IEEE WCNC*, 2012.

[43] G. Auer, O. Blume, V. Giannini, I. Gódor, *et al.*, “D 2.3: Energy efficiency analysis of the reference systems, areas of improvements and target breakdown,” *EARTH*, Jan. 2012.

[44] C. She and C. Yang, “Energy efficiency and delay in wireless systems: Is their relation always a tradeoff?” *IEEE Trans. on Wireless Commun.*, vol. 15, no. 11, pp. 7215–7228, Nov. 2016.

[45] 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.

[46] 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.

[47] J. Gregory, *Constrained Optimization In The Calculus Of Variations and Optimal Control Theory*. Chapman and Hall/CRC, 2018.

[48] A. Goldsmith, *Wireless Communications*. Cambridge University Press, 2005.

[49] S. Boyd and L. Vandanberghe, *Convex Optimization*. Cambridge Univ. Press, 2004.

[50] Video Trace Library. [Online]. Available: http://trace.eas.asu.edu/videotraces2/cgs/cif/Sony_G16B15_CIF_DQP6_5EL_48_42_36_30_24_18/

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

[52] H. A. Omar, W. Zhang, A. Abdrabou, 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.

[53] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski, *Robust Optimization*. Princeton University Press, 2009.