Predictive Resource Allocation with Coarse-grained Mobility Pattern and Traffic Load Information

Jia Guo*, Changyang She† and Chenyang Yang*
*Beihang University, Beijing, China
†Singapore University of Technology and Design, Singapore
Emails: guojia@buaa.edu.cn, shechangyang@gmail.com, cyyang@buaa.edu.cn

***Abstract*—Predictive resource allocation can exploit residual resources in wireless networks to support high throughput, improve user experience, and enhance energy efficiency. Most priori works assume that fine-grained knowledge for user trajectory and/or traffic load is known, which is hard to predict in practice. In this paper, we investigate predictive resource allocation to achieve high throughput for mobile users requesting video-on-demand (VoD) services, which employs cell-level coarse grained information. In the start of a prediction window, we only need to predict the cells the users to be associated with, the sojourn time of each user in each cell, the loads of VoD traffic and realtime traffic at each base station (BS). These information is translated into two thresholds, which are introduced to help each BS to determine when and how much data to transmit. Two-threshold-based algorithms are provided. Simulation results show that the algorithms perform closely to the optimal predictive resource allocation with perfect fine-grained information in terms of supporting high request arrival rate and improving user experience, and one algorithm even outperforms the optimal method with prediction errors.**

***Index Terms*—Predictive resource allocation, coarse-grained information, high throughput, quality of service**

I. Introduction

To support the explosively growing traffic demands, one of the main trend techniques for the fifth generation (5G) cellular networks is to improve spectral efficiency (SE) by network densification. While improving SE is always beneficial, in reality the resources are often under-utilized in many base stations (BSs) because of the temporal-spatial traffic variation.

The dynamic nature of the traffic comes from user behavior, which has long been regarded as random. With the recent flourish of big data analysis, the user behavior is shown predictable. For example, the traffic load and user trajectory can be predicted [1], [2], from which the future average resource usage status of a network and the average channel gains of a user (with the help of radio map [3]) can be derived [4], [5]. As a consequence, predictive resource allocation emerges as a promising way to exploit residual resources [6].

For non-realtime traffic such as video on demand (VoD), it is possible to serve mobile stations (MSs) when they are in good channel condition [5] and/or located in light-loaded cells [4], i.e., with higher data rate. Assumed that the MS's future instantaneous data rate can be predicted accurately, several problems were formulated and solved in [5] to optimize a resource allocation plan to determine which BSs along the trajectory of a MS serves the MS with how much resources. To deal with the prediction errors, robust optimization frameworks were proposed in [7], [8], where the prediction errors on future rates are either modeled as Gaussian noises or bounded noises.

While existing results show remarkable gain in supporting high throughput [5], [9] or reducing power consumption of the BSs [10], all of them require the predicted information in the granularity of seconds [9]. However, as far as the authors known, no existing works can predict the trajectories of outdoor MSs and the traffic loads of BSs in such a granularity. Besides, in order to construct a fine-grained radio map, large amount of drive tests are required, which is expensive. Moreover, predicting trajectories of large number of MSs in such a fine grain incurs high complexity. Fortunately, some rough predictions for mobility pattern, such as which cells a MS will enter and how long the MS will stay in each cell, are available [11]. This inspires us to investigate how to allocate radio resources with coarse-grained prediction.

In this paper, we propose predictive resource allocation method with cell-level prediction for mobility pattern and traffic load. To achieve high throughput while satisfy the quality of service (QoS) for the MSs with VoD requests, we find two thresholds based on the predicted information for making the resource allocation plan. In particular, the thresholds are respectively used to determine whether or not a user is in good channel condition and will travel to a BS with heavy load. Two robust predictive resource allocation algorithms are presented. Surprisingly, though heuristic, the two algorithms are able to perform closely to relevant optimal solution with perfect prediction, in terms of supporting high throughput given user satisfaction rate and tolerance of QoS, as demonstrated by simulations.

II. System Model

Consider a multi-cell network, where each cell is with radius $R_{\rm b}$ and each BS is with height $h_{\rm b}$. Each BS may serve two kinds of traffic with bandwidth $W_{\max}$, transmit power $P_{\max}$ and $N_{\rm t}$ antennas. The first kind is realtime traffic, such as phone calls and video conference. The other is VoD traffic. Because realtime traffic has higher priority, the VoD traffic can be served by the residual resources of the network after the QoS of realtime users is guaranteed. For the MSs in the network that request VoD service, we call them VoD users

This work is supported by National Natural Science Foundation of China (NSFC) under Grant 61731002 and 61429101.

or simply MSs in the sequel. Assume that there is a central processor (CP), which can gather data from BSs and MSs to predict information within a prediction window and then share the information to BSs for predictive resource allocation.

Time is discretized to frames each with duration Δ (say 1 s), and each frame includes T_s time slots each with duration of unit time (say 1 ms). The durations are defined according to the channel variation, i.e., the variation of large scale fading (i.e., path-loss and shadowing) and small scale fading due to user mobility. Assume that large scale channel gain remains constant within each frame and may vary among frames, and the small scale channel gain remains constant within each time slot and follows independent and identical distribution (i.i.d.) among time slots. The prediction window includes T_f frames.

To exploit residual resource in the network to support high throughput, a MS is only associated to the BS with the highest average channel gain. To focus on illustrating how to use coarse prediction for predictive resource allocation, we consider time division multiple access to avoid multi-user interference. In particular, each BS serves only one MS with all residual bandwidth and transmit power in each time slot, and serves multiple MSs in the same cell in different time slots. Then, maximal ratio transmission is the optimal beamforming. If MS_k is associated with BS_m (denoted as BS_m) in the jth frame, then the achievable rate of the MS in the tth time slot of the jth frame can be expressed as

$$R_{j,t}^k = W_{j,t}^m \log_2 \left(1 + \frac{\alpha_j^k \|\mathbf{h}_{j,t}^k\|^2}{N_0 W_{j,t}^m} p_{j,t}^m\right), \quad (1)$$

where $W_{j,t}^m$ and $p_{j,t}^m$ are respectively the residual bandwidth and transmit power in the tth time slot of the jth frame, N_0 is the noise power spectrum density, $\mathbf{h}_{j,t}^k \in \mathbb{C}^{N_t \times 1}$ is the small scale Rayleigh fading channel vector with i.i.d. elements, $\alpha_j^k = (d_j^k)^\beta 10^{X_j^k/10}$ is the large scale channel gain (also called average channel gain in the rest of the paper) in the jth frame. Here, d_j^k is the distance between MS_k and its associated BS in the jth frame, β is the path-loss exponent, X_j^k reflects shadowing, which is a Gaussian distributed random variable (i.e., shadowing follows log-normal distribution) and is i.i.d. among different users, and $\|\cdot\|$ denotes magnitude.

To reflect the residual bandwidth after serving randomly arrived RT traffic with random service time, we model $W_{j,t}^m$ as i.i.d. random variables among the time slots in the jth frame [4]. Assume that the residual transmit power is proportional to the residual bandwidth [10], i.e., $p_{j,t}^m = W_{j,t}^m P_{\max}/W_{\max}$. Then, the time-average achievable rate in the jth frame (called average rate for short) of MS_k can be expressed as,

$$R_j^k = \frac{1}{T_s}\sum_{t=1}^{T_s} R_{j,t}^k = \frac{1}{T_s}\sum_{t=1}^{T_s} W_{j,t}^m \log_2\left(1+\frac{\alpha_j^k \|\mathbf{h}_{j,t}^k\|^2}{\sigma^2} P_{\max}\right), \quad (2)$$

where $\sigma^2 = N_0 W_{\max}$.

In a prediction window, the VoD requests of the MSs randomly arrive at BS_m with average arrival rate of λ_m (requests/frame). Each video file is divided into multiple segments and then coded. Each segment is a stand-alone unit with B_{seg} bits and playback time of T_{seg} frames. Once a segment is completely received by a MS, it can be decoded and played. To avoid playback interruption due to empty playout buffer (i.e., stalling), a segment should be sent to the MS before the end of playing previous segment.

III. Predictive Resource Allocation with Coarse-grained Information

Existing works have demonstrated large performance gain of predictive resource allocation by either assuming known future data rate or large scale channel gains and residual bandwidths in a prediction window, in the grain of a frame duration [10] or even a time slot duration [5]. In this section, we strive to answer the following question: is it possible to achieve the promising gain with coarse-grained information?

To this end, we go back to the basic idea of predictive resource allocation: to improve network performance without degrading the QoS for a delay tolerant mobile user, the required data should be transmitted the MS when it can be served with higher rate [5].

Such an idea can be translated into the following intuitions. In order to support high throughput, the BSs should transmit more bits to the MSs with higher large scale channel gains. In order to avoid stalling, each BS needs to decide how many video segments should be transmitted in each frame according to the buffer status at each MS in its coverage as well as the residual resources of the BSs each MS to be served. Because a MS may experience long stalling time when it is served by a busy BS, more segments should be transmitted by a BS with light traffic load to the MSs heading to heavy-loaded cells. These intuitions seem simple, but how we connect them with cell-level prediction is not straightforward. In what follows, we find two "rulers" for predictive resource allocation: one for average channel gain and the other for average residual bandwidth, to judge whether or not they are high.

A. Predicting the Threshold for Average Channel Gain

To judge in what channel condition a BS should transmit to a MS, a simple "ruler" is the mean value of the average channel gains in a prediction window. Then, we do not need the prediction of fine-grained trajectory, which is a sequence of location and time pairs with sampling period of seconds.

To see how such a mean value is obtained and provide the intuition to find a more effective "ruler" that employs cell-level mobility pattern prediction, i.e., the associated BSs and sojourn time in each cell for a MS, we start from a finer prediction.

Consider MS_k, who will travel across the 1st, $\cdots$, Mth cells successively in the prediction window, and will be associated with BS_m during $T_{m,k}$ frames. Assume that the route that MS_k will travel is predicted with the method proposed in [2] (this does not mean that the location in each frame is predicted). If the arrival and departure locations of MS_k in each cell can be further predicted, and the MS moves in a constant speed, then the mean value of the average channel gains in a prediction

window can be predicted as

$$\alpha_{\mathrm{ave},k} = \frac{\sum_{n=1}^{M} \alpha_{\mathrm{ave},k}^{m} T_{m,k}}{\sum_{m=1}^{M} T_{m,k}}, \tag{3}$$

where $\alpha_{\mathrm{ave},k}^{m}$ is the average value of the large scale fading gains of MS_k when it is on a road in the mth cell. However, mean value is sensitive to outliers, which are the average channel gains with large prediction errors for the problem at-hand. Besides, predicting the particular route and these two locations need more computing resources.

To find a "ruler" that is robust to prediction errors, we can use median. As shown in [12], 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. Unfortunately, it is hard to derive the median of average channel gains $\alpha_{\mathrm{med},k}$ as to deriving the mean value $\alpha_{\mathrm{ave},k}$. Inspired by the expression of (3), we introduce a threshold for average channel gains as

$$\alpha_{\mathrm{th}}^{k} \triangleq \frac{\sum_{m=1}^{M} \alpha_{\mathrm{med}}^{m} \overline{T}_m}{\sum_{m=1}^{M} \overline{T}_m}, \tag{4}$$

where $\alpha_{\mathrm{med}}^{m}$ is the median of the large scale fading gains in the mth cell, and $\overline{T}_m = \sum_{k \in \mathcal{K}_{j,m}} T_{m,k}/card(\mathcal{K}_{j,m})$ is the average number of frames that the MSs are associated with BS_m, $\mathcal{K}_{j,m}$ is the set of MSs within the coverage of BS_m in the jth frame, and $card(\cdot)$ is the cardinality of a set.

In practice, each BS can compute and store the median with the average channel gains of the MSs it served and update the medians when new MSs arrive.

Such a threshold seems too rough. Nonetheless, as we will explain via numerical result in the next section, even such a coarse "ruler" can achieve good performance for predictive resource allocation.

B. Predicting the Threshold for Average Residual Bandwidth

To find a "ruler" for average residual bandwidth, we first find the residual bandwidth that can exactly ensure the QoS of all the MSs associated with a BS (say BS_m) in a frame (say the jth frame), which is denoted as $W_{\mathrm{th},j}^{m}$. With $W_{\mathrm{th},j}^{m}$, BS_m is able to transmit $B_{\mathrm{seg}}/T_{\mathrm{seg}}$ bits to each MS associated with it in the jth frame, and then a segment can be conveyed in the duration of T_{seg} frames before playback the segment. Hence, to avoid stalling, $W_{\mathrm{th},j}^{m}$ should satisfy

$$\frac{W_{\mathrm{th},j}^{m}}{T_s} \sum_{k=1}^{K} s_j^k \sum_{t=1}^{T_s} \log_2\left(1 + \frac{\alpha_j^k \|\mathbf{h}_{j,t}^k\|^2}{\sigma^2} P_{\max}\right) = K \frac{B_{\mathrm{seg}}}{T_{\mathrm{seg}}\Delta}, \tag{5}$$

where K is the number of MSs who are associated with BS_m in the jth frame, Δ is the frame duration, $s_j^k \in [0,1]$ is the percentage of the time resources assigned to MS_k in the jth frame, and $\sum_{k=1}^{K} s_j^k \leq 1$. Then, $W_{\mathrm{th},j}^{m}$ can be obtained as,

$$W_{\mathrm{th},j}^{m} = \frac{K B_{\mathrm{seg}} T_s / T_{\mathrm{seg}}}{\Delta \sum_{k=1}^{K} s_j^k \sum_{t=1}^{T_s} \log_2\left(1 + \frac{\alpha_j^k \|\mathbf{h}_{j,t}^k\|^2}{\sigma^2} P_{\max}\right)}. \tag{6}$$

$W_{\mathrm{th},j}^{m}$ depends on the information hard to predict, say future small scale channel gains and large scale channel gain of the MSs in each frame. Hence, it is not viable to used as a threshold to determine whether a MS is heading to a cell with high or low residual resource. Nevertheless, by introducing several assumptions and approximations, we can find a practical threshold as shown in the following proposition.

Proposition 1: If (i) $s_j^k = 1/K$, (ii) N_{t} is large, (iii) $\frac{\alpha_j^k \|\mathbf{h}_{j,t}^k\|^2}{\sigma^2} P_{\max} \gg 1$, (iv) $d_j^k \sim \mathbb{U}(h_{\mathrm{b}}, \sqrt{h_{\mathrm{b}}^2 + R_{\mathrm{b}}^2})$ and is i.i.d. among MSs, then $W_{\mathrm{th},j}^{m}$ can be approximated as,

$$W_{\mathrm{th}}^{m} = \frac{\lambda_m \overline{T}_m B_{\mathrm{seg}}}{\log_2\left(1 + \frac{\bar{d}^{\beta} N_{\mathrm{t}}}{\sigma^2} P_{\max}\right) T_{\mathrm{seg}}}, \tag{7}$$

where $\bar{d} = (h_{\mathrm{b}} + \sqrt{R_{\mathrm{b}}^2 + h_{\mathrm{b}}^2})/2$ is the average distance, and $\mathbb{U}(\cdot,\cdot)$ denotes uniform distribution.

Proof: See Appendix A. ∎

Both the average arrival rate λ_m and sojourn time $T_{m,k}$ are predictable as reported in [1] and [11]. Besides, the path-loss exponents in typical scenarios (say macro cells, pico cells, and small cells) have been estimated in existing literatures or 3GPP reports. Hence, W_{th}^{m} can serves as a bandwidth threshold.

C. Threshold-based Predictive Resource Allocation

At the beginning of a prediction window, the CP only needs to compute α_{th}^{k} and W_{th}^{m}, with the prediction on the cells each MS to be associated with, the sojourn time of each MS in each cell, the average arrival rate of VoD traffic, as well as average residual bandwidth at each BS (which can be derived from the average arrive rate of realtime traffic [4]). Then, CP informs the two "rulers" to each relevant BS.

At the beginning of each frame, a BS can determine how many segments should be transmitted to each MS in its coverage by the following two steps. Step 1): To avoid stalling, the BS should convey a segment to a MS until a segment is in the MS's buffer before the end of each frame. Step 2): If there are still resources available in the frame after the first step, then the BS will transmit more segments to the MSs who meet the following conditions with the two "rulers: (i) experiencing good channels, (ii) heading to the BSs with low residual bandwidth.

To support high throughput, each BS should transmit more segments to the MSs with higher average rate in each frame. This suggests that the number of extra segments transmitted to a MS can be designed as proportional to the average rate of the MS in each frame. Denote the percentage of residual time resources at BS_m after the first step in the jth frame as S_j^m. Then, BS_m can transmit $\frac{S_j^m}{K_j^m} \cdot \frac{R_j^k \Delta}{B_{\mathrm{seg}}}$ segments to MS_k to exploit the residual resource, where MS_k is among the K_j^m MSs who satisfy conditions (i) and (ii) in the jth frame.

After $\overline{W}_j^m$ and α_j^k are estimated at BS_m at the beginning of the jth frame, it is easy to show that the average rate of MS_k in the jth frame can be accurately approximated as

$$R_j^k \approx \overline{W}_j^m \log_2\left(1 + \frac{\alpha_j^k N_{\mathrm{t}}}{\sigma^2} P_{\max}\right). \tag{8}$$

when N_{t} is large, where $\overline{W}_j^m$ is the mean value of residual bandwidth of BS_m in the jth frame.

The threshold-based predictive resource allocation is summarized in Algorithm 1, where $\overline{W}^m$ is the average residual bandwidth within the prediction window for BS$_m$, $D_{j,k}$ is the amount of data transmitted to MS$_k$ in the jth frame.

Algorithm 1 Predictive Resource Allocation

Input: λ_m, $\overline{T}_m$, α^m_{med} and $\overline{W}^m$

At the beginning of a prediction window, the CP predicts W^m_{th} for BS$_m$ with (7) and α^k_{th} for MS$_k$ with (4), and informs the two "rulers" to each relevant BS.

At the beginning of the jth frame, BS$_m$ first estimates $\overline{W}^m_j$ and α^k_j and then estimates R^k_j with (8), and MS$_k$, $k \in \mathcal{K}_{j,m}$ reports the amount of data in its buffer, D_k, to BS$_m$. Then, BS$_m$ operates in each time slot with following two steps, where at the start of each time slot, BS$_m$ estimates $\mathbf{h}^k_{j,t}, \forall k \in \mathcal{K}_{j,m}$, schedules the MS with maximal $R^k_{j,t}$ and transmits with maximal ratio transmission.

Step 1:

$D_{j,k} = 0, \forall k \in \mathcal{K}_{j,m}$
$t := 1$
while $t \le T_s$ **do**
  $\tilde{\mathcal{K}} := \{k | k \in \mathcal{K}_{j,m}$ and $D_k < B_{\text{seg}}\}$
  if $\tilde{\mathcal{K}} \neq \emptyset$ **then** $k := \arg\max\{R^k_{j,t} | k \in \tilde{\mathcal{K}}\}$
  else break
  end if
  $D_k := D_k + R^k_{j,t}\Delta/T_s$
  $t := t + 1$
end while

Step 2:

if $t \le T_s$ **then**
  $S^m_j := 1 - (t-1)/T_s$
  $\hat{\mathcal{K}} := \{k | k \in \mathcal{K}_{j,m}$ and $\alpha^k_j > \alpha^k_{\text{th}}$ and $\overline{W}^{n_k} < W^{n_k}_{\text{th}}\}$
  ($\overline{W}^{n_k}$ and $W^{n_k}_{\text{th}}$ are the predicted average residual bandwidth and the bandwidth threshold for the next BS that MS$_k$ will associate to, respectively)
  $K^m_j := card(\hat{\mathcal{K}})$
  while $t \le T_s$ **do**
    $\check{\mathcal{K}} := \{k | k \in \hat{\mathcal{K}}$ and $D_{j,k} < \frac{S^m_j}{K^m_j} R^k_j \Delta\}$
    if $\check{\mathcal{K}} \neq \emptyset$ **then**
      $k := \arg\max\{R^k_{j,t} | k \in \check{\mathcal{K}}\}$
      $D_{j,k} := D_{j,k} + R^k_{j,t}\Delta/T_s$
    else $k := \arg\max\{R^k_{j,t} | k \in \mathcal{K}_{j,m}\}$
    end if
    $D_k := D_k + R^k_{j,t}\Delta/T_s$
    $t := t + 1$
  end while
end if

When the residual bandwidth varies significantly among BSs, the MSs can benefit from Algorithm 1 because they can receive more data when they associate with a BS with abundant residual resource before associating with a BS with less residual resource. When the residual bandwidth of BSs in the network are with little difference, the prediction of the bandwidth threshold is of little use. Then, Algorithm 1 can be degenerated into a simpler algorithm, called **Algorithm 2**, which operates as follows.

At the beginning of a prediction window, the CP predicts α^k_{th} for each MS in the network, and then informs the "ruler" to each relevant BS. At the beginning of the jth frame, BS$_m$ estimates α^k_j, and MS$_k$, $k \in \mathcal{K}_{j,m}$ reports the amount of data in it buffer to BS$_m$. Then, BS$_m$ operates in each time slot of the jth frame with previous two steps, where in Step 2) only the condition (i) is considered. In other words, the BS transmits extra segments to MSs when their large scale fading gains exceed the threshold.

IV. Simulation and Numerical Results

In this section, we evaluate the performance of the proposed algorithms with simulations.

Consider a cellular network with cell radius $R_{\text{b}} = 250$ m. We consider six BSs, each with height $h_{\text{b}} = 20$ m and equipped with $N_{\text{t}} = 8$ antennas, which are located along a straight line. The maximal transmit power of each BS is 40 W and cell-edge SNR is set as 5 dB, where the intercell interference is implicitly reflected. The path loss model is $36.8 + 36.7\log_{10}(d)$, where d is the distance between the BS and MS in meter. The standard derivation and decorrelation distance of shadowing are respectively 8 dB and 120 m [13]. The MSs move along three roads of straight lines with minimum distance from the BSs as 50 m, 100 m and 150 m, respectively. Each MS requests a video with size of $B = 20$ Mbytes and playback duration of 100 s. Each video consists of $N = 10$ segments, i.e., each segment with size of 1 Mbytes is played out for $T_{\text{seg}} = 10$ s. Each frame is with duration of 1 s, and each time slot is with duration $\Delta = 10$ ms, i.e., each frame contains $T_s = 100$ time slots. The prediction window contains $T_f = 300$ frames, i.e., 5 min.

To characterize the different resource usage status of the BSs by serving the RT traffic in an under-utilized network, we consider three types of BSs: (i) busy BS with average residual bandwidth in the prediction window as $\overline{W}^m = 1$ MHz, (ii) not-so-busy BS with $\overline{W}^m = 3$ MHz, (iii) idle BS with $\overline{W}^m = 10$ MHz, which are located along the line as idle, not-so-busy, busy, busy, not-so-busy, idle BS. The results are obtained from 100 Monte Carlo trails. In each trail, the MSs start to initiate requests randomly at a location in one of the BSs, the trajectories change randomly with speed uniformly distributed in $(10, 20)$ m/s and directions uniformly selected from 0 or +180 degree, the requests of the MSs arrive from the 1st to the 100th frame in the prediction window according to Poisson process. In each time slot of each trail, the small-scale channel changes independently according to Rayleigh fading, and the instantaneous residual bandwidth $W^m_{j,t}$ changes randomly according to Gaussian distribution with mean value of $\overline{W}^m$ and deviation $0.2\overline{W}^m$. With this setting, the sojourn time of each MS in each cell is random.

We compare the two algorithms with the following methods.

- **Optimal**: This is a modified version of the optimal methods proposed in [5], [9], which minimizes the total transmission time without causing stalling. In [5], [9], the

resource allocation plan is only made once because the requests of MSs are assumed known at the beginning of prediction window. For a fair comparison, we modify the optimal methods such that a resource allocation plan is made for a MS only after it requests a video file.

- **Baseline**: This is a non-predictive method [14], where each BS serves the MS with the earliest deadline in each time slot. If several MSs have the same deadline, then the MS with most bits to be transmitted is served firstly.

We first show the accuracy of using α_{th}^k with (4) as an estimate, measured by $\alpha_{\text{th}}^k/\alpha_{\text{med},k}$, where $\alpha_{\text{med},k}$ is obtained by computing the median of perfect average channel gains in a prediction window with granularity of one second. Here, the MSs initiate their requests in the first BS and arrive at the sixth BS at the end of the prediction window. We validate the accuracy of (4) in the following two cases where different information in the prediction window need to predict: (i) the BSs that MSs will associate, (ii) the roads that the MSs will travel as well as their locations of arriving and departing each cell. The sojourn time of each MS in each cell is assumed known for both cases. Fig. 1 shows the cumulative distribution function (CDF) of the accuracy after 10000 Monte Carlo trails. As expected, α_{th}^k is more accurate when more information is predicted. In case (i), we can see that $0.35 < \alpha_{\text{th}}^k/\alpha_{\text{med},k} < 2.4$.

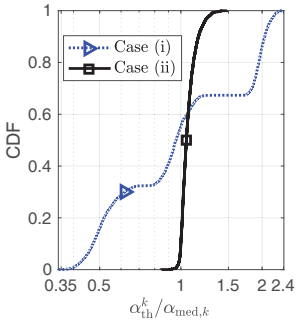

Fig. 1. CDF of $\alpha_{\text{th}}^k/\alpha_{\text{med},k}$.

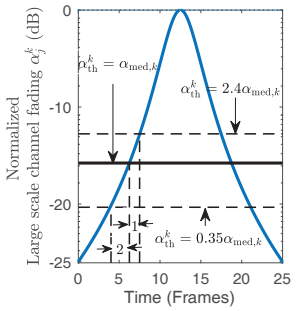

Fig. 2. Average channel gain of MS_k with a speed of 20 m/s.

Then, we explain why such a very inaccurate value of α_{th}^k can still serve as an effective "ruler" for the proposed algorithms via numerical results. In Fig. 2, we provide the average channel gain of MS_k when it travels across a cell, where the gain is normalized by its maximal value. For easy understanding, only path loss is considered. We can observe that the MS uses 25 frames to traverse across a cell while the dynamic range of average channel gains is 25 dB. Compared with such a large dynamic range, the prediction error on α_{th}^k is very small. As a result, the BS makes wrong decision in Step 2) (i) only in two or four frames with α_{th}^k.

Next, we validate the accuracy of approximating (6) with (7). Assume that λ_m and $T_{m,k}$ can be predicted accurately, and $\overline{T}_m$ is obtained by taking the average of sojourn time of all MSs in the mth cell. The true values are random variables obtained from (6) with 1000 trails, where in each trail the MSs are randomly distributed in the mth cell and s_j^k is set as proportional to the average rate of MS_k in (8). Note that this simulation setup differs from the assumptions used in deriving (7). This is because in simulation $s_j^k \neq 1/K$, (ii) $N_{\text{t}} = 4$ is not large, $\frac{\alpha_j^k \|\mathbf{h}_{j,t}^k\|^2}{\sigma^2} P_{\max}$ may not far exceed one, and the distances of the uniformly located MSs do not follow uniform distribution. Table I shows the statistics of the errors between the true values and W_{th}^m computed with (7). The results show that W_{th}^m can be approximated with a bias no more than 0.1 MHz and standard deviation no more than 1 MHz when the average arrival rate of VoD users $\lambda_m \leq 0.9$ requests/s. This implies that Algorithm 1 is applicable when the difference of residual bandwidths of adjacent cells exceeds 1 MHz. Otherwise, we can simply employ Algorithm 2.

TABLE I
MEAN VALUE AND STANDARD DEVIATION OF APPROXIMATION ERRORS

λ_m (requests/s)	0.1	0.3	0.5	0.7	0.9
Mean value (MHz)	-0.03	-0.06	-0.07	-0.07	-0.01
Standard deviation (MHz)	0.40	0.46	0.61	0.70	0.89

Finally, we evaluate the performance of the proposed algorithms. To show the traffic carrying ability of the network for supporting the MSs with given tolerance on QoS, we evaluate the maximal request arrival rate of VoD users given maximal stalling time expected by 99.8% of all MSs. All the methods are simulated with both perfect prediction and with errors. For perfect prediction, optimal method employs perfect values of α_j^k and W_j^m in the window, and the proposed algorithms employ perfect values of α_j^k and W_j^m to compute the two thresholds. For imperfect prediction, in the optimal method, the prediction errors of α_j^k and W_j^m are set as Gaussian distribution with standard derivation $0.2\alpha_j^k$ and $0.2W_j^m$, respectively. In Algorithm 1 and Algorithm 2, α_{th}^k is computed by (4) with the cell-level prediction. To reflect the prediction error of W_{th}^m, we model the predicted errors of λ_m and $\overline{T}_m$ as Gaussian distributed random variables with standard derivation $0.2\lambda_m$ and $0.2\overline{T}_m$, respectively, then compute W_{th}^m with (7).

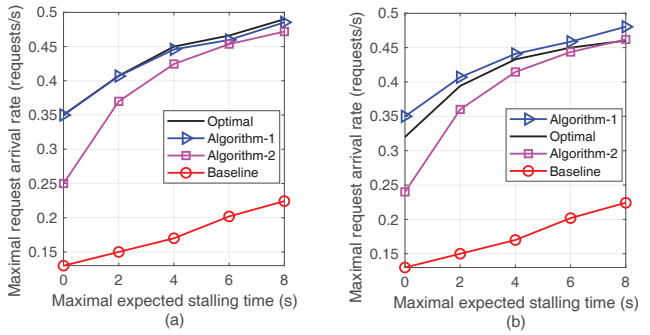

Fig. 3. Maximal traffic carrying ability to support MSs with given QoS, where in (a) prediction is perfect, (b) predicted information is imperfect.

The results show that when the predicted information is perfect in a fine-grained level, the performance of Algorithm 1 is very close to the optimal method. As expected, Algorithm 1 outperforms Algorithm 2 when the MSs tolerate low stalling time. Surprisingly, the performance of the proposed algorithms degrades little when the prediction is very coarse and with larger errors. By contrast, the performance of the optimal method is sensitive to prediction errors. With prediction errors, Algorithm 1 even outperforms the optimal method. All the predictive methods have dramatically performance gain over the non-predictive baseline.

In previous setup, we set the average residual bandwidth in the prediction window fixed during simulations, and the path-loss exponents of all cells identical. Nonetheless, more simulations by setting randomly values of $\overline{W}^m$ in each trail and random values of β among different cells show similar trends as previous results, which are not provided for conciseness.

V. Conclusions

In this paper, we studied predictive resource allocation with coarse grained mobility pattern and traffic load information. We first found two "rulers" to determine whether a MS is experiencing good channel and is heading to a busy BS. In order to support high arrival rate of VoD requests while avoid long stalling for MSs, we proposed a simply way to make resource allocation plan with the "rulers" predicted at the start of a prediction window. Simulation results demonstrated that the performance of our method is close to the optimal solution with fine-grained prediction and exhibits dramatic gain over the non-predictive resource allocation. Essentially, such a surprising result comes from a largely-overlooked fact: the average channel gain varies in a very large scale, which makes its accurate and fine-grained prediction unnecessary.

Appendix A
Proof of Proposition 1

If $s_j^k = 1/K$, then (6) can be written as,

$$W_{\mathrm{th},j}^m = \frac{KB_{\mathrm{seg}}/T_{\mathrm{seg}}}{\frac{1}{K}\frac{\Delta}{T_s}\sum_{k=1}^{K}\sum_{t=1}^{T_s}\log_2\left(1+\frac{\alpha_j^k\|\mathbf{h}_{j,t}^k\|^2}{\sigma^2}P_{\max}\right)}. \quad \text{(A.1)}$$

Denote $\xi_{j,t}^k = \log_2\left(1+\frac{\alpha_j^k\|\mathbf{h}_{j,t}^k\|^2}{\sigma^2}P_{\max}\right)$, which is a random variable. By taking the average over small scale channel, shadowing, and distance, we can derive that

$$\mathbb{E}\{\xi_{j,t}^k\} = \int_{\bar{d}-\delta/2}^{\bar{d}+\delta/2}\Big\{\int_{-\infty}^{\infty}\Big(\int_0^{\infty}\log_2\left(1+\frac{d^\beta 10^{X_\sigma/10}\|\mathbf{h}\|^2}{\sigma^2}P_{\max}\right) f^{(H)}(\|\mathbf{h}\|^2)\mathrm{d}\|\mathbf{h}\|^2\Big) f^{(X)}(X)\mathrm{d}X\Big\} f^{(d_j)}(d)\mathrm{d}d, \quad \text{(A.2)}$$

where $\bar{d} = (h_{\mathrm{b}} + \sqrt{h_{\mathrm{b}}^2+R_{\mathrm{b}}^2})/2$ is the average distance, $\delta = h_{\mathrm{b}} - \sqrt{h_{\mathrm{b}}^2+R_{\mathrm{b}}^2}$ is the deviation from the average, $f^{(H)}(\|\mathbf{h}\|^2) = \frac{(\|\mathbf{h}\|^2)^{N_{\mathrm{t}}-1}e^{-\|\mathbf{h}\|^2}}{\Gamma(N_{\mathrm{t}})}$ is the probability density function (PDF) of $\|\mathbf{h}_{j,t}^k\|^2$, $f^{(X)}(X) = \exp(-X^2/2\sigma_X^2)/\sqrt{2\pi}\sigma_X$ is the PDF of X_j^k, $f^{(d_j)}(d) = 1/\delta$ is the PDF of d_j^k, $\Gamma(\cdot)$ is the Euler gamma function, and $\mathbb{E}\{\cdot\}$ denotes mean value.

When $\frac{(d_j^k)^\beta 10^{X_j^k/10}\|\mathbf{h}\|^2}{\sigma^2}P_{\max} \gg 1$, we have $\log_2\left(1+\frac{(d_j^k)^\beta 10^{X_j^k/10}\|\mathbf{h}\|^2}{\sigma^2}P_{\max}\right) \approx \log_2\left(\frac{(d_j^k)^\beta 10^{X_j^k/10}\|\mathbf{h}\|^2}{\sigma^2}P_{\max}\right)$.

From [15], we can obtain the following integral result,

$$\int_0^{\infty} a\ln(bx)x^{N-1}e^{-cx}\mathrm{d}x = ac^{-N}\Gamma(N)\left\{\ln\left(\tfrac{b}{c}\right)+\psi(N)\right\},$$

where $a>0$, $b>0$, $c>0$, $\psi(\cdot)$ is the digamma function. Then, we have

$$\begin{aligned}&\int_0^{\infty}\log_2\left(1+\frac{d^\beta 10^{X/10}\|\mathbf{h}\|^2}{\sigma^2}P_{\max}\right)f^{(H)}(\|\mathbf{h}\|^2)\mathrm{d}\|\mathbf{h}\|^2\\ \approx\ &\log_2\left(\frac{d^\beta 10^{X/10}}{\sigma^2}P_{\max}\right)+\frac{\psi(N_{\mathrm{t}})}{\ln 2}\\ =\ &\frac{X}{10}\log_2(10)+\beta\log_2(d)+\frac{\psi(N_{\mathrm{t}})}{\ln 2}+\log_2(P_{\max}/\sigma^2).\end{aligned}$$

Upon substituting into (A.2), we can obtain

$$\begin{aligned}\mathbb{E}\{\xi_{j,t}^k\} \approx\ &\int_{\bar{d}-\delta/2}^{\bar{d}+\delta/2}\Big\{\int_{-\infty}^{\infty}\left(\frac{X}{10}\log_2(10)+\beta\log_2(d)+\frac{\psi(N_{\mathrm{t}})}{\ln 2}\right)\\ &\cdot f^{(X)}(X)\mathrm{d}X\Big\}f^{(d_j)}(d)\mathrm{d}d\\ =\ &\frac{\beta}{\ln 2}\left(\frac{\bar{d}}{\delta}\ln\left(\frac{\bar{d}+\delta/2}{\bar{d}-\delta/2}\right)+\frac{1}{2}\ln\left(\bar{d}^2-\frac{\delta^2}{4}\right)-1\right)+\frac{\psi(N_{\mathrm{t}})}{\ln 2}+C.\end{aligned}$$

When $\delta \ll \bar{d}$, $\frac{\bar{d}}{\delta}\ln\left(\frac{\bar{d}+\delta/2}{\bar{d}-\delta/2}\right) \approx 1$, and $\ln\left(\bar{d}^2-\delta^2/4\right) \approx \ln(\bar{d})$. Moreover, when N_{t} is large, $\psi(N_{\mathrm{t}}) \approx \ln(N_{\mathrm{t}})$. Thus, we have

$$\mathbb{E}\{\xi_{j,t}^k\} \approx \beta\log_2(d)+\log_2(N_{\mathrm{t}})+C \approx \log_2\left(1+\frac{\bar{d}^\beta N_{\mathrm{t}}}{\sigma^2}P_{\max}\right).$$

Then, the mean value of $\xi_j = \frac{1}{K}\frac{1}{T_s}\sum_{k=1}^{K}\sum_{t=1}^{T_s}\xi_{j,t}^k$ is $\mathbb{E}\{\xi_j\} = KT_s\mathbb{E}\{\xi_{j,t}^k\}/KT_s = \mathbb{E}\{\xi_{j,t}^k\}$, and (A.1) can be approximated as $W_{\mathrm{th}}^m \approx \frac{KB_{\mathrm{seg}}/T_{\mathrm{seg}}}{\Delta\log_2\left(1+\frac{\bar{d}^\beta N_{\mathrm{t}}}{\sigma^2}P_{\max}\right)}$. Since d_j^k, X_j^k and $\|\mathbf{h}_{j,t}^k\|^2$ are i.i.d. among MSs, ξ_j are i.i.d. among MSs, too. Thus, when K grows, the variance of ξ_j decreases. This means that we can use $\mathbb{E}\{\xi_j\}$ to approximate ξ_j. By further using the average number of MS $\lambda_m\overline{T}_m\Delta$ to approximate K, from (A.1) we obtain (7).

References

[1] L. Nie, D. Jiang, S. Yu, and H. Song, "Network traffic prediction based on deep belief network in wireless mesh backbone networks," in *IEEE WCNC*, 2017.

[2] A. Nadembega, A. Hafid, and T. Taleb, "A destination and mobility path prediction scheme for mobile networks," *IEEE Trans. Vehi. Tech.*, vol. 64, no. 6, pp. 2577–2590, 2015.

[3] M. Kasparick, R. L. G. Cavalcante, S. Valentin, S. Staczak, and M. Yukawa, "Kernel-based adaptive online reconstruction of coverage maps with side information," *IEEE Trans. on Vehi. Tech.*, vol. 65, no. 7, pp. 5461–5473, 2016.

[4] C. Yao, C. Yang, and I. Chih-Lin, "Data-driven resource allocation with traffic load prediction," *Journal of Commun. & Info. Net.*, vol. 2, no. 1, pp. 52–65, 2017.

[5] H. Abou-zeid, H. Hassanein, and S. Valentin, "Optimal predictive resource allocation: Exploiting mobility patterns and radio maps," in *IEEE GLOBECOM*, 2013.

[6] L. Zheng and G. D. Veciana, "Optimizing stored video delivery for mobile networks: The value of knowing the future," in *IEEE INFOCOM*, 2013.

[7] R. Atawia, H. Abou-Zeid, H. S. Hassanein, and A. Noureldin, "Joint chance-constrained predictive resource allocation for energy-efficient video streaming," *IEEE Journal on Sel. Areas in Commun.*, vol. 34, no. 5, pp. 1389–1404, 2016.

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

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

[10] C. Yao, C. Yang, and Z. Xiong, "Energy-saving predictive resource allocation planning and allocation," *IEEE Trans. Commun.*, vol. 64, no. 12, pp. 5078–5095, Dec. 2016.

[11] J. K. Lee and J. C. Hou, "Modeling steady-state and transient behaviors of user mobility:formulation, analysis, and application," in *ACM MAHNC*, 2006.

[12] E. R. Ziegel, "Probability and statistics for engineering and the sciences," *Technometrics*, vol. 101, no. 4, pp. 497–498, 2004.

[13] E. U. T. R. Access, "Further advancements for e-utra physical layer aspects," *3GPP Technical Specification TR*, vol. 36, p. V2, 2010.

[14] D. Su and C. Yang, "User-centric downlink cooperative transmission with orthogonal beamforming based limited feedback," *IEEE Trans. on Commu.*, vol. 63, no. 8, pp. 2996–3007, 2015.

[15] E. Zeidler, *Oxford users' guide to mathematics*. Oxford University Press, 2004.