SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

December 2023, Vol. 66 222302:1–222302:15
https://doi.org/10.1007/s11432-022-3769-9

When the gain of predictive resource allocation for content delivery is large?

Chenzuo ZHANG, Jia GUO* & Chenyang YANG

School of Electronic and Information Engineering, Beihang University, Beijing 100191, *China*

Received 14 November 2022/Revised 10 March 2023/Accepted 15 May 2023/Published online 30 October 2023

Abstract By predicting future information such as data rate with sensory wireless data, radio resources can be pre-allocated for content delivery. Such an integrated sensing and communications technique can help improve network performance and user experience. To justify the cost paid for predicting future information, it is important to understand in which scenarios predictive resource allocation yields a large gain over the non-predictive counterpart. In this paper, we strive to identify the key factors that affect the gain of predictive resource allocation by deriving the closed-form expression of the gain. We are concerned with minimizing the transmission time required for content delivery such as file downloading to users with an expected deadline, where the resources of base stations are shared with real-time services. Then, the performance gain is measured by the difference of the average time required by predictive and non-predictive resource allocation. Inspired by the solution of the optimization problem, we resort to the theory of order statistics for deriving the performance gain. We find that the gain depends on the statistics (i.e., mean value and standard deviation) of the user's average data rates in the prediction window. Then, we separately analyze how the statistics of the bandwidth available for content delivery and user mobility affect the gain. We use simulation with a real dataset of traffic load to validate the analysis and quantify the impact of the key factors. Our results show that predictive resource allocation can reduce the transmission time even for non-moving users. The performance gain is high when the network is busy or the cell radius is large.

Keywords predictive resource allocation, performance analysis, average data rates, residual bandwidth, user mobility

Citation Zhang C Z, Guo J, Yang C Y. When the gain of predictive resource allocation for content delivery is large? Sci China Inf Sci, 2023, 66(12): 222302, https://doi.org/10.1007/s11432-022-3769-9

1 Introduction

Recently, integrated sensing and communications is emerging as a research focus of next-generation wireless networks (such as beyond 5G and 6G) [1, 2], which aims to improve communication performance with the aid of sensory data or to use the communication network to implement sensing tasks. Assisted by the sensory data such as the location, speed, and heading direction of users, communication performance can be improved by optimizing beamforming or resource allocation [1, 3, 4]. As one of the techniques of sensing-assisted communications, predictive resource allocation (PRA) is proposed in recent years, where future resource allocation is optimized by predicting future information based on the observed past data sequence. By exploiting the radio resources in the wireless network that have long been observed as under-utilized, PRA can significantly boost network performance such as spectral/energy efficiency with information predicted from sensory data.

The radio resource under-utilization issue originates partially from the conservative and reactive design for wireless systems, since user behaviors have long been regarded as random and non-predictable. Thanks to the introduction of artificial intelligence to wireless edge [5], the behavior-related information such as traffic load and mobility patterns has been recognized as predictable with machine learning recently. In [6–8], the traffic load of wireless networks was predicted with sensory data in different resolutions, from which the future resource usage status of a network can be derived. In [9], the high-resolution

* Corresponding author (email: guojia@buaa.edu.cn)

© Science China Press 2023

trajectory of a mobile user was predicted by deep learning, from which the future average channel gains of a user can be obtained with the help of a pre-stored radio map [10, 11].

Aided by the predicted information, radio resources can be assigned in advance for adapting to future network dynamics. This provides a possible way to fully use radio resources for boosting network performance and quality of service (QoS) of users. In particular, PRA has been shown to provide remarkable gains in terms of reducing energy consumption [12–14], increasing network throughput [10, 15, 16], improving energy efficiency [17] and user experience [18] over the non-predictive counterparts. While the value of knowing the future has been evaluated by simulations and experiments with a variety of optimizations [10, 12–15, 17–19], the future information cannot be predicted without expense. To make predictions, a central processor (CP) needs to gather the sensory data from users and train predictors with the data, which consumes computation and storage resources. Since the performance gain brought by information prediction and data sensing depends on the system settings and application scenarios [12], it is important to identify the key impacting factors on the performance gain and how these factors affect the gain, which can provide guidance for implementing PRA in practice.

1.1 Related work

PRA, also referred to as anticipatory or proactive resource management in [20], has been optimized for either real-time (RT) service [21–25] or non-real-time (NRT) service [10, 12–16, 18], or the hybrid of both services [17]. Since we are concerned with PRA for NRT services, the existing studies of optimizing PRA for RT services such as phone calls, video conferences, and ultra-low latency communications are not introduced in detail due to limited space.

For NRT services such as video on demand or file downloading (i.e., content delivery), by predicting future information such as average channel gains or average data rates in each frame of a prediction window with a minute-level horizon, a PRA plan can be made to determine how much radio resource is assigned to each user in each frame of the window. The PRA plan can be optimized to improve the QoS of each user and the performance of a network. For example, it was shown in [13, 14] that by minimizing the transmission time to deliver contents for a given number of NRT requests, the energy consumption of the base stations (BSs) can be reduced. It was demonstrated in [15] that network throughput can be improved by PRA without consuming more power or bandwidth.

To show the potential of leveraging the future information for delay-tolerant services, existing studies optimize the PRA policies towards various objectives under the assumption of knowing perfect future data rates [10, 13, 18] or average channel gains [12]. Considering that prediction is never perfect, a robust optimization method was proposed in [14] by modeling the prediction errors as random variables, and the impact of prediction errors on the performance of PRA policy was evaluated via simulations in [12] and analyzed in [15]. Recently, machine learning methods were employed in [8, 15, 19] to predict future information with real-measured data in cellular networks. Simulation results in [15] showed that the proposed policy with predicted information performs closely to the corresponding optimal solution with perfect predictions, both providing large gains over a non-predictive counterpart in supporting high throughput for content delivery.

All prior studies in the literature evaluate the performance gain of PRA policy for NRT service either by simulations or by experiments [10, 12–18]. The evaluation results of these studies show that the performance gain of the PRA policy is higher when the prediction window is longer, the number of NRT users is larger, and the requested file is a larger size. However, to the best of our knowledge, no existing studies identified the key factors impacting the gain through theoretical analyses, which are important to justify the cost paid for making the prediction and sensing. This is because the closed-form solution of a PRA problem is hard to obtain. Moreover, no prior studies considered non-mobile users, even though the majority of content delivery services are requested by indoor users.

1.2 Motivation and contributions

In this paper, we strive to answer the following question: what are the impacting factors for the large gain of PRA in terms of improving resource usage efficiency for content delivery?

To this end, we take the transmission time required for delivering a content (i.e., a file) as the performance metric, and analyze the performance gain characterized by the difference between the required

transmission time of predictive and non-predictive resource allocation (non-PRA). The shortened transmission time can be translated to the reduced energy consumption [14] and the increased network throughput [15].

We consider a cellular network, where each BS serves both NRT and RT traffics. Due to stringent delay requirements of RT services, the NRT users are served by the residual resources after reserving the resource for guaranteeing the QoS of RT users. We are concerned with the PRA for the users requesting NRT services, referred to as user equipments (UEs) in the sequel.

Since the performance of policy with predicted information is close to the optimal policy with perfect future information [15], we assume that future information is perfect. To gain useful insight, we consider a practical scenario in the derivations where the requests arrival rate of content delivery is low, such that only one UE initiates a request in each cell in the prediction window [26][1)]. We then show the impact of multiple UEs in each cell via simulation.

The major contributions are summarized as follows.

• We resort to the theory of order statistics [27] for deriving the closed-form expression of the performance gain of the PRA policy over the non-PRA policy, inspired by the solution of the PRA problem. We analyze several core impacting factors on the gain, including the relative randomness of residual bandwidth, cell radius, and user mobility. As far as the authors know, this is the first work analyzing the gain of the PRA policy for NRT service theoretically.

• We find that the PRA policy yields performance gain even for non-mobile UEs. The upper bound of the gain with the infinite length of the prediction window is inversely proportional to a ζ-coefficient, which reflects the variance of residual bandwidth relative to its mean value. From a real dataset of traffic load, we observe that this coefficient is small when the network is busy with RT traffic.

• For mobile UEs, we find that the performance gain increases slightly with the number of cells that the UEs traverse within the prediction window. Our analytical and simulation results show that the upper bound of the performance gain is high when the network is busy (i.e., the RT traffic load is heavy, the file size requested by each UE is large, and the number of UEs is large) and the cell radius is large.

The rest of this paper is organized as follows. In Section 2, we introduce the system model, the optimal PRA policy, and the non-PRA policy. In Section 3, we analyze the impact of the mean value and standard deviation of the average data rate of each user on the performance gain. We proceed to analyze the impact of the statistics of the residual bandwidth and the mobility of NRT users on the gain, respectively. Simulation results and conclusion are given in Sections 4 and 5, respectively.

Notations. $\|\cdot\|^2$ denotes two-norm, $\mathrm{E}\{\cdot\}$ denotes expectation, $\mathbb{D}\{\cdot\}$ denotes variance, $\lceil\cdot\rceil$ denotes the ceiling function.

2 System model and PRA

Consider a cellular network, where a CP is connected with M BSs. Each BS equipped with N_{tx} antennas serves both RT traffic and NRT traffic with bandwidth $W_{\max}$ and transmit power $P_{\max}$. Since the RT traffic has a lower delay tolerance and higher priority than the NRT traffic, a fraction of the bandwidth should be reserved for RT users to ensure their QoS. Then, the residual bandwidth is employed to serve NRT users (i.e., UEs). Each UE requests to download a file, which should be conveyed before an expected deadline to guarantee its QoS.

Time is discretized into frames each with a duration of Δ, which is defined according to the coherence time of large-scale channel gains (i.e., average channel gains). Each frame includes N_s time slots each with a duration of unit time, which is defined according to the coherence time of small-scale channels (i.e., instantaneous channels). In practice, the timescales of a frame and a time slot are respectively in the second level and millisecond level for mobile users, and hence the value of N_s is large.

Assume that each UE is associated with the BS with the highest large-scale channel gain. To avoid multi-user interference, time division multiple access is applied, i.e., each BS serves only one UE with all the residual bandwidth and residual transmit power in each time slot, and serves multiple UEs associated with it in different time slots. For the tractability of analysis, we treat inter-cell interference as noise as in most of the previous studies for PRA [10, 12, 14, 15, 17, 18]. Then, maximal ratio transmission is the optimal beamforming.

1) According to the analysis in [26] for a real dataset gathered in a campus covering eight cells each with a radius of 250 m, the maximal request arrival rate for video delivery in each cell in each minute of a week is 0.625.

Since RT requests arrive randomly and RT users may move, the resource occupied by RT traffic and hence the residual resources are time-varying. Denote $W_{j,t}$ and $p_{j,t}$ respectively as the instantaneous residual bandwidth and residual transmit power available for NRT traffic in the tth time slot of the jth frame. The residual bandwidth is the system bandwidth $W_{\max}$ minus the reserved bandwidth for RT traffic. The residual transmit power can be assumed as proportional to the residual bandwidth [12], i.e., $p_{j,t} = W_{j,t}P_{\max}/W_{\max}$.

When the kth UE, denoted as UE_k, is associated with a BS in the jth frame and served by the BS with all its residual resources, the instantaneous data rate of UE_k in the tth time slot of the jth frame can be expressed as $R_{j,t}^k = W_{j,t}\log_2(1+\frac{\alpha_j^k\|\boldsymbol{h}_{j,t}^k\|^2}{N_0W_{j,t}}p_{j,t}) = W_{j,t}\log_2(1+\frac{\alpha_j^k\|\boldsymbol{h}_{j,t}^k\|^2}{\sigma^2}P_{\max})$, where $\boldsymbol{h}_{j,t}^k \in \mathbb{C}^{N_{\mathrm{tx}}\times 1}$ is the small-scale channel vector with independently and identically distributed (i.i.d.) elements, $\alpha_j^k = (d_j^k)^\beta 10^{X_j^k/10}$ is the large-scale channel gain in the jth frame, d_j^k is the distance between UE_k and its associated BS in the jth frame, β is the path-loss exponent, X_j^k reflects shadowing, N_0 is the noise power spectrum density, and $\sigma_0^2 = N_0W_{\max}$ is the noise power.

The time-average data rate of UE_k in the jth frame (called the average rate for short in the sequel) is

$$R_j^k = \frac{1}{N_s}\sum_{t=1}^{N_s} R_{j,t}^k = \frac{1}{N_s}\sum_{t=1}^{N_s} W_{j,t}\log_2\left(1+\frac{\alpha_j^k\|\boldsymbol{h}_{j,t}^k\|^2}{\sigma_0^2}P_{\max}\right) \overset{\text{(a)}}{\approx} W_j\log_2\left(\frac{\alpha_j^kN_{\mathrm{tx}}}{\sigma^2}P_{\max}\right) \triangleq W_j\gamma_j^k, \quad (1)$$

where N_s is the number of time slots in a frame, $W_j \triangleq \frac{1}{N_s}\sum_{t=1}^{N_s} W_{j,t}$, and $\gamma_j^k \triangleq \log_2(\frac{\alpha_j^kN_{\mathrm{tx}}}{\sigma^2}P_{\max})$. (a) comes from two approximations $\|\boldsymbol{h}_{j,t}^k\|^2 \approx N_{\mathrm{tx}}$ and $1+\frac{\alpha_j^kN_{\mathrm{tx}}}{\sigma^2}P_{\max} \approx \frac{\alpha_j^kN_{\mathrm{tx}}}{\sigma^2}P_{\max}$, which are accurate when N_{tx} is large and when the average signal-to-noise ratio (SNR) $\frac{\alpha_j^kN_{\mathrm{tx}}}{\sigma^2}P_{\max}$ is high, respectively. When $W_{j,t}$ is i.i.d. among time slots and N_s is large, $W_j \approx \mathrm{E}\{W_{j,t}\}$.

Assume that the CP can gather data from the BSs and UEs to predict future information. When a UE initializes a request for a file that should be downloaded before an expected deadline, the CP establishes a prediction window for the UE, whose duration is assumed to be the same as the duration before the deadline for notational simplicity. The prediction window is with the length of T_p and contains N_p frames, i.e., $T_p = N_p\Delta$. T_p is usually in the timescale of tens of seconds or several minutes [10, 12, 14, 15, 17–19].

At the beginning of the prediction window, the CP predicts the average rate in each frame of the window, i.e., $R_1^k, \ldots, R_{N_p}^k$, $k = 1, \ldots, K$, with the method proposed in [15]. Specifically, we can first predict the trajectory of each UE (say UE_k) as well as the RT traffic loads (i.e., number of RT requests) of its associated BSs in each frame (say the jth frame) of the prediction window, and convert them to the large scale fading α_j^k and the residual bandwidth W_j, respectively. Then, the average rate can be obtained in (1). After prediction, the CP makes the resource allocation plan for conveying the files that the UEs requested, and informs the plan to the BSs that will serve the UEs. Since the performance loss of PRA brought by prediction errors is small [15], in order to focus on analyzing the potential of PRA policy, we assume that the future average rates in the prediction window are perfectly predicted.

The resource allocation plan is optimized to determine the percentage of time slots assigned to each UE in each frame. Since a frame contains hundreds of time slots, the percentage can be regarded as continuous. Denote $\boldsymbol{s}^k = [s_1^k, \ldots, s_{N_p}^k]^{\mathrm{T}}$ as the plan for UE_k, where $s_j^k \in [0,1]$ is the percentage of the time slots assigned to the UE in the jth frame. The plan of every UE can be optimized to minimize the total transmission time for serving the K UEs in the network from the following problem:

$$\min_{\boldsymbol{s}^1,\ldots,\boldsymbol{s}^K} \sum_{j=1}^{N_p}\sum_{k=1}^{K} s_j^k, \quad (2a)$$

$$\text{s.t.} \sum_{j=1}^{N_p} s_j^kR_j^k\Delta \geqslant B,\ k = 1, \ldots, K, \quad (2b)$$

$$\sum_{k\in\mathcal{K}_{j,i}} s_j^k \leqslant 1,\ i = 1, \ldots, M,\ j = 1, \ldots, N_p, \quad (2c)$$

where B (in bits) is the size of the file, Eq. (2b) is the QoS requirement of each UE (i.e., the amount of data able to be transmitted in the N_p frames should exceed the file size), Eq. (2c) ensures the transmission time assigned to all UEs of each cell in each frame not exceeding the frame duration, and $\mathcal{K}_{j,i}$ is the set of UEs located in the ith cell in the jth frame.

From this linear programming problem, the plan for every UE can be obtained by numerical algorithms. After being informed by the CP, the BSs along the trajectory of each UE can download the files to the UEs according to the plan. If multiple UEs in a cell need to be served in a frame according to the plan, then the BS transmits to the UE with the maximal instantaneous data rate in each time slot of the frame using the maximal-ratio transmission. This is the PRA policy.

With traditional non-PRA policy, the BSs deliver files to the UEs with the best effort. Specifically, each BS selects the UE with the largest instantaneous channel gain in each time slot for transmission from all associated UEs, until the requested file of every UE is completely downloaded.

When the NRT requests arrival rate is low such that there is only one UE in a cell in the prediction window, the PRA policy with the optimal plan obtained from the problem in (2) always transmits to the UE in the frames with the largest average rates within the window until the file is completely downloaded, which exploits a kind of "macro-time diversity". When exploiting conventional "micro" time diversity in non-PRA policy, a BS transmits in the time slots with the largest small-scale channel gains of a mobile user. By contrast, when exploiting the "macro-time diversity", the frames with high average rates from the prediction window are selected for transmission. Since the high average rate in a frame comes from either large residual bandwidth or large average channel gain or both, it is unnecessary to transmit to a UE under its best channel conditions.

Remark 1. The minimal total transmission time achieved by the PRA policy only depends on the values of the average rates in the assigned frames, independent of the indexes of the frames with large average rates.

Remark 2. To implement the optimal plan obtained from problem (2), the average rates within the prediction window are first sorted from the largest to the smallest values, and then the frames with the largest average rates are selected for transmission until the file is completely conveyed. This suggests that the performance of the PRA policy depends on the statistic of the sorted average rates, which can be analyzed by the theory of order statistics [28].

3 Gain of PRA

In this section, we first briefly introduce the preliminary of order statistics. Then, we derive the performance gain of the PRA policy over the non-PRA policy and its upper bound. Finally, we analyze the impact of several key factors on the upper bound respectively for non-moving and mobile UE.

3.1 Preliminary of order statistics [28]

The theory of order statistics deals with the properties of ordered random variables and the functions involving these variables. If random variables $X_1, \ldots, X_n$ are arranged in the order of magnitude as $X_{(1)} \leqslant \cdots \leqslant X_{(n)}$, then $X_{(i)}$ is called the ith order statistic, $i = 1, \ldots, n$.

Suppose that $X_1, \ldots, X_n$ are i.i.d. random variables with cumulative distribution function (CDF) denoted as $F(x)$. Let $F_{(i)}(x)$ denote the CDF of the ith order statistic $X_{(i)}$. Then, the CDF of the largest order statistic $X_{(n)}$ is given by $F_{(n)}(x) = P\{X_{(n)} \leqslant x\} = P\{X_i \leqslant x, \forall i = 1, \ldots, n\} = F^n(x)$, where $P\{A\}$ denotes the probability of a random event A, and $F^n(x)$ is the nth power of $F(x)$.

The general result of the CDF for the order statistic is $F_{(i)}(x) = P\{X_{(i)} \leqslant x\} = \sum_{r=i}^{n} \frac{r!}{n!(n-r)!} F^r(x)[1 - F(x)]^{(n-r)}$. With $F_{(i)}(x)$, the probability density function (PDF) and the mean value of the order statistic can be derived. The results are given in [28] and are not provided due to limited space.

3.2 Performance gain of PRA and its upper bound

The performance gain is measured by the difference of the required minimal total transmission time of the two policies, which can be expressed as $\Delta t = t_{\mathrm{NPRA}} - t_{\mathrm{PRA}}$, where t_{PRA} and t_{NPRA} are the average minimal time to transmit the requested file by the PRA and non-PRA policies, respectively.

When there is one UE in each cell, the performance gain is shown in Proposition 1. For notational simplicity, we remove the superscript of the UE index in the sequel.

Proposition 1. If the average rates are independent among the frames within the prediction window and follow a uniform distribution with mean value $\overline{R}$ and standard deviation σ_R, then the minimal total

transmission time of the PRA policy is $t_{\mathrm{PRA}} \approx \frac{(T_p+\Delta)B}{(T_p+\Delta)\overline{R}+\sqrt{3}T_p\sigma_R}$, and the performance gain is

$$\Delta t \approx \frac{\sqrt{3}T_pB}{(T_p+\Delta)\frac{\overline{R}^2}{\sigma_R}+\sqrt{3}T_p\overline{R}}, \tag{3}$$

where the approximations are accurate when $\overline{R}$ is large, σ_R or B is small.

Proof. See Appendix A.

From Proposition 1, we can obtain a corollary as follows.

Corollary 1. The derivative of Δt in (3) over T_p is $\frac{\mathrm{d}(\Delta t)}{\mathrm{d}T_p} \approx \frac{\sqrt{3}B\sigma_R}{((T_p+\Delta)\overline{R}+\sqrt{3}T_p\sigma_R)^2} > 0$.

This indicates that the performance gain is larger for a longer prediction window, which agrees with the simulation results in [12, 18, 19]. Yet the growth rate of the performance gain becomes slower by increasing T_p, as empirically observed in [12], since the value of $\frac{\mathrm{d}(\Delta t)}{\mathrm{d}T_p}$ decreases with T_p. When $T_p \to \infty$, the gain in (3) approaches an upper bound as

$$\lim_{T_p\to\infty} \Delta t \approx \frac{\sqrt{3}B}{\frac{\overline{R}^2}{\sigma_R}+\sqrt{3}\,\overline{R}}. \tag{4}$$

We define $\Delta t^{\mathrm{UB}} \triangleq \frac{\sqrt{3}B}{\frac{\overline{R}^2}{\sigma_R}+\sqrt{3}\,\overline{R}}$, which is higher when $\overline{R}$ is smaller or σ_R larger. Nonetheless, from the following derivative

$$\frac{\mathrm{d}(\Delta t^{\mathrm{UB}})}{\mathrm{d}\sigma_R} = \frac{\sqrt{3}B\overline{R}^2}{(\overline{R}^2+\sqrt{3}\sigma_R\overline{R})^2}, \tag{5}$$

we know that the growth rate of the upper bound becomes slower as σ_R increases.

As shown in (1), the average rate of a UE in a frame depends on the residual bandwidth and large-scale channel gain. The residual bandwidth decreases with the RT traffic load as analyzed in [15], and the large-scale channel gain depends on the location of the UE in the frame. If the BSs are busy with RT traffic or the UE is not in good channel conditions in the prediction window, the value of $\overline{R}$ is small. If the RT traffic load changes significantly or the large-scale channel gains of the UE change considerably in the window, the value of σ_R is large. By summarizing the proposition and the corollary, we can obtain the following observation.

Observation 1 (dynamic average rates of each UE). The transmission time required by the PRA policy is short when $\overline{R}$ or σ_R is large. Both the increasing speed of the performance gain with T_p and the upper bound of the gain are larger when $\overline{R}$ is smaller, B or σ_R is larger.

The changes in the large-scale channel gains depend on the mobility of the UE and the radius of a cell. In the following, we separately analyze the impact of the statistics of residual bandwidth and user mobility on the upper bound of the performance gain.

3.3 Impact of the statistics of residual bandwidth

To analyze the performance gain solely from the fluctuation of the residual bandwidth, in this subsection we consider the scenario where the UE does not move. Then, the average SNR in every frame is a constant within the prediction window and hence the subscript j in γ_j is removed. The average rates only change with the residual bandwidth of a BS, which is time-varying due to the random arrival of RT requests.

Proposition 2. If the residual bandwidth of a BS is independent among the frames in the prediction window and follows a uniform distribution with mean value $\overline{W}$ and standard deviation σ_W, then the upper bound of the performance gain can be approximated as

$$\lim_{T_p\to\infty} \Delta t \approx \frac{\sqrt{3}B}{\frac{\overline{W}^2\gamma}{\sigma_W}+\sqrt{3}\,\overline{W}\gamma} \triangleq \Delta t_s^{\mathrm{UB}}, \tag{6}$$

which is accurate when the average SNR is high, $\overline{W}$ is large, σ_W or B is small.

Proof. See Appendix B.

To gain useful insight, we further approximate the upper bound in the following corollary.

Corollary 2. If $\overline{W} \gg \sigma_W$, the upper bound of the performance gain in (6) can be approximated as $\Delta t_s^{\mathrm{UB}} \approx \frac{\sqrt{3}B}{\frac{\overline{W}^2\gamma}{\sigma_W}} = \frac{\sqrt{3}B}{\zeta\gamma}$, where $\zeta \triangleq \frac{\overline{W}^2}{\sigma_W}$.

The parameter ζ is referred to as ζ-coefficient in the sequel, which reflects the fluctuation of residual bandwidth relative to the square of its mean value.

Remark 3. ζ-coefficient is similar to but with different expression from the amount of fading, which is defined as $[\mathrm{E}\{h\}]^2/\mathbb{D}\{h\}$ for wireless channel h [29]. The notion of the amount of fading has been used as the measure for the severity of channel fading, and hence has also been used as a simple measure for the performance of channel diversity (i.e., a kind of "micro" diversity). Different from the amount of fading, the mean value $\overline{W}$ dominates the magnitude of the ζ-coefficient.

Since the residual bandwidth decreases with the RT traffic load as analyzed in [15], we can obtain Observation 2.

Observation 2 (dynamic RT traffic loads). The PRA policy provides a large gain for a non-mobile UE when the relative randomness of RT traffic load is large (i.e., the ζ-coefficient is small).

If the value of $\overline{W}$ is fixed, the observation indicates that the PRA policy provides a large gain when the residual bandwidth fluctuates drastically within the prediction window (i.e., σ_W is large), which agrees to intuition. This is because in this scenario the values of the residual bandwidth in the selected frames by the PRA policy are larger, which leads to the reduction of the transmission time.

If the value of σ_W is fixed, however, the observation implies that the PRA policy provides a large gain when the network is busy with RT traffic (i.e., $\overline{W}$ is small), which may be counter-intuitive. At first glance, the gain from serving the UE in a predictive manner should come from exploiting the excess resource in the network after serving the RT traffic with higher priority and thus should be large when the residual resource is abundant. Nonetheless, the gain is not high in a non-busy network because the transmission time required by the non-PRA policy for delivering a file is also short. In a busy network, the performance gain of the PRA policy comes from the "macro-time diversity", i.e., by selecting the frames with the largest values of W_j for transmission.

3.4 Impact of user mobility and cell radius

When a UE moves, the average rates in the prediction window depend on the locations in the trajectory of the UE. By using the PRA policy, the BSs can deliver the requested file to the UE with less time by choosing the frames with better channels for transmission.

To analyze the performance gain from time-varying large-scale channel gains, we consider the scenario where the UE moves, while the residual bandwidth does not change within the prediction window (i.e., $W = W_j$). Then, the average rates only change with the large-scale channel gains.

To obtain useful insight, we further assume that the UE moves across m cells with a constant speed along a road, and do not consider shadowing in this subsection. Then, the time-varying pattern of large-scale channel gains is the same among cells. The UE needs $N_c = \lceil \frac{D_b}{d_f} \rceil$ frames to traverse a cell, where d_f is the distance that the UE travels in a frame and D_b is the diameter of the cell. In each cell, the average rate achieves the maximal value $R_{\max}$ when the UE moves to the location with minimum distance from the BS, and achieves the minimal value $R_{\min}$ when the UE is at the cell-edge (i.e., $d_j = D_b/2$).

Since the minimal total transmission time required by the PRA policy does not rely on the ordering of the frames in the prediction window as mentioned in Remark 1, randomizing the frames in the window does not change the minimal transmission time. Therefore, it is reasonable to assume that the average rates are uniformly distributed in $[R_{\min}, R_{\max}]$.

Proposition 3. If the average rates of a UE in each cell follow uniform distribution within $[R_{\min}, R_{\max}]$, then the performance gain can be approximated as

$$\Delta t \approx \frac{B}{\overline{R}} - \frac{2B(N_c+1)}{W(\kappa_{\min}+(2N_c+1)\kappa_{\max})} - \frac{4B^2(\kappa_{\max}-\kappa_{\min})(N_c+1)^2N_c}{T_pW^2(\kappa_{\min}+(2N_c+1)\kappa_{\max})^3}, \tag{7}$$

where $\kappa_{\min} = \log_2(D_b^\beta \frac{N_{\mathrm{tx}}P_{\max}}{2^\beta\sigma^2})$, $\kappa_{\max} = \log_2(d_{\min}^\beta \frac{N_{\mathrm{tx}}P_{\max}}{\sigma^2})$, and $d_{\min}$ is the minimum distance between the UE and the BS. The approximation is accurate when the average SNR is high, B is small or W is large.

Proof. See Appendix C.

Considering that $R_{\min} = W\kappa_{\min}$ and $R_{\max} = W\kappa_{\max}$, it is not hard to obtain the following corollary.

Corollary 3. If $N_c \gg 1$, the upper bound of the performance gain in (7) can be approximated as

$$\lim_{N_c\to\infty} \Delta t \approx \frac{B}{W}\left(\frac{2}{\kappa_{\max}+\kappa_{\min}} - \frac{1}{\kappa_{\max}}\right) \triangleq \Delta t_{\mathrm{m}}^{\mathrm{UB}}.$$

The corollary suggests that the upper bound is higher when the UE moves across a cell with a larger radius (i.e., $\kappa_{\min}$ is smaller).

It is worth noting that m is not in (7). When serving a mobile UE with the PRA policy and without considering shadowing, we can obtain Observation 3.

Observation 3 (dynamic UE locations). The performance gain does not grow with the number of cells that the UE traverses within the prediction window. The gain is large when the cell radius is large.

The observation that the gain is not larger for a UE moving across more cells within the prediction window may contradict our intuition, since the BSs seem to have more chances to deliver data to the UE with higher average rates. However, for a given duration of the prediction window, a UE moving with a higher speed can traverse more cells, while the duration it experiences good channels in each cell is shorter. Then, the data amount that the BSs deliver to the UE in "good" frames reduces. Moreover, the number and magnitudes of the good channels in each cell are identical for a UE with constant speed, if shadowing is not considered. In this scenario, a UE does not experience higher average rates by traversing more cells, and the gain depends on the dynamic range of the large-scale channels in a single cell.

4 Simulation and numerical results

In this section, we first compute the ζ-coefficients with a real dataset of traffic load. Then, we evaluate the accuracy of the approximations and validate the previous analyses by comparing the numerical and simulation results. Finally, we demonstrate the impact of several factors on the performance gain via numerical results.

Consider a cellular network of six cells, where $D_b = 500$ m, $N_{\mathrm{tx}} = 8$, $P_{\max} = 40$ W, and $W_{\max} = 20$ MHz. The cell-edge SNR is set as 5 dB, where the inter-cell interference is implicitly reflected. The path loss model is $36.8 + 36.7\log_{10}(d)$, where d is the BS-UE distance in meter. The standard deviation and decorrelation distance of shadowing are 8 dB and 50 m, respectively [30]. The BSs are deployed alongside a road, and the minimal distance between each BS and the road is $d_{\min} = 50$ m. The UEs are located on the road. Each frame is with duration of $\Delta = 1$ s, and each time slot is with duration of 10 ms, i.e., each frame contains $N_s = 100$ time slots. The prediction window is with duration $T_p = 60$ s [14]. The file size B is set as 15, 30, or 45 MB (i.e., Mbytes), which corresponds to a standard definition video, a high-definition video, or a full high-definition video with a one-minute duration [31]. This system setup is used in the following unless otherwise specified.

All the simulations are implemented on Matlab 2018a.

4.1 ζ-coefficient in a real dataset

To observe the magnitude of the ζ-coefficient, we consider a real dataset. The dataset is collected in a campus covered by eight cells each with a radius of 250 m, composed of the traffic volume (i.e., requested data amount) for video-on-demand service from five video websites in each second of seven days.

Since the data of RT traffic is unavailable, we synthesize the RT traffic load and the residual bandwidth of each BS from this dataset using the method in [15]. To this end, the packet arrival process of each RT user is set as Poisson with an average arrival rate of 1000 packets/s, the packet size follows an exponential distribution with a mean value of 4 kbits, and the RT users are uniformly distributed in each cell.

In Figure 1, we show the synthesized data of RT traffic volume and the residual bandwidth, from which we can observe the peak time and off-peak time each day.

We use relative residual bandwidth $\overline{W}/W_{\max}$ to measure how busy a BS is with RT traffic in the duration of a prediction window. A small value of $\overline{W}/W_{\max}$ indicates a busy network with RT traffic.

In Table 1, we provide the empirically obtained values of $\overline{W}/W_{\max}$, $\overline{W}$ and σ_W, as well as the ζ-coefficients computed with $\zeta = \overline{W}^2/\sigma_W$ in different time of a day. The values of $\overline{W}$ and σ_W are computed with the residual bandwidth in every second within the one-minute-long prediction windows of the same time in the seven days.

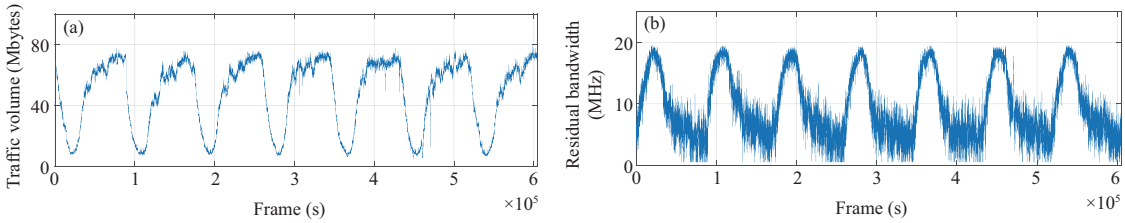

Figure 1 (Color online) (a) RT traffic volume; (b) residual bandwidth in each second.

Table 1 ζ-coefficients in different time of a day, $T_p = 60$ s

Time of the day	23:00	21:30	11:30	8:30	8:00
$\overline{W}/W_{\max}$ (%)	24	30	50	70	80
$\overline{W}$ (MHz)	4.8	6	10	14	16
σ_W (MHz)	2.4	2.1	1.78	1.3	1.05
ζ ($\times 10^7$)	0.9	1.7	5.6	15.1	24.4

We can see from the table that the network is busiest around 23:00 and is idlest around 8:00. When the network is busy, the average residual bandwidth is low and its standard deviation is high (hence the ζ-coefficient is small). This comes from the fact that the number of RT users is large in the busy time of the day, and each RT user occupies a random amount of resources [15], hence the standard deviation of the residual resource is large.

4.2 Validating the analyses and impact of key factors

In what follows, we first evaluate the accuracy of the approximations with numerical results. Then, we show the impact of the assumptions and analyze the impact of key factors via numerical and simulation results in three scenarios.

In the simulation, the performance gain Δt is obtained by calculating the difference between the total transmission time required by the PRA and non-PRA policies for delivering the files and then taking the average over 500 Monte Carlo trails. In each trial, the UEs initiate requests at random locations on the road, and the small-scale fading channels are randomly generated according to Rayleigh distribution. The PRA policy is obtained by solving problem (2), and the non-PRA policy is the resource allocation mentioned in Section 2.

4.2.1 *Accuracy of approximations*

In the proof of Proposition 1, we use Taylor's formula to obtain an approximation result of N_d (i.e., the number of frames to download a file with PRA policy) from (A6) as (A7). In Figures 2(a)–(c), we respectively show the impact of the values of $\overline{R}$, σ_R and B on the approximate accuracy of the performance gain, where the curves are obtained from N_d computed with (A6) without approximation and the markers are obtained from the approximated N_d computed with (A7). The performance gain is obtained by $\Delta t = \frac{B}{\overline{R}} - N_d\Delta$. From the figures, we can see that when $B = 45$ MB, the approximated performance gain is accurate for $\overline{R} \in [8, 160]$ Mbits/s and $\sigma_R \in [0, 40]$ Mbits/s. When the values of B are smaller, the ranges of $\overline{R}$ and σ_R for accurately approximating the gain are wider than when $B = 45$ MB (not shown for conciseness).

In the proof of Proposition 2, we introduce another approximation $\overline{R} \approx \overline{W}\gamma$ in (B1), which is accurate for high SNR. In Figure 2(d), we show the impact of the SNR on the approximation accuracy for the upper bound of performance gain. From the figure, we can see that when SNR is higher than 5 dB, the approximated upper bound is accurate.

4.2.2 *Non-mobile UEs under varying residual resources*

In this scenario, three UEs do not move within the prediction window. The residual bandwidth of the BS varies among frames according to Gaussian distribution with mean value $\overline{W}$ and standard deviation

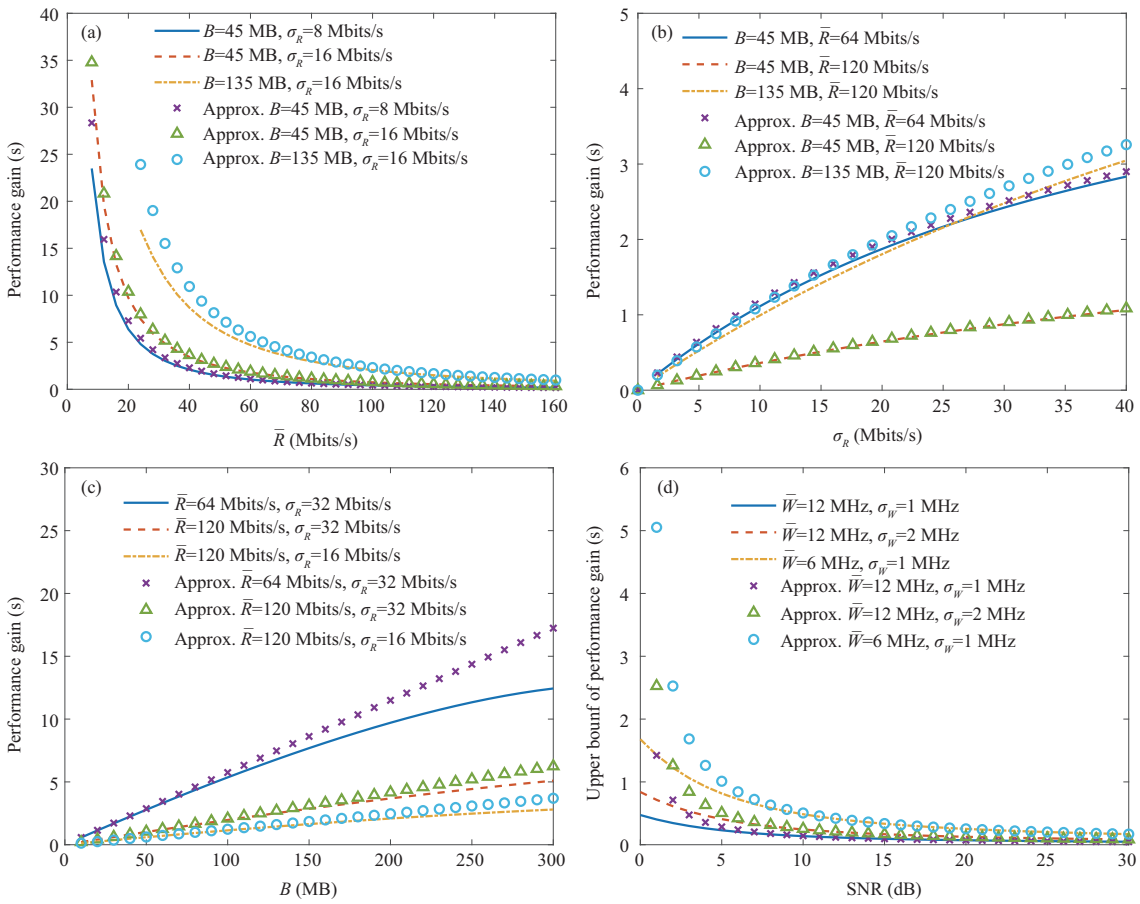

Figure 2 (Color online) Approximation accuracy versus (a) $\overline{R}$, (b) σ_R, (c) B, and (d) SNR.

σ_W[2)], where the values of $\overline{W}$ and σ_W are obtained from Table 1 according to the ζ-coefficient.

In Figure 3, we show the numerical and simulation results of the performance gain under different values of ζ. The numerical results are obtained by multiplying the result from (6) with three, since the gain is in terms of reducing the total transmission time for three UEs. The simulation results are obtained by averaging the values of Δt achieved by the PRA policy with $T_p = 60$ s, without the assumptions in Proposition 2. Specifically, the residual bandwidth no longer follows a uniform distribution, the average SNR is not always high, and there is more than one UE in each cell. Besides, the values of B are set according to H.264 standard [31], which are not artificially set as small to satisfy the assumption in the proposition. The minimum residual bandwidth in the simulation is set as 30% of the total bandwidth, because the non-PRA policy cannot complete the transmission within 60 s if $\overline{W}/W_{\max} < 30\%$. It is shown that the numerical results are very close to the simulation results, which validates Proposition 2 (and also Proposition 1, since for non-mobile UEs, the average rate follows the same distribution as the residual bandwidth). We can observe that the performance gain is larger when the ζ-coefficient is smaller. According to Table 1, this implies that the gain is large when a BS is busy.

To show the impact of the assumption $\overline{W} \gg \sigma_W$ on the result in Corollary 2, we further compare the numerical and simulation results of the upper bound under different values of ζ, where the numerical results are obtained from the approximated Δt_s^{UB} in Corollary 2. To this end, we consider $T_p = 180$ s in the simulation, and set $B = 45$ MB. The results are similar to Figure 3, hence are not provided for conciseness. According to Table 1, when the residual bandwidth is 30%, $\zeta = 1.7\times10^7$, and $\overline{W}/\sigma_W = 6/2.1 \approx 3$. In this case where the value of $\overline{W}/\sigma_W$ is the smallest, the simulated upper bound is 5.2 s, whereas the numerical result obtained from the approximated Δt_s^{UB} is 6.1 s, i.e., the maximal gap caused by the assumption is only 0.9 s.

2) By analyzing the PDF from the real dataset, we find that the residual bandwidth approximately follows Gaussian distribution. By analyzing the nonlinear autocorrelation of the traffic load recorded with one-second resolution, we find that the residual bandwidth is nearly independent among frames.

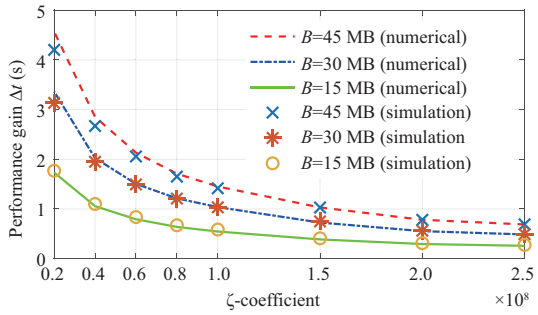

Figure 3 (Color online) Validation of the results in Propositions 1 and 2.

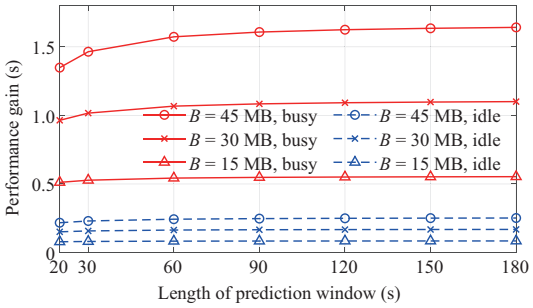

Figure 4 (Color online) Impact of T_p on the upper bound of the performance gain.

In Figure 4, we provide the numerical results of the performance gain versus T_p obtained from (3). We consider a busy network where $\zeta = 2\times10^7$ ($\overline{W}/W_{\max}$ is around 32%) and an idle network where $\zeta = 2.5\times10^8$ ($\overline{W}/W_{\max}$ is around 80%). It is shown that the upper bound of the performance gain by delivering a large file in a busy network (where $\overline{R}$ is small) is much higher than the upper bound achieved by delivering a small file. In the non-busy network, the gaps between the upper bounds for different values of B are much smaller. When $T_p = 180$ s, the performance gain can achieve 98% of the upper bound for all cases, i.e., PRA cannot provide more gain with a prediction window longer than three minutes.

4.2.3 *Moving UE under constant residual resources*

In this scenario, a UE moves across cells along the road. The residual bandwidth does not vary within the prediction window, and $B = 45$ MB. Considering that the frame duration is defined according to the coherence time of the large-scale channels and the coherence distance of shadowing is determined by the propagation environment, we fix the distance that the UE moves in a frame as $d_f = 10$ m, hence a UE needs $N_c = \lceil \frac{D_b}{d_f} \rceil = 50$ frames to traverse one cell.

In Figure 5, we show the numerical and simulation results of the performance gain when a UE moves across different numbers of cells with different speeds in a prediction window with $T_p = 60$ s. Specifically, the UE moves across one cell, two cells, and three cells with singled, doubled, and tripled speeds, respectively, and the lengths of a frame are respectively $\Delta = 1.2$, 0.6, and 0.4 s. We consider a busy network, where $\overline{W}/W_{\max} = 30\%$. The numerical results are obtained from (7). The corresponding simulation results with shadowing are shown in Figure 6.

We can see from Figure 5(a) that the performance gain is almost not affected by m, which agrees with Observation 3 (dynamic UE locations). However, the simulation results in Figure 6 show that the transmission time of the PRA policy decreases with m and the transmission time of the non-PRA policy does not change with m, and hence the performance gain increases with m. This is because when shadowing is considered, the fluctuation pattern of the average rates is not identical among cells, which does not agree with the assumptions in Proposition 3. In this case, the UE has more chances to experience good channels when it traverses more cells, and the value of σ_R is larger. Nonetheless, the increased gain is not significant. This is because the gain grows slowly with σ_R according to (5). Specifically, the value of σ_R caused by the path loss variation between the maximal and minimal average SNR (corresponding to $\kappa_{\max}$ and $\kappa_{\min}$) is about 25 dB. Compared to path loss variation, the gain brought by shadowing with the standard deviation of 8 dB is marginal.

To evaluate the performance gain under different network settings, we simulate the scenarios with different types of BSs, where a UE moves with a constant speed and the system parameters are listed in Table 2 [32]. The corresponding values of κ are provided in Table 3. It can be seen from Figure 7 that the performance gain in the UMa scenario is higher. This is because the value of $\kappa_{\min} = \log_2(D_b^\beta \frac{N_{\text{tx}} P_{\max}}{2^\beta \sigma^2})$ in the UMa scenario is smaller than the UMi-Street scenario, due to larger D_b in the UMa scenario. The results validate Observation 3 (dynamic UE locations) that the PRA policy can provide larger gain in a sparse cellular network, where each cell has a large coverage.

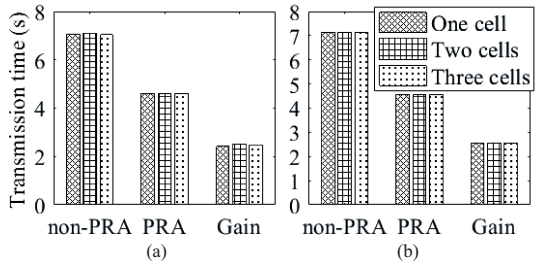

Figure 5 Impact of the number of cells a UE traversed, without shadowing. (a) Simulation results; (b) numerical results.

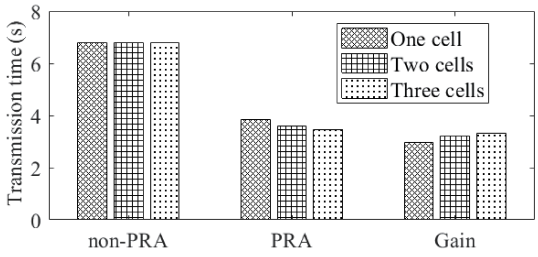

Figure 6 Impact of the number of cells a UE traversed, with shadowing.

Table 2 Parameters of different types of BSs [32]

Scenarios	$P_{\max}$ (W)	D_b (m)	N_{tx}	$W_{\max}$ (MHz)
UMi-Street	40	500	4	20
UMa	80	1000	4	20

Table 3 Values of κ in simulation, $d_{\min} = 50$ m

Scenarios	$\kappa_{\max}$	$\kappa_{\min}$	$\overline{\kappa}$
UMi-Street	13.22	4.65	8.93
UMa	14.21	2.25	8.23

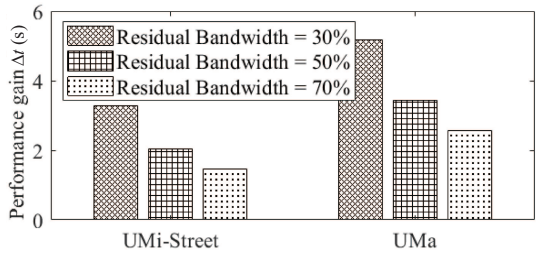

Figure 7 Impact of the types of BSs, a UE moves across two cells with the speed of 60 km/h, with shadowing.

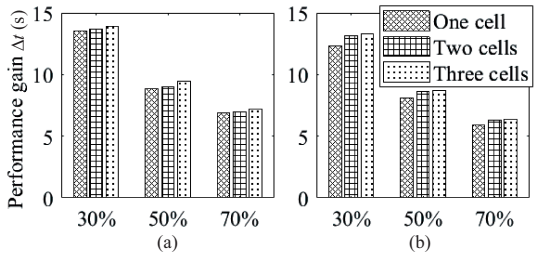

Figure 8 Impact of all factors, three moving UEs, with shadowing. (a) Dynamic residual bandwidth; (b) constant residual bandwidth.

4.2.4 *Moving UEs under time-varying residual resources*

In this scenario, three UEs move across the cells along the road, and the residual bandwidth of each BS in every frame is either identical (i.e., constant residual bandwidth) or generated independently with Gaussian distribution (i.e., dynamic residual bandwidth). When the residual bandwidth is dynamic, the values of ζ for $\overline{W}/W_{\max} = 30\%$, 50%, and 70% are 1.7×10^7, 5.6×10^7, and 15.1×10^7, respectively, as shown in Table 1. Again, $T_p = 60$ s, $B = 45$ MB, $d_f = 10$ m, $N_c = 50$ frames, and we control the speed of each UE and the duration of each frame such that the UEs can travel through one, two or three cells within an identical duration of the prediction window, the same as the previous scenario.

Figure 8 shows the simulation results of the performance gain Δt in the network with three mobile UEs. It can be seen that the gain decreases with the average residual bandwidth in the prediction window, and slightly increases with the number of cells (since shadowing is considered).

By comparing Figures 8(a) and (b), we can see that the value of Δt achieved by time-varying residual bandwidth in the prediction window is larger than the value with constant residual bandwidth, given the same relative residual bandwidth. Yet the gain from time-varying residual bandwidth is much lower than the gain from the average residual bandwidth. This can be explained from Remark 3, i.e., $\overline{W}$ has a larger impact on the ζ-coefficient (and also the performance gain) than σ_W. Moreover, as shown in Table 1, the values of σ_W decrease with $\overline{W}/W_{\max}$, hence have less impact on ζ and the performance gain when the value of $\overline{W}/W_{\max}$ is large (i.e., the network is idle with RT traffic).

By comparing the result with 30% residual bandwidth in Figure 8(b) and the value of Δt in Figure 6 multiplied by three, we can observe an extra gain of 4.3 s in the scenario of "transverse three cells". this

is the additional "macro-multi-user diversity gain" from coordinating the UEs for transmission in the frames of the predictive window with their largest average rates.

5 Conclusion

In this paper, we analyzed the performance gain of PRA for UEs requesting content delivery over the non-predictive counterpart. We derived the closed-form expressions of the performance gain and its upper bound in reducing the transmission time required by delivering a file when there is one UE in each cell in the prediction window. We resorted to a real dataset to observe the magnitude of a critical parameter, ζ-coefficient, in practical systems. We validated the analyses and showed the impact of the dynamic of residual bandwidth, cell radius, the number of cells a UE traversed, shadowing, and the number of UEs on the performance gain via simulation and numerical results. Our results demonstrated that the performance gain of PRA is high when the network is busy or the cell radius is large. Even for non-mobile users, PRA can provide large performance gain, which comes from the fluctuation of the residual resource and is inversely proportional to the ζ-coefficient. In future studies, we will take prediction errors into account when analyzing the performance gains of PRA, and extend this work to 6G-related scenarios, such as analyzing the performance gain of PRA for transmitting virtual reality videos in cell-free networks.

Acknowledgements This work was supported by Key Project of National Natural Science Foundation of China (Grant No. 61731002).

References

1 Liu F, Cui Y H, Masouros C, et al. Integrated sensing and communications: toward dual-functional wireless networks for 6G and beyond. IEEE J Sel Areas Commun, 2022, 40: 1728–1767
2 Wang Z Q, Du Y, Wei K J, et al. Vision, application scenarios, and key technology trends for 6G mobile communications. Sci China Inf Sci, 2022, 65: 151301
3 Chen W R, Li L X, Chen Z, et al. Enhancing THz/mmWave network beam alignment with integrated sensing and communication. IEEE Commun Lett, 2022, 26: 1698–1702
4 Liu F, Yuan W J, Masouros C, et al. Radar-assisted predictive beamforming for vehicular links: communication served by sensing. IEEE Trans Wireless Commun, 2020, 19: 7704–7719
5 Restuccia F, Melodia T. Deep learning at the physical layer: system challenges and applications to 5G and beyond. IEEE Commun Mag, 2020, 58: 58–64
6 Xu Y, Xu W J, Yin F, et al. High-accuracy wireless traffic prediction: a GP-based machine learning approach. In: Proceedings of the IEEE Global Communications Conference (GlobeCom), Singapore, 2017
7 Wang J, Tang J, Xu Z Y, et al. Spatiotemporal modeling and prediction in cellular networks: a big data enabled deep learning approach. In: Proceedings of the IEEE Conference on Computer Communications (ICC), Atlanta, 2017
8 Nagib A M, Abou-Zeid H, Hassanein H S, et al. Deep learning-based forecasting of cellular network utilization at millisecond resolutions. In: Proceedings of the IEEE International Conference on Communications (ICC), Montreal, 2021
9 Altché F, Fortelle A L. An LSTM network for highway trajectory prediction. In: Proceedings of the IEEE International Conference on Intelligent Transportation Systems (ICITS), Yokohama, 2017
10 Abou-zeid H, Hassanein H, Valentin S. Optimal predictive resource allocation: exploiting mobility patterns and radio maps. In: Proceedings of the IEEE Global Communications Conference (GlobeCom), Atlanta, 2013
11 Chen J T, Yatnalli U, Gesbert D. Learning radio maps for UAV-aided wireless networks: a segmented regression approach. In: Proceedings of the IEEE International Conference on Communications (ICC), Paris, 2017
12 Yao C T, Yang C Y, Xiong Z X. Energy-saving predictive resource planning and allocation. IEEE Trans Commun, 2016, 64: 5078–5095
13 Abou-zeid H, Hassanein H S, Valentin S. Energy-efficient adaptive video transmission: exploiting rate predictions in wireless networks. IEEE Trans Veh Technol, 2014, 63: 2013–2026
14 Atawia R, Hassanein H S, Ali N A, et al. Utilization of stochastic modeling for green predictive video delivery under network uncertainties. IEEE Trans Green Commun Netw, 2018, 2: 556–569
15 Guo J, Yang C Y. Impact of prediction errors on high throughput predictive resource allocation. IEEE Trans Veh Technol, 2020, 69: 9984–9999
16 Yang W T, Chi X F, Zhao L L, et al. Predictive two-timescale resource allocation for VoD services in fast moving scenarios. IEEE Trans Veh Technol, 2021, 70: 10002–10017
17 She C Y, Yang C Y. Energy efficient resource allocation for hybrid services with future channel gains. IEEE Trans Green Commun Netw, 2020, 4: 165–179
18 Lu Z, Veciana G D. Optimizing stored video delivery for mobile networks: the value of knowing the future. In: Proceedings of the IEEE INFOCOM, Turin, 2013
19 Bui N, Widmer J. Data-driven evaluation of anticipatory networking in LTE networks. IEEE Trans Mobile Comput, 2018, 17: 2252–2265
20 Bui N, Cesana M, Hosseini S A, et al. A survey of anticipatory mobile networking: context-based classification, prediction methodologies, and optimization techniques. IEEE Commun Surv Tut, 2017, 19: 1790–1821
21 Soh W S, Kim H S. A predictive bandwidth reservation scheme using mobile positioning and road topology information. IEEE ACM Trans Netw, 2006, 14: 1078–1091
22 Lin C Y, Chen K C, Wickramasuriya D, et al. Anticipatory mobility management by big data analytics for ultra-low latency mobile networking. In: Proceedings of the IEEE International Conference on Communications (ICC), Kansas City, 2018
23 Choi S, Shin K G. Adaptive bandwidth reservation and admission control in QoS-sensitive cellular networks. IEEE Trans Parallel Distrib Syst, 2002, 13: 882–897

24 Nadembega A, Hafid A, Taleb T. Mobility-prediction-aware bandwidth reservation scheme for mobile networks. IEEE Trans Veh Technol, 2015, 64: 2561–2576
25 Lee C, Cho H, Song S, et al. Prediction-based conditional handover for 5G mm-Wave networks: a deep-learning approach. IEEE Veh Technol Mag, 2020, 15: 54–62
26 Han S Q, Tan X F, Qi K Q, et al. Rethinking the gain of multicasting and proactive caching for VoD service. IEEE Wireless Commun, 2020, 27: 133–139
27 David H A, Nagaraja H N. Order Statistics. Hoboken: Wiley, 2004
28 Eberhard Z. Oxford Users' Guide to Mathematics. Oxford: Oxford University Press, 2004
29 Simon M K, Alouini M S. Digital Communication over Fading Channels: A Unified Approach to Performance Analysis. New York: John Wiley, 2000
30 Access E. Further Advancements for E-UTRA Physical Layer Aspects. 3GPP Technical Specification TR 36.814, 2010
31 Rec B I. H.264: advanced video coding for generic audiovisual services. http://www.itu.int
32 3rd Generation Partnership Project. Study on channel model for frequencies from 0.5 to 100 GHz (release 15). 3GPP TR 38.901, 2018

Appendix A Proof of Proposition 1

Since the average rates follow uniform distribution with mean value $\overline{R}$ and standard deviation σ_R, the PDF of the average rates is $f(R) = \frac{1}{2\sqrt{3}\sigma_R}$ if $R \in [\overline{R} - \sqrt{3}\sigma_R, \overline{R} + \sqrt{3}\sigma_R]$, and $f(R) = 0$ otherwise [28]. Then, the CDF of the average rates is $F(R) = \int_{-\infty}^{R} r \cdot f(r)\mathrm{d}r = \frac{R-\overline{R}+\sqrt{3}\sigma_R}{2\sqrt{3}\sigma_R}$ when $R \in [\overline{R} - \sqrt{3}\sigma_R, \bar{R} + \sqrt{3}\sigma_R]$, and $F(R) = 0$ and 1 respectively when $R \in (-\infty, \overline{R} - \sqrt{3}\sigma_R)$ and $R \in (\overline{R} + \sqrt{3}\sigma_R, \infty)$.

Denote $R_1, R_2, \ldots, R_{N_p}$ as the average rates of the UE in the N_p frames of the prediction window, which are random variables. By sorting $R_1, R_2, \ldots, R_{N_p}$ in ascending order, we can obtain $R_{(1)} \leqslant R_{(2)} \leqslant \cdots \leqslant R_{(N_p)}$, where $R_{(i)}$ is the ith lowest data rate. The PDF of $R_{(i)}$ is [27],

$$f_{(i)}(R) = \frac{N_p!}{(i-1)!\,(N_p-i)!} F^{i-1}(R)[1-F(R)]^{N_p-i} f(R)\,. \tag{A1}$$

Then, the mean value of $R_{(i)}$ can be derived as

$$\begin{aligned}\mathrm{E}\{R_{(i)}\} &= \textstyle\int_{-\infty}^{\infty} R \cdot f_{(i)}(R)\mathrm{d}R \overset{\text{(a)}}{=} C_{i,N_p} \int_{-\infty}^{\infty} R \cdot F^{i-1}(R)[1-F(R)]^{N_p-i} f(R)\mathrm{d}R \\ &\overset{\text{(b)}}{=} C_{i,N_p} \textstyle\int_0^1 R \cdot F^{i-1}(R)[1-F(R)]^{N_p-i}\mathrm{d}F(R),\end{aligned} \tag{A2}$$

where $C_{i,N_p} \triangleq \frac{N_p!}{(i-1)!(N_p-i)!}$, (a) is obtained by substituting (A1), and (b) comes from $f(R) = \mathrm{d}F(R)/\mathrm{d}R$.

By replacing $F(R)$ by u, (A2) can be rewritten as

$$\mathrm{E}\{R_{(i)}\} = C_{i,N_p} \textstyle\int_0^1 F^{-1}(u) u^{i-1}(1-u)^{N_p-i}\mathrm{d}u, \tag{A3}$$

where $R = F^{-1}(u)$ is the inverse function of $F(R)$.

From the expression of $F(R)$ above, we can obtain that $F^{-1}(u) = 2\sqrt{3}\sigma_R u + \overline{R} - \sqrt{3}\sigma_R$, $0 \leqslant u \leqslant 1$. By substituting it into (A3), we have

$$\begin{aligned}\mathrm{E}\{R_{(i)}\} &= C_{i,N_p}\left(2\sqrt{3}\sigma_R \textstyle\int_0^1 u^i(1-u)^{N_p-i}\mathrm{d}u + (\overline{R} - \sqrt{3}\sigma_R)\int_0^1 u^{i-1}(1-u)^{N_p-i}\mathrm{d}u\right) \\ &\overset{\text{(a)}}{=} 2\sqrt{3}\sigma_R \frac{C_{i,N_p}}{C_{i+1,N_p+1}} + \overline{R} - \sqrt{3}\sigma_R = \overline{R} - \sqrt{3}\sigma_R + \frac{2\sqrt{3}\sigma_R i}{N_p+1},\end{aligned} \tag{A4}$$

where (a) comes from $\int_0^1 x^n(1-x)^m\mathrm{d}x = \frac{1}{C_{n+1,m+n+1}} = \frac{n!m!}{(m+n+1)!}$ [28].

Then, by selecting the N_d frames with the largest average rates for transmission, on average the amount of data that the PRA policy can be transmitted is

$$\mathrm{E}\{\Delta R_{(N_p)} + \cdots + \Delta R_{(N_p-N_d+1)}\} = \Delta \sum_{i=N_p-N_d+1}^{N_p} \mathrm{E}\{R_{(i)}\} = N_d\Delta\left(\overline{R} + \tfrac{\sqrt{3}\sigma_R(N_p-N_d)}{N_p+1}\right). \tag{A5}$$

Since the file of size B should be delivered within the prediction window, we have $B = N_d\Delta(\overline{R} + \frac{\sqrt{3}\sigma_R(N_p-N_d)}{N_p+1})$. By solving this quadratic equation of N_d, we obtain

$$N_d = \frac{(N_p+1)\frac{\bar{R}}{\sigma_R} + \sqrt{3}N_p}{2\sqrt{3}}\left(1 - \left(1 - \frac{4\sqrt{3}B(N_p+1)}{\sigma_R\Delta\big((N_p+1)\frac{\overline{R}}{\sigma_R} + \sqrt{3}N_p\big)^2}\right)^{\frac{1}{2}}\right). \tag{A6}$$

When $\overline{R}$ is large or B is small, $\frac{4\sqrt{3}B(N_p+1)}{\sigma_R\Delta((N_p+1)\frac{\overline{R}}{\sigma_R}+\sqrt{3}N_p)^2} \to 0$. Then, the $(\cdot)^{\frac{1}{2}}$ term in (A6) can be accurately approximated with the Taylor's formula [28], i.e., $(1-x)^c = 1 + \sum_{i=1}^{n} \frac{(-1)^i x^i \prod_{j=0}^{i-1}(c-j)}{i!} + o(x^n) \approx 1 - cx$. With the formula, Eq. (A6) can be approximated as

$$N_d \approx \frac{(N_p+1)B}{\Delta((N_p+1)\overline{R} + \sqrt{3}N_p\sigma_R)}. \tag{A7}$$

The transmission time of the PRA policy is $t_{\mathrm{PRA}} = N_d\Delta$ and the duration of the prediction window is $T_p = N_p\Delta$. Then, from (A7) we can obtain the expression of t_{PRA} in Proposition 1. The transmission time used by the non-PRA policy to convey B bits is $t_{\mathrm{NPRA}} = \frac{B}{\overline{R}}$. Then, the performance gain in (3) can be obtained.

Appendix B Proof of proposition 2

In this appendix, we omit subscripts j and k in R_j^k, W_j^k and γ_j^k for notational simplicity.

It is shown in (1) that the average rate $R \approx W\gamma$, where the approximation is accurate when the average SNR is high. Since the UE does not move within the prediction window, γ is a constant. Moreover, since the residual bandwidth W is with mean value $\overline{W}$ and standard deviation σ_W in the prediction window, we can obtain the mean value of the average rate as $\overline{R} \approx \overline{W}\gamma$, and the standard deviation as $\sigma_R \approx \sigma_W \gamma$. Then, from (4) the performance gain can be obtained as

$$t_s^{\mathrm{UB}} \overset{\text{(a)}}{\approx} \frac{\sqrt{3}B}{\frac{\overline{R}^2}{\sigma_R} + \sqrt{3\overline{R}}} \overset{\text{(b)}}{\approx} \frac{\sqrt{3}B}{\frac{\overline{W}^2\gamma}{\sigma_W} + \sqrt{3\overline{W}\gamma}}, \tag{B1}$$

where the approximation (a) is from (4), which is accurate when $\overline{W}$ is large, σ_W or B is small, and the approximation (b) is accurate when the average SNR is high.

Appendix C Proof of Proposition 3

Since the time-varying pattern of large-scale fading is identical among all the m cells, the resource allocation pattern of the PRA policy is also identical among cells. Hence, we can analyze the time required by the PRA policy in one cell and then multiply it with m to obtain the total time to deliver a file. Since a UE uses N_c frames to traverse one cell, the PRA policy will sequentially choose the N_d frames with the largest average rates from the N_c frames for transmission.

Again, we sort the average rates of the UE in a single cell in ascending order, i.e., $R_{(1)} \leqslant R_{(2)} \leqslant \cdots \leqslant R_{(N_c)}$, where $R_{(i)}$ is the ith lowest data rate in the N_c frames. Similar to the derivation for (A4), when the average rates of the UE follow uniform distribution within range $[R_{\min}, R_{\max}]$, the mean value of $R_{(i)}$ can be obtained as $\mathrm{E}\{R_{(i)}\} = R_{\max} - \frac{N_c-i+1}{N_c+1}(R_{\max} - R_{\min})$. Then, the average amount of data that can be transmitted in the N_d frames by the PRA policy can be obtained as

$$\mathrm{E}\{\Delta R_{(N_c)} + \cdots + \Delta R_{(N_c-N_d+1)}\} = N_d\Delta\left(R_{\min} + \frac{2N_c - N_d + 1}{2(N_c+1)}(R_{\max} - R_{\min})\right). \tag{C1}$$

Since the time-varying pattern of the average rates is identical among cells and the file with B bits should be downloaded within the prediction window after the UE traverses m cells, the number of frames selected for transmission in one cell can be obtained by letting (C1) equal to $\frac{B}{m}$, i.e., $\frac{B}{m} = N_d\Delta(R_{\min} + \frac{(2N_c-N_d+1)}{2(N_c+1)}(R_{\max} - R_{\min}))$. By solving N_d from this quadratic equation and further using Taylor's formula $(1-x)^c = 1 - cx + \frac{c(c-1)x^2}{2!} + o(x^2)$, we can approximate the average number of frames required by the PRA policy for delivering the file in one cell as

$$N_d \approx \frac{2B\,(N_c+1)}{m\Delta(R_{\min} + (2N_c+1)R_{\max})} + \frac{4B^2(R_{\max} - R_{\min})(N_c+1)^2}{(m\Delta)^2(R_{\min} + (2N_c+1)R_{\max})^3}, \tag{C2}$$

which is accurate when B is small or W is large.

Since a UE traverses m cells within the prediction window, the total transmission time of the PRA policy is $t_{\mathrm{PRA}} = mN_d\Delta$. Recall that a UE uses N_c frames to traverse a cell, the duration of the prediction window is $T_p = mN_c\Delta$. Then, from (C2) the average transmission time of the PRA policy in m cells can be obtained as

$$t_{\mathrm{PRA}} \approx \frac{2B\,(N_c+1)}{R_{\min} + (2N_c+1)R_{\max}} + \frac{4B^2(R_{\max} - R_{\min})(N_c+1)^2N_c}{T_p(R_{\min} + (2N_c+1)R_{\max})^3}. \tag{C3}$$

Since $\alpha_j = (d_j)^\beta$, from (1) we have $R_{\min} \approx W\log_2(D_b^\beta \frac{N_{\mathrm{tx}}}{2^\beta\sigma^2}P_{\max})$ and $R_{\max} \approx W\log_2(d_{\min}^\beta \frac{N_{\mathrm{tx}}}{\sigma^2}P_{\max})$. Upon substituting into (C3), we have $t_{\mathrm{PRA}} \approx \frac{2B(N_c+1)}{W(\kappa_{\min}+(2N_c+1)\kappa_{\max})} + \frac{4B^2(\kappa_{\max}-\kappa_{\min})(N_c+1)^2N_c}{T_pW^2(\kappa_{\min}+(2N_c+1)\kappa_{\max})^3}$, where $\kappa_{\min} \triangleq \log_2(D_b^\beta \frac{N_{\mathrm{tx}}P_{\max}}{2^\beta\sigma^2})$, and $\kappa_{\max} \triangleq \log_2(d_{\min}^\beta \frac{N_{\mathrm{tx}}P_{\max}}{\sigma^2})$.

Considering that the transmission time of the non-PRA policy is $t_{\mathrm{NPRA}} = \frac{B}{R}$, the performance gain can be obtained.