Journal of Communications and Information Networks, Vol.2, No.1, Mar. 2017
DOI: 10.1007/s41650-017-0005-y

Research paper

Special Issue on Wireless Big Data

Data-driven resource allocation with traffic load prediction

Chuting Yao[1], Chenyang Yang[1]*, Chih-Lin I[2]

1. Beihang University, Beijing 100191, China
2. China Mobile Research Institute, Beijing 100053, China

* Corresponding author, email: cyyang@buaa.edu.cn

Abstract: Wireless big data is attracting extensive attention from operators, vendors and academia, which provides new freedoms in improving the performance from various levels of wireless networks. One possible way to leverage big data analysis is predictive resource allocation, which has been reported to increase spectrum and energy resource utilization efficiency with the predicted user behavior including user mobility. However, few works address how the traffic load prediction can be exploited to optimize the data-driven radio access. We show how to translate the predicted traffic load into the essential information used for resource optimization by taking energy-saving transmission for non-real-time user as an example. By formulating and solving an energy minimizing resource allocation problem with future instantaneous bandwidth information, we not only provide a performance upper bound, but also reveal that only two key parameters are related to the future information. By exploiting the residual bandwidth probability derived from the traffic volume prediction, the two parameters can be estimated accurately when the transmission delay allowed by the user is large, and the closed-form solution of global optimal resource allocation can be obtained when the delay approaches infinity. We provide a heuristic resource allocation policy to guarantee a target transmission completion probability when the delay is no-so-large. Simulation results validate our analysis, show remarkable energy-saving gain of the proposed predictive policy over non-predictive policies, and illustrate that the time granularity in predicting traffic load should be identical to the delay allowed by the user.

Keywords: predictive resource allocation, big data, traffic load, energy saving

Citation: C. T. Yao, C. Y. Yang, C.-L. I. Data-driven resource allocation with traffic load prediction [J]. Journal of communications and information networks, 2017, 2(1): 52-65.

1 Introduction

To support the explosively growing traffic demands in big data era, the main trend techniques for the 5G (fifth Generation) cellular networks are to update network architecture, boost throughput by network densification, and explore more spectrum[1]. Yet in reality the resources are often under-utilized in most BSs (Base Stations), due to the time-varying traffic pattern. It has been observed from prevalent networks that in average less than 15% resource blocks are truly used in practice. To deal with such a dilemma, which comes from the inherent human routine activity, managing radio resources by exploiting predictable human behavior is a possible solution.

Recently, wireless big data is attracting extensive attention from operators, vendors and academia[2], which is expected to provide new freedom in optimiz-

Manuscript received Jan. 15, 2017; accepted Feb. 8, 2017
This work is supported by the National Natural Science Foundation of China (No. 61671036).

ing various performance of wireless networks. However, it remains elusive about whether or not big data analysis can improve resource utilization efficiency, and how to optimize data-driven radio access networks.

More or less inspired by the recent report that human behavior is highly predictable[3], optimizing resource allocation by leveraging the prediction ability endowed by big data is emerging as a new research area[4-8]. With big data analytics, the traffic volume and the mobility pattern can be predicted[9-11], at least within a prediction window. From the traffic volume prediction, the average network resource usage status can be estimated. From the trajectory of a mobile user, the average channel gains can be predicted with the help of a radio map[8]. With various machine learning algorithms such as collaborative filtering that has long been studied for recommendation problems[12], the content popularity and even the preferred content of an individual user are possible to be known before the user(s) truly initiates the request. Undoubtedly, predicting the behavior related information is challenging, considering that the numbers of users and contents are extraordinarily huge. This naturally raises the following questions: What performance of wireless access networks can be improved by exploiting the predictable information and with how much gain? How to translate the predictable knowledge to the essential information that can be used to optimize radio resource allocation and how to exploit the information?

Most existing works of predictive resource allocation assume that there exists only one class of mobile users in the network either requesting RT (Real-Time) or NRT (Non-Real-Time) services, and preserve or pre-allocate radio resources according to their channel variation in the future[4-8]. Yet a real-world cellular network needs to support different kinds of services with various priorities. While jointly allocating resources to different services can maximize the resource utilization efficiency, it is very hard to predict trajectories of all mobile users if not impossible, considering the prohibitive computational complexity to process such a huge amount of data.

Fortunately, in real-world systems a large percentage of users requesting NRT services are not mobile. If we can predict the traffic load of the BS where these NRT users are located, which are dynamic due to the random requests from the RT users in the cell, then we can pre-allocate resources to the NRT users to exploit the excess resources. Then, the data to be processed will be reduced drastically. As reported in Ref. [13] and references therein, the dynamics of traffic load exhibit periodical characteristic among days and even hours. This implies that the traffic load is highly predicable. With the predicted traffic loads for different services, we can infer the resource occupation status for different classes of services. However, the prediction granularity for different purposes of resource management are quite different. For example, in order to close idle BSs for energy saving during the off-peak time, the traffic volume in every hour of the next day was predicted in Ref. [9], while the traffic volume in future 5 min was predicted in Ref. [10]. In order to assist predictive resource allocation, which usually operates in milliseconds, what time granularity in predicting traffic load is necessary? How to translate the predicted traffic load to the useful information in resource allocation? These questions are important but still open.

In this paper, we strive to answer these questions by taking energy saving as an example performance metric. We first illustrate how to derive the residual bandwidth probability from the traffic volume prediction. Then, by formulating and solving an energy minimizing problem with future instantaneous bandwidth information, we reveal that the optimal resource allocation only depends on two parameters, which can be estimated accurately when the transmission delay allowed by the NRT user is large. Next, we derive the closed-form expression of the optimal resource allocation when the delay approaches infinity, and provide a heuristic policy to guarantee a target transmission completion probability with finite delay. The simulation results validate our analysis, and show remarkable energy save gain from the proposed policy.

2 System model

Consider a cellular system where a BS equipped with N_t antennas serves multiple users with RT traffic and a user with NRT traffic. The total bandwidth and transmit power of the BS are $W_{\max}$ and $p_{\max}$, respectively.

Due to the random behavior and possible mobility of the RT users, the instantaneous request arrival rate of the RT service is time varying. To capture the essence of the problem at hand and simplify the analysis, we assume that the instantaneous arrival rate of the RT traffic keeps constant in a time slot with duration Δ, and may change among time slots.

Assume that the NRT user is equipped with a single antenna. We model the request of the NRT user as to download a file from the BS with B bits before a deadline with T time slots, i.e., T is the preferred transmission delay of the NRT user. To emphasize the role of traffic load prediction on radio resource allocation, assume that the NRT user is stationary. For conciseness, in the rest of this paper a/the "user" is referred to as the NRT user unless otherwise specified.

Since RT traffic is with high priority, only part of resources can be used for the NRT user. In particular, the BS serves the NRT user with residual resources W^t and $p^t_{\max}$, which may vary among different time slots. Denote $m^t \in \{0,1\}$ as a time slot scheduling indicator of the user. When $m^t = 1$, the user is scheduled by the BS in the tth time slot, otherwise it is not. Then, in the tth time slot, the achievable rate of serving the NRT user (in nats) is

$$R^t = m^t W^t \ln\left(1 + \frac{h}{W^t N_0} p^t\right), \tag{1}$$

where $p^t \leqslant p^t_{\max}$ is the transmit power allocated to the user, h is the channel gain including path loss, shadowing and small scale channel, and N_0 is the noise power spectrum density.

Assume that the BS can be switched into sleep mode when the BS dose not serve any traffic. Then, the power consumed at BS by serving the NRT user contains not only the transmit power but also the extra circuit power for BS operation when there is no RT traffic[1]. The power consumed at the BS for the NRT user in the tth time slot can be modeled as[14]

$$p^t_{\mathrm{Ex}} = \frac{1}{\xi} p^t + \mathbf{1}(W^t = W_{\max}) m^t (p_{\mathrm{act}} - p_{\mathrm{sle}}), \tag{2}$$

where ξ is the power amplifier efficiency, p_{act} and p_{sle} are respectively the circuit power consumption when the BS is in active and sleep modes, and $\mathbf{1}(c) = 1$ when c is true and $\mathbf{1}(c) = 0$ otherwise.

3 Resource allocation with traffic volume prediction

In this section, we first show how to translate the predictable traffic volume into the information to be exploited in the resource allocation, i.e., residual bandwidth probability. Then, by formulating and finding the solution of an energy minimizing problem with future instantaneous residual bandwidth, we show how to apply the residual bandwidth probability for radio resource allocation.

3.1 From traffic volume to residual bandwidth probability

With big data analysis, the traffic volumes for different services are reported predictable at least within a time window[9,10]. In this paper, the traffic volume is the traffic load (i.e., average request arrival rate) of the RT traffic multiplying the average service time of each RT request.

To illustrate how to translate the predicted traffic volume into the statistics of residual bandwidth available for NRT user, we make the following assumptions: (1) The request arrival of the RT traffic is Poisson process with average rate λ. (2) Serving each request occupies $1/L \cdot 100\%$ of the total bandwidth in each time slot, where L is the maximal number

[1] If the BS is busy with RT traffic when serving the NRT user, its circuit power is not taken into account for the NRT user. Otherwise, the BS needs to consume extra circuit power to operate the BS.

of RT requests the BS able to serve simultaneously. (3) The service time of each RT request follows exponential distribution with mean time V. (4) A new RT request will not be admitted to the BS if all the bandwidth is occupied, i.e., when $W^t = W_{\max}$. For mathematical tractability and notational simplicity, the transmit power reserved for the RT traffic is assumed in proportion to the occupied bandwidth[2], i.e., $p^t_{\mathrm{RT}} = (1 - W^t/W_{\max})p_{\max}$. Since the BS can at most serve L requests simultaneously, the total bandwidth can be regarded as L servers and each request can be viewed as a customer with different service time. Then, the transmission of the RT traffic follows a queuing M/M/C/C discipline[15], and the occupation probability of the servers can reflect the statues for the excess resources at the BS. Then, the probability that $(1 - l/L) \cdot 100\%$ of the total bandwidth is occupied by the RT traffic can be obtained from Ref. [16] as

$$P_l \triangleq P\left(W^t = \frac{l}{L}W_{\max}\right) = \frac{\frac{(\lambda V)^{L-l}}{(L-l)!}}{\sum_{l=0}^{L}\frac{(\lambda V)^l}{l!}}. \quad (3)$$

P_l can reflect the average resource utilization status of the BS, which is called residual bandwidth probability.

In practice, the above-mentioned four assumptions may not hold, then the relation between P_l and traffic volume λV will be no longer the expression in Eq. (3). Yet the basic principle to translate the predicted traffic volume into residual bandwidth probability is still applicable. In fact, it is no need to first predict λV and then convert it to P_l. Instead, we can predict P_l from traffic data directly, say simply estimate P_l in a specific time (e.g., 12:00 a.m.) each day and then use the prediction in the past days to predict the value in the same time of the next day.

3.2 Resource allocation with future instantaneous residual bandwidth

To provide a performance upper bound and gain insight into how the translated information is exploited to allocate resources for conveying the B bits (i.e., $B\ln 2$ nats) before the deadline of T for the NRT user, we first formulate an energy minimization problem with bandwidth information in all future time slots, i.e., in the first time slot, the instantaneous residual bandwidths in the future, W^t, $t = 1, \cdots, T$, are known. The optimal predictive resource allocation that minimizes the average total energy consumption in T time slots for the NRT user can be obtained from the following problem,

$$\min_{m^1,\cdots,m^T,p^1,\cdots,p^T} \ \frac{1}{T}\sum_{t=1}^{T} p^t_{\mathrm{Ex}}\Delta \quad (4a)$$

$$\text{s.t.} \quad \sum_{t=1}^{T} R^t = \frac{B\ln 2}{\Delta} \triangleq T\bar{R} \quad (4b)$$

$$p^t \geqslant 0, \ t = 1, \cdots, T, \quad (4c)$$

where $\bar{R}$ is the required time-average rate for the user, reflecting the demand of the NRT user.

According to whether or not the BS serves the RT traffic, we divide the T time slots into busy time and idle time. Denote $\mathcal{T}_{\mathrm{ac}} = \{t|W^t < W_{\max}\}$ with cardinality T_{ac} as the index set of time slots, where in these time slots the BS is active since it needs to serve the RT traffic, and $\mathcal{T}_{\mathrm{id}} = \{t|W^t = W_{\max}\}$ with cardinality $T_{\mathrm{id}} = T - T_{\mathrm{ac}}$ as the index set of the idle time slots. Then, $\mathcal{T}_{\mathrm{id}}$ is the complementary set of $\mathcal{T}_{\mathrm{ac}}$ as shown in Fig. 1.

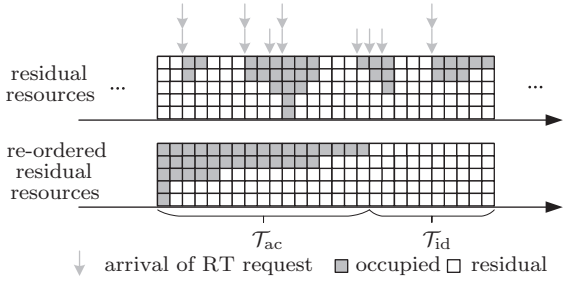


Figure 1 Illustration of the idle and busy time slots. To help understand, we have re-ordered the time slots in the lower sub-figure.

To save energy consumed by the circuits, the BS can be switched into sleeping mode in idle time slots.

[2] Simulation results show that this assumption has minor impact on the performance of the proposed policy, which are not shown for space limitation.

If the NRT user needs to exploit some of the idle time slots, i.e., $\{t|m^t = 1, t \in \mathcal{T}_{\text{id}}\}$, the BS needs to activate these time slot. Considering Eq. (2) and the definition of the idle time slot set, the objective function of problem (4) can be rewritten as,

$$\begin{aligned}&\frac{1}{T}\frac{1}{\xi}\sum_{t=1}^{T} p^t \Delta + \frac{1}{T}\sum_{t\in\mathcal{T}_{\text{id}}} m^t (p_{\text{act}} - p_{\text{sle}})\Delta \\ &\triangleq \frac{1}{T}\frac{1}{\xi}\sum_{t=1}^{T} p^t \Delta + \frac{\kappa T_{\text{id}}}{T}(p_{\text{act}} - p_{\text{sle}})\Delta, \end{aligned} \tag{5}$$

where $\kappa \triangleq \frac{1}{T_{\text{id}}}\sum_{t\in\mathcal{T}_{\text{id}}} m^t$ is the active ratio of idle time slots.

Since the channel gain h is the same among the T time slots, randomly selecting any κT_{id} idle time slots to serve the NRT user consumes the same energy. Denote the κT_{id} idle time slots set activated for the NRT user as $\mathcal{N}$. Since for the idle time slots that are not activated, the transmit power allocated to the NRT users is zero, i.e., $p^t = 0, t \in \mathcal{T}_{\text{id}}$ and $t \notin \mathcal{N}$, problem (4) can be equivalently transformed to

$$\min_{\substack{\kappa, p^t, \\ t\in\mathcal{T}_{\text{ac}}\cup\mathcal{N}}} \Psi_p \triangleq \frac{1}{T\xi}\sum_{t\in\mathcal{T}_{\text{ac}}\cup\mathcal{N}} p^t + \frac{\kappa T_{\text{id}}}{T}(p_{\text{act}} - p_{\text{sle}}) \tag{6a}$$

$$\text{s.t.} \sum_{t\in\mathcal{T}_{\text{ac}}\cup\mathcal{N}} W^t \ln\left(1 + \frac{h}{W^t N_0} p^t\right) = T\bar{R} \tag{6b}$$

$$p^t \geqslant 0,\ t \in \mathcal{N}\cup\mathcal{T}_{\text{ac}}, \tag{6c}$$

where we omit the duration of each time slot Δ, which does not affect the solution.

From problem (6), we can see that the objective function is a sum of transmit power and circuit power, and the constraint only affects the transmit power. Therefore, to find the global optimal solution, we can first fix the circuit power by fixing κ to find the minimal average transmit power. Then, the optimal power allocation can be obtained to minimize the average total power by searching $\kappa \in [0, 1]$. In the following, we solve problem (6) with two steps.

When κ is given, the second term of Eq. (6a) is fixed. Then, the problem can be reduced into a transmit power minimizing problem as follows,

$$\min_{p^t, t\in\mathcal{T}_{\text{ac}}\cup\mathcal{N}} \sum_{t\in\mathcal{T}_{\text{ac}}\cup\mathcal{N}} p^t \tag{7a}$$

$$\text{s.t.} \quad \text{Eqs. (6b), (6c).} \tag{7b}$$

By relaxing Eq. (6b) as $\sum_{t\in\mathcal{T}_{\text{ac}}\cup\mathcal{N}} W^t \ln\left(1 + \frac{h}{W^t N_0} p^t\right) \leqslant T\bar{R}$, problem (7) becomes a convex problem without losing optimality[3]. Then, the optimal solution can be solved from the KKT (Karush-Kuhn-Tucker) conditions[17] as

$$p^t = W^t\left(\nu - \frac{N_0}{h}\right), \tag{8}$$

where ν is a power level, which is

$$\begin{aligned}\nu &= \frac{N_0}{h}\exp\left(\frac{T\bar{R}}{\sum_{t\in\mathcal{T}_{\text{ac}}\cup\mathcal{N}} W^t}\right) \\ &= \frac{N_0}{h}\exp\left(\frac{T\bar{R}}{\sum_{t\in\mathcal{T}_{\text{ac}}} W^t + \kappa T_{\text{id}} W_{\max}}\right). \end{aligned} \tag{9}$$

The solution suggests that more power should be allocated to the time slots with more residual bandwidth.

Considering that the residual bandwidth for the idle time slots in $\mathcal{N}$ is $W_{\max}$, the objective function in problem (6) can be expressed as a function of κ, which is

$$\begin{aligned}\Psi_p = &\frac{1}{T\xi}\sum_{t\in\mathcal{T}_{\text{ac}}} W^t \left(\frac{N_0}{h}\exp\left(T\bar{R}\right.\right. \\ &\left.\left.\times\left(\sum_{t\in\mathcal{T}_{\text{ac}}} W^t + \kappa T_{\text{id}} W_{\max}\right)^{-1}\right) - \frac{N_0}{h}\right) \\ &+ \frac{\kappa T_{\text{id}}}{T\xi} W_{\max}\left(\frac{N_0}{h}\exp\left(\frac{N_0}{h}T\bar{R}\right.\right. \\ &\left.\left.\times\left(\sum_{t\in\mathcal{T}_{\text{ac}}} W^t + \kappa T_{\text{id}} W_{\max}\right)^{-1}\right) - \frac{N_0}{h}\right) \\ &+ \frac{\kappa T_{\text{id}}}{T}(p_{\text{act}} - p_{\text{sle}}). \end{aligned} \tag{10}$$

Since the number of activated idle time slots κT_{id} should be an integer, by searching κT_{id} from 0 to T_{id}, it is easy to find the optimal solution of κ^* that

[3] The equality constraint and the inequality constraint give rise to the same optimal solution, because if the optimal power p^{t*} satisfies $\sum_{t\in\mathcal{T}_{\text{ac}}\cup\mathcal{N}} W^t \ln\left(1 + \frac{h}{W^t N_0} p^{t*}\right) < T\bar{R}$, we can always find a new power allocation $\tilde{p}^{t*} = cp^{t*}$ with $c < 1$, which can meet the user's requirement $\bar{R}$ with properly selected c but consuming less transmit power and the same circuit power.

minimizes the average total power Ψ_p. Then, the optimal power allocation of problem (6) is

$$p^{t*} = \begin{cases} W^t \left(\nu^* - \dfrac{N_0}{h}\right), & \mathcal{T}_{\text{ac}} \cup \mathcal{N}^*, \\ 0, & \text{else}, \end{cases} \tag{11}$$

where ν^* is obtained by substituting κ^* into Eq. (9), and $\mathcal{N}^*$ can be any subset of $\mathcal{T}_{\text{id}}$ with cardinality $\kappa^* T_{\text{id}}$.

It is worthy to note that in the 1st time slot when the NRT user initiates its request, only two parameters, the active ratio of idle time slots κ^* and the power level ν^*, contain the future instantaneous residual bandwidth $W^t, t = 2, \cdots, T$.

3.3 Predicting two parameters from residual bandwidth probability

The solution obtained in previous section is not viable in practice, because when optimizing the resource allocation for the NRT user in the 1st time slot, the future instantaneous residual bandwidth W^t for $t > 1$ is hard to predict if not impossible, owing to the random arrival of RT traffic.

Fortunately, only the two parameters ν^* and κ^* depend on the future instantaneous information, which do not change across the T time slots, and can serve as the resource planning parameters after predicted in the first time slot. With these two parameters, the BS can determine the operation mode (active or sleep) and allocate resource (time and power) for the upcoming T time slots, i.e., make a resource allocation plan for the NRT user: when to transmit and with how much radio resources. In what follows, we show how to estimate these two parameters from the residual bandwidth probability P_l and the corresponding estimation accuracy.

We first estimate the power level ν with given κ, and then estimate κ^*. We start from the following proposition.

Proposition 1 When T is large, $\dfrac{\bar{R}}{\ln(h\nu/N_0)}$ asymptotically follows Gaussian distribution, i.e.,

$$\frac{\bar{R}}{\ln(h\nu/N_0)} \sim \mathbb{N}(\mu_w, \sigma_w^2), \tag{12}$$

where $\mu_w = \sum_{l=1}^{L-1} P_l \dfrac{l}{L} W_{\max} + \kappa P_L W_{\max}$, and

$$\sigma_w^2 = \frac{1}{T}\left(\sum_{l=1}^{L-1} P_l \frac{l^2}{L^2} W_{\max}^2 + \kappa P_L W_{\max}^2 - \left(\sum_{l=1}^{L-1} P_l \frac{l}{L} W_{\max} + \kappa P_L W_{\max}\right)^2\right),$$

with P_L denoting the probability that the residual bandwidth equals to $W_{\max}$.

Proof See Appendix A).

From the expression of σ_w^2, we have $\lim_{T\to\infty} \sigma_w^2 = 0$. This indicates that for a given κ, $\dfrac{\bar{R}}{\ln(h\nu/N_0)}$ can be estimated by its mean value μ_w without errors when $T \to \infty$. Then, ν can be estimated asymptotically as follows,

$$\nu = \frac{N_0}{h} \exp\left(\frac{\bar{R}}{\mu_w}\right). \tag{13}$$

Since κ^* is obtained by minimizing the average total power Ψ_p in Eq. (6b), if Ψ_p is estimated accurately, then κ^* can be estimated accurately. The following proposition indicates that Ψ_p can be estimated asymptotically as its mean value $\mu_{\Psi_p}^{\infty}$ without errors when $T \to \infty$.

Proposition 2 When $T \to \infty$, the expectation of Ψ_p approaches to

$$\lim_{T\to\infty} \mu_{\Psi_p} \triangleq \mu_{\Psi_p}^{\infty} = \frac{1}{\xi}\frac{N_0}{h}\mu_w\left(\mathrm{e}^{\frac{\bar{R}}{\mu_w}} - 1\right) + \kappa P_L (p_{\text{act}} - p_{\text{sle}}), \tag{14}$$

and the variance is

$$\lim_{T\to\infty} \sigma_{\Psi_p}^2 = 0. \tag{15}$$

Proof See Appendix B).

Propositions 1 and 2 indicate that with the unbiased estimation of the total average power Ψ_p when $T \to \infty$, the optimal active ratio κ^* and hence the optimal power level ν^* can be estimated accurately.

In the following, we find the optimal estimate κ^* and ν^* by using $\mu_{\Psi_p}^{\infty}$ as the estimate of the power consumption Ψ_p.

To find κ^* that minimizes the average total power, we first derive the first order derivative of $\mu_{\Psi_p}^{\infty}$ with respect to κ as

$$\frac{\partial \mu_{\Psi_p}^{\infty}}{\partial \kappa} = \frac{1}{\xi}\frac{N_0}{h}\left(\mathrm{e}^{\frac{\bar{R}}{\mu_w}} - 1 - \frac{\bar{R}}{\mu_w}\mathrm{e}^{\frac{\bar{R}}{\mu_w}}\right)\frac{\partial \mu_w}{\partial \kappa} + P_L(p_{\text{act}} - p_{\text{sle}}),$$

where $\frac{\partial \mu_w}{\partial \kappa} = P_L W_{\max}$ is obtained from the expression of μ_w in Proposition 1.

From Eq. (14), the second order derivative of $\mu_{\Psi_p}^{\infty}$ with respect to κ can be derived as

$$\frac{\partial^2 \mu_{\Psi_p}^{\infty}}{\partial \kappa^2} = \frac{1}{\xi}\frac{N_0}{h}\frac{P_L^2 W_{\max}^2 \bar{R}^2}{\mu_w^3}\mathrm{e}^{\frac{\bar{R}}{\mu_w}}, \tag{16}$$

which is a positive value with easy transformations. This indicates that $\partial \mu_{\Psi_p}^{\infty}/\partial \kappa$ is an increase function of κ.

According to the value of $\partial \mu_{\Psi_p}^{\infty}/\partial \kappa$, $\kappa^* \in [0,1]$ can be found in the following three cases as shown in Fig. 2.

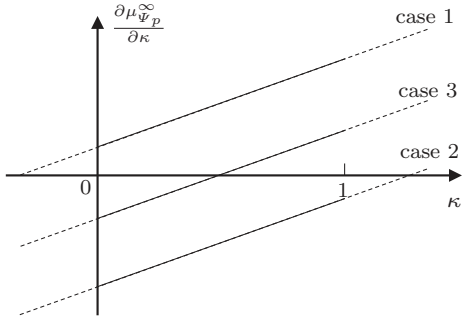


Figure 2 Example of $\frac{\partial \mu_{\Psi_p}^{\infty}}{\partial \kappa}$

1. $\frac{\partial \mu_{\Psi_p}^{\infty}}{\partial \kappa}|_{\kappa=0} > 0$, the average total power increases with κ. Then, the optimal estimate $\kappa^* = 0$.
2. $\frac{\partial \mu_{\Psi_p}^{\infty}}{\partial \kappa}|_{\kappa=1} < 0$, the average total power decreases with κ. Then, $\kappa^* = 1$.
3. The average total power first increases and then decreases with κ, the optimal estimate κ^* satisfies

$$\frac{\partial \mu_{\Psi_p}^{\infty}}{\partial \kappa}|_{\kappa=\kappa^*} = 0, \tag{17}$$

from which κ^* can be obtained as

$$\kappa^* = \frac{\frac{\bar{R}}{\mathcal{L}_w(x)+1} - \sum_{l=1}^{L-1} P_l \frac{l}{L} W_{\max}}{P_L W_{\max}}, \tag{18}$$

where $\mathcal{L}_w(x)$ is principle branch of Lambert-W function[18], and $x = \frac{\xi(p_{\text{act}} - p_{\text{sle}})h}{N_0 W_{\max} e} - \frac{1}{e}$.

With κ^*, we can obtain the optimal estimate ν^* from Eq. (13) and the expression of μ_w in Proposition 1 as follows,

$$\nu^* = \begin{cases} \frac{N_0}{h}\exp\left(\frac{\bar{R}}{\sum_{l=1}^{L-1} P_l \frac{l}{L} W_{\max}}\right), & \text{if } \kappa^* = 0, \\ \frac{N_0}{h}\exp\left(\mathcal{L}_w(x)+1\right), & \text{if } \kappa^* \in (0,1), \\ \frac{N_0}{h}\exp\left(\frac{\bar{R}}{\sum_{l=1}^{L} P_l \frac{l}{L} W_{\max}}\right), & \text{if } \kappa^* = 1. \end{cases} \tag{19}$$

It is shown from Eq. (18) that κ^* is a non-decreasing function of the user demand $\bar{R}$. It means that if the NRT user requests more data before the deadline T, more idle time slots should be activated to transmit. From the numerical results of κ^* and ν^* computed from Eqs. (18) and (19) in Fig. 3, we can find that with the increase of $\bar{R}$, κ^* and ν^* will not change simultaneously. When the user demand is lower than a value such as R_1, the BS allocates power to all the busy time slots, and when $\bar{R} > R_1$, the BS begins to activate idle time slots. When the user demand is too high (i.e., $\bar{R} > R_2$) such that all the idle time slots have been activated, the BS will increase the power level again.

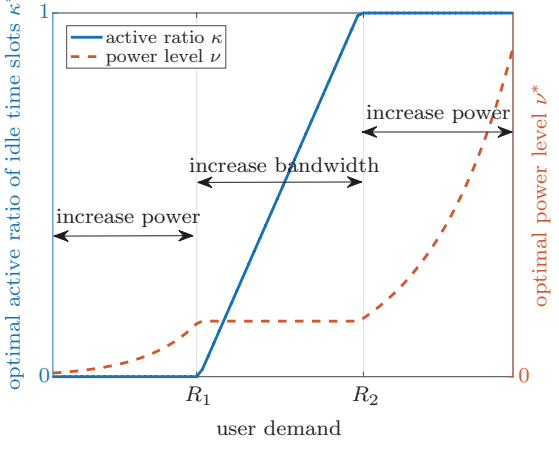


Figure 3 Illustration of Eqs. (18) and (19)

3.4 Predictive resource allocation for a finite deadline

In this section, we propose a predictive resource allocation for practical use where the value of T is finite.

When T is not infinity, using ν^* as the power level may fail to transmit the B bits before the deadline. Intuitively, if the estimated power level is higher, the completion probability, defined as the probability that the B bits can be completely transmitted before the Tth time slot, will be higher. Such a completion probability can reflect how much percent NRT users are satisfied, i.e, the satisfactory rate of the NRT user. Therefore, to transmit B bits before deadline, we first estimate the active ratio $\hat{\kappa}^* = \kappa^*$, and increase the completion probability only by adjusting the power level.

From Proposition 1, $\dfrac{\bar{R}}{\ln(h\nu/N_0)}$ follows Gaussian distribution $\mathbb{N}(\mu_w, \sigma_w^2)$ when $T \to \infty$. To ensure the user satisfactory rate, we can make a conservative resource allocation by increasing transmit power, we adjust the power level as follows,

$$\hat{\nu}^* \triangleq \frac{N_0}{h} \exp\left(\frac{\bar{R}}{\mu_w|_{\kappa=\kappa^*} - \eta\sigma_w|_{\kappa=\kappa^*}}\right) \overset{\Pr}{\geqslant} \nu^*, \quad (20)$$

where $\overset{\Pr}{\geqslant}$ means greater in probability of Pr, and η reflects the degree of the conservation. Since for a normal distributed random variable, 95% of its values are less than two-fold standard deviations from its mean value, when $\eta = 2$, $\Pr = 95\%$. Similarly, when $\eta = 0$, $\Pr = 50\%$, and when $\eta = 1$, $\Pr = 68\%$. Here Pr is the target completion probability predetermined for the predictive resource allocation, that can be controlled by η.

The proposed predictive resource allocation policy is implemented in the following two time scales.

• When the NRT user initiates its request in the 1st time slot, the BS makes a resource usage plan for the user according to its demand (i.e., B and T), the target completion probability Pr, and the predicted residual bandwidth probability P_l, by estimating the power level $\hat{\nu}^*$ with Eq. (20) and the active ratio $\hat{\kappa}^* = \kappa^*$ with Eq. (18). These two parameters will serve as a "ruler" in the subsequent on-line transmission: $\hat{\kappa}^*$ determines whether a time slot should be scheduled to serve the user, and $\hat{\nu}^*$ determines how much power needs to be allocated to the time slot.

• In the tth time slot, the BS computes the optimal transmit power with Eq. (11). If the BS has no RT traffic to serve, then the BS switches into sleep mode in probability $1 - \kappa^*$. Otherwise, the BS transmits to the user with transmit power $p^t = W^t(\hat{\nu}^* - \dfrac{N_0}{h})$.

If there are some bits of the NRT user failing to transmit before the deadline, they will be transmitted with $p^t_{\max}$.

3.5 Extension to multiple NRT users

When there exists multiple NRT users, a user scheduler needs to be introduced. Then, the NRT users can be scheduled in different time slots randomly selected, and the number of time slots can be determined under various criteria, say fairness among users. Then, the resource allocation parameters for each NRT user can be predicted by using the proposed method, and the power allocation to each user can use the same way as in the single NRT user scenario.

4 Simulation and numerical results

In this section, we validate our analysis, evaluate the performance of the proposed predictive resource allocation, and show the impact of the granularity of traffic volume prediction.

We consider a single micro BS with cell radius of 100 m, $N_t = 4$. The maximal transmit power and bandwidth are $p_{\max} = 13$ W and $W_{\max} = 10$ MHz, respectively. The duration of each time slot is set as $\Delta = 10$ ms.

The BS serves the NRT user with the residual resources left by serving randomly arrived RT requests. In the simulation, the RT traffic arrival is Poisson process with average rate $\lambda = 0.2$ requests/time slot, and each request of RT traffic occupies the same bandwidth of 2 MHz (i.e., 20% of 10 MHz) with

transmit power 2.6 W (i.e., 20% of 13 W). To reflect the difference in requests and channel conditions of different RT users, their service time follows exponential distribution with mean $V = 2$ time slots[15]. Each BS can serve at most $L = 5$ requests of RT traffic in one time slot, and the newly arrived RT request will not be admitted if the BS has been fully occupied. In this simulation setup, about 10% resources are occupied by the RT traffic.

The NRT user is located at $d = 50$ m from the BS, and the path-loss model is $30.6 + 36.7 \lg(d)$[19]. The noise power spectrum density is $N_0 = -165$ dBm/Hz. The circuit power consumption of the BS in active mode is $p_{\text{act}} = 40.14$ W, and that in sleeping mode is $p_{\text{sle}} = 15.5$ W[14]. The power amplifier efficiency is $\xi = 26\%$[14]. The results are obtained from 1 000 Monte Carlo trails, in each trail the random request arrival and service time of RT traffic vary. Unless otherwise specified, this simulation setup is used for all results.

4.1 Validation of the analysis

We first validate the estimation accuracy of ν when given κ, and then the estimation accuracy of average power consumption that is used to estimate κ^*. To show the impact of finite number of T, we compare different numbers of time slots $T = 1\,000$ and $T = 10\,000$ with the same user demand $\bar{R}$.

In Fig. 4, we compare the simulated mean values of power level ν given different κ with the numerical results of power level obtained from Eq. (13) when $\bar{R} = 1, 4, 7$ Gbit/s, which represent the three typical cases in Fig. 2, where a larger value of $\bar{R}$ means more data need to be transmitted to the user before deadline. With other values of $\bar{R}$, the trends are similar and hence are not provided. We also provide the simulated standard deviations of the estimated power level, as shown with curves within the magnified window. It can be seen that the deviation from the mean value is small. This shows that the estimate of the power level is very accurate when the required deadline is 100 s or even 10 s.

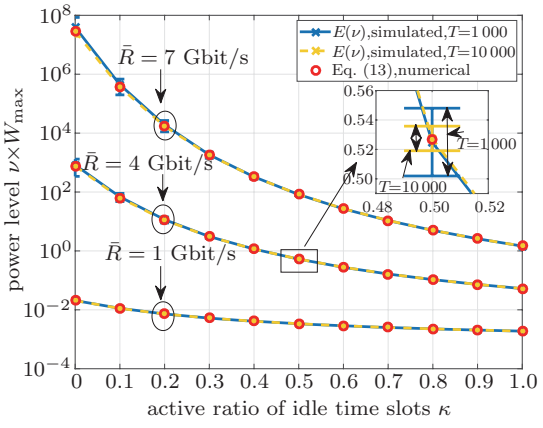


Figure 4 Power level ν vs. κ

In Fig. 5, we simulate the expectation of the average total power consumption over T time slots with different κ. We also provide the simulated standard deviations of the estimated total power consumption in the magnified window. It can be seen that the deviation from the mean value is small, which indicates that when the deadline is 100 s or even 10 s, the estimation of the average total power consumption (and hence κ) can be very accurate.

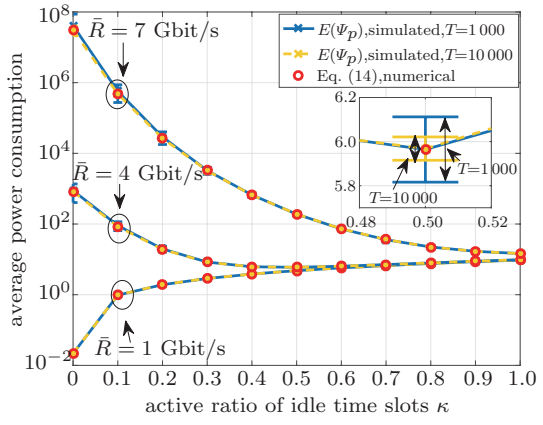


Figure 5 Average total power vs. κ

4.2 Performance evaluation

To evaluate the performance of the proposed predictive resource allocation, we compare the following strategies (the first two strategies are non-predictive) via simulation, first considering perfect value of P_l.

- SE-maximizing (with legend "SE"): the BS transmits with maximal residual resources W^t and

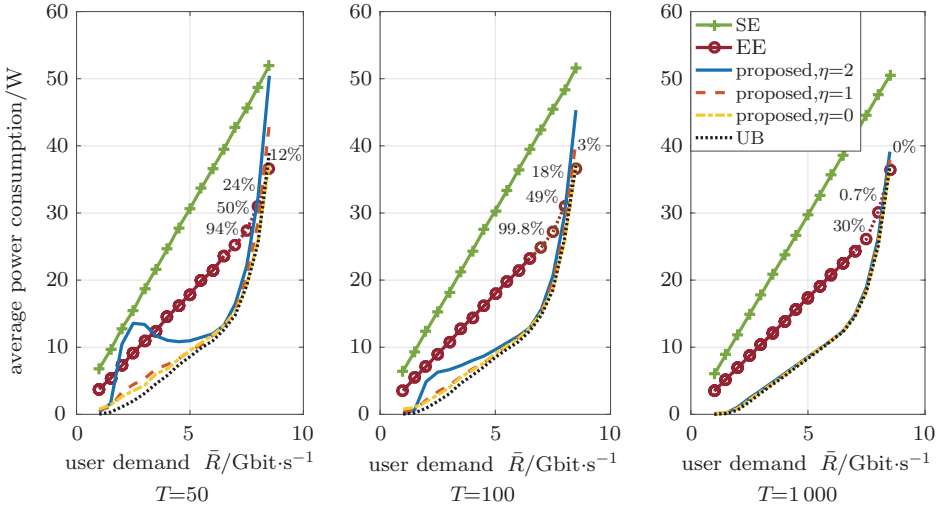


Figure 6 Average power consumption vs. $\bar{R}$. P_l is perfectly known. The probability marked is the completion rate of "EE"

$p^t_{\max}$, then the spectral efficiency in each time slot is maximized.

- EE-maximizing (with legend "EE"): According to the deadline and the remaining data to be transmitted of the user, the BS can obtain an expected time-average transmit rate in the remaining duration as R^t_0. To fully use the residue bandwidth, the BS serves the user with $p^t_{\max}$ when $W^t < W_{\max}$. In the case where the BS needs to consume extra circuit power (i.e., $W^t = W_{\max}$), the BS serves the user according to the principle of EE-optimal resource allocation (i.e., maximizing the instantaneous EE in each time slot while ensuring the QoS required by the user), whose optimal transmit power can be found from $p^{t*} = \arg\max \dfrac{R^t}{p^t_{\text{Ex}}}, R^t \geqslant R^t_0, p^t \geqslant 0, p^t \leqslant p^t_{\max}$.
- Resource allocation with future instantaneous residual bandwidth information (with legend "UB"): The BS transmits with the optimal resource allocation obtained from problem (4).
- Resource allocation with residual bandwidth probability (with legend "Proposed"): The predictive resource allocation proposed in subsection 3.4.

In Fig. 6, we compare average total power consumption of different strategies. It can be seen that the proposed policy almost overlaps with "UB", i.e., it can almost achieve all the energy saving potential when $T = 1\,000$. The fluctuation in the result for the proposed policy with $\eta = 2$ comes from the inaccurate estimate on the variance σ_w when T is short. When $\bar{R}$ is not very large, the energy saving gain of predictive resource allocation is remarkable over both "SE" and "EE" strategies that are non-predictive (say, about 100% gain when $\bar{R} < 7$ Gbit/s).

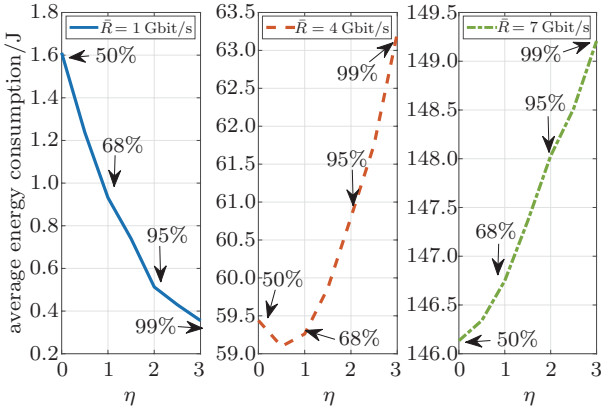


Figure 7 Average energy consumption vs. η. P_l is known, $T = 1\,000$. The probabilities marked on the curves are corresponding target completion probability

In Fig. 7, we show the average total energy consumption versus η. When $\bar{R}$ is low, high completion probability and low energy consumption can be achieved at the same time. This is because in this case the bits failed to be transmitted before the deadline consume more energy since circuit power consumption dominates the total consumption. Increasing the power level can efficiently reduce the

amount of bits failing to transmit before deadline, which can also reduce the circuit power consumption for delivering these bits left. However, when $\bar{R}$ is high, the transmit power dominates, there is a tradeoff between completion probability and energy consumption.

4.3 Impact of prediction granularity and errors

The reported prediction horizon for traffic volume in the literature ranges from 5 min[10] to one day[9], and the corresponding prediction granularity ranges from 5 min to 1 h. The prediction granularity is the duration of a period that a traffic load predictor outputs a predicted value, which reflects the prediction resolution. While the big data analysis indicates that the overall traffic volume variation of multiple BSs exhibits strong periodic trend in each day, the traffic loads at each BS and at a fine time granularity are observed as random and non-stationary. Considering that an implicit assumption in deriving the residual bandwidth probability P_l is that the traffic is stationary within the T time slots, and within a sufficient small prediction granularity, a non-stationary random process can be approximated as stationary, a high prediction resolution will yield a more accurate prediction. However, it also leads to high computational complexity due to frequently making the prediction and requires large storage space at the BS. This naturally raises the following question: how large the prediction granularity should be for a given duration of T time slots?

We first show whether or not a finer prediction granularity than T is necessary. To reflect the impact of non-stationary traffic load within the T time slots, we artificially generate two extreme cases from the original simulation setup, as shown in Fig. 8, where the residual bandwidth probabilities for the three cases are all equal to P_l.

Case 1: Residual bandwidth in the T time slots changes randomly according to the Poisson arrival of RT traffic as in the original simulation setup.

Case 2: Residual bandwidth of case 1 are reordered in an ascending order, while keeping the same P_l.

Case 3: Residual bandwidth of case 1 are reordered in a descend order, while keeping the same P_l.

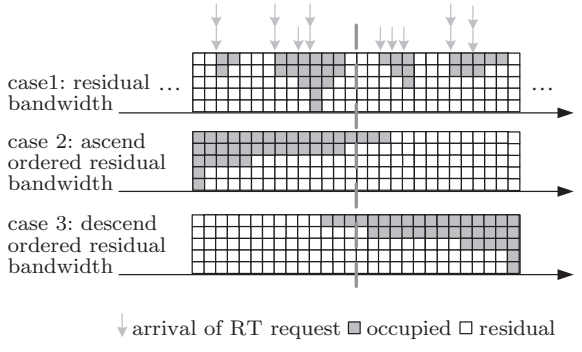


Figure 8 Illustration of one snap-shot of the residual bandwidth for the three cases

The traffic loads in cases 2 and 3 are highly non-stationary. If the prediction granularity is much smaller than T, then the traffic loads can be approximately regarded as stationary within the period of the prediction granularity.

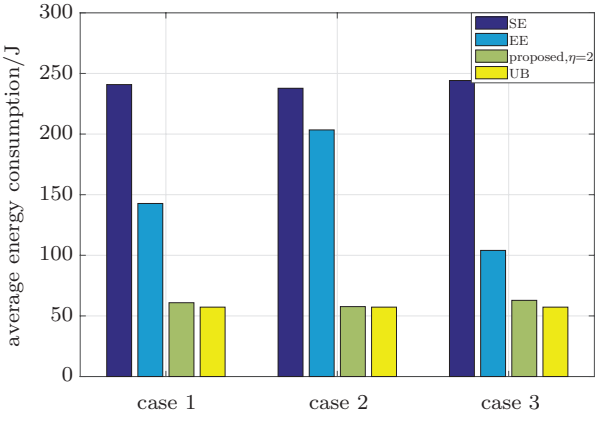


Figure 9 Average total energy consumption in the three cases. $\bar{R} = 4$ Gbit/s

To show if predicting in a smaller granularity than T is necessary, in Fig. 9 we simulate the average total energy consumption achieved by the four strategies for the three cases when the prediction granularity is set as T time slots. It can be observed that the non-stationary traffic has little impact on the per-

formance. This suggests that the predictive resource allocation does not need higher prediction resolution.

On the other hand, if the prediction granularity is longer than the duration of making the predictive resource allocation, say the prediction granularity is $2T$ or even hundred times of the T time slots, then there may be large deviation between the actual traffic load and the predicted traffic load. Moreover, even if the prediction granularity is identical to T, there inevitably exist prediction errors in practice.

In Fig. 10, we simulate the average total energy consumption achieved by the four strategies when the predicted traffic volume is inaccurate, which may be caused by either larger prediction granularity than T or prediction errors or both. The actual average arrival rate is $\lambda = 0.2$ requests per time slot, and the predicted average arrival rate is a constant changing from 0 to 0.4 requests per time slot, i.e., the largest prediction error is 100%. Considering that the maximal prediction error of traffic volume in every 5 min is within 20% as reported in Ref. [10], i.e., from 0.16 to 0.24 requests per time slot, the results show that the energy saving gain of the proposed policy is robust to the prediction errors on traffic loads.

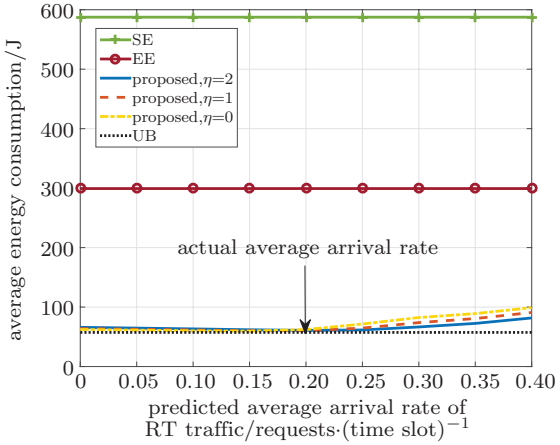


Figure 10 Average total energy consumption vs. prediction deviation of traffic load. $\bar{R} = 4$ Gbit/s, $T = 1\,000$

The results in Figs. 9 and 10 indicate that the prediction granularity should be identical to the required deadline of the NRT user for predictive resource allocation.

5 Conclusion

In this paper, we showed how to translate the predictable traffic load to the useful information in resource allocation by taking an energy-saving predictive resource allocation as an example. By formulating and solving an energy minimizing problem for a non-real-time user with future instantaneous information, we revealed that the optimal resource allocation only depends on two key parameters, which can be estimated from the residual bandwidth probability accurately when the transmission delay allowed by the NRT user is large. We derived the closed-form expression of the optimal resource allocation when the delay approaches infinity, and provided a heuristic resource allocation policy to guarantee a target transmission completion probability for practice use. The simulation results validated our analysis, showed remarkable energy save gain from the predictive policy, and illustrated the impact of of prediction granularity and errors in traffic load on the performance of the predictive resource allocation.

Appendix

A) Proof of Proposition 1.

Eq. (9) can be rewritten as

$$\frac{\bar{R}}{\ln(h\nu/N_0)} = \frac{1}{T}\sum_{\mathcal{N}\cup\mathcal{T}_{\mathrm{bu}}} W^t. \tag{21}$$

When $T \to \infty$, according to central limit theorem, $\frac{1}{T}\sum_{\mathcal{N}\cup\mathcal{T}_{\mathrm{bu}}} W^t = \frac{1}{T}\sum_{t=1}^{T} m^t W^t$ asymptotically follows Gaussian distribution $\mathbb{N}(\mu_w, \sigma_w^2)$, where μ_w and σ_w^2 are respectively the mean value and the variance of $\frac{1}{T}\sum_{t=1}^{T} m^t W^t$.

From Eq. (8), it can be seen that $p^t > 0$, i.e., $m^t = 1$ for all the busy time slots with non-zero residual bandwidth. For the κT_{id} activated idle time slots, $m^t = 1$. Then, the mean value of $\frac{1}{T}\sum_{t=1}^{T} m^t W^t$ can be derived as

$$\mu_w = \mathbb{E}\{m^t W^t\} = \sum_{l=1}^{L-1} P_l \frac{l}{L} W_{\max} + \kappa P_L W_{\max} \tag{22}$$

and the variance can be derived as

$$\sigma_w^2 = \mathbb{V}ar\Big\{\frac{1}{T}\sum_{t=1}^{T} m^t W^t\Big\} = \frac{1}{T}\mathbb{V}ar\{m^t W^t\}$$
$$= \frac{1}{T}\Big(\sum_{l=1}^{L-1} P_l \frac{l^2}{L^2} W_{\max}^2 + \kappa P_L W_{\max}^2\Big)$$
$$-\frac{1}{T}\Big(\Big(\sum_{l=1}^{L-1} P_l \frac{l}{L} W_{\max} + \kappa P_L W_{\max}\Big)^2\Big). \quad (23)$$

B) Proof of Proposition 2.

Denote $X \triangleq \frac{1}{T}\sum_{\mathcal{N}\cup\mathcal{T}_{\text{bu}}} W^t$ and $Y \triangleq \frac{1}{T}\sum_{\mathcal{N}} 1$. Then, Ψ_p can be written as follows,

$$\Psi_p = \frac{1}{\xi}\frac{N_0}{h} X\left(\exp\left(\frac{R}{X}\right) - 1\right) + Y(p_{\text{act}} - p_{\text{sle}}). \quad (24)$$

When $T \to \infty$, according to central limit theorem, X asymptotically follows Gaussian distribution $\mathbb{N}(\mu_w, \sigma_w^2)$ as in Eq. (12) and Y asymptotically follows Gaussian distribution $\mathbb{N}(\kappa P_L, \frac{1}{T}(\kappa P_L - (\kappa P_L)^2))$. Denote the probability density functions of X as $f_x(X)$. Then, the expectation of Ψ_p can be obtained as

$$\mu_{\Psi_p} = \int_{-\infty}^{\infty} \frac{1}{\xi}\frac{N_0}{h} X\left(\exp\left(\frac{R}{X}\right) - 1\right) f_x(X)\mathrm{d}X$$
$$+\kappa P_L(p_{\text{act}} - p_{\text{sle}}).$$

Since X converges to μ_w when $T \to \infty$, the average power consumption when $T \to \infty$ can be obtained as

$$\lim_{T\to\infty} \mu_{\Psi_p} = \frac{1}{\xi}\frac{N_0}{h}\mu_w\left(\exp\left(\frac{R}{\mu_w}\right) - 1\right)$$
$$+\kappa P_L(p_{\text{act}} - p_{\text{sle}}). \quad (25)$$

In a similar way, we can show that the variance of Ψ_p satisfying $\lim_{T\to\infty}\sigma_{\Psi_p}^2 = 0$.

References

[1] N. Bhushan, J. Y. Li, D. Malladi, et al. Network densification: the dominant theme for wireless evolution into 5G [J]. IEEE commun. mag., 2014, 52(2): 82-89.

[2] C. Jardak, P. Mähönen, J. Riihijärvi, et al. Spatial big data and wireless networks: experiences, applications, and research challenges [J]. IEEE network, 2014, 28(4): 26-31.

[3] C. M. Song, Z. H. Qu, N. Blumm, et al. Limits of predictability in human mobility [J]. Science, 2010, 327(5968): 1018-1021.

[4] R. Atawia, H. Abou-zeid, H. S. Hassanein, et al. Robust resource allocation for predictive video streaming under channel uncertainty [C]//IEEE Global Communications Conference, Austin, USA, 2014: 4683-4688.

[5] Z. Lu, G. De Veciana. Optimizing stored video delivery for mobile networks: the value of knowing the future [C]//The 32nd IEEE International Conference on Computer Communications, Turin, Italy, 2013: 2706-2714.

[6] N. Bui, J. Widmer. Mobile network resource optimization under imperfect prediction [C]//IEEE 16th International Symposium on A World of Wireless, Mobile and Multimedia Networks, Boston, USA, 2015: 1-9.

[7] C. T. Yao, C. Y. Yang, Z. X. Xiong. Energy-saving predictive resource planning and allocation [J]. IEEE Transactions on Communications, 2016, 64(12): 5078-5095.

[8] H. Abou-Zeid, H. S. Hassanein. Predictive green wireless access: exploiting mobility and application information [J]. IEEE wirel. commun., 2013, 20(5): 92-99.

[9] R. P. Li, Z. F. Zhao, X. Zhou, et al. Energy savings scheme in radio access networks via compressive sensing-based traffic load prediction [J]. Transactions on emerging telecommunications technologies, 2014, 25(4): 468-478.

[10] M. Mardani, G. B. Giannakis. Estimating traffic and anomaly maps via network tomography [J]. IEEE/ACM transactions on networking, 2016, 24(3): 1533-1547.

[11] A. Nadembega, A. Hafid, T. Taleb. A destination and mobility path prediction scheme for mobile networks [J]. IEEE transactions on vehicular technology, 2015, 64(6): 2577-2590.

[12] Y. Shi, M. Larson, A. Hanjalic. Collaborative filtering beyond the user-item matrix: a survey of the state of the art and future challenges [J]. ACM computing surveys, 2014, 47(1): 3.

[13] E. Oh, B. Krishnamachari, X. Liu, et al. Toward dynamic energy-efficient operation of cellular network infrastructure [J]. IEEE communications magazine, 2011, 49(6): 56-61.

[14] G. Auer, V. Giannini, C. Desset, et al. How much energy is needed to run a wireless network? [J]. IEEE wireless communications, 2011, 18(5): 40-49.

[15] S. K. Das, S. K. Sen, K. Basu, et al. A framework for bandwidth degradation and call admission control schemes for multiclass traffic in next-generation wireless networks [J]. IEEE journal on selected areas in communications, 2003, 21(10): 1790-1802

[16] A. O. Allen. Probability, statistics, and queueing theory: with computer science applications [M]. San Diego: Gulf Professional Publishing, 1990.

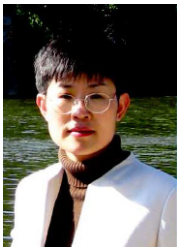
[17] S. Boyd, L. Vandenberghe. Convex optimization [M]. Cambridge: Cambridge University Press, 2004.

[18] R. M. Corless, G. H. Gonnet, D. E. Hare, et al. On the Lambert W function [J]. Advances in computational mathematics, 1996, 5(1): 329-359.

[19] TR 36.814 V1.2.0. Further advancements for E-UTRA physical layer aspects (Release 9) [S]. 3GPP, 2009.

About the authors

Chuting Yao received the B.S. degree from the School of Advanced Engineering, Beihang University (BUAA), Beijing, China, in 2011. She is currently pursuing the Ph.D. degree with the School of Electronics and Information Engineering, BUAA. Her research interests lie in green radio and big data. (Email: ctyao@buaa.edu.cn)



Chenyang Yang [corresponding author] is a full professor with the School of Electronic and Information Engineering, Beihang University. Her recent research interests include wireless big data, green radio, and tactile internet. (Email: cyyang@buaa.edu.cn)

Chih-Lin I received her Ph.D. degree in electrical engineering from Stanford University. She is the China Mobile Chief Scientist of Wireless Technologies, in charge of advanced wireless communication R and D effort of China Mobile Research Institute. (Email: icl@chinamobile.com)