Proactive Resource Allocation Planning with Three-levels of Context Information

Jia Guo, Chuting Yao and Chenyang Yang

Beihang University, Beijing, China
Email: {guojia, ctyao, cyyang}@buaa.edu.cn

***Abstract*—Big data analysis makes predicting the application, network and user levels of context information possible. Yet it is unclear how to exploit these information to utilize the wireless resources more efficiently. In this paper, we attempt to illustrate the potential of using these information to improve the spectrum usage efficiency. To this end, we assume that a central unit in the multicell network can predict the mobile users' request, base stations' congestion status, and user mobility pattern within a prediction window. To fully use the excess resource within the window and leave more resources for the unpredictable traffic arrived after the window, we formulate a proactive resource allocation planning problem to minimize the maximal transmission completion time. A heuristic low complexity algorithm is introduced to find the transmission plan from the problem, which determines where, when and what to transmit to the users. We use two representative scenarios to demonstrate the performance gain of the proactive resource allocation using context information over the reactive scheme that the transmission starts after the users' requests truly arrive.**

I. Introduction

To support the explosively growing traffic demands, various new techniques are under investigation for the fifth generation (5G) cellular networks. Except the update of network architecture, another main trend is to provide higher throughput, say by network densification [1]. While further improving spectral efficiency is always beneficial, in real-world networks the bandwidth resource is often not fully used because of the time-varying traffic pattern. According to the measurement on available spectrum and traffic load, many base stations (BS) have a large amount of excess resource during the off-peak time, while the BSs in the hotpot are very busy and even become congested during the peak time.

A recent report in *Science* magazine indicates that human behavior is highly predictable [2]. With big data analysis, network resource usage status can be estimated by predicting a traffic map [3, 4]. Besides, the mobility pattern can be predicted by analyzing the historic user behavior [5, 6], at least within a prediction window, from which the average channel gains can be obtained with the help of a radio map [7]. Moreover, the content popularity, and even the preferred content of an individual user, is possible to be known before the user(s) truly initiates the request by using the collaborative filtering, which has long been studied in various recommendation systems [8]. Undoubtedly, predicting human behavior with wireless big data is rather challenging. This naturally raises the following question: can such valuable information gained from the prediction be exploited to improve the usage efficiency of wireless network resources, say energy and bandwidth?

With predicted content popularity, local caching at the wireless edge can reduce the backhaul cost, offload the traffic in core and access networks, improve user experience and energy efficiency [9–11]. Yet how other information that are able to be predicted could impact the wireless resource management is largely unexplored. If big data analysis can be made in a central unit (CU) connected to BSs with strong computing capability for predicting the network resource usage state, mobile user trajectory, the content to be requested and the request arrival time, then how these information can be exploited for improving the performance of wireless networks? In fact, with the ability of prediction endowed by big data, many factors affecting the network performance such as request arrival, network status and user locations that have long been regarded as random will become deterministic to a large extent. With these information, called application level, network level, and user level context information [12], and noticing the fact that today's smart phones have large storage size for caching requested files, a resource allocation plan can be made for each user before transmission, including which BSs along the trajectory of a mobile user should pre-download a file to the user, in which duration and with how many resources. In this way, the excess resources in the network can be exploited. Such a concept has been proposed to save energy consumed at the BSs, with the notion of *predictive* or *proactive* resource allocation in the literature [7, 13–15].

In this paper, we attempt to show the possible gain of improving spectrum utilization efficiency by leveraging the prediction ability of big data analysis. To this end, we assume that the three levels of context information is perfectly known, although the prediction is never perfect. We formulate a resource allocation planning optimization problem for pre-downloading the files to be requested to users, which optimizes the transmission duration at each BS along the trajectory of multiple mobile users to minimize the maximal transmission completion time. To find the solution with affordable computational complexity, a heuristic algorithm is then proposed. By

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

providing simulation results for two representative scenarios, we illustrate that the proactive resource allocation can provide a promising new way to support the explosive traffic demands, alternative to increasing the network capacity by deploying more bandwidth and antenna resources.

II. System Model

Consider a multi-cell network with N_b BSs, where each BS is equipped with N_t antennas and transmits in a time-slotted fashion. The BSs serve two classes of users, one requesting real-time (RT) service such as phone call, the other requesting content delivery such as file downloading. The requests of both classes of users arrive randomly. The RT service is served with high priority and hence with reserved resources, and the content delivery can only use the residual resources at each BS, which are random and time-varying.

In this paper, we are concerned with the resource allocation for the mobile users (MSs) that demand for content delivery, each requests one file with size of B bits.

A. Context Information

All BSs are connected to a CU. Assume that the CU can predict some *statistical* information within a prediction window with length of T_f frames as follows. (1) The request arrival time and the requested file of every MS, i.e., the *application level context information*. (2) The trajectory of each MS. With radio map [7], the CU can obtain the large scale channel gains for each MS, i.e., the *user level context information*. (3) The average residual resources (say bandwidth) remained at each BS after serving the RT traffic, i.e., the *network level context information*.

After predicting the context information, the CU can make a resource allocation plan for conveying the files that the MSs will request, called a *transmission plan*, which determines where, when, what, and with how much resources to transmit. Then, the CU informs the BSs along the trajectory of each MS. The BSs can pre-download the required files to the MSs before they initiate requests, and continue to transmit the remaining files (if some files have not been conveyed completely) after the MSs's requests arrive, according to the plan. In this way, the experience of the MSs can be dramatically improved, and the excess resource of the network can be fully used to alleviate the congestion in peak time. This is sharply different from the traditional transmission mechanism, where the MSs are served with best effort after their requests arrive.

B. Channel Model and Achievable Rate

To reflect the variation of the path-loss and shadowing due to user mobility, we assume that the large scale fading gains remain constant within each frame and may vary among frames. Each frame includes T_s time slots. The small scale fading is assumed as block fading, which remains constant in each time slot and varies among time slots independently, as shown in Fig. 1.

For mathematical tractability, assume that only the closest BS to a MS pre-downloads (or transmits) the file to be

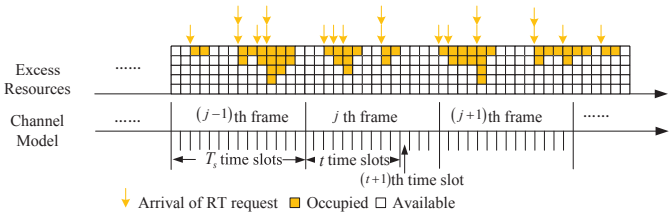

Fig. 1. Illustration of traffic and channel model.

requested (or having been requested) to the MS. The residual bandwidth and transmit power available for the kth MS (denoted as MS_k) in the tth time slot of jth frame at the closest BS are denoted as $W_{j,k}^t$ and $p_{\max,j,k}^t$, respectively.

Denote $m_{j,k}^t \in \{1, 0\}$ as an indicator of transmission plan made for the kth MS. When $m_{j,k}^t = 1$ or 0, the file to be requested by MS_k *is* or *is not* pre-downloaded by its closest BS in the tth time slot of the jth frame. According to the plan, a BS may need to pre-download files to multiple MSs. To avoid multi-user interference, the BS transmits to the MSs in different time slots. Then, the received signal of MS_k in the tth time slot of the jth frame is

$$y_{j,k}^t = m_{j,k}^t \sqrt{\alpha_k^j} (\mathbf{h}_{j,k}^t)^H \mathbf{w}_{j,k}^t \sqrt{p_{\max,j,k}^t} x_{j,k}^t + n_{j,k}^t, \quad (1)$$

where $x_{j,k}^t$ is the transmit symbol with $\mathbb{E}\{|x_{j,k}^t|^2\} = 1$, $\mathbf{h}_{j,k}^t \in \mathbb{C}^{N_t \times 1}$ is the independent and identically distributed (i.i.d.) Rayleigh fading channel vector, α_k^j is the corresponding large scale fading gain between the MS and the closest BS, $\mathbf{w}_{j,k}^t \in \mathbb{C}^{N_t \times 1}$ is the beamforming vector, $n_{j,k}^t$ is the noise with variance σ^2, and $\mathbb{E}\{\cdot\}$ represents expectation. Since MS_k is scheduled only by one BS in each time slot, maximum ratio transmission (MRT) is optimal, i.e., $\mathbf{w}_{j,k}^t = \mathbf{h}_{j,k}^t / \|\mathbf{h}_{j,k}^t\|$, where $\|\cdot\|$ denotes Euclidean norm.

In the tth time slot of the jth frame, the achievable rate of MS_k is

$$R_{j,k}^t = W_{j,k}^t \log_2(1 + g_{j,k}^t p_{\max,j,k}^t), \quad (2)$$

where $g_{j,k}^t \triangleq \alpha_k^j \|\mathbf{h}_{j,k}^t\|^2 / \sigma^2$ is the equivalent channel gain.

III. Making the Transmission Plan

Denote the start time instant of the prediction window as the 1st time slot of the 1st frame. At this start time, the CU can predict the requests of multiple MSs that will randomly arrive at different cells in the network by the end of the window, i.e., the T_sth time slot of the T_fth frame. Then, the CU can make the resource allocation plan for pre-downloading files to these MSs and inform the corresponding BSs. Note that the CU does not know the traffic load and network resource usage status after the prediction window. To leave more resources to the upcoming MSs whose requests may arrive after the window, the plan is made to minimize the maximal transmission completion time, as detailed later. In this way, the network throughputs can be improved by proactively avoiding the possible congestion in the future time.

For easy understanding, in what follows, we first formulate a minimax problem to optimize the resource allocation plan, under the assumption that the small scale channels gains and

instantaneous residual resources within the prediction window are known, in addition to the context information. Then, we provide an algorithm to solve a more viable minimax problem, assuming that only the context information is known.

A. Problem Formulation

With the informed transmission plan, each BS pre-downloads the files to the MSs entered into its coverage at the start time of the prediction window, before they initiate their requests. We call the duration from the start time of the window to the time instant that a file has been completely conveyed as the *transmission completion time* for a MS. To exploit the excess resources in the network, we minimize the maximal transmission completion time of all the MSs, whose requests arrive within the predicted window.

Denote K as the number of MSs who will initiate a request during the predicted window, and the t_kth time slot of the J_kth frame as the time instant that the transmission procedure for MS$_k$ is finished. Then, the transmission completion time for MS$_k$ is $(J_k-1)T_s+t_k$. The minimax resource allocation planning problem can be formulated as follows,

$$\min_{\mathbf{M}_1,\ldots,\mathbf{M}_K} \max \ (J_k-1)T_s+t_k \tag{3a}$$

$$s.t. \ \sum_{j=1}^{J_k-1}\sum_{t=1}^{T_s} m_{j,k}^t R_{j,k}^t \Delta_t + \sum_{t=1}^{t_k} m_{J_k,k}^t R_{J_k,k}^t \Delta_t \geq B, \tag{3b}$$

$$J_k \leq T_f, t_k \leq T_s, k=1\ldots,K, \tag{3c}$$

$$\sum_{k\in\mathcal{K}_i^t} m_{j,k}^t \leq 1, i=1,\ldots,N_b, \tag{3d}$$

where $\mathbf{M}_k=[\mathbf{m}_{1,k},\ldots,\mathbf{m}_{T_f,k}]$, $\mathbf{m}_{j,k}=[m_{j,k}^1,\ldots,m_{j,k}^{T_s}]^H$ is the indicator vector of the transmission plan for MS$_k$ in the jth frame, (3b) ensures that the amount of data able to be transmitted within the planned transmission time exceed the size of the file, (3c) reflects the requirement that the transmission should be completed within the predicted window, and (3d) is the interference-free constraint that each BS only transmits to a single MS in each time slot, $\mathcal{K}_i^t$ is the set of MSs that enter the coverage of the ith BS in the tth time slot, $i=1,\ldots,N_b$, and Δ_t is the duration of each time slot.

Problem (3) is an integer programming with prohibitive search space of $\mathcal{O}(K^{T_sT_f})$. Moreover, the normalized channel gain and residual bandwidth in future time slots are required to solve the problem, which can not be predicted accurately if not impossible. To make the problem viable that requires only average channel and network status information, we introduce an extra constraint $m_{j,k}^t=m_{j,k}, t=1,\ldots,T_s$, which means that the BS schedules MSs over frames instead of time slots. Then the indicator matrix $\mathbf{M}_k$ is simplified to a vector as $\mathbf{m}_k=[m_{1,k},\ldots,m_{T_f,k}]^H$. In addition, by assuming $T_s\to\infty$ and the small scale channel gains, residual bandwidth and transmit power in time slots as ergodic, we can approximate the left hand side of (3b) as

$$\sum_{j=1}^{J_k} m_{j,k}\sum_{t=1}^{T_s} R_{j,k}^t\Delta_t \approx \sum_{j=1}^{J_k} m_{j,k}\mathbb{E}\{R_{j,k}^t\}T_s\Delta_t \tag{4}$$

where the average is taken over small scale channel gains, residual bandwidth and transmit power. For notational simplicity, denote $\bar{R}_{j,k}\triangleq\mathbb{E}\{R_{j,k}^t\}$.

Then, the minimax transmission planning can be found from the following problem,

$$\min_{\mathbf{m}_1,\ldots,\mathbf{m}_K} \max \ J_kT_s \tag{5a}$$

$$s.t. \ \sum_{j=1}^{J_k} m_{j,k}\bar{R}_{j,k}T_s\Delta_t \geq B, \tag{5b}$$

$$J_k \leq T_f, k=1\ldots,K, \tag{5c}$$

$$\sum_{k\in\mathcal{K}_{i,j}} m_{j,k} \leq 1, i=1,\ldots,N_b, \tag{5d}$$

where $\mathcal{K}_{i,j}$ is the set of MSs that enter the coverage of the ith BS in the jth frame. Problem (5) only needs the user and network levels of context information in the prediction window.

The search space of this problem is $\mathcal{O}(K^{T_f})$, which however is still too large to find the optimal solution.

B. Transmission Planning Algorithm

Introduce a variable $J\triangleq\max J_k$. Then, finding the solution of problem (5) is equivalent to finding the minimal value of J that makes the following problem feasible,

$$\min_{J,\mathbf{m}_1,\ldots,\mathbf{m}_K} J \tag{6a}$$

$$s.t. \ \sum_{j=1}^{J} m_{j,k}\bar{R}_{j,k}T_s\Delta_t \geq B, \tag{6b}$$

$$J \leq T_f, \tag{6c}$$

$$\sum_{k\in\mathcal{K}_{i,j}} m_{j,k} \leq 1, i=1,\ldots,N_b. \tag{6d}$$

Since the values of J is finite within $\{1,\ldots,T_f\}$, the key of solving problem (5) is to find whether problem (6) is feasible for a given J. If a group of planning indicators for all users, $\mathbf{m}_1,\ldots,\mathbf{m}_K$, can be obtained for a given value of J, problem (6) is feasible for the given J, and then the feasible solution with minimal value of J is the final solution.

The transmission plans of multiple MSs are coupled, since allocating a frame to one MS may affect the plans of other MSs. To find the feasibility of problem (6) efficiently, in what follows we propose a heuristic algorithm, where the transmission plan of each MS is sequentially designed and the order to make the plans is founded by a sort of branch and bound method.

For a given value of J, after obtaining the feasibility solutions of the former $k-1$ MSs, the CU finds the feasi-

ble solution of the following problem, which minimizes the number of frames occupied by the kth MS, i.e.,

$$\min_{\mathbf{m}_k} \sum_{j=1}^{J} m_{j,k} \tag{7a}$$

$$s.t. \sum_{j=1}^{J} m_{j,k} \bar{R}_{j,k} T_s \Delta_t \geq B, \tag{7b}$$

$$\sum_{[k] \in \mathcal{K}_{i,j} \bigcap \{[1],\ldots,k\}} m_{j,k} \leq 1, i = 1, \ldots, N_b. \tag{7c}$$

where constraint (7b) ensures that B bits can be conveyed within J frames, and constraint (7c) ensures that the kth MS does not use the resources already occupied by the former $k-1$ MSs. After finding the feasible solutions from a series of K problems like this for every J among $\{1, \ldots, T_f\}$, the CU can obtain the plan for the kth user from the feasible solutions with minimal value of J.

When problem (7) is feasible, its solution is easy to find with closed-form expression, which however has complicated form and hence is not shown. In fact, the solution is simply transmitting at the remaining frames not selected by the former $k-1$ MSs with largest achievable rates, as illustrated as Fig. 2. The problem may become infeasible when too few resources left by the former MSs. In this case, the transmission of the B bits for the kth MS cannot be completed.

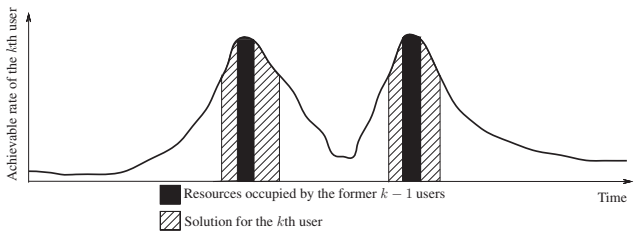

Fig. 2. Transmission plan for MS_k after some resources are occupied by the former $k-1$ MSs.

For any given value of J, the order of MSs for sequentially finding feasible solutions of problem (7) may lead to different feasible states. By exhaustive searching among all possible orders for the MSs, whether J is feasible for the K problems of (7) can be found. However, in the worst case, $K!$ orders need to be tried, which leads to high computational complexity.

To obtain a viable algorithm to judge the feasibility with given value of J, we introduce a branch-and-bound like method. The intuition behind the algorithm is that the selection of former MSs is more important than the later MSs in order to improve the completion rate. For example, if the 1st MS is properly selected, then more resources can be left for the 2nd MS and so on, and hence the transmission completion rate within the prediction window will be high. The completion rate is equal to the number of bits delivered by the feasible solutions of the K sequentially solved problem (7) divided by the overall KB bits.

We define the sequences of MSs starting with the same MS as a branch, and the corresponding sequences as subbranches (i.e., an order of MSs to sequentially solve problem (7)). Since some orders of MSs may make the problem infeasible while others may not, we introduce the completion rate as the bound to remove branches (and hence the corresponding subbranches). Instead of computing the completion rates of all its sub-branches, the bound of the completion rates of the subbranches of one branch is approximated by that of a randomly picked subbranch of the branch (because it is more important to select former MSs). After finding the branch with the largest completion rate, the starting MS is selected as the 1st MS in the order of MSs. Then, we treat the subbranches of this branch with the same 2nd MS as a "new" branch, and repeat previous steps to select the 2nd MS in the order. When the order for all the K MSs has been found, if the completion rate is less than one, the CU will regard problem (7) with this order of MSs as infeasible with given J.

To make the algorithm easy understanding, we provide an example in Fig. 3, where the requests of four MSs, user A, user B, user C and user D, arrive the network within the prediction window. There are four branches in total, each starting with different MSs, and each with six subbranches. Hence, there are overall 24 possible orders (sequences) of MSs. The algorithm is then implemented by the following steps.

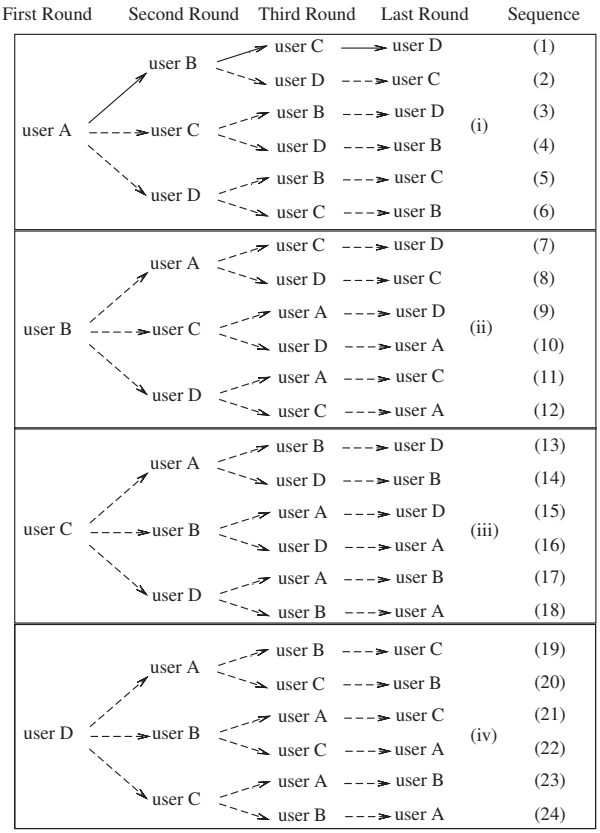

Fig. 3. Example of selecting order of MSs for any given value of J.

- **Step 1**: Randomly pick one subbranch in each branch, say subbranch (1), (7), (13) and (19), respectively from

the four branches. Then, judge whether problem (7) is feasible for the given value of J for each of the four orders of MSs.

- **Step 2**: If problem (7) is feasible for a branch, say branch (i), then J is feasible. If problem (7) is infeasible for all branches, then compute the completion rate for every branch, and pick up the branch with the largest completion rate.
- **Step 3**: After the branch with the largest completion rate is selected in the first round, say branch (i), go back to Steps 1 and 2, where there are three "new" branches each starting with user B, user C and user D, respectively. Iterate through the two steps 1 and 2 until the final order of MSs is obtained, say the 1st user sequence marked with "(1)" in the figure.

C. Transmission Policy for Pre-downloading

After the transmission plan is made for every MS in the 1st time slot of the first frame, the CU informs the corresponding BSs along the trajectory of each MS. When a mobile MS enters the coverage of a BS who is planned to serve the MS in the predetermined frames, the BS starts to estimate the instantaneous channel information of the MS, and transmits the file to the MS with MRT using the instantaneous residual transmit power and residual bandwidth subsequently in every time slot of every frames.

IV. Simulation Results

In this section, we evaluate the performance of the proposed resource allocation planning by simulations.

Consider a N_b-cell system with cell radius $D = 250$ m, where $N_b = 13$, and all BSs each equipped with $N_t = 6$ antennas are located along a straight line. Nine mobile MSs with speed v_k^t uniformly distributed in $(5, 25)$ m/s move along three roads of straight lines with minimum distance from the BSs as 50 m, 100 m and 200 m, respectively, as shown in Fig. 4. Each MS will request a file with $B = 30$ Mbytes [9] within a prediction window containing $T_f = 200$ frames. Each frame is with duration of one second, and contains $T_s = 100$ time slots, i.e., each time slot is with duration $\Delta_t = 10$ milliseconds. The MSs separately initiate their requests every 10 seconds starting from the 100th second in the prediction window. The maximal transmit power of each BS is 40 W and cell-edge SNR is set as 5 dB, where the intercell interference is implicitly reflected. The path loss model is $36.8 + 36.7 \log_{10}(d)$, where d is the distance between the BS and user in meter. To reflect the different resource usage status of the BSs by serving the RT traffic, we consider two types of BSs: idle BS with average bandwidth $\overline{W} = 1$ MHz and busy BS with average bandwidth $\overline{W} = 10$ MHz, which are alternately located along the line as idle, idle, busy, busy, idle, idle, and so on. The results are obtained from 100 Monte Carlo trails, where in each trail the trajectory of each user stays the same, while the small-scale fading channel is subject to i.i.d Rayleigh block fading and the bandwidth in each time slot uniformly varies with average value of $\overline{W}$.

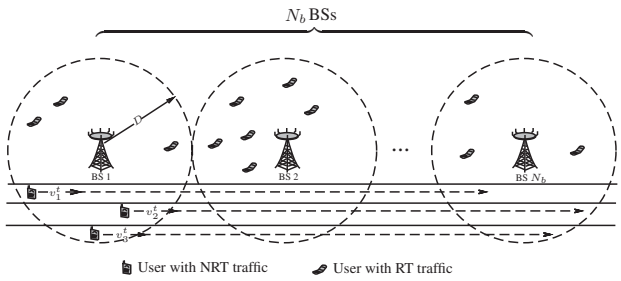

Fig. 4. Simulation setups, only three users with NRT traffic are shown in this figure for simplicity.

We use the minimal transmission completion time of all MSs as the performance metric. When this duration is short, the residual resources in the prediction window are fully used. Since context information is not obtained for free, a nature question is whether only one level of information can obtain most of the performance gain. Since user level and network level context information should be employed jointly to predict future data rate, the following transmission schemes are simulated.

- *Proactive resource allocation with three levels of context information* (with legend "All Context"): The CU makes a resource allocation plan for each MS arrived in the prediction window with three levels of context information by using the proposed algorithm. Then, the BSs transmit to the MSs according to the plan informed by CU.
- *Proactive resource allocation only with application level context information* (with legend "A Context"): The CU knows that several MSs will initiate their requests within the prediction window and knows the files that each MS will request. With this information at the start time of the window, the CU informs the BSs who are closest to the MSs to pre-download the files to the MSs before their requests actually arrive and continue to transmit if some files have not been completely conveyed after the requests arrive. The transmission before and after the requests arrive can be performed with best efforts, i.e., using all the instantaneous residual bandwidth and transmit power of the BS. When several MSs are in the same cell at the same time slot, the BS transmits to the MS who can achieve the highest data rate in the time slot.
- *Reactive resource allocation without context information* (with legend "No Context"): This is the traditional transmission scheme, where the transmission begins right after the requests truly arrive, again with best efforts.

To observe the gain brought by the prediction, we consider two representative application scenarios.

- *Scenario 1*: At the start time of the prediction window, the MSs who will initiate requests are located in the same cell, where the BS (called the 1st BS) may be in idle or busy state. When the MSs actually send their requests, they move into different cells due to the different speeds.
- *Scenario 2*: At the start time of the prediction window,

the MSs who will initiate requests are located in different cells. When the MSs actually send their requests, they enter into the same cell, where the BS (called the last BS) may be in idle or busy state.

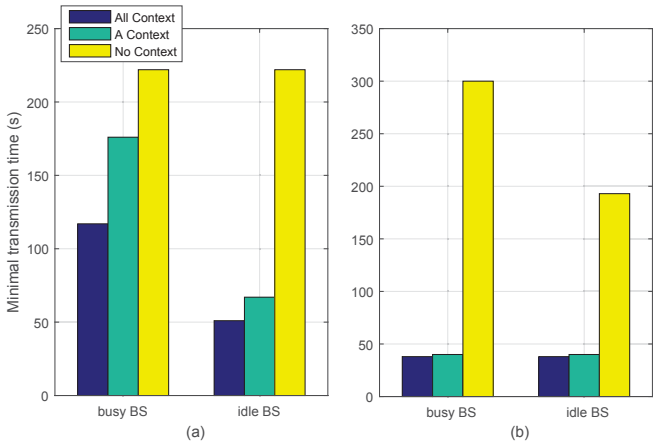

Fig. 5. Minimal transmission completion time of different transmission schemes. (a) Scenario 1, (b) Scenario 2. The legends "idle BS" and "busy BS" indicate the resource usage status of the 1st or last BS when multiple MSs are gathered.

The simulation results are provided in Fig. 5. It is shown from Fig. 5(a) that in scenario 1 "All Context" performs better than "A Context". This is because only with the application level context information the 1st BS always transmits to the MSs with good channels, which may miss the good channels or large residual bandwidth for other MSs. Both proactive resource allocation schemes exhibit remarkable gains over the reactive scheme, where the gains are even remarkable when the 1st BS is idle (more than three-fold performance gain in terms of reducing the transmission completion time). The gain comes from "serving" the MSs in advance before they start to request. This indicates that in this scenario, knowing all the context information is beneficial.

It is shown from Fig. 5(b) that "A Context" performs almost the same as "All Context" and both are much better than "No Context". The performance gain comes from offloading the traffic in the last BS where congestion will occur, by exploiting the context information for transmitting in advance. The proactive resource allocation schemes provide more than six-fold performance gain in terms of reducing the transmission completion time when the last BS is busy. Since the MSs naturally distributed in different cells at the start of the prediction window, the traffic is in fact automatically offloaded to different BSs simply by using the application level information. This indicates that only exploiting application level context information is sufficient in this scenario.

V. Conclusions

In this paper, we investigated the performance gain of proactive resource allocation by exploiting the application level, network level and user level information. We first formulated a resource allocation planning problem to minimize the maximal transmission completion time within a prediction window. We then provided a heuristic algorithm to find the solution. Simulation results illustrated that the proactive transmission mechanism can improve the spectrum usage efficiency by exploiting the residual resources in the network and provide substantial gain over the reactive transmission mechanism that starts after the users' requests truly arrive.

References

[1] N. Bhushan, J. Li, D. Malladi, R. Gilmore, D. Brenner, A. Damnjanovic, R. Sukhavasi, C. Patel, and S. Geirhofer, "Network densification: the dominant theme for wireless evolution into 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 82–89, Feb. 2014.

[2] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, Feb. 2010.

[3] Y. Zang, F. Ni, Z. Feng, S. Cui, and Z. Ding, "Wavelet transform processing for cellular traffic prediction in machine learning networks," in *IEEE ChinaSIP*, 2015.

[4] M. Mardani and G. B. Giannakis, "Estimating traffic and anomaly maps via network tomography," *IEEE Trans. Netw.*, vol. Early access, 2016.

[5] J. Froehlich and J. Krumm, "Route prediction from trip observations," Soc. Automotive Eng. World Congress, Tech. Rep., 2008.

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

[7] H. Abou-Zeid and H. S. Hassanein, "Toward green media delivery: location-aware opportunities and approaches," *IEEE Wireless Commun.*, vol. 21, no. 4, pp. 38–46, Aug. 2014.

[8] Y. Shi, M. Larson, and A. Hanjalic, "Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges," *ACM Comput. Surveys*, vol. 47, no. 1, pp. 1–45, 2014.

[9] N. Golrezaei, A. F. Molisch, A. G. Dimakis, and G. Caire, "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution," *IEEE Commun. Mag.*, vol. 51, no. 4, pp. 142–149, 2013.

[10] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5g wireless networks," *IEEE Commun. Mag.*, vol. 52, no. 8, pp. 82–89, Aug. 2014.

[11] D. Liu and C. Yang, "Energy efficiency of downlink networks with caching at base stations," *IEEE JSAC*, Apr. 2016.

[12] C. Park, Y. Seo, K. Park, and Y. Lee, "The concept and realization of context-based content delivery of NGSON," *IEEE Commun. Mag.*, vol. 50, no. 1, pp. 74–81, Jan. 2012.

[13] H. Abou-zeid and H. Hassanein, "Predictive green wireless access: exploiting mobility and application information," *IEEE Wireless Commun.*, vol. 20, no. 5, pp. 92–99, Oct. 2013.

[14] C. Yao, C. Yang, and Z. Xiong, "Power-saving resource allocation by exploiting the context information," in *IEEE PIMRC*, 2015.

[15] C. Yao, B. Chen, C. Yang, and G. Wang, "Energy-saving pushing based on personal interest and context information," in *IEEE VTC Spring, accepted*, 2016.