Virtual Spectrum Hole: Exploiting User Behavior-Aware Time-Frequency Resource Conversion

Hangguan Shan, *Member, IEEE,* Zhifeng Ni, Weihua Zhuang, *Fellow, IEEE,* Aiping Huang, *Senior Member, IEEE,* and Wei Wang, *Member, IEEE*

***Abstract*—In this paper, to address network congestion stemmed from traffic generated by advanced user equipments, we propose a novel network resource allocation strategy, time-frequency resource conversion (TFRC), via exploiting user behavior, a specific kind of context information. The key idea is to use radio resources mainly on the traffic/connection to which a user pays attention. The TFRC withdraws spectrum resources strategically from connection(s) not focused on by the user, providing re-useable spectrum called "virtual spectrum hole". Considering an LTE-type cellular network, a double-threshold guard channel policy is proposed to facilitate the implementation of TFRC. An analytical model is established to study benefits of exploiting TFRC in terms of call-level performance, including new call blocking, handoff call dropping, and recovering call dropping probabilities. Numerical results demonstrate the effectiveness of the proposed approach, in increasing the cell capacity (maximum user number per cell) while limiting potential service quality degradation introduced by the newly proposed technique.**

***Index Terms*—Time-frequency resource conversion, context-aware resource allocation, user behavior, virtual spectrum hole, call admission control, call-level performance.**

I. Introduction

THE rapid development of mobile broadband technologies such as WiFi and Long-Term-Evaluation (LTE) has offered great convenience and comfort to our daily life. With an increasing demand for various services using limited radio resources, mobile networks can be overloaded from time to time [1], thus degrading user quality of experience (QoE). One of the main reasons for the situation is related to the rapidly growing usage of advanced user equipment (UE) (e.g., smartphones, netbooks, and tablet devices) [2,3]. It has been reported that, in 2012, 92 percent of total global handset traffic was already represented by smartphones [1]. Furthermore, users of advanced UE have been observed not only generating a higher average load but also spending a longer time with their devices than with traditional cellphones [4,5]. Thus, there is a trend that the more the users of advanced UEs, the heavier the radio resource consumption, leading to a more frequent occurrence of spectrum starvation and thus network congestion. It is therefore important to develop radio resource management strategies tailored for advanced UE proliferated scenarios.

To manage radio resources in a congested network, *differential service strategies* are adopted popularly. To increase the overall network bandwidth efficiency, when the network load tends to be heavy and high-priority services demand more radio resources, such a strategy usually imposes quality-of-service (QoS) degradation to low-priority services or the users with a specific traffic type. For example, to accommodate more calls in an overloaded system, differential service oriented call admission control schemes (e.g., subrating [6], service degradation [7], and bandwidth stealing [8]) force some ongoing low-priority calls to operate under a degraded service mode. Similarly, when the network is congested due to some users with heavy traffic, differential service oriented congestion control can penalize these users by imposing high fees on them, thus alleviating the problem of congestion and potentially generating higher revenue for network operators (e.g., [9,10]). The benefits of differential service strategies can be ascribed to an appropriate compromise between radio resource limitation and service demand.

However, directly applying differential service strategies in an advanced UE proliferated scenario can be ineffective and/or inefficient. There is an important difference between an advanced UE proliferated network and a network only having traditional cellphones. A device in the former network usually supports multitasking functions by relevant platforms (e.g., iOS and Android), while a device in the latter network does not. A user of advanced UE can generate multiple requests and be in service for applications (either homogeneous or heterogeneous) simultaneously. However, in a practical system, not all applications have the same priority at the UE side. From the perspective of user perception, the application of focus usually plays a key role in the perceived QoS at the UE side. As such, directly applying differential service strategies by

Manuscript received December 3, 2013; revised April 1, 2014; accepted May 29, 2014. The associate editor coordinating the review of this paper and approving it for publication was Z. Han.

H. Shan, A. Huang, and W. Wang are with the Department of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, 310027 China (e-mail: {hshan, aiping.huang, wangw}@zju.edu.cn).

Z. Ni is with MicroStrategy, Hangzhou, 310012 China (e-mail: zni@microstrategy.com).

W. Zhuang is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, N2L 3G1 Canada (e-mail: wzhuang@uwaterloo.ca).

This research was partially supported by National Hi-Tech R&D Program of China (No. 2014AA01A702), National Key Basic Research Program of China (No. 2012CB316104), National Natural Science Foundation of China (No. 61201228, 60972058), Specialized Research Fund for the Doctoral Program of Higher Education (No. 20120101120077). Zhejiang Provincial Natural Science Foundation of China (No. LY12F01021, LR12F01002, LY13F010003), and a research grant from the Natural Science and Engineering Research Council (NSERC) of Canada.

Digital Object Identifier ******

for example simply reducing radio resource supply according to service type, without knowing the specific application that a user focuses on, can pose extra yet unnecessary QoE degradation. Take web browsing as an example. When people using netbook surf on the YouTube website, they can open multiple video clips of interest simultaneously. Except for the most interested one, they would pause the other clips and let them buffer packets. Here, the clip of user interest and thus in the foreground of the screen is the one of current user focus, while the others are not. If the network happens to be in congestion and differential service strategies (such as those in [7]) are applied by simply imposing QoS degradation on all web applications in the system, the user tends to suffer from poor quality of experience. Yet, the user QoE can be improved, if the network is aware of the user behavior and allocates sufficient network resources to the application of focus, using a certain amount of resources taken from the rest applications that the user does not focus on.

As a result, by utilizing the information on which application is of current user focus, new network resource management strategies are promising to improve both network spectrum utilization and user quality of experience. The key idea of the work is thus to use radio resources mainly on the application to which a user pays attention. In general, the information identifying the application of focus is a specific kind of context information (CI) introduced in [11,12]. Context information defined there can be any knowledge about data transmissions collected by UE and utilized by a resource manager (e.g., a base station in [11,12]). It includes knowledge not only about traffic features such as data delivery deadline, application type (e.g., interactive applications or system applications without user interface), and request type (e.g., for caching or for immediate display), but also about user behaviors reflected for example in the set of active applications (i.e., the applications of focus). The knowledge of UEs' CI leads to new perspectives for resource management (named context-aware resource allocation) in wireless networks [11]–[18].

In this work, we explore benefits of context-aware resource allocation by investigating context-aware call admission control, and evaluate its call-level performance analytically. Specifically, we propose a call admission control strategy for LTE-type cellular networks utilizing user behavior information that is reflected in the set of active applications. With the user behavior information, a resource manager implements a newly proposed mechanism, *time-frequency resource conversion* (TFRC), to strategically yet independently manage each user's resource allocation. It recycles the radio resources step-by-step from the applications that are not temporarily focused on by the user, which produces *virtual spectrum hole*[1]. The strategy helps a congested network to not only accommodate more users but also allocate sufficient radio resources to application(s) that dominates users' QoE. The main contributions and significance of this work are summarized as follows.

[1]We call the spectrum hole produced by time-frequency resource conversion "virtual spectrum hole", to distinguish the spectrum hole obtained by cognitive radio techniques [19].

First, we propose a new context-aware resource allocation strategy, time-frequency resource conversion, to withdraw spectrum resources from the application(s) that is not temporarily of interest to the user, thus producing extra re-usable spectrum in a congested network. Second, based on the proposed TFRC, we propose a context-aware resource management strategy with a TFRC-oriented call admission control scheme named *double-threshold guard channel policy*, through which the correlation of resource allocation among all admitted users or the complexity of context-aware resource allocation can be reduced. Third, an analytical model is presented to evaluate the impact of the proposed context-aware technique on call-level performance. In addition to new call blocking and handoff call dropping probabilities, we analyze the drawback of TFRC on an admitted user, measured by the probability (named *recovering call dropping probability*) that the user cannot recover his/her recycled spectrum resources if the network admits new users too aggressively. Fourth, an optimization framework to balance the performance tradeoff among the three types of calls (i.e., new calls, handoff calls, and recovering calls) is proposed to fine tune the system parameters. Simulation results demonstrate the accuracy of the analysis, and show that under the settings of simulation study and compared with a user behavior-unaware call admission control scheme, the proposed approach can increase system capacity by as large as 90%, in terms of the admitted user number. However, the potential drawback in terms of recovering call dropping probability can be limited to a much smaller order of magnitude as compared with the new call blocking or handoff call dropping probability.

The remainder of this paper is organized as follows. We discuss the related works in Section II. The system model and an overview of user behavior-driven TFRC are described in Section III. Details of the TFRC strategy and double-threshold guard channel policy for a TFRC-based cellular network are discussed in Section IV. An analytical model to study the impact of the new technique and an optimization framework to fine tune system parameters are presented in Section V. Performance evaluation is given in Section VI. Finally, conclusions are presented in Section VII.

II. Related Work

User behavior-driven or context-aware resource allocation has recently been attracting increasing attention from the research community [11]–[18]. We discuss the relevant references next, and highlight the difference of our work from the existing studies.

In [11,12], by exploiting the feedback context information about data delivery deadline and data amount, a context-aware scheduler is proposed to facilitate efficient resource allocation among traffic flows, according to a time-utility function-based throughput-delay tradeoff. The scheduling algorithm relies on future channel quality information, with the impact of channel prediction error studied in [13]. The scheduling algorithm is further applied at small cell base stations [14], to better steer delay-tolerant data traffic between cellular and WiFi radio access technologies. However, the scheduler proposed

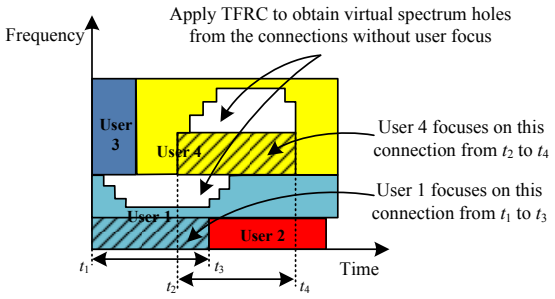


Fig. 1. Virtual spectrum hole by time-frequency resource conversion.

in [11,12] does not deal with whether a scheduled flow (connection) is focused on by the user or not. A background connection may thus gain a higher priority than a foreground connection which a user pays attention to. Further, for the optimal schedule, the radio resource management strategy suffers from high computational complexity, since traffic flows of all users are scheduled jointly in a large decision space. By making the network better informed of its UE hardware type (screen size) and the QoS requirements (e.g., data rate, delay, and packet error ratio) from all alive applications, the context-aware cell association scheme proposed in [15] helps small cell networks to not only better balance traffic load in the network but also improve QoS provisioning for applications of each user. Nevertheless, the differentiation between foreground and background applications of a user is not utilized in the proposed approach. Proposed in [16]–[18] is an inter-cell predictive resource allocation framework for mobile video delivery, exploiting the context information about user trajectories and radio maps.

To the best of our knowledge, very few such attempts have been made for resource allocation utilizing context information about user focus and studying it from the perspective of context-aware call admission control. Further, there is no analytical framework available to analyze the performance of context-aware resource allocation. The motivation behind this work is thus to explore the potential approach for context-aware call admission control and identify the benefits of context-aware resource allocation analytically.

III. User Behavior-Driven TFRC and System Model

A. Overview of User Behavior-Driven TFRC

The basic idea of time-frequency resource conversion can be explained using Fig. 1. Consider four users in a cell, sharing the available spectrum resources. The time-frequency resources allocated to these users are distinguished by different colors. If a user has multiple connections[2] (e.g., user 1 and user 4 in Fig. 1), each of the user's connections is allocated a certain amount of spectrum resources, associated with a specific time-frequency block in the figure.

When a user is focusing on the service that one of the user's multiple connections provides, the bandwidth resources consumed by the user's other connections contribute little to

[2]Connection and application with data transmission are interchangeable thereafter in this paper, since we assume that each application with data transmission corresponds to a single connection.

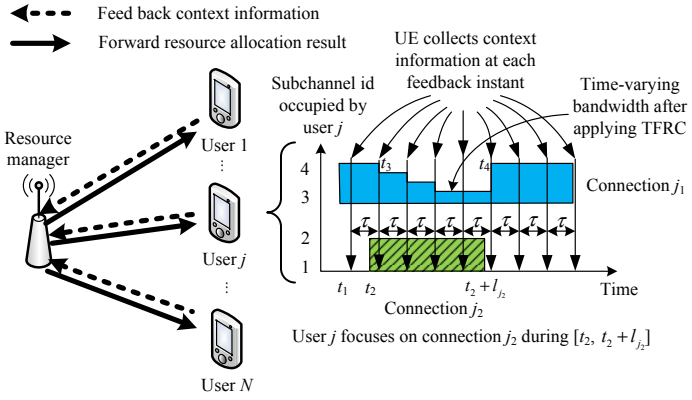


Fig. 2. TFRC by utilizing context information fed back from UE.

the quality of user's current experience. The bandwidth resources of the user's other connections can thus be strategically reduced and reused in the network. On the other hand, when a change of user focus occurs, the bandwidth supply to the newly focused connection should be guaranteed (i.e., increased to a normal supply) accordingly. For instance, when user 1 pays attention to the shadow represented connection from time t_1 to time t_3, the bandwidth of the other connection of the user can be reduced and maintained at a low level until the user switches his/her focus at time t_3 (see Fig. 1). In this work, we name such user behavior-driven process time-frequency resource conversion. The detailed TFRC strategy is presented in Section IV.

To identify the connection(s) that a user is focusing on, the UE's operating system and the platform libraries can be utilized, since they detect and handle interface events (e.g., touchscreen, screen off/lock), thus are aware of the set of active applications which are currently of user's concern [12]. Also, applying TFRC to shift bandwidth utilization in the time dimension, we can create a virtual spectrum hole to facilitate spectrum reuse. However, it is noteworthy that, for a single user, though the spectrum hole gained by TFRC can last for serval seconds, even serval minutes or hours, it remains a shift of spectrum usage in time dimension. To achieve a stable network performance improvement, one needs to multiplex multiple virtual spectrum holes (e.g., two virtual spectrum holes can be utilized continually from time t_1 to t_4 in the system depicted in Fig. 1), and the network load should alternate between heavy and light, which has been observed in many networks [20].

B. System Model for TFRC-based Cellular Network

To study the impact of utilizing user behavior-driven TFRC, we consider an LTE-type cellular network as shown in Fig. 2 and focus on call-level performance. Suppose that in a single cell there are C subchannels, each with average data rate R_b. In general, a subchannel can be a subcarrier or a bundle of subcarriers in an orthogonal frequency-division multiplexing (OFDM)-based system. Each user can initiate multiple connections, and the network thus may need to serve multiple connections simultaneously for a single user. The base station in the cell plays a role of resource manager, which

periodically collects UE-side context information, namely, the connection(s) that the user is currently focusing on among all connections. Similar to [21], for simplicity, we categorize the connections in the network into two types: wide-band and narrow-band connections, represented by T_1 and T_2, respectively, such as for high and low-definition video transmissions. We let r_1 and r_2 respectively denote the average spectrum resource amount (in subchannel number) requested by each connection of types T_1 and T_2, clearly $r_1 > r_2$.

The user locations in the network are assumed to be distributed according to a two-dimensional Poisson point process with density ρ. Thus, on average, there are $\bar{K}_A = \rho A$ users in a cell of area A [22]. The call arrival process in the cell is considered to be Poisson. Assuming that the call generating processes of all the users in the cell are independent, the call arrival process from every user is also Poisson [23]. Without loss of generality, we assume that any user in the cell initiates a new connection according to a Poisson process with rate λ_u, and the new connection belongs to T_k with probability P_k, where $k = 1, 2$ and $P_1 + P_2 = 1$. For a call connection of class T_k and without applying TFRC (i.e., normal spectrum supply when in data transmission), its duration l_k follows an exponential distribution with mean $1/\mu_k$. Assuming that the movement of each user is a random walk and is a stationary process, the cell residual time is exponentially distributed with mean $1/\eta$.

C. User Behavior Model

Each user in the network can have multiple simultaneous connections (either homogenous or heterogeneous) in data transmission. To the best of our knowledge, there is no existing model describing user behavior (i.e., user focus) with respect to multiple coexisting connections. To capture the main feature of user focus with traceability, we model user behaviors in terms of connection state as follows.

As shown in Fig. 3, a connection is in one of three states: 1) the connection is in data transmission and in the foreground of the screen, denoted by S_1; 2) the connection is in data transmission but in the background of the screen, denoted by S_2; and 3) the connection ends, denoted by S_3. A foreground (background) connection is referred to as the application with (without) user focus. For analysis traceability, we assume that, at each instant, a user only focuses on one application and thus there is at most a single connection in the foreground of the screen, but multiple applications can remain in the background. Except for a new arrival of a call connection, the user will keep focusing on the foreground application until it ends. A newly arrived call connection is always in state S_1, until it ends (thus transferring to state S_3), or until another new call connection arrives and transfers to state S_2 if it is still in data transmission. A background connection in state S_2 transfers to a foreground connection in state S_1 if the following conditions satisfy: when the current foreground connection ends, no other new call connection arrives and the user chooses this background connection as the next foreground connection from the background connection set. Any connection completed data transmission is assumed to be transferred to state S_3.

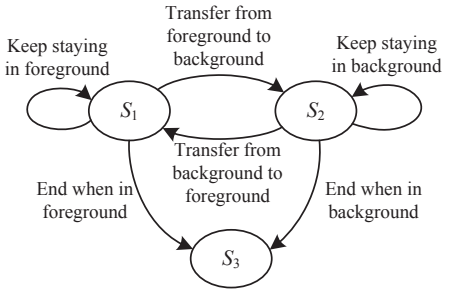


Fig. 3. User behavior model in terms of connection state.

IV. TFRC Strategy and TFRC-Oriented CAC Policy

A. TFRC Strategy

To accommodate more users especially when the network load is high, the time-frequency resource conversion aims to withdraw spectrum resources from the connections not temporarily being focused on by their owners. However, if the network admits new users too aggressively with the recycled spectrum resources, it can degrade services to users already in system, as there can be insufficient spectrum resources when they request to recover their connections. Hence, we propose a time-frequency resource conversion strategy to produce more virtual spectrum holes and to ensure service quality for the users applied TFRC.

1) Notation: Let $S^f_{j,t}$ and $S^b_{j,t}$ denote the set of applications in state S_1 (foreground of the screen) and in state S_2 (background of the screen) respectively, which user j's UE detects and feeds back to the resource manager as context information at time t. Clearly, $|S^f_{j,t}| \leq 1$ and $|S^b_{j,t}| \geq 0$. Take user j in Fig. 2 as an example. Two connections, j_1 and j_2 ($j_1, j_2 \in \{T_1, T_2\}$), with durations l_{j_1} and l_{j_2}, are initiated by the user at t_1 and t_2, respectively. Before t_2, the user only has one connection (j_1) in data transmission and suppose the related application is detected in state S_1, the two sets fed back at time $t_3 - \tau$ are therefore $S^f_{j,t_3-\tau} = \{j_1\}$ and $S^b_{j,t_3-\tau} = \emptyset$, respectively, where τ is the feedback period of context information (see Fig. 2). However, after the user opens a new application at time t_2 and focuses on the new application rather than the previous one (j_1), the context information is updated as $S^f_{j,t_3} = \{j_2\}$ and $S^b_{j,t_3} = \{j_1\}$, respectively.

2) Date Reception Model: Take user j in Fig. 2 as an example. Denote by $r^{(i)}_{j_1,j_2}$ the decreased number of subchannels that connection j_1 occupies after the i^{th} report about j_2's state of S_1. Here, the report (feedback) index is counted after connection j_1 transfers from state S_1 to state S_2, i.e., $i = 1$ at time t_3 in Fig. 2. Then, the virtual spectrum hole that user j contributes between the i^{th} and $(i+1)^{th}$ reports is given by $r^{(i)}_{j_1,j_2}$ and, accordingly, the number of subchannels supplied to connection j_1 is reduced to $r_{j_1} - r^{(i)}_{j_1,j_2}$ after the i^{th} report. If user j turns to focus on connection j_1 again between the m^{th} and $(m+1)^{th}$ reports, the total data amount that connection j_1 received between t_2 (i.e., the time that user j pays attention to new connection j_2) and t_4 (i.e., the time the resource manager detects that the user focuses on connection j_1 again) is

$$R_{rv}(j_1, j_2, m) = [r_{j_1}\Delta t + \tau \sum\nolimits_{i=1}^{m} (r_{j_1} - r^{(i)}_{j_1,j_2})]R_b \quad (1)$$

where $\Delta t = t_3 - t_2$ is a detection delay, r_{j_1} ($\in \{r_1, r_2\}$) is the number of subchannels required by j_1 at state S_1.

3) Connection State Transition Probability: Mathematically, the probability (denoted by P_{rf}) that user j turns to focus on connection j_1 again between the m^{th} and $(m+1)^{th}$ reports can be measured by the transition probability that connection j_1 transfers from state S_2 to state S_1. Let E_1, E_2, and E_3 respectively denote the event that a new call connection of user j arrives, the event that the connection in $S^f_{j,t_3+m\tau}$ (i.e., connection j_2) ends, and the event that user j chooses connection j_1 from background connection set $S^b_{j,t_3+m\tau}$, between the m^{th} and $(m+1)^{th}$ reports. Let E_4 denote the event that connection j_1 is still in data transmission at the $(m+1)^{th}$ report. The probability, P_{rf}, is given by

$$\begin{aligned} P_{rf} &= P(\bar{E}_1 \cap E_2 \cap E_3 \cap E_4) \\ &= P(\bar{E}_1)P(E_2)P(E_3)P(E_4) \end{aligned} \tag{2}$$

as the four events are independent. Probabilities $P(\bar{E}_1)$ and $P(E_2)$ can be found respectively according to the Poisson arrival process of call connections and exponentially distributed call duration, given by

$$P(\bar{E}_1) = e^{-\lambda_u \tau},\ P(E_2) = 1 - e^{-\mu_{j_2}\tau}. \tag{3}$$

Suppose all the connections in the background connection set are equally likely to be selected to foreground, $P(E_3)$ in (2) is given by

$$P(E_3) = 1/|S^b_{j,t_3+m\tau}|. \tag{4}$$

From Appendix A, the service rate of a call connection with reduced spectrum supply is scaled proportionally according to the remaining spectrum supply. When connection j_1 is selected as the foreground connection between the m^{th} and $(m+1)^{th}$ reports, its service rate is equal to $\mu_{j_1}(1-r^{(m)}_{j_1,j_2}/r_{j_1})$. Whereby, we have

$$P(E_4) = e^{-\mu_{j_1}(1-r^{(m)}_{j_1,j_2}/r_{j_1})\tau}. \tag{5}$$

4) QoE Degradation Model: Once user j requests a spectrum recovery for connection j_1 (at t_4 in Fig. 2), the resource manager should resume full spectrum supply for the requested connection as soon as possible. A quick spectrum recovery helps the user to maintain experience quality, but reduces the opportunity to serve more new users. To reduce the likelihood of incomplete spectrum recovery, we consider a "*recovery protection mechanism*": If a user has multiple connections, the spectrum resources of the user's previously ended connection will be reserved for his/her other connections with a potential recovery request. Thus, when connection j_2 of user j ends, its spectrum resources are reserved for connection j_1.

However, even with a successful spectrum recovery, user j may suffer from a certain extent of QoE degradation as compared with full spectrum supply to all his/her connections. According to [24], the QoE degradation (denoted by $Q(j_1, j_2, m)$) of user j due to the reduced spectrum supply to connection j_1 can be modeled by an exponential relationship with delivered data for the connection in spectrum reduction period, as follows

$$Q(j_1, j_2, m) = \alpha e^{-\beta R^d_{rv}(j_1,j_2,m)} + \gamma \tag{6}$$

where $R^d_{rv}(j_1, j_2, m) = R_{rv}(j_1, j_2, m) - r_{j_1}\Delta t R_b$ is the delivered data amount for connection j_1 in spectrum reduction period as compared with that of full spectrum supply, representing service disturbance; m indicates that connection j_1 requests to resume full spectrum supply at the $(m+1)^{th}$ report. Here, α, β, and γ are non-negative parameters depending on the measured QoE metric (e.g., cancelation rate of HTTP objects studied in [24]).

Based on the probability (P_{rf}) of a changed user focus between the m^{th} and $(m+1)^{th}$ feedbacks and the impact ($Q(j_1, j_2, m)$) of spectrum recovery exactly at the $(m+1)^{th}$ feedback, we model the QoE degradation of user j if continuing to cut off his/her spectrum resources after the m^{th} report by

$$G(j_1, j_2, m) = P_{rf} \cdot Q(j_1, j_2, m). \tag{7}$$

Clearly, the maximal QoE degradation (denoted by $G_{\max}(j_1, j_2, m)$) occurs when we cut off all spectrum supply at the first instant of time-frequency resource conversion (i.e., $r^{(1)}_{j_1,j_2} = r_{j_1}$), leading to $G_{\max}(j_1, j_2, m) = P(\bar{E}_1) \cdot P(E_2) \cdot P(E_3) \cdot 1 \cdot (\alpha + \gamma)$. Thus, the normalized QoE degradation of user j at each time-frequency resource conversion instant can be represented as

$$\begin{aligned} J(j_1, j_2, m) &= G(j_1, j_2, m)/G_{\max}(j_1, j_2, m) \\ &= P(E_4)\frac{\alpha e^{-\beta R^d_{rv}(j_1,j_2,m)} + \gamma}{\alpha + \gamma}. \end{aligned} \tag{8}$$

The normalized QoE degradation depends directly on the probability whether or not the newly selected foreground connection is in data transmission. Provided that a connection will be selected as a foreground connection (from the background connection set) at the next feedback instant, only if the connection is in data transmission, there is QoE degradation.

In the case that j_1 is a streaming video and the QoE degradation is represented by the video stopping probability (i.e., the probability that the playback freezes in the middle of a media playout), (8) can be analytically derived according to [25] as

$$J(j_1, j_2, m) = P(E_4)\exp(-\frac{2R^d_{rv}(j_1, j_2, m)}{\alpha_T}\beta_T) \tag{9}$$

where $\alpha_T > 0$ and $\beta_T > 0$ are predefined coefficients to characterize the network dynamics and the expectation of net packet growth rate in the user-side buffer. In general, (9) captures that, given network dynamics and net packet growth rate, the more the pre-buffered data, the smaller the probability that the video will be frozen, thus the better the user's quality of experience. By comparing (8) and (9), we observe that, in the video streaming scenario, parameters in (8) satisfy $\alpha = 1$, $\beta = 2\beta_T/\alpha_T$, and $\gamma = 0$.

5) TFRC Strategy: In contrast to QoE degradation, the normalized virtual spectrum hole that the user contributes between the m^{th} and the $(m+1)^{th}$ reports about connection j_2 is given by $H(j_1, j_2, m) = r^{(m)}_{j_1,j_2}/r_{j_1}$. Aiming at maximizing a user's contribution to better reuse the spectrum resources of the entire system while taking the user's own QoE into account, we formulate the following optimization problem

(OP) to manage the decreased subchannel number for each background connection (say connection j_1 of user j) at each report (say the m^{th} report about connection j_2):

$$\begin{aligned} &\max_{r^{(m)}_{j_1,j_2}} \ \phi(j_1,j_2,m) = H(j_1,j_2,m) - w \cdot J(j_1,j_2,m) \\ &\text{s.t.} \quad 0 \le r^{(m)}_{j_1,j_2} \le r_{j_1},\ r^{(m)}_{j_1,j_2} \in \mathbb{Z} \end{aligned} \tag{10}$$

where $w(>0)$ is a weighting factor. Obviously, a larger value of w favors solutions with smaller QoE degradation.

By noting in Appendix B that ϕ is a concave function of $r^{(m)}_{j_1,j_2}$, the OP can be transformed into a formal convex one, as long as we consider $\min_{r^{(m)}_{j_1,j_2}} -\phi(x_{j1}, x_{j2}, m)$ and relax it by removing the integrality constraint. According to Appendix B, the optimal TFRC strategy is given by (11), where $\hat{r}^{(m)}_{j_1,j_2}$ (the root of (41)) is the optimal setting of $r^{(m)}_{j_1,j_2}$ without constraint $0 \le r^{(m)}_{j_1,j_2} \le r_{j_1}$, and $\lfloor x \rfloor$ ($\lceil x \rceil$) is the floor (ceiling) function. Specifically, for the video stream case associated with (9), we have

$$\hat{r}^{(m)}_{j_1,j_2} = r_{j_1} + \frac{\theta}{\theta+\xi} \sum_{i=1}^{m-1} (r_{j_1} - r^{(i)}_{j_1,j_2}) - \frac{1}{\theta+\xi} \ln(w r_{j_1}(\theta+\xi)) \tag{12}$$

where $\theta = \beta R_b \tau$ and $\xi = \mu_{j_1} \tau / r_{j_1}$. As shown in Appendix B, with an increase of m, $\hat{r}^{(m)}_{j_1,j_2}$ increases gradually until it reaches the maximal (i.e., when all subchannels of a connection are withdrawn). That is, after certain rounds of reporting, $r^{(m)^*}_{j_1,j_2}$ in (11) will be finally set as r_{j_1}. The rational is that, after the buffered data amount for connection j_1 is large enough to ensure the user's QoE requirement, to produce more virtual spectrum holes, the connection should keep frozen thereafter until it is awakened (i.e., the user focuses on it again). Denoting m_{j_1,j_2} as the maximum number of report times before the number of subchannels is decreased to 0, we have $r_{j_1} - r^{(m_{j_1,j_2}-1)}_{j_1,j_2} > 0$, $r_{j_1} - r^{(m_{j_1,j_2})}_{j_1,j_2} = 0$.

In practice, to implement the proposed TFRC strategy, the resource manager or the base station needs to periodically collect UE-side context information (see Fig. 2) based on which it can implement TFRC individually for each user. As the resource allocation here is connection oriented, the potential scalability issue, which is outside of the scope of the paper, should be carefully addressed[3]. On the other hand, as the strategy is per user based, compared with a joint schedule for all admitted users and/or connections (e.g., [11,12]), the correlation of resource allocation among all admitted users or the complexity of context-aware resource allocation is reduced. However, QoE unbalance can occur among different users. Finally, the signaling of context information from each UE to the base station can rely on packet gateway-based approach [12], if a dedicated control channel and direct forwarding in the data plane are difficult to implement.

B. TFRC-Oriented CAC Policy

When implementing the proposed TFRC strategy, as shown in Fig. 4, three types of calls coexist in a cell: new calls, handoff calls, and recovering calls. The first two have been frequently studied in the literature, while the last one is newly introduced here. A recovering call is an existing connection which requests to resume its full spectrum supply after being in state S_2 with reduced spectrum allocation. The recovering calls should have higher priority over the other two types of calls in resource allocation. Further, a handoff call has higher priority over a new call, since interrupting an ongoing call usually brings more negative experience than blocking a new call.

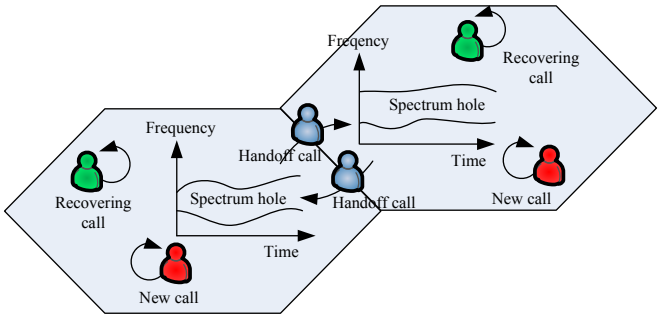


Fig. 4. Three types of calls coexist in TFRC-applied cells.

According to the priority order, we propose a double-threshold guard channel policy tailored for a TFRC-based LTE-type cellular network. Within the total C subchannels in a cell, C_R subchannels are reserved only for recovering calls and C_{HR} ($> C_R$) subchannels are reserved for both recovering and handoff calls. Namely, a new (handoff) call cannot be admitted if the number of the used subchannels is no less than $C - C_{HR}$ ($C - C_R$). With the highest priority, a recovering call can use any available channel(s). In particular, the policy is beneficial when the available resources are insufficient to admit a user with multiple connections, but sufficient to support at least the connection that an existing user is focusing on.

V. Performance Analysis and Parameter Optimization

A. System State Description

For presentation clarity, we limit the number of connections of a single user to two. However, the following analysis can be directly extended to a case of the number of connections larger than two. We have six types of users in the network:

- Users with only one T_1 connection in state S_1, denoted by U_{T_1};
- Users with only one T_2 connection in state S_1, denoted by U_{T_2};
- Users with two simultaneous T_1 connections, one in state S_1 and the other in state S_2, denoted by U_{T_1,T_1};
- Users with two simultaneous connections in which a T_1 connection is established first and is currently in state S_2, while a T_2 connection is currently in state S_1, denoted by U_{T_1,T_2};
- Users with two simultaneous connections in which a T_2 connection is established first and is currently in state S_2, while a T_1 connection is currently in state S_1, denoted by U_{T_2,T_1};

[3]Addressing the scalability issue of TFRC will be our future work, to be studied from the perspective of scheduling.

$$r_{j_1,j_2}^{(m)^*} = \begin{cases} 0, & \hat{r}_{j_1,j_2}^{(m)} \le 0 \\ \arg \max\limits_{r_{j_1,j_2}^{(m)^*} \in \{\lfloor \hat{r}_{j_1,j_2}^{(m)} \rfloor, \lceil \hat{r}_{j_1,j_2}^{(m)} \rceil\}} \{\phi(j_1,j_2,m)\}, & 0 < \hat{r}_{j_1,j_2}^{(m)} \le r_{j_1} \\ r_{j_1}, & \hat{r}_{j_1,j_2}^{(m)} > r_{j_1} \end{cases} \tag{11}$$

- Users with two simultaneous T_2 connections, one in state S_1 and the other in state S_2, denoted by U_{T_2,T_2}.

For the first two types of users, each user occupies either r_1 or r_2 subchannels. Yet, for the other four types of users, the number of subchannels that a user occupies depends not only on the connection type but also the progress of TFRC applied to the user. Specifically, user $j \in U_{j_1,j_2}$ ($j_1, j_2 \in \{T_1, T_2\}$), can be in one of $m_{j_1,j_2}+1$ states characterized by the number of withdrawn subchannels $r_{j_1,j_2}^{(i)}, i = 0, 1, ..., m_{j_1,j_2}$, where $r_{j_1,j_2}^{(0)} = 0$. For notation simplicity, in the following we use $r_I^{(i)}$, $r_{II}^{(i)}$, $r_{III}^{(i)}$, and $r_{IV}^{(i)}$ (m_I, m_{II}, m_{III}, and m_{IV}) to respectively denote $r_{T_1,T_1}^{(i)}$, $r_{T_1,T_2}^{(i)}$, $r_{T_2,T_1}^{(i)}$, and $r_{T_2,T_2}^{(i)}$ (m_{T_1,T_1}, m_{T_1,T_2}, m_{T_2,T_1}, and m_{T_2,T_2}). Hence, the number of subchannels occupied by a user in U_{T_1,T_1} is in the set given by

$$\left\{2r_1, 2r_1 - r_I^{(1)}, 2r_1 - r_I^{(2)}, ..., 2r_1 - r_I^{(m_I-1)}, r_1\right\} \tag{13}$$

based on which we can define the system state associated with users in U_{T_1,T_1} as follows

$$\mathcal{N}_I := (\mathcal{N}_I^{(0)}, \mathcal{N}_I^{(1)}, ..., \mathcal{N}_I^{(m_I-1)}, \mathcal{N}_I^{(m_I)}) \tag{14}$$

where $\mathcal{N}_I^{(i)}$, $i = 0, 1, ..., m_I$, represents the number of users who use $2r_1 - r_I^{(i)}$ subchannels. Similarly, we can define the system state associated respectively with users in U_{T_1,T_2}, U_{T_2,T_1}, and U_{T_2,T_2} as follows

$$\begin{aligned} \mathcal{N}_{II} &:= (\mathcal{N}_{II}^{(0)}, \mathcal{N}_{II}^{(1)}, ..., \mathcal{N}_{II}^{(m_{II}-1)}, \mathcal{N}_{II}^{(m_{II})}) \\ \mathcal{N}_{III} &:= (\mathcal{N}_{III}^{(0)}, \mathcal{N}_{III}^{(1)}, ..., \mathcal{N}_{III}^{(m_{III}-1)}, \mathcal{N}_{III}^{(m_{III})}) \\ \mathcal{N}_{IV} &:= (\mathcal{N}_{IV}^{(0)}, \mathcal{N}_{IV}^{(1)}, ..., \mathcal{N}_{IV}^{(m_{IV}-1)}, \mathcal{N}_{IV}^{(m_{IV})}) \end{aligned} \tag{15}$$

where $\mathcal{N}_{II}^{(i)}$, $\mathcal{N}_{III}^{(i)}$, and $\mathcal{N}_{IV}^{(i)}$ represent the numbers of users who use $r_1 - r_{II}^{(i)} + r_2$, $r_2 - r_{III}^{(i)} + r_1$, and $2r_2 - r_{IV}^{(i)}$ subchannels, respectively. For presentation simplicity, we define $U_{T_{j_1}^{(i)},T_{j_2}}$ ($j_1, j_2 \in \{1, 2\}$, $i \in \{0, 1, ..., m_{T_{j_1},T_{j_2}}\}$) to represent the set of users who have two simultaneous connections in which the T_{j_1}-type connection is in time-frequency resource conversion thus being taken away $r_{T_{j_1},T_{j_2}}^{(i)}$ subchannels. Clearly, $U_{T_{j_1},T_{j_2}} = \bigcup_{i=0}^{m_{T_{j_1},T_{j_2}}} U_{T_{j_1}^{(i)},T_{j_2}}$.

With the system states associated with each user type, we can characterize the whole system state (denoted by S_T) as (16), where n_1 and n_2 are the numbers of users in U_{T_1} and U_{T_2}, respectively. The total cell traffic load can be expressed as

$$\begin{aligned} R(S_T) = n_1 r_1 + n_2 r_2 &+ \sum_{i=0}^{m_I} (2r_1 - r_I^{(i)}) \cdot \mathcal{N}_I^{(i)} \\ &+ \sum_{i=0}^{m_{II}} (r_1 - r_{II}^{(i)} + r_2) \cdot \mathcal{N}_{II}^{(i)} \\ &+ \sum_{i=0}^{m_{III}} (r_2 - r_{III}^{(i)} + r_1) \cdot \mathcal{N}_{III}^{(i)} \\ &+ \sum_{i=0}^{m_{IV}} (2r_2 - r_{IV}^{(i)}) \cdot \mathcal{N}_{IV}^{(i)}. \end{aligned} \tag{17}$$

According to the relation between the cell traffic load and cell capacity, the state space can be defined as

$$\Upsilon = \{ S_T | \, R(S_T) \le C \}. \tag{18}$$

Further, given state S_T, the number of admitted users and the number of idle users (i.e., user without any connection) are respectively given by

$$U(S_T) = n_1 + n_2 + \sum_{j=I}^{IV} \sum_{i=0}^{m_j} \mathcal{N}_j^{(i)} \tag{19}$$

$$U_{idle}(S_T) = \bar{K}_A - U(S_T). \tag{20}$$

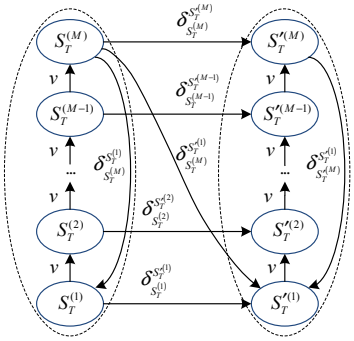
B. State Transitions

The events that trigger a transition of the system state can be classified into five categories: 1) a new call arrival; 2) a handoff user arrival; 3) a handoff user departure; 4) a call termination; and 5) a periodic time-frequency resource conversion. Consider a new T_1-type call arrival. When the call arrives in the cell, it will be accepted if $R(S_T) + r_1 \le C - C_{HR}$; otherwise, the call is blocked. Depending on the user who initiates the new call, there are three types of state transitions:

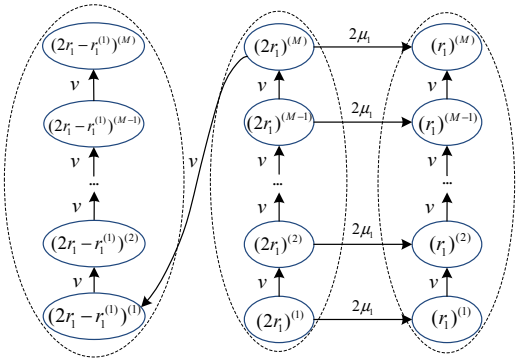
- When the call is initiated by a user without any connection, $n_1 \to n_1 + 1$, with a rate of $U_{idle}(S_T)\lambda_u P_1$;
- When the call is initiated by a user with only a single T_1-type connection, $n_1 \to n_1 - 1$, $\mathcal{N}_I^{(0)} \to \mathcal{N}_I^{(0)} + 1$, with a rate of $n_1 \lambda_u P_1$;
- When the call is initiated by a user with only a single T_2-type connection, $n_2 \to n_2 - 1$, $\mathcal{N}_{III}^{(0)} \to \mathcal{N}_{III}^{(0)} + 1$, with a rate of $n_2 \lambda_u P_1$.

Similarly, we can identify state transitions for a new T_2-type call arrival and all the other events. The detailed analysis of state transition is omitted due to space limitation. All the system state transitions are summarized in Table I, where λ_{h_k}, $\lambda_{h(I)}^{(i)}$, $\lambda_{h(II)}^{(i)}$, $\lambda_{h(III)}^{(i)}$, and $\lambda_{h(IV)}^{(i)}$ are the handoff rates of users in U_{T_k}, $U_{T_1^{(i)},T_1}$, $U_{T_1^{(i)},T_2}$, $U_{T_2^{(i)},T_1}$, and $U_{T_2^{(i)},T_2}$, respectively, with $k = 1, 2$. We analyze all state transitions and derive the aforesaid handoff rates in [26]. The interested reader can refer to [26] for details.

$$S_T = \left(\begin{array}{l} n_1; n_2; \overbrace{\mathcal{N}_I^{(0)}, \mathcal{N}_I^{(1)}, \ldots \mathcal{N}_I^{(m_I-1)}, \mathcal{N}_I^{(m_I)}}^{\mathcal{N}_I}; \overbrace{\mathcal{N}_{II}^{(0)}, \mathcal{N}_{II}^{(1)}, ..., \mathcal{N}_{II}^{(m_{II}-1)}, \mathcal{N}_{II}^{(m_{II})}}^{\mathcal{N}_{II}}; \\ \underbrace{\mathcal{N}_{III}^{(0)}, \mathcal{N}_{III}^{(1)}, \ldots \mathcal{N}_{III}^{(m_{III}-1)}, \mathcal{N}_{III}^{(m_{III})}}_{\mathcal{N}_{III}}; \underbrace{\mathcal{N}_{IV}{}^{(0)}, \mathcal{N}_{IV}^{(1)}, \ldots \mathcal{N}_{IV}^{(m_{IV}-1)}, \mathcal{N}_{IV}^{(m_{IV})}}_{\mathcal{N}_{IV}} \end{array} \right) \tag{16}$$



(a) The general multiple-stair Markov approximation model.



(b) A substate transition example when applying the multiple-stair Markov approximation model.

Fig. 5. Multiple-stair Markov approximation model.

C. Multiple-Stair Markov Approximation

Notice that, in the five classes of state transitions, the first four can occur at any time and the time between adjacent transitions is exponentially distributed according to our traffic assumptions. However, the last one associated with time-frequency resource conversion is carried out with a period of τ. As this state transition, which changes with the deterministic rule designed in Section IV, only possibly takes place on discrete and periodic time spots, the dwell time at those relevant states is not exponentially distributed. The technique of Markov chain/process is thus invalid to analyze the system performance. To address the non-memoryless state transitions, similar to [27], we adopt a multiple-stair Markov model [28] to approximate the mixed continuous-discrete Markov process as follows.

Each state in the system is divided into multiple inter-connected substates, i.e., S_T is divided into $S_T^{(1)}, S_T^{(2)}, ..., S_T^{(M)}$. By assuming a large enough value of the substate number M, all intra-state and inter-state transitions can be modeled memoryless. As shown in Fig. 5(a), substates $S_T^{(1)}$ to $S_T^{(M)}$ inside an identical state are arranged in a temporal sequence with an exponential transition from each to the next, each with a rate $v = M/\tau$. For the last substate $S_T^{(M)}$, if the state S_T remains unchanged definitely at the end of a period of time-frequency resource conversion (e.g., because no user with multiple connections exists in the network), substate $S_T^{(M)}$ returns to $S_T^{(1)}$ with rate $\delta_{S_T^{(M)}}^{S_T^{(1)}} = v$; otherwise, $\delta_{S_T^{(M)}}^{S_T^{(1)}} = 0$, and a deterministic state transition accompanied with TFRC leads to a state change from substate $S_T^{(M)}$ to the first substate of the next state (say state S_T') to which S_T must transfer according to the TFRC strategy, with a rate of $\delta_{S_T^{(M)}}^{S_T'^{(1)}} = v$. In addition, each substate can transfer to its counterpart (the substate with the same index) of the target state, for example, from $S_T^{(m)}$ to $S_T'^{(m)}$, with a rate (denoted by $\delta_{S_T^{(m)}}^{S_T'^{(m)}}$ for $m = 1, 2, ..., M$ in Fig. 5(a)) that is derived in [26] and listed in Table I, if the two involved states (i.e., S_T and S_T') have state transition in the absence of the periodic time-frequency resource conversion (i.e., belonging to the first four classes of state transitions). To clarify the approximation model, we give an example as follows.

Assume that in current state a single user exists in the cell and has two simultaneous T_1 connections without TFRC, then the state transition is shown in Fig. 5(b). For notation simplicity, in the figure we use $(2r_1)^{(m)}$ to denote the m^{th} substate of the referred state. As long as one of the user's connections terminates, the current state transfers to the state (denoted by (r_1)) with a single T_1 connection; therefore, substate $(2r_1)^{(m)}$ can transfer to substate $(r_1)^{(m)}$ with a rate of $2\mu_1$, where $m = 1, 2, ..., M$. When the current period of time-frequency resource conversion ends, substate $(2r_1)^{(M)}$ transfers directly to the first substate of the state associated with the user who has applied the first round of time-frequency resource conversion, represented by $(2r_1 - r_1^{(1)})^{(1)}$, with a rate of v.

After applying the multiple-stair Markov model, the system state can be expressed as $S_T^{(m)} = (S_T, m)$. The effectiveness of the approximation model has been demonstrated in [27] in terms of the mean and the variance of approximated state dwell time[4]. Using the system state transitions in Table I, the set of steady-state probabilities, $\{\pi(S_T^{(m)})\}$, can be obtained with the normalization condition $\sum_{\forall S_T^{(m)}} \pi(S_T^{(m)}) = 1$. Note that we have $\pi(S_T) = \sum_{m=1}^{M} \pi(S_T^{(m)})$.

[4]For any state S_T with a state change due to periodic TFRC, let t_m and t respectively denote the dwell time at substate $S_T^{(m)}$ and the total dwell time at state S_T, satisfying $t = \sum_{m=1}^{M} t_m$. As shown in [27], the expectation and variance of t are $E[t] = \tau$ and $\mathrm{Var}(t) = \tau^2/M$, respectively. Obviously, $\mathrm{Var}(t)$ decreases as M increases.

TABLE I
SYSTEM STATE TRANSITIONS.

State transition	Relevant event	Condition	Transition rate
$n_k \to n_k + 1$ $k = 1, 2$	A new T_k connection is generated by a user without any connection	$R(S_T) + r_k \leq C - C_{HR}$	$U_{idle}(S_T)\lambda_u P_k$
$n_1 \to n_1 - 1$ $\mathcal{N}_I^{(0)} \to \mathcal{N}_I^{(0)} + 1$	A new T_1 connection is generated by a user with one T_1 connection	$R(S_T) + r_1 \leq C - C_{HR}$	$n_1\lambda_u P_1$
$n_2 \to n_2 - 1$ $\mathcal{N}_{III}^{(0)} \to \mathcal{N}_{III}^{(0)} + 1$	A new T_1 connection is generated by a user with one T_2 connection	$R(S_T) + r_1 \leq C - C_{HR}$	$n_2\lambda_u P_1$
$n_1 \to n_1 - 1$ $\mathcal{N}_{II}^{(0)} \to \mathcal{N}_{II}^{(0)} + 1$	A new T_2 connection is generated by a user with one T_1 connection	$R(S_T) + r_2 \leq C - C_{HR}$	$n_1\lambda_u P_2$
$n_2 \to n_2 - 1$ $\mathcal{N}_{IV}^{(0)} \to \mathcal{N}_{IV}^{(0)} + 1$	A new T_2 connection is generated by a user with one T_2 connection	$R(S_T) + r_2 \leq C - C_{HR}$	$n_2\lambda_u P_2$
$n_k \to n_k + 1$ $k = 1, 2$	A handoff requester is accepted in U_{T_k}	$R(S_T) + r_k \leq C - C_R$	λ_{h_k}
$\mathcal{N}_I^{(i)} \to \mathcal{N}_I^{(i)} + 1$ $i = 0, 1, ..., m_I$	A handoff requester is accepted in $U_{T_1^{(i)},T_1}$ since his two T_1 connections are accepted	$R(S_T) + 2r_1 - r_I^{(i)} \leq C - C_R$	$\lambda_{h(I)}^{(i)}$
$\mathcal{N}_I^{(m_I)} \to \mathcal{N}_I^{(m_I)} + 1$	A handoff requester is accepted in $U_{T_1^{(m_I)},T_1}$ since his one T_1 connection with (without) user focus is accepted (frozen)	$R(S_T) + r_1 \leq C - C_R < R(S_T) + 2r_1 - r_I^{(i)}$ $i = 0, 1, ..., m_I - 1$	$\lambda_{h(I)}^{(i)}$
$\mathcal{N}_{II}^{(i)} \to \mathcal{N}_{II}^{(i)} + 1$ $i = 0, 1, ..., m_{II}$	A handoff requester is accepted in $U_{T_1^{(i)},T_2}$ since his two connections (one T_1 and one T_2) are accepted	$R(S_T) + r_1 - r_{II}^{(i)} + r_2 \leq C - C_R$	$\lambda_{h(II)}^{(i)}$
$\mathcal{N}_{II}^{(m_{II})} \to \mathcal{N}_{II}^{(m_{II})} + 1$	A handoff requester is accepted in $U_{T_1^{(m_{II})},T_2}$ since his T_2 (T_1) connection with (without) user focus is accepted (frozen)	$R(S_T) + r_2 \leq C - C_R < R(S_T) + r_1 - r_{II}^{(i)} + r_2$ $i = 0, 1, ..., m_{II} - 1$	$\lambda_{h(II)}^{(i)}$
$\mathcal{N}_{III}^{(i)} \to \mathcal{N}_{III}^{(i)} + 1$ $i = 0, 1, ..., m_{III}$	A handoff requester is accepted in $U_{T_2^{(i)},T_1}$ since his two connections (one T_2 and one T_1) are accepted	$R(S_T) + r_2 - r_{III}^{(i)} + r_1 \leq C - C_R$	$\lambda_{h(III)}^{(i)}$
$\mathcal{N}_{III}^{(m_{III})} \to \mathcal{N}_{III}^{(m_{III})} + 1$	A handoff requester is accepted in $U_{T_2^{(m_{III})},T_1}$ since his T_1 (T_2) connection with (without) user focus is accepted (frozen)	$R(S_T) + r_1 \leq C - C_R < R(S_T) + r_2 - r_{III}^{(i)} + r_1$ $i = 0, 1, ..., m_{III} - 1$	$\lambda_{h(III)}^{(i)}$
$\mathcal{N}_{IV}^{(i)} \to \mathcal{N}_{IV}^{(i)} + 1$ $i = 0, 1, ..., m_{IV}$	A handoff requester is accepted in $U_{T_2^{(i)},T_2}$ since his two T_2 connections are accepted	$R(S_T) + 2r_2 - r_{IV}^{(i)} \leq C - C_R$	$\lambda_{h(IV)}^{(i)}$
$\mathcal{N}_{IV}^{(m_{IV})} \to \mathcal{N}_{IV}^{(m_{IV})} + 1$	A handoff requester is accepted in $U_{T_2^{(m_{IV})},T_2}$ since his one T_2 connection with (without) user focus is accepted (frozen)	$R(S_T) + r_2 \leq C - C_R < R(S_T) + 2r_2 - r_{IV}^{(i)}$ $i = 0, 1, ..., m_{IV} - 1$	$\lambda_{h(IV)}^{(i)}$
$n_k \to n_k - 1$ $k = 1, 2$	A handoff user in U_{T_k} departs	NA	$n_k\eta$
$\mathcal{N}_I^{(i)} \to \mathcal{N}_I^{(i)} - 1$ $i = 0, 1, ..., m_I$	A handoff user in $U_{T_1^{(i)},T_1}$ departs	NA	$\mathcal{N}_I^{(i)}\eta$
$\mathcal{N}_{II}^{(i)} \to \mathcal{N}_{II}^{(i)} - 1$ $i = 0, 1, ..., m_{II}$	A handoff user in $U_{T_1^{(i)},T_2}$ departs	NA	$N_{II}^{(i)}\eta$
$\mathcal{N}_{III}^{(i)} \to \mathcal{N}_{III}^{(i)} - 1$ $i = 0, 1, ..., m_{III}$	A handoff user in $U_{T_2^{(i)},T_1}$ departs	NA	$\mathcal{N}_{III}^{(i)}\eta$
$\mathcal{N}_{IV}^{(i)} \to \mathcal{N}_{IV}^{(i)} - 1$ $i = 0, 1, ..., m_{IV}$	A handoff user in $U_{T_2^{(i)},T_2}$ departs	NA	$\mathcal{N}_{IV}^{(i)}\eta$
$n_k \to n_k - 1$ $k = 1, 2$	A connection of users in U_{T_k} ends	NA	$n_k\mu_k$
$\mathcal{N}_I^{(i)} \to \mathcal{N}_I^{(i)} - 1$ $n_1 \to n_1 + 1$ $i = 0, 1, ..., m_I$	If $i \neq m_I$, a connection of users in $U_{T_1^{(i)},T_1}$ ends; otherwise, a connection without TFRC of users in $U_{T_1^{(i)},T_1}$ ends	NA	$N_I^{(i)}\mu_1(2 - \frac{r_I^{(i)}}{r_1})$
$\mathcal{N}_{II}^{(i)} \to \mathcal{N}_{II}^{(i)} - 1$ $n_2 \to n_2 + 1$ $i = 0, 1, ..., m_{II} - 1$	A T_1 connection of users in $U_{T_1^{(i)},T_2}$ ends	NA	$N_{II}^{(i)}\mu_1(1 - \frac{r_{II}^{(i)}}{r_1})$
$\mathcal{N}_{II}^{(i)} \to \mathcal{N}_{II}^{(i)} - 1$ $n_1 \to n_1 + 1$ $i = 0, 1, ..., m_{II}$	A T_2 connection of users in $U_{T_1^{(i)},T_2}$ ends; his T_1 connection resumes full spectrum supply	$R(S_T) + r_{II}^{(i)} - r_2 \leq C$	$\mathcal{N}_{II}^{(i)}\mu_2$
$\mathcal{N}_{III}^{(i)} \to \mathcal{N}_{III}^{(i)} - 1$ $n_1 \to n_1 + 1$ $i = 0, 1, ..., m_{III} - 1$	A T_2 connection of users in $U_{T_2^{(i)},T_1}$ ends	NA	$N_{III}^{(i)}\mu_2(1 - \frac{r_{III}^{(i)}}{r_2})$
$\mathcal{N}_{III}^{(i)} \to \mathcal{N}_{III}^{(i)} - 1$ $n_2 \to n_2 + 1$ $i = 0, 1, ..., m_{III}$	A T_1 connection of users in $U_{T_2^{(i)},T_1}$ ends; his T_2 connection resumes full spectrum supply	NA	$\mathcal{N}_{III}^{(i)}\mu_1$
$\mathcal{N}_{IV}^{(i)} \to \mathcal{N}_{IV}^{(i)} - 1$ $n_2 \to n_2 + 1$ $i = 0, 1, ..., m_{IV}$	If $i \neq m_{IV}$, a connection of users in $U_{T_2^{(i)},T_2}$ ends; otherwise, a connection without TFRC of users in $U_{T_2^{(i)},T_2}$ ends	NA	$N_{IV}^{(i)}\mu_2(2 - \frac{r_{IV}^{(i)}}{r_2})$
$S_T^{(m)} \to S_T^{(m+1)}, \forall S_T \in \Upsilon$ $m = 1, 2, ..., M - 1$	Intra-state transition of any system state occurs in a temporal sequence in a period of TFRC but before the end of a period of TFRC	NA	M/τ
$S_T^{(M)} \to S_T'^{(1)}, \mathcal{N}_j^{(0)} \to 0,$ $\mathcal{N}_j^{(i)} \to \mathcal{N}_j^{(i-1)}, \mathcal{N}_j^{(m_j)} \to$ $\mathcal{N}_j^{(m_j)} + \mathcal{N}_j^{(m_j-1)}, j \in \{I, II,$ $III, IV\}, i = 1, 2, ..., m_j - 1$	Inter-state transition occurs at the beginning of a new TFRC period	NA	M/τ

$$b_0(S_T^{(m)}) = \underbrace{\lambda_n(S_T)}_{\text{New call arrival}} + \underbrace{\lambda_{h1}+\lambda_{h2}+\sum_{j=I}^{IV}\sum_{i=1}^{m_j}\lambda_{h(j)}^{(i)}}_{\text{Handoff user arrival}} + \underbrace{\left(n_1+n_2+\sum_{j=I}^{IV}\sum_{i=1}^{m_j}\mathcal{N}_j^{(i)}\right)\eta}_{\text{Handoff user departuare}}$$
$$+ \underbrace{n_1\mu_1+n_2\mu_2+\sum_{i=1}^{m_I}\mathcal{N}_I^{(i)}\mu_1\left(2-r_I^{(i)}/r_1\right)+\sum_{i=1}^{m_{II}}\mathcal{N}_{II}^{(i)}\left(\mu_1\left(1-r_{II}^{(i)}/r_1\right)+\mu_2\right)}_{\text{Call termination}}$$
$$+ \underbrace{\sum_{i=1}^{m_{III}}\mathcal{N}_{III}^{(i)}\left(\mu_2\left(1-r_{III}^{(i)}/r_2\right)+\mu_1\right)+\sum_{i=1}^{m_{IV}}\mathcal{N}_{IV}^{(m)}\mu_2\left(2-r_{IV}^{(i)}/r_2\right)}_{\text{Call termination}} + \underbrace{v}_{\substack{\text{Periodic time-frequence}\\ \text{resource conversion}}} \tag{31}$$

D. Performance Measures

In this subsection, we derive new call block, handoff call dropping, and recovering call dropping probabilities for the newly proposed TFRC-based scheme. In general, a new call can be initiated by a user with any number of connections. However, given a finite call duration, the probability that a user has multiple simultaneous connections decreases with an increase of the number of simultaneous connections. Therefore, we approximate the new call arrival rate generated by all users in a cell by the new call arrival rate generated together by users without any connection and by those with a single connection. Then, given state S_T, the new call arrival rate (denoted by $\lambda_n(S_T)$) is given by

$$\lambda_n(S_T) \doteq (U_{idle}(S_T)+n_1+n_2)\lambda_u. \tag{21}$$

It is easy to derive that the approximation error of the average new call arrival rate in the cell is given by $1-\sum_{S_T}\lambda_n(S_T)\pi(S_T)/\bar{K}_A\lambda_u$. Then, the new call blocking probability for the T_k-type connection, $P_{NB}^{(k)}$, can be expressed as

$$P_{NB}^{(k)} = \frac{\sum_{S_T\in\Omega_k}\pi(S_T)\lambda_n(S_T)}{\sum_{S_T\in\Upsilon}\pi(S_T)\lambda_n(S_T)},\ k=1,2 \tag{22}$$

where

$$\Omega_k = \{S_T|\, R(S_T)+r_k > C-C_{HR}\} \tag{23}$$

defines the state set in which a new call will be blocked due to insufficient idle channels. The overall new call blocking probability P_{NB} thus is

$$P_{NB} = P_{NB}^{(1)}P_1 + P_{NB}^{(2)}P_2 \tag{24}$$

where P_1 (P_2) is the probability that a new call is a T_1 (T_2)-type connection. Similarly, the handoff call dropping probability for users in $U_{T_1}\cup U_{T_1,T_1}\cup U_{T_2,T_1}$ (denoted by $P_{HD}^{(1)}$) or those in $U_{T_2}\cup U_{T_1,T_2}\cup U_{T_2,T_2}$ (denoted by $P_{HD}^{(2)}$) is

$$P_{HD}^{(k)} = \sum_{S_T\in\Phi_k}\pi(S_T) \tag{25}$$

$$\Phi_k = \{S_T|\, R(S_T)+r_k > C-C_R\} \tag{26}$$

where $k=1,2$. As a result, the overall handoff call dropping probability is given by

$$P_{HD} = P_{HD}^{(1)}Q_1 + P_{HD}^{(2)}Q_2 \tag{27}$$

where Q_1 and Q_2 represent the probabilities that a handoff call is triggered by a user in $U_{T_1}\cup U_{T_1,T_1}\cup U_{T_2,T_1}$ and in $U_{T_2}\cup U_{T_1,T_2}\cup U_{T_2,T_2}$, respectively, and are given by

$$Q_1 = \frac{\lambda_{h_1}+\sum_{i=0}^{m_I}\lambda_{h(I)}^{(i)}+\sum_{i=0}^{m_{III}}\lambda_{h(III)}^{(i)}}{\lambda_{h_1}+\lambda_{h_2}+\sum_{j=I}^{IV}\sum_{i=0}^{m_j}\lambda_{h(j)}^{(i)}} \tag{28}$$

$$Q_2 = \frac{\lambda_{h_2}+\sum_{i=0}^{m_{II}}\lambda_{h(II)}^{(i)}+\sum_{i=0}^{m_{IV}}\lambda_{h(IV)}^{(i)}}{\lambda_{h_1}+\lambda_{h_2}+\sum_{j=I}^{IV}\sum_{i=0}^{m_j}\lambda_{h(j)}^{(i)}}. \tag{29}$$

For a recovering call, based on our recovery protection mechanism, a spectrum recovery failure happens only to users in U_{T_1,T_2}. Specifically, for users in $U_{T_1^{(i)},T_2}$, where $i\in\{1,2,...,m_{II}\}$, the recovering call dropping probability is

$$P_{RD}^{(i)} = \frac{\sum_{S_T\in\Psi_i}\sum_{m=1}^{M}\pi(S_T^{(m)})\cdot b_1(S_T^{(m)})/b_0(S_T^{(m)})}{\sum_{S_T\in\Psi_i'}\sum_{m=1}^{M}\pi(S_T^{(m)})\cdot b_1(S_T^{(m)})/b_0(S_T^{(m)})} \tag{30}$$

where $b_0(S_T^{(m)})$ defined in (31) indicates the rate at which the system leaves state $S_T^{(m)}$, $b_1(S_T^{(m)}) = \mathcal{N}_{II}^{(i)}\mu_2$ indicates the rate at which a call recovery happens, $\Psi_i = \left\{S_T|\, R(S_T)+r_{II}^{(i)}-r_2>C,\ \mathcal{N}_{II}^{(i)}>0\right\}$ is the state set in which a recovering call from $U_{T_1^{(i)},T_2}$ drops, and $\Psi_i' = \left\{S_T|\mathcal{N}_{II}^{(i)}>0\right\}$ defines the state set in which $U_{T_1^{(i)},T_2}$ is non-empty. For all other types of users, the recovering call dropping probabilities are equal to zero, due to the recovery protection mechanism. As a result, the total recovering call dropping probability is

$$P_{RD} = \sum_{S_T\in\Psi''}\sum_{m=1}^{M}\frac{\pi(S_T^{(m)})\sum_{i=1}^{m_{II}}\mathcal{N}_{II}^{(i)}\mu_2\cdot q(S_T,i)}{b_2(S_T^{(m)})} \tag{32}$$

$$q(S_T,i) = \begin{cases} 0, & R(S_T)+r_{II}^{(i)}-r_2\le C \\ 1, & R(S_T)+r_{II}^{(i)}-r_2>C \end{cases} \tag{33}$$

$$b_2(S_T^{(m)}) = \mu_1 \left(\sum_{i=1}^{m_I} \mathcal{N}_I^{(i)} + \sum_{i=1}^{m_{III}} \mathcal{N}_{III}^{(i)} \right) + \mu_2 \left(\sum_{i=1}^{m_{II}} \mathcal{N}_{II}^{(i)} + \sum_{i=1}^{m_{IV}} \mathcal{N}_{IV}^{(i)} \right) \quad (34)$$

$$\Psi'' = \left\{ S_T | \sum\nolimits_{j=I}^{IV} \sum\nolimits_{i=1}^{m_j} \mathcal{N}_j^{(i)} > 0 \right\} \quad (35)$$

where $q(S_T, i)$ is an indicator variable for whether or not the recovering call drops, $b_2(S_T^{(m)})$ denotes the transition rate due to call recovery, and Ψ'' defines the state set in which the system always has user(s) with multiple simultaneous connections.

E. Parameter Optimization

For the performance measures, we propose an optimization framework to balance the tradeoff among three types of calls (i.e., new calls, handoff calls, and recovering calls), thus fine tuning system parameters. Specifically, for a given number C of channels, we minimize the new call blocking probability subject to hard constraints respectively on the handoff call dropping and recovering call dropping probabilities. Mathematically, the system parameters are set according to the following optimization problem

$$\begin{aligned} \min_{C_R, C_{HR}} \quad & P_{NB} \\ \text{s.t.} \quad & P_{HD} \leq P_h \\ & P_{RD} \leq P_r \end{aligned} \quad (36)$$

where P_h and P_r represent the upper bounds of the handoff call dropping probability and recovering call dropping probability, respectively. By exploiting the monotonic relation between the new call blocking probability and the guard channel numbers (i.e., C_{HR} and C_R), a search algorithm can be developed to solve (36) efficiently. In a case that handoff call dropping probability should be minimized as well, for example, in a dense urban covered by a hyper-dense small cell network, other optimization objectives such as weighted linear combination of recovering call dropping and handoff call dropping probabilities can be considered similarly.

VI. Performance Evaluation

In this section, we evaluate the performance of the proposed resource management strategy and verify the theoretical analysis. The simulation has been implemented on a custom matlab-based platform which bases on the LTE-A system's physical layer parameter settings [29]. By considering the downlink spectrum efficiency of 3.7 bit/s/Hz/cell and subchannel bandwidth of 15 kHz, the average data rate per subchannel is set to be 55.5 kbps. Each wide-band (T_1-type) connection and each narrow-band (T_2-type) connection in the simulations occupy 16 and 8 subchannels and are generated with probabilities $P_1 = 0.3$ and $P_2 = 0.7$, respectively. Their mean call durations ($1/\mu_1$ and $1/\mu_2$) are set as 200 and 100 seconds, respectively. Here we study a video streaming case and thus adopt video stopping probability studied in [25] for QoE degradation when applying the TFRC strategy. Similar to

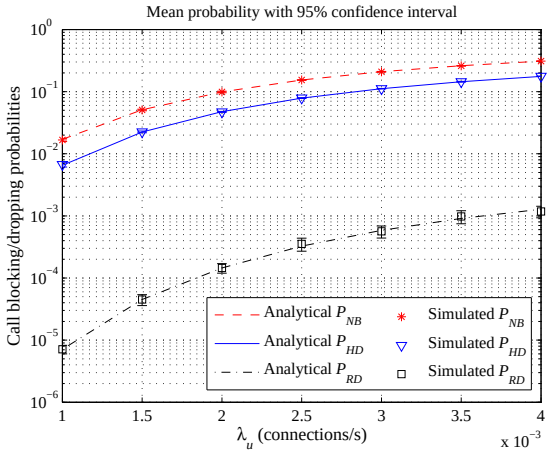


Fig. 6. The probabilities of new call blocking, handoff call dropping, and recovering call dropping.

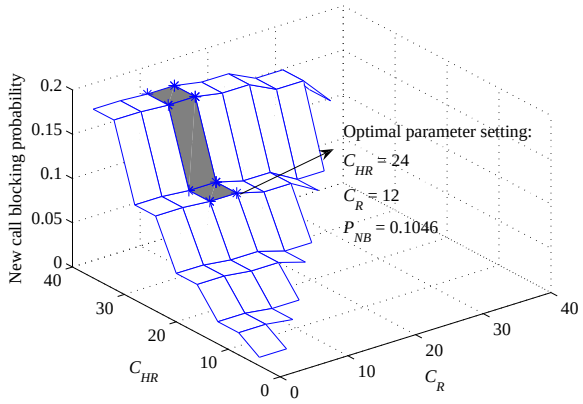


Fig. 7. The search for optimal parameter setting of double-threshold guard channel policy.

[25], α_T and β_T used to characterize the network dynamics are set according to the transmission of video clip named "Susi & Strolch". Thus, in the network scenario under consideration, α_T for both types of connections is set to 1.484×10^6, and β_T for wide-band and narrow-band connections is set to 53.564 and 12.97, respectively. Further, the resource manager collects every user's context information once per second (i.e., $\tau = 1$ second) and weights equally between increasing virtual spectrum hole and reducing QoE degradation when formulating the TFRC strategy (i.e., $w = 1$). In the simulation, we limit the number of simultaneous connections per user to not larger than two, which actually underestimates the total amount of virtual spectrum holes and thus the performance improvement ascribed to TFRC. The mean of cell residual time ($1/\eta$) and the substate number (M) are set to be 100 seconds and 4, respectively. We perform the simulations for 10 runs, each with 10^6 state transitions, and average the simulation results.

Fig. 6 shows the relationship between the traffic load (characterized by traffic density per user λ_u (connection/second)) and call-level performance measured by call blocking/dropping probabilities P_{NB}, P_{HD}, and P_{RD}, at

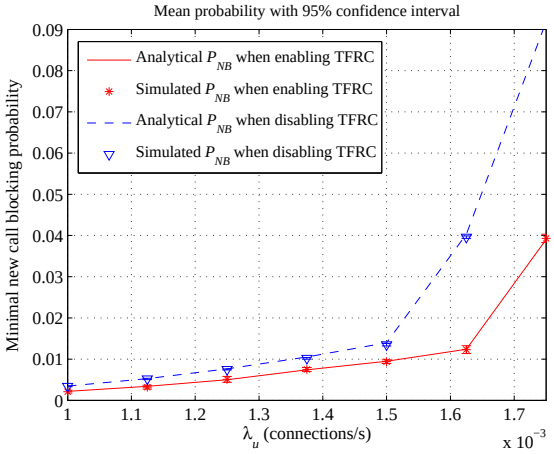


Fig. 8. Comparison of the minimal new call blocking probability between TRFC-disabled and TRFC-enabled systems given $P_{HD} \leq 2\%$ and $P_{RD} \leq 0.1\%$.

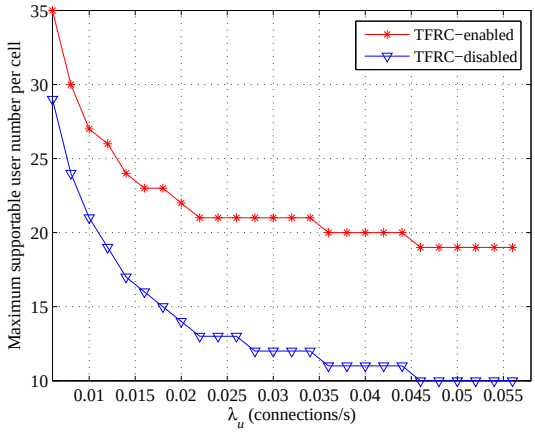


Fig. 9. Comparison of the maximal supportable user number between TRFC-disabled and TRFC-enabled systems.

$C = 78$, $\bar{K}_A = 10$, $C_{HR} = 16$, and $C_R = 8$. It is observed that the analytical results match well with the simulation results. Further, with an increase of traffic load, all the three probabilities increase. However, the recovering call dropping probability maintains a much smaller order of magnitude as compared with the new call blocking and handoff call dropping probabilities. For example, at $\lambda_u = 1.5 \times 10^{-3}$ connection/second, the recovering call dropping probability is much less than 10^{-4} while the other two are larger than 10^{-2}, demonstrating the low QoE degradation when applying the newly proposed technique.

Taking the scenario of Fig. 6 as an example, in Fig. 7 we show the optimal parameter setting of the double-threshold guard channel policy. Here, λ_u is fixed to 1.5×10^{-3} connection/second, and the upper bounds for the handoff call dropping probability (P_h) and the recovering call dropping probability (P_r) are 2% and 0.1%, respectively. Within the valid change interval (i.e., $0 < C_R < C_{HR} < C$), feasible solutions satisfying the hard constraints on the dropping probabilities are searched. For clarity, Fig. 7 depicts the search results with a step of 4 subchannels, in which feasible solutions exist in the shadowed area and are marked with $*$. It is observed that, given the reserved channel number (C_{HR}) for handoff and recovering calls, the new call blocking probability decreases with an increase of the reserved channel number (C_R) for recovering calls, due to the benefit of reducing the capability to serve handoff calls (consistent with the results in [30]). On the other hand, given a reserved channel number for recovering calls, the new call blocking probability decreases with a decrease of the reserved channel number for handoff and recovering calls, due to more channels available to new calls. As such, there exists an optimal parameter setting ($C_{HR} = 24$ and $C_R = 12$) in terms of the minimal new call blocking probability with satisfied handoff call and recovering call performance.

With the same procedure of parameter optimization, we also compare the call-level performance in terms of the minimal new call blocking probability between a system enabling TFRC and that disabling this function. Fig. 8 compares both simulation and analytical results. The analytical results for the system without TFRC have been derived following the approach proposed in [30], which is omitted here due to space limitation. As observed in Fig. 8, the new call blocking probability of the system with TFRC is much lower than its counterpart. Further, the performance gap increases with the traffic load. The main reason is that, the more the traffic per user generates, the more the opportunities we can apply TFRC to produce virtual spectrum holes, thus effectively reducing chances of blocking new calls. However, it is noteworthy that users in a TFRC-based system suffer from QoE degradation, with the recovering call dropping probability bounded by 0.1% in the simulations.

The performance improvement of time-frequency resource conversion is further studied by simulation and in terms of the number of users per cell, as shown in Fig. 9. Here, to increase the probability of accommodating a relatively large number of users, we increase the subchannel number per cell C to 320. It is clear that, for a system either with or without TFRC, as the traffic load of each user increases, the supportable user number per cell decreases, due to the increased spectrum resources required by each user. However, the system with TFRC always accommodates more users than its counterpart, especially when each user's traffic load is large, and the improvement of supportable user number per cell here can be as large as 90%.

VII. Conclusions

The exponential growth in mobile data traffic - typical predictions give an approximately 1000x increase over the next decade - poses huge challenges and requires great efforts to overcome the limitations of radio resources. Applying user behavior-aware and/or context-aware techniques to customize radio resource allocation to match the experience or preference of the individual user is one of promising solutions.

To improve the service quality that users experience in a congested network, a new resource management strategy named time-frequency resource conversion has been presented in this paper. The basic idea is to take account of user behaviors in resource allocation, by allocating radio resources

mainly to the connection that a user pays attention to. A double-threshold guard channel policy tailored for the new strategy has been proposed, and an analytical model has been established to evaluate its impact on the call-level performance of an LTE-type cellular network. Simulation results verify the accuracy of the analysis, and show that the new strategy can help the network accommodate more users in a heavy traffic condition and the potential drawback of the new technique measured by the recovering call dropping probability can be well controlled. Further works to refine time-frequency resource conversion will be carried out to address issues such as time-frequency resource conversion strategy design with a consideration of QoE balance among all users, and scheduling for interference management and QoE improvement.

APPENDIX A
DERIVATION OF SERVICE RATE OF A CONNECTION IN TFRC

Take user j in Fig. 2 as an example and assume $j_1 = j_2 = T_1$. Firstly, consider the case that the spectrum resource supplied to connection j_1 has been reduced by $r^{(m)}_{j_1,j_2}$, which however is still less than r_{j_1}. Denote by $l^{(m)}_{j_1}$ the call holding time of connection j_1 when it is in this state, where m implies the progress of TFRC. Since the data amount of connection j_1 remains the same with and without TFRC, we have

$$r_1 R_b(t_2-t_1)+R_{rv}(j_1,j_2,m-1)+(r_1-r^{(m)}_{j_1,j_2})R_b l^{(m)}_{j_1} = r_1 R_b l_{j_1} \tag{37}$$

where the three items of the left side are referred to as the delivered data amounts of connection j_1 in the following three disjointed durations, namely, before the user gives his/her attention to the new connection j_2, and before and after the spectrum resource supplied to connection j_1 has been reduced by $r^{(m)}_{j_1,j_2}$. The second term $R_{rv}(j_1,j_2,m-1)$ is given by (1). With some manipulation, we have

$$l^{(m)}_{j_1} = [l_{j_1} - \sum_{i=1}^{m-1}(1-r^{(i)}_{j_1,j_2}/r_1)\tau-(t_3-t_1)]r_1/(r_1-r^{(m)}_{j_1,j_2}). \tag{38}$$

Then we have the following result by applying the memoryless property of the exponential distribution

$$\begin{aligned}
&P\{l^{(m)}_{j_1} > x | l^{(m)}_{j_1} > 0\}\\
&= P\{(1-r^{(m)}_{j_1,j_2}/r_1) l^{(m)}_{j_1} > (1-r^{(m)}_{j_1,j_2}/r_1)x | l^{(m)}_{j_1} > 0\}\\
&= P\{l_{j_1} - \sum_{i=1}^{m-1}(1-r^{(i)}_{j_1,j_2}/r_1)\tau-(t_3-t_1) > (1-\\
&\quad r^{(m)}_{j_1,j_2}/r_1)x | l_{j_1} - \sum_{j=1}^{m-1}(1-r^{(j)}_{j_1,j_2}/r_1)\tau-(t_3-t_1) > 0\}\\
&= P\{l_{j_1} > (1-r^{(m)}_{j_1,j_2}/r_1)x\}\\
&= \exp[-\mu_1(1-r^{(m)}_{j_1,j_2}/r_1)x]
\end{aligned} \tag{39}$$

which implies that the service rate of connection j_1 in TFRC is $\mu_1(1-r^{(m)}_{j_1,j_2}/r_1)$. Secondly, if connection j_1 is frozen, i.e., $r^{(m)}_{j_1,j_2} = r_{j_1}$, it is clear that the service rate of connection j_1 is reduced to zero. In this case, the connection can terminate only after the user focuses on it again.

APPENDIX B
DERIVATION OF THE OPTIMAL TFRC STRATEGY

To solve optimization problem (10), we first remove the integrality constraint and obtain the second derivative of its objective function as follows

$$\begin{aligned}
\frac{d^2\phi}{d(r^{(m)}_{j_1,j_2})^2} = &-\frac{\alpha w(\theta+\xi)^2}{\alpha+\gamma} e^{-(\theta+\xi)(r_{j_1}-r^{(m)}_{j_1,j_2})-\theta\sum_{i=1}^{m-1}(r_{j_1}-r^{(i)}_{j_1,j_2})}\\
&-\frac{\gamma w\xi^2}{\alpha+\gamma} e^{-\xi(r_{j_1}-r^{(m)}_{j_1,j_2})}
\end{aligned} \tag{40}$$

where $\theta = \beta R_b \tau$ and $\xi = \mu_{j_1}\tau/r_{j_1}$. It is easy to check that $\frac{d^2\phi}{d(r^{(m)}_{j_1,j_2})^2} < 0$. That is, ϕ is a concave function of $r^{(m)}_{j_1,j_2}$. Then, by setting $\frac{d\phi}{dr^{(m)}_{j_1,j_2}} = 0$, we have

$$\begin{aligned}
\frac{1}{r_{j_1}} = &\frac{\alpha w(\theta+\xi)}{\alpha+\gamma}\cdot e^{-(\theta+\xi)(r_{j_1}-r^{(m)}_{j_1,j_2})}\cdot e^{-\theta\sum_{i=1}^{m-1}(r_{j_1}-r^{(i)}_{j_1,j_2})}\\
&+\frac{\gamma w\xi}{\alpha+\gamma}\cdot e^{-\xi(r_{j_1}-r^{(m)}_{j_1,j_2})},
\end{aligned} \tag{41}$$

whose root denoted by $\hat{r}^{(m)}_{j_1,j_2}$ maximizes the objective function when taking no variable constraint into account. As m increases, to make (41) holds, $r^{(m)}_{j_1,j_2}$ must increase until it reaches r_{j_1}. In general, a numerical approach such as Newton's method is required to derive the root of (41). However, if $\gamma = 0$ (e.g., in the video stream case associated with (9)), $\hat{r}^{(m)}_{j_1,j_2}$ has the following closed-form solution

$$\hat{r}^{(m)}_{j_1,j_2} = r_{j_1} + \frac{\theta}{\theta+\xi}\sum_{i=1}^{m-1}(r_{j_1}-r^{(i)}_{j_1,j_2}) - \frac{1}{\theta+\xi}\ln(w r_{j_1}(\theta+\xi)). \tag{42}$$

Finally, by studying the relation between $\hat{r}^{(m)}_{j_1,j_2}$ and variable constraint in (10), we can derive the optimal TFRC strategy, given by

$$r^{(m)^*}_{j_1,j_2} = \begin{cases} 0, & \hat{r}^{(m)}_{j_1,j_2} \le 0\\ \arg\max\limits_{r^{(m)^*}_{j_1,j_2}\in\{\lfloor\hat{r}^{(m)}_{j_1,j_2}\rfloor,\lceil\hat{r}^{(m)}_{j_1,j_2}\rceil\}} \{\phi(j_1,j_2,m)\}, & 0 < \hat{r}^{(m)}_{j_1,j_2} \le r_{j_1}\\ r_{j_1}, & \hat{r}^{(m)}_{j_1,j_2} > r_{j_1}. \end{cases} \tag{43}$$

REFERENCES

[1] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update, 2012-2017", USA, Feb. 2013.

[2] S.-W. Lee, J.-S. Park, H.-S. Lee, and M.-S. Kim, "A study on smartphone traffic analysis," in *Proc. 13th Asia-Pacific Network Operations and Management Symposium*, 2011.

[3] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das, "Understanding traffic dynamics in cellular data networks," in *Proc. IEEE/ACM INFOCOM'11*, pp. 882-890, 2011.

[4] G. Maier, F. Schneider, and A. Feldmann, "A First look at mobile handheld device traffic," in *Proc. ACM 11th International Conference on Passive and Active Network Measurement*, pp. 161-170, 2010.

[5] H. Falaki, R. Mahajan, S. Kandula, D. Lymberopoulos, R. Govindan, and D. Estrin, "Diversity in smartphone usage," in *Proc. ACM 8th international conference on Mobile systems, applications, and services*, pp. 179-194, 2010.

[6] A. Sgora and D. Vergados, "Handoff prioritization and decision schemes in wireless cellular networks: A survey," *IEEE Communications Surveys & Tutorials*, vol. 11, no. 4, pp. 57-77, 2009.

[7] Y. Ge and G. S. Kuo, "An efficient admission control scheme for adaptive multimedia services in IEEE 802.16e networks," in *Proc. IEEE VTC-2006 Fall*, 2006.

[8] D. Niyato and E. Hossain, "Radio resource management games in wireless networks: An approach to bandwidth allocation and admission control for polling service in IEEE 802.16," *IEEE Wireless Commun.*, vol. 14, no. 1, pp. 27-35, Feb. 2007.
[9] B. Al-Manthari, N. Nasser, and H. Hassanein, "Congestion pricing in wireless cellular networks," *IEEE Communications Surveys & Tutorials*, vol. 13, no. 3, pp. 358-371, 2011.
[10] H. Zhou, K. Sparks, N. Gopalakrishnan, P. Monogioudis, F. Dominique, P. Busschbach, and J. Seymour, "Deprioritization of heavy users in wireless networks," *IEEE Commun. Mag.*, vol. 49, no. 10, pp. 110-117, Oct. 2011.
[11] M. Proebster, M. Kaschub, and S. Valentin, "Context-aware resource allocation to improve the quality of service of heterogeneous traffic," in *Proc. IEEE ICC'11*, 2011.
[12] M. Proebster, M. Kaschub, T. Werthmann, and S. Valentin, "Context-aware resource allocation for cellular wireless networks," *EURASIP J. Wireless Commun. and Networking (WCN)*, no. 216, July 2012.
[13] T. Werthmann, M. Kaschub, M. Proebster, and S. Valentin, "Simple channel predictors for lookahead scheduling," in *Proc. IEEE 75th Vehicular Technology Conference (VTC Spring)*, 2012.
[14] M. Bennis, M. Simsek, A. Czylwik, W. Saad, S. Valentin, and M. Debbah, "When cellular meets wifi in wireless small cell networks," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 44-50, June 2013
[15] F. Pantisano, M. Bennis, W. Saad, S. Valentin, and M. Debbah, "Matching with externalities for context-aware user-cell association in small cell networks," in *Proc. IEEE GLOBECOM'13*, 2013.
[16] H. Abou-zeid, H. S. Hassanein, and S. Valentin, "Optimal predictive resource allocation exploiting mobility patterns and radio maps," in *Proc. IEEE GLOBECOM'13*, 2013.
[17] H. Abou-zeid and H. S. Hassanein, "Predictive green wireless access exploiting mobility and application information," *IEEE Wireless Communications*, vol. 20, no. 5, pp. 92-99, Oct. 2013.
[18] H. Abou-zeid and H. S. Hassanein, "Efficient lookahead resource allocation for stored video delivery in multi-cell networks," in *Proc. IEEE WCNC'14*, 2014.
[19] S. Haykin, "Cognitive radio: Brain-empowered wireless communications," *IEEE J. Sel. Areas Commun.*, vol. 23, no. 2, pp. 201-220, Feb. 2005.
[20] D. Willkomm, S. Machiraju, J. Bolot, and A. Wolisz, "Primary user behavior in cellular networks and implications for dynamic spectrum access," *IEEE Commun. Mag.*, vol. 47, no. 3, pp. 88-95, Mar. 2009.
[21] Y. Zhang, Y, Xiao, and H. Chen, "Queueing analysis for OFDM subcarrier allocation in broadband wireless multiservice networks," *IEEE Trans. Wireless Commun.*, vol. 7, no. 10, pp. 3951-3961, Oct. 2008.
[22] P. Tran-Gia, N. Jain, and K. Leibnitz, "Code division multiple access wireless network planning considering clustered spatial customer traffic," in *Proc. 8th International Telecommunication Network Planning Symposium*, Italy, Oct. 1998.
[23] S. M. Ross, *Introduction to Probability Models*, 9th ed., Academic Press, 2007, pp. 310-311.
[24] M. Fiedler, T. Hossfeld, and P. Tran-Gia, "A generic quantitative relationship between qualtiy of experience and quality of service," *IEEE Network*, vol. 24, no. 2, pp. 36-41, March/April 2010.
[25] T. H. Luan, L. X. Cai, and X. Shen, "Impact of network dynamics on users' video quality: Analytical framework and QoS provision," *IEEE Trans. Multimedia*, vol. 12, no. 1, pp. 64-78, Jan. 2010.
[26] H. Shan, Z. Ni, W. Zhuang, A. Huang, and W. Wang, "State transition analysis of time-frequency resource conversion-based call admission control for LTE-type cellular network," Available at: arXiv:1312.0333
[27] R. Yu, Y. Zhang, M. Huang, and S. Xie, "Cross-layer optimized call admission control in cognitive radio networks," *Mobile Networks and Applications*, vol. 15, no. 5, pp 610-626, Oct. 2010.
[28] Timed Markov Models. http://www.mathpages.com/home/kmath589/kmath589.htm
[29] M. Sawahashi, Y. Kishiyama, A. Morimoto, D. Nishikawa, and M. Tanno, "Coordinated multipoint transmission/reception techniques for LTE-Advanced," *IEEE Wireless Communications*, pp. 26-34, June 2010.
[30] R. Ramjee, D. Towsley, and R. Nagarajan, "On optimal call admission control in cellular networks," *Wireless Networks*, vol. 3, no. 1, pp. 29-41, Mar. 1997.

Hangguan Shan (M'10) received his B.Sc. and Ph.D. degrees respectively from Zhejiang University and Fudan University, in 2004 and 2009, all in electrical engineering. He was a postdoctoral research fellow in University of Waterloo from 2009 to 2010. In February 2011, he joined the Department of Information Science and Electronic Engineering, Zhejiang University, as an Assistant Professor. He is a co-recipient of the Best Industry Paper Award from IEEE Wireless Communications and Networking Conference (WCNC) 2011, Quintana-Roo, Mexico. His current research focuses on cross-layer protocol design, resource allocation and QoS provisioning in wireless networks. Dr. Shan has served on the Technical Program Committee (TPC) as member in various international conferences including for example IEEE GLOBECOM, IEEE ICC, IEEE WCNC, IEEE VTC. He has also served as the Publicity Co-Chair for the third and fourth IEEE International Workshops on Wireless Sensor, Actuator and Robot Networks (WiSARN), and the fifth International Conference on Wireless Communications and Signal Processing (WCSP).

Zhifeng Ni received the B.Sc. degree in Communication Engineering from Hangzhou Dianzi University, and M.Sc. degree in Information and Communication Engineering from Zhejiang University, in 2011 and 2014, respectively. He is with MicroStrategy, Hangzhou, China, working as a software engineer.

Weihua Zhuang (M'93-SM'01-F'08) has been with the Department of Electrical and Computer Engineering, University of Waterloo, Canada, since 1993, where she is a Professor and a Tier I Canada Research Chair in Wireless Communication Networks. Her current research focuses on resource allocation and QoS provisioning in wireless networks. She is a co-recipient of the Best Paper Awards from the IEEE International Conference on Communications (ICC) 2007 and 2012, IEEE Multimedia Communications Technical Committee in 2011, IEEE Vehicular Technology Conference (VTC) Fall 2010, IEEE Wireless Communications and Networking Conference (WCNC) 2007 and 2010, and the International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness (QShine) 2007 and 2008. She received the Outstanding Performance Award 4 times since 2005 from the University of Waterloo, and the Premiers Research Excellence Award in 2001 from the Ontario Government. Dr. Zhuang is a Fellow of the IEEE, a Fellow of the Canadian Academy of Engineering (CAE), a Fellow of the Engineering Institute of Canada (EIC), and an elected member in the Board of Governors of the IEEE Vehicular Technology Society. She was the Editor-in-Chief of IEEE Transactions on Vehicular Technology (2007-2013), the Technical Program Symposia Chair of the IEEE Globecom 2011, and an IEEE Communications Society Distinguished Lecturer (2008-2011).

Aiping Huang (SM'08) graduated from Nanjing Institute of Post and Telecommunications, China in 1977, received M.Sc. degree from Nanjing Institute of Technology, China in 1982, and received Licentiate of Tech. degree from Helsinki University of Technology (HUT), Finland in 1997. She worked from 1977 to 1980 as an engineer at Design and Research Institute of Post and Telecom. Ministry, China. From 1982 to 1994, she was with Zhejiang University (ZJU), China as an assistant professor and then associate professor in the Dept. of Scientific Instrumentation. She was a visiting scholar and then a research scientist at HUT from 1994 to 1998. From 1998, she is a full professor of the Dept. of Information and Electronics Engineering at ZJU. She serves as the director of Zhejiang Provincial Key Laboratory of Info. Network Tech., and the vice chair of IEEE ComSoc Nanjing Chapter. She published a book and more than 160 papers in refereed journals and conferences on signal processing, communications and networks. Her current research interests include cognitive radio networks and cross-layer design, planning and optimization of cellular mobile communication networks and heterogeneous networks.

Wei Wang (S'08-M'10) received his B.S. degree in Communication Engineering and Ph.D. degree in Signal and Information Processing in 2004 and 2009, respectively, both from Beijing University of Posts and Telecommunications, China. Now, he is an associate professor with the Department of Information Science and Electronic Engineering, Zhejiang University, China. He visited University of Michigan, Ann Arbor, USA from 2007 to 2008, and began to visit Hong Kong University of Science and Technology, Hong Kong since 2013. His research interest mainly focuses on cognitive radio networks, spectrum aggregation, green communications, and radio resource management for wireless networks. He is the editor of the book "Cognitive Radio Systems" (Intech, 2009) and serves as an editor for Transactions on Emerging Telecommunications Technologies. He also serves as a co-chair or TPC member for several international conferences.