Traffic-Aware QoS Provisioning and Admission Control in OFDMA Hybrid Small Cells

Ravikumar Balakrishnan, *Student Member, IEEE*, and Berk Canberk, *Member, IEEE*

***Abstract*—Recent problems in wireless cellular networks, such as network capacity and indoor coverage, have been addressed by the orthogonal frequency-division multiple-access (OFDMA) small-cell deployments of next-generation Long-Term Evolution Advanced (LTE-A) cellular systems. In this new paradigm, the deployment of hybrid-access small cells can be seen as an optimal solution since they serve both the registered indoor users and unregistered neighboring users in the small-cell coverage area. However, effective quality-of-service (QoS) provisioning and fair admission control pop up as two crucial challenges in these hybrid accesses. Motivated by these challenges, in this paper, we propose a traffic-aware OFDMA hybrid small-cell deployment for QoS provisioning and an optimal admission control strategy for next-generation cellular systems. The traffic awareness in the proposed framework is provided by deriving a novel traffic-aware utility function, which differentiates the user QoS levels with the user's priority indexes, channel conditions, and traffic characteristics. An optimization procedure is formulated, and a novel heuristic is also developed to solve the traffic-aware scheduling problem under transmitted power constraints. To further enhance the proposed scheme, an admission control algorithm based on the utility function is also proposed. The proposed QoS awareness and admission control mechanism are evaluated by thorough simulations, and we show that our proposed framework achieves an optimum QoS performance in terms of total throughput and traffic delay.**

***Index Terms*—Admission control, heterogeneous traffic, hybrid access, power constraint, quality of service (QoS), scheduling, small cells, utility.**

I. Introduction

THE GROWING capacity needs for real-time traffic and delay-sensitive quality-of-service (QoS) requirements in current cellular systems are caused by recent technological advances in wireless end-user devices. The Long-Term Evolution Advanced (LTE-A), which is a part of the next-generation cellular system deployments, can be seen as an effective solution to fulfil these sophisticated requirements such as large amount of multimedia and data traffic. More specifically, the existing wireless cellular architecture that consists of a single macrocell layer can be underlaid by several LTE-A-based small cells, which are orchestrated by low-power small-cell base stations. These small cells are classified with the aid of their coverage areas as picocells, macrocells, and femtocells [2]. These LTE-A small cells increase network capacity through the spatial reuse of spectrum and improve indoor cellular coverage [2], [5].

To come up with effective solutions for the indoor cellular coverage challenge, small cells are deployed by considering different access schemes. The *Open-Access Small Cells*, which serve all mobile users, i.e., both small-cell users (SUs) and external users (EUs) in the coverage area without any limitation, offer the largest increase in network capacity while degrading the QoS of SUs. The QoS degradation is particularly large when the number of EUs increases or when the EUs are running bandwidth-hungry applications [2]. The *Closed-Access Small Cells* reserve exclusive access for SUs. This approach is capable of providing better QoS performance for the SUs; however, the performance can also be significantly affected if there are nearby EUs that cause strong interference to the small-cell network. The *Hybrid-Access Small Cells* utilize ad hoc schemes to achieve QoS guarantees for the SUs in the presence of unregistered EUs. The hybrid-access scheme can provide differentiable service to SUs and EUs and, thus, is the most optimal solution for QoS provisioning. Consequently, being the extended work of our previous study in [2], in this paper, we also consider the hybrid-access small cells.

There are several studies that investigate the hybrid-access small-cell schemes in different perspectives. Choi *et al.* in [6] consider hybrid access in femtocells where they propose a fixed probability p for EUs to be able to connect to a femtocell based on the computation of the carrier-to-interference (C/I) ratio at the location of the EUs. In [16], Valcarce *et al.* propose a hybrid-access scheme for orthogonal frequency-division multiple-access (OFDMA) small cells where a limited number of subchannels v is reserved for EU access. Although the outage probability is shown to notably decrease for EUs in this scheme, increasing v can affect the throughput achieved by SUs. In addition, a lower outage probability does not necessarily equal the QoS performance of both SUs and EUs. Further work has been conducted under a hybrid-access approach; for example, in [19], Xia *et al.* propose an adaptive access control strategy based on the average cellular user density. It is shown that the ergodic rate for EUs is notably increased under the low-user-density case, whereas under the high-user-density case, the rate gain for EUs is not significant. There are also several scheduling policies in the current literature. The EXP rule proposed in [12] is shown to offer improved QoS performance in terms of throughput and delay over proportionally fair (PF) and Modified Largest Weighted Delay First (M-LWDF) scheduling schemes when there is a mixture of real-time and nonreal-time

Manuscript received February 12, 2013; revised July 12, 2013; accepted August 18, 2013. Date of publication August 29, 2013; date of current version February 12, 2014. The review of this paper was coordinated by Prof. Y. Cheng.

R. Balakrishnan is with the Broadband Wireless Networking Laboratory, School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: rbalakrishnan6@ece.gatech.edu).

B. Canberk is with the Department of Computer Engineering, Computer and Informatics Faculty, Istanbul Technical University, Istanbul 34469, Turkey (e-mail: canberk@itu.edu.tr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVT.2013.2280124

users in the system. Another popular approach to QoS scheduling is the utility-based approach. Scheduling rules based on maximizing utility, which represents the amount of satisfaction that can be obtained by scheduling a resource for a user, have been proposed in [10], [14], and [15]. The utility functions here are defined as decreasing functions of the packet delay in the queue. In [8], [14], and [15], although the scheduling rule is shown to achieve throughput optimality, the utility function does not provide strict bounds on delay. In addition, many of the given policies only consider subcarrier allocation without any power constraints. Since small cells are limited by hardware on the total transmit power and by the interference they cause to the overlapping macrocell layer, power constraints become significant in our problem.

The ultimate objective of having a hybrid-access scheme is that QoS for SUs is provisioned while the EUs will also be served to maximize resource utilization. In all of the aforementioned studies, the traffic characteristics are not considered while doing the scheduling and access controls in LTE-A small cells. However, such considerations are crucially needed to obtain most QoS optimal small-cell deployment. Here, for the scheduling problem, one of the fundamental performance metrics is network stability, which guarantees the queue size to be bounded for all packet arrivals within the capacity region. The capacity region is defined as the convex hull of the m-element set of all arrival rate vectors $\vec{\lambda}(m)$ as

$$\vec{\lambda}(m) = [\lambda_1(m), \ldots, \lambda_N(m)] \qquad \forall m \tag{1}$$

for a system containing $N(m)$ users that the system can support without making the queues unstable. If the channel state can be represented by a finite number of states M, the capacity region depends on the link transmission rate vector $\vec{r}(m)$ as

$$\vec{r}(m) = [r_1(m), \ldots, r_N(m)] \qquad \forall m \tag{2}$$

under state m, where $m \in M$.

Scheduling policies such as Maximum Delay Scheduling can stabilize the queues for admissible arrival rates. At the same time, these policies can result in poor delay performance and unfair allocation for the SUs if the EU traffic is bursty. One of the objectives of this extended version paper of [2] is to propose and evaluate an optimal scheduling scheme that accounts for the higher layer traffic characteristics for QoS provisioning in hybrid small cells. In addition, when the number of users in the cell increases or when the traffic arrivals are outside the capacity region, the scheduler cannot handle fair allocation toward achieving end-user QoS. Therefore, in this paper, an admission control procedure tightly coupled to the scheduling policy is also proposed.

Being the extended work of our previous study in [2], we summarize the main contributions of this paper as follows.

- By adopting the specific OFDMA-based hybrid small cell in [2], we enhance the QoS provisioning of the traffic-aware utility function for a large amount of heterogeneous users.

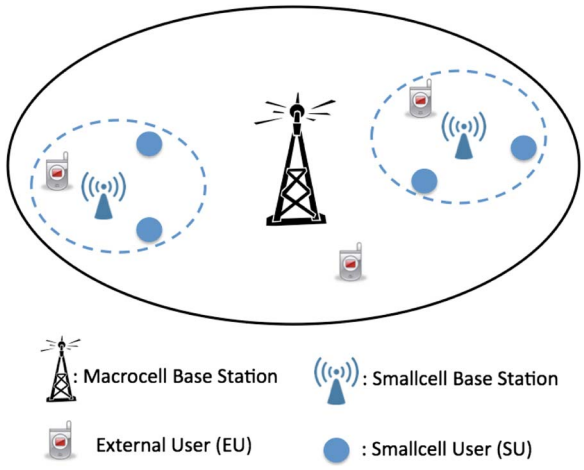

Fig. 1. Considered network topology.

- We propose a novel optimal subcarrier-allocation algorithm to perform QoS-based scheduling using traffic utility as the cost function.
- To enhance the scheduling policy, we develop an admission control algorithm based on the proposed traffic utility function.

The rest of this paper is organized as follows: In Section II, the network model for OFDMA hybrid small cells and the heterogeneous traffic models are presented. Section III details the traffic-aware utility function. Here, we define and solve the problem of constrained QoS scheduling using traffic-aware utility-based optimization. Here, an optimal scheduling heuristic to solve the optimization is also given. Moreover, a traffic-aware utility-based admission control algorithm is explained in this section. We provide thorough performance evaluations in Section IV in terms of delay and variances of the traffic. We conclude this paper by summarizing the achievements in Section V.

II. Network and Heterogeneous Traffic Modeling

A. Network Model

In this paper, we consider a macrocell base station that orchestrates the EUs in a given coverage area, as shown in Fig. 1. There are some predefined hybrid-access OFDMA small-cell networks in the network. These small cells serve their preregistered users (SUs) and some EUs that are under the small-cell coverage area, as shown in the figure. Moreover, we consider a time-varying, bursty, and location-dependent wireless channel, which also poses a major challenge in achieving optimal QoS performance and scheduling. To control all these challenges, we design our proposed mechanism taking into consideration the exchange of system dynamics such as channel conditions, location, queue state, and application layer requirements to maintain QoS satisfaction. Under such a setup, the time is slotted, and the wireless channel is assumed to be unvarying during the slot length. At the beginning of each slot, the scheduler obtains the channel gain from the lower layers through user feedback. Using this information, the data rates achievable and power

required for the user in the time slot is determined. Based on these parameters, the scheduling algorithm performs resource scheduling to achieve the QoS objectives [2].

The downlink of an OFDMA hybrid small-cell network overlaid on a macrocell coverage area is considered, as shown in Fig. 1, with a small-cell access point (SAP) serving $\mathcal{N}$ users $\{1, 2, \ldots, N\}$. Out of this, $\mathcal{F}$ represents the set of all SUs $\{1, 2, \ldots, F\}$, and $\mathcal{E}$ represents the set of all EUs $\{1, 2, \ldots, E\}$, and therefore, $\mathcal{N} = \mathcal{F} \cup \mathcal{E}$. B represents the total system bandwidth consisting of K subcarriers. Hence, the bandwidth of each subcarrier is represented as $\Delta B = B/K$. The time is slotted, and each slot has a duration of T_s equivalent to the coherent time of the channel.

The signal-to-noise ratio (SNR) gap [3] is defined as

$$\text{SNR}_{gap} = \frac{-1.5}{\ln(5 * \text{BER})} \tag{3}$$

where BER is a given bit error rate for user n to transmit on subcarrier k. The transmission rate $r_{n,k}$ for user n on subcarrier k is given as

$$r_{n,k} = \frac{B}{K} * \log\left(1 - \frac{1.5 * |h_{n,k}|^2 * p_{n,k}}{\ln(5 * \text{BER}) * \sigma^2}\right) \tag{4}$$

where $h_{n,k}$ represents the channel gain of user n transmitting on subcarrier k. $p_{n,k}$ represents the required power for user n to transmit on subcarrier k for a given BER. The noise power over a subcarrier is represented as σ^2. Each user can be assigned several subcarriers with the constraint that the same subcarrier cannot be assigned to different users in the same slot. This is represented by the binary variable $s_{n,k}(t)$ indicating whether subcarrier k is assigned to user n or not in slot t. Hence, the subcarrier assignment constraint is given as $\sum_{n=1}^{N} s_{n,k}(t) = 1$. Therefore, the maximum achievable data rate to user n in slot t is given by

$$\mu_n(t) = \sum_{k=1}^{K} s_{n,k}(t) r_{n,k}(t). \tag{5}$$

Inserting (4) into (5), we obtain the maximum achievable data rate per user as

$$\mu_n(t) = \sum_{k=1}^{K} s_{n,k}(t) * \frac{B}{K} * \log\left(1 - \frac{1.5 * |h_{n,k}(t)|^2 * p_{n,k}(t)}{ln(5 * \text{BER}) * \sigma^2}\right). \tag{6}$$

B. Heterogeneous Traffic Models

The SAP has queues corresponding to each of the n user types it serves. The arrival process $\Lambda_n(t)$ represents the number of packet arrivals at queue n in time t. Here, the mean arrival rate is given by $\lambda_n \triangleq E[\Lambda_n(t)]$, and the mean arrival rate vector is given as $\overrightarrow{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_N)$. $\overrightarrow{Q(t)} = (Q_1(t), Q_2(t), \ldots, Q_N(t))$ represents the queue length vector. The waiting time of a packet in the queues is represented by vector $\overrightarrow{W(t)} = (W_1(t), W_2(t), \ldots, W_N(t))$. The queue evolves according to the Discrete-Time Queueing Law as

$$Q_n(t+1) = \max\left(Q_n(t) - \mu_n(t) T_s, 0\right) + \Lambda_n(t). \tag{7}$$

By Little's Law, the waiting time of user n in slot t is given by

$$W_n(t) = \frac{\overline{Q_n}(t)}{\lambda_n} \tag{8}$$

where $\overline{Q_n}(t)$ is the average queue length.

The users served under SAP are grouped into three classes using three different queuing disciplines as in [4]. These classes are as follows.

- Constant Bit Rate (CBR) Users: These users have deterministic behaviors and are modeled by a D/G/1 queuing system. The average waiting time is calculated as

$$W_{CBR,n} = \frac{\lambda_{CBR,n} \sigma^2_{CBR,\overline{X_n}}}{2(1 - \lambda_{CBR,n} \overline{X_n})}, \qquad n \in N \tag{9}$$

  where $\lambda_{CBR,n}$, $\sigma^2_{CBR,\overline{X_n}}$, and $\overline{X_n}$ are the mean arrival rate, the variance of the service time, and the mean service time, respectively, and $\overline{X_n} = E[1/\mu_n]$ for OFDMA.
- Video-Streaming Users: These users are modeled using *Gamma Distribution* with shape parameter s and a G/G/1 queuing system where the average waiting time is

$$W_{Vid,n} = \frac{\lambda_{Vid,n}\left(\sigma^2_{Vid,\overline{X_n}} + s/\lambda_{Vid,n}\right)}{2(1 - \lambda_{Vid,n}\overline{X_n})}, \qquad n \in N. \tag{10}$$

- Best Effort (BE) Users: The BE users can be modeled using an M/G/1 queueing system where the average queue waiting time is expressed as

$$W_{BE,n} = \frac{\lambda_{BE,n}\left(\sigma^2_{BE,\overline{X_n}} + \sigma^2_T\right)}{2(1 - \lambda_{BE,n}\overline{X_n})}, \qquad \forall n \in N \tag{11}$$

  where σ^2_T is the variance of the interarrival time, and $\rho_{BE} = \overline{\lambda_{BE}}/\overline{X_m}$ is the utilization.

Our objective is to stabilize the queues of all SUs and EUs when the arrivals are inside the capacity region. In addition, we want to offer QoS performance for different SU traffic types in terms of maximizing throughput and minimizing delay. These objectives together present an interesting case of QoS provisioning. Scheduling policies such as PF scheduling and M-LWDF are not suitable in the presence of heterogeneous traffic since they do not provide bounded delay performance.

III. Proposed Framework

The proposed framework is embedded into each SAP and is shown in Fig. 2. It has four main parts: *the QoS Classification of Heterogeneous Traffic*, *the Calculation of Utility Function*, *Traffic-Aware Admission Control*, and *Power Constraint Scheduling*.

The QoS classification of heterogeneous traffic part uses the SU and EU requests, as shown in Fig. 2, to calculate the average waiting time of each user types. These calculations are aforementioned in the previous sections as BE, Video-Streaming, and CBR traffic [see (13)–(15)]. The need for achieving diverse QoS requirements for heterogeneous traffic classes calls for

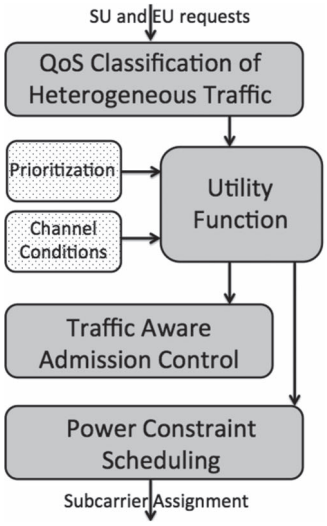

Fig. 2. Proposed framework.

an improved scheduling rule that can deal with their unique attributes. In particular, under the hybrid small-cell setup, the packet delay for SUs must be bounded in the presence of EUs. Wang and Akyildiz, in [17] and [18], show that if the message size of users exhibit heavy-tail characteristics with an index α, then the delay has an infinite mean and infinite variance for $\alpha < 1$ and $\alpha < 2$, respectively. The authors also propose a modified maximum weight-α scheduling policy that allocates channels for users based on queue size raised to the power α to guarantee bounded delay mean and variance.

In this paper, we propose a novel traffic-aware utility-based scheduling policy (TA-Utility) for hybrid small cells to effectively provision QoS [2]. The scheduler is fed with the information of the channel state and the traffic information to make scheduling decisions at every time slot t based on the computation of the utility function. Our scheduling policy not only is weighted as in [9] but considers several heterogeneous traffic types that are modeled by different queuing disciplines as well. In the original weighted alpha scheduler [9], the only parameter is the weight of the waiting queue size; however, in our model, we have also the different channel conditions (i.e., different queue-server models) and different utilities. The three different utility functions for BE, Video-Streaming, and CBR traffic [see (13)–(15)] are derived in the following paragraphs.

The utility function, as shown in Fig. 2, associated with the allocation of subcarrier k to user n is defined as

$$U_{n,k}(t) = \gamma_n W_n^{\alpha}(t) r_{n,k}(t) \tag{12}$$

where $\gamma_n = a_n/\bar{r}_n$. Here, a_n represents the *priority index* and can be tuned for SUs and EUs to achieve the required QoS for each user type. $\bar{r}_n$ represents the average transmission rate for user n over all subcarriers measured over a time window. α is the exponent of the average waiting time W_n for packets in queue n. This is the *traffic coefficient* that takes unique values for different traffic classes.

Finally, by inserting the transmission rate $r_{n,k}(t)$ obtained in (4) and each of the average waiting times $W_n(t)$ of (9)–(11) into our utility function (12), we obtain the utility functions for heterogeneous user types for OFDMA small cells as follows.

The traffic-aware utility function for CBR users, i.e.,

$$U_{n,k}^{CBR}(t) = \gamma_n * \frac{B}{K} * \left(\frac{\lambda_{CBR,n}\sigma^2_{CBR,\overline{X_n}}}{2(1-\lambda_{CBR,n}\overline{X_n})} \right)^{\alpha} * \log\left(1 - \frac{1.5 * |h_{n,k}|^2 * p_{n,k}}{\ln(5 * \mathrm{BER}) * \sigma^2}\right). \tag{13}$$

The traffic-aware utility function for Video-Streaming users, i.e.,

$$U_{n,k}^{vid}(t) = \gamma_n * \frac{B}{K} * \left(\frac{\lambda_{Vid,n}\left(\sigma^2_{Vid,\overline{X_n}} + s/\lambda_{Vid,n}\right)}{2(1-\lambda_{Vid,n}\overline{X_n})} \right)^{\alpha} * \log\left(1 - \frac{1.5 * |h_{n,k}|^2 * p_{n,k}}{\ln(5 * \mathrm{BER}) * \sigma^2}\right). \tag{14}$$

The traffic-aware utility function for BE users, i.e.,

$$U_{n,k}^{BE}(t) = \gamma_n * \frac{B}{K} * \left(\frac{\lambda_{BE,n}\left(\sigma^2_{BE,\overline{X_n}} + \sigma^2_T\right)}{2(1-\lambda_{BE,n}\overline{X_n})} \right)^{\alpha} * \log\left(1 - \frac{1.5 * |h_{n,k}|^2 * p_{n,k}}{\ln(5 * \mathrm{BER}) * \sigma^2}\right). \tag{15}$$

A. Intuition

The utility function defined in (12) is aimed at achieving the heterogeneous objectives of QoS perceived by different user types. In (12), the utility function is proportional to the waiting time of user n's packet raised to the power α. This implies that as the waiting time of the packet of a user becomes large, the QoS requirement of that user is high. Hence, this user has a high priority during scheduling.

Parameter α is specified in the exponent to enforce QoS differentiation between services. For the real-time users, the delay performance is critical, and they have a strict deadline on the waiting time of the packet. For CBR user types, the throughput and delay performance are important. Since different users have varying degrees of delay bounds, varying the values of α can impact the strictness of the QoS requirement. In other words, a larger value of α (considering $W > 1$ units) specifies that the queue needs to be served more urgently. It can be observed that by setting the value of α to 1 for all traffic types, the utility function shows similarity to the M-LWDF rule. The relationship between the choice of α and stability conditions will be discussed later in this section.

In addition to considering packet delay, (12) also captures the transmission rate for user n to transmit on subcarrier k. This enables users that have better channel quality than other users to have higher priority during scheduling. In the γ_n parameter, we use the average data rate for user n over all subcarriers. Therefore, $U_{n,k}(t)$ is not simply a function of the instantaneous channel quality but the average channel quality. To provide fair allocation for SUs in the presence of bursty EU traffic, parameter a_n can be set to a higher value for SUs. In other words, a_n is used as a bias factor to increase the priority of the SUs compared with the EUs. In this paper, the choice of the priority index a_n we utilize for the simulations is based on the results provided in [1], where $a_n = (-log\delta_n/T_n)$.

$\delta_n = Prob\{W_n > T_n\}$ indicates the maximum delay violation probability, and T_n is the delay threshold for user n. To differentiate between EUs and SUs, we utilize a higher value of δ_n for the EUs compared with the SUs. The value of T_n depends on the traffic type. The values of the parameters utilized in the simulations are explained in Section IV.

Discussion on Stability and Delay Bounds: Markakis *et al.* in [9] prove that the mean of the queue length under steady state becomes infinite when scheduling policies such as Maximum Weight Scheduling are utilized, where the tail index of the arrival process is not considered in the scheduling decisions, considering that at least one of the arrival processes follows a heavy-tailed distribution (where the tail coefficient ≤ 2). Under the presence of heavy-tailed distribution, the relation between the α parameter used for a maximum weight-α is given in [9] such that if $\alpha_{heavy} + 1 < C_{heavy}$ and $\alpha_{light} + 1 < C_{light}$, then the system of queues containing such a mix of heavy- and light-tailed arrivals is stable. Here, C is the tail coefficient. In such a case, it is also shown that $\bar{Q}_{tail}^{\alpha_{tail}} < \infty$, where $\bar{Q}_{tail}$ is the steady-state mean queue length of queue type $tail$, such that $tail \in \{heavy, light\}$.

In our scheduling policy, we propose that the scheduler depends on the waiting time of the packet raised to the power of α. Using Little's Law, the relationship between the average queue size $\bar{Q}_{tail}$ and the (average) waiting time W_{tail} is given as $W_{tail} = \bar{Q}_{tail}/\lambda_{tail}$. Hence, with the proper choices of α for the different traffic types, it can be deduced that the bounded queue size and, hence, the bounded average waiting time can be achieved under the proposed utility-based scheduling approach. Such an approach requires the knowledge of the α values of different traffic arrival processes. The mathematical proof for the bounded average waiting time for the proposed approach is left for future work.

B. Power-Constrained Utility-Based Scheduling

The subcarrier allocation with power constraints using the proposed utility function is performed based on the following optimization objective:

$$\max_{S} \quad \sum_{n=1}^{N}\sum_{k=1}^{K} U_{n,k}(t)s_{n,k}(t) \tag{16}$$

$$\text{subject to} \quad \sum_{n=1}^{N} s_{n,k}(t) = 1 \tag{17}$$

$$s_{n,k}(t) \in \{0,1\} \tag{18}$$

$$\sum_{n=1}^{N} p_{n,k}s_{n,k}(t) \leq P_s \tag{19}$$

$$\sum_{n=1}^{N}\sum_{k=1}^{K} p_{n,k}s_{n,k}(t) \leq P_{tot} \tag{20}$$

where the optimization variable S is the subcarrier allocation matrix with order $N \times K$. $S_i = \{s_{i,1}, s_{i,2}, \ldots, s_{i,K}\}$ is the set of subcarriers allocated to node i, P_s is the maximum allowed subcarrier power, and P_{tot} is the total transmission power available at the SAP.

C. Minimal Algorithm for Utility-Based Subcarrier Assignment

The given optimization problem can be classified into the multiple-choice knapsack problem (MCKP) with additional constraints on the maximum weight of each item. The MCKP is defined as a binary knapsack problem with additional disjoint multiple-choice constraints [13]. The constraints are such that the items are divided into multiple classes and only one item is to be selected from each of the classes. The MCKP has been shown to be NP-hard since the KP problem needs to be solved in the process; nevertheless, through dynamic programming, it is shown to be solved in pseudopolynomial time [7]. A minimal algorithm for solving the MCKP is presented in [11]. First, the integrality constraint $s_{n,k}(t) \in \{0,1\}$ is relaxed to $0 \leq s_{n,k}(t) \leq 1$ to obtain the linear MCKP (LMCKP). A simple partitioning algorithm is proposed for solving the LMCKP and obtaining a feasible solution. Using the initial solution, dynamic programming is used to solve the MCKP. The partitioning algorithm can compute in $O(n)$ time a small subset of items called as the core of classes to be considered for the optimal value. New classes are then added to the core by need.

Applied to the utility-based subcarrier assignment problem, the classes correspond to the set of subcarriers $\mathcal{K}$. Each item corresponds to a node n to be assigned for a subcarrier k. In our problem, we have an additional constraint in the form of maximum per-subcarrier power. Only one node among N nodes is assigned for subcarrier k, given that it satisfies the global and local power constraints. $U_{n,k}(t)$ indicates the profit obtained when $p_{n,k}(t)$ is the power allocated. The output of the algorithm is the matrix S of dimension $N \times K$ with assignment indicators $s_{n,k}(t)$. The algorithm is presented in Algorithm 1. The procedure partitionalgo() provides an LP-optimal solution for the relaxed LMCKP. The function reduceclass() uses upper-bound computation and dominance tests to prune nodes for each subcarrier, whereas reduceset() checks and updates the CurrentBestSolution if a state improves the lower bound. The computational complexity of the 1-D MCKP is shown to be $O(n + P_{tot}\sum_{R_k \in C} num_k)$, where R_k is the set of subcarriers in core C, and num_k is the number of nodes considered for each subcarrier in the core. The algorithm obtains the optimal solution in linear time for a small core and in pseudopolynomial time when the core is large. When the number of users is not considerably large, it can be also shown that performing adaptive modulation combined with subcarrier assignment does not increase the algorithm complexity significantly. This is a reasonable assumption since small cells, on average, support a few tens of users.

D. Traffic-Aware Utility-Based Admission Control

Admission control algorithms focus on balancing the load on the system based on a set of rules. It is possible that in a time-varying multicarrier system, some users experience poor channel gain for a significant amount of time over all channels. As a result, the scheduler needs to allocate a large amount of resources to these users. The rest of the users experience a significant drop in the data rate achieved. Furthermore, their head-of-line packet delay, which is the delay experienced by

the packet at the head of the queue, can potentially become unbounded.

Similarly, when a new user requests for resources with the SAP, the admission control procedure must evaluate if scheduling can be performed for the new user without affecting the QoS of the existing users. If the new user is a high-priority user, the algorithm must decide which among the existing connections need to be released or assigned lesser resources. Due to this dependence between the scheduler and the admission control procedures, we advocate the joint functioning of scheduling and admission control procedures. The objective of the admission control procedure, therefore, must be to *admit as many users as possible while preserving the feasibility of the scheduler*.

Algorithm 1: Minimal Algorithm for TA-Utility based Subcarrier Assignment

input : Utility matrix U of dimension $N \times K$, Required Power matrix p of dimension $N \times K$

1 $\{a, b_k, s_{b_a,a}, s_{b'_a,a}\} \Leftarrow$ `partitionalgo` (U, p, P_{tot});
where $a :=$ fractional subcarrier, $\{b_k\} :=$ LP Optimal soln., $s_{b_a,a}, s_{b'_a,a} :=$ fractional variables in a;
2 Calculate $\lambda = (U_{b'_a,a} - U_{b_a,a})/(p_{b'_a,a} - p_{b_a,a})$;
3 Calculate $\lambda_k^+ = \max_{n \in N, p_{n,k} > p_{b_k,k}} (U_{n,k} - U_{b_k,k})/(p_{n,k} - p_{b_k,k})$, $k = 1, ..., K$, $k \neq a$;
4 Calculate $\lambda_k^- = \min_{n \in N, p_{n,k} < p_{b_k,k}} (U_{b_k,k} - U_{n,k})/(p_{b_k,k} - p_{n,k})$, $k = 1, ..., K$, $k \neq a$;
5 Gradients $L^+ = \{\lambda_k^+\}$ and $L^- = \{\lambda_k^-\}$ for $k = 1, ..., K, k \neq a$;
6 `sortascen` (L^+);
7 `sortdescen` (L^-);
8 Current Best Solution $z := 0$; Initial Core $C := N_a$;
9 Current Set of States $Y_C :=$ `reduceclass` (N_a);
10 Vectors in Y_C represented by states (θ_k, π_k, ν_k); where $\theta_k := \sum_{k \in C} p_{y_k,k} + \sum_{k \notin C} p_{b_k,k}$, $\pi_k := \sum_{k \in C} U_{y_k,k} + \sum_{k \notin C} U_{b_k,k}$, $\nu_k :=$ partial representation of vector $\vec{y_i}$;
11 **repeat**
12 `reduceset` (Y_C);
13 **if** $(Y_C = \emptyset)$ **then** break ;
14 Choose $nextsubcarrier\ k$ from L^+ s.t. $R_k :=$ `reduceclass` (k)
15 **if** $|R_k| > 1$ **then** `addclass` (Y_C, R_k) ;
16 Repeat steps 14 - 15 for L^-;
17 `reduceset` (Y_C);
18 **if** $(Y_C = \emptyset)$ **then** break ;
19 **until** *forever*;
20 Find optimal allocation S;

We specify the rules for performing admission control in a hybrid small cell with heterogeneous user traffic. There are two types of priorities, i.e., one for the user type and the other for the traffic type. The SUs have a higher priority over the EUs. The traffic types with their decreasing priorities include *CBR*, *Video Conferencing*, and *BE*. When a new user with a certain traffic type is requesting for resources, the admission control procedure has, broadly, three choices:

1) The user is admitted without affecting the QoS of the rest of the users; *(or)*
2) the user is admitted provided there is a lower priority user session that can be released or its resources reduced; *(or)*
3) the user is not admitted.

The utility function used for the scheduling problem in Section III is such that a low utility value for a user's session at a given time instant implies that the user has one or more of the following characteristics:

- deep channel fade for the duration;
- lower priority (being an EU);
- lower priority traffic type (BE);
- lesser packets waiting in the queue.

From the previous discussion, it can be observed that the utility function can provide a sense of how the user is performing in a given time slot. At the same time, the instantaneous utility can be misleading because the same user with lesser utility in a given slot may have a greater utility due to better channel conditions in the next slot. Therefore, we define a new performance metric that will be used in making admission control decisions. First, based on the outcome of the *traffic-aware utility-based subcarrier assignment* procedure, the utility of each user over the assigned subcarriers is computed. Thus

$$U_n(t) = \sum_{k=1}^{K} U_{n,k}(t) s_{n,k}(t). \tag{21}$$

Then, we define a *normalized utility* parameter for each user given by

$$\hat{U}_n(t) = \frac{U_n(t)}{U_{\max}(t)} \tag{22}$$

where $U_{\max}(t) = \max_n\{U_n(t)\}$. We also define the time-average normalized utility over T slots as

$$E\left[\hat{U}_n(t)\right] = \frac{1}{T} \sum_{\tau=t-T}^{t} \hat{U}_n(\tau). \tag{23}$$

The time-average normalized utility provides a long-term view of how the user session is performing. A low value of $E[\hat{U}_n(t)]$ indicates that the user is most likely the candidate to be released as the scheduler is not able to meet the QoS requirements of the user. We represent the utility threshold as U_{th} and is defined in the region [0, 1]. The normalized utility is used since the utilities can widely vary over different slots, and it is necessary to have a relative value that can be compared with U_{th}. It can be argued that the ratio between the average normalized utilities of different users and the ratio of the average absolute utilities are fairly comparable. Based on this rationale, we utilize $E[\hat{U}_n(t)]$ to provide a long-term view of the QoS performance. The admission control algorithm is presented in Algorithm 2.

IV. Performance Evaluation

The performance of the constrained utility-based scheduling is evaluated, where the utility function is modeled based on OFDMA system parameters and queue models using MATLAB. The minimal algorithm routine implemented in C is used through the MATLAB mex file to perform subcarrier allocation decisions. The traffic types modeled for our problem are described as follows.

A. Simulation Setup

We simulate the downlink of a small cell with the SAP serving N users. The user distances d_n are randomly modeled with a mean of 10 m and a variance of 4 m. The path loss L_n^p for all users is modeled based on the following indoor propagation model:

$$L_n^p = 37 + 30\log(d_n) + L_n^{wall} \tag{24}$$

where L_n^{wall} is the penetration loss due to walls for user n and is also randomly modeled. The noise power is obtained from the following equation:

$$p^{noise} = -174 + 10\log(\Delta B) + f \tag{25}$$

where f is the noise figure in decibels. All users experience lognormal shadowing L_n^{shadow} with a mean of 10 dB. The SNR distribution φ_n for unit power is obtained using

$$\varphi_n = 10^{L_n^p - L_n^{shadow} + p^{noise}}. \tag{26}$$

Algorithm 2: Traffic-aware Utility based Admission Control Algorithm

```
for each time slot t do
    for each user n do
        Determine E[Û_n(t)] based on the subcarrier allocation algorithm
        if E[Û_n(t)] < U_th then
            Preempt user n
        end
    end
    if a new user m requests for resources then
        Compute Û_m(t)
        if Û_m(t) < U_th then
            Do not admit the user
        else
            Admit user m and schedule subcarriers according to U_m(t)
        end
    else
        continue
    end
end
```

The SNR distribution is obtained as $\Phi_{n,k} = \varphi_n p_{n,k}$. The system parameters considered are shown in Table I. For the computation of throughput, we have five users from each of the traffic classes described in the previous section with a mix of SUs and EUs. For this purpose, the arrival rates of CBR users are randomly distributed with a range of 75–125 kb/s. The Video-Streaming users have arrival rates randomly distributed within a range of 200–256 kb/s. The BE users have randomly distributed arrival rates within a range of 150–180 kb/s. The shape parameter for Video-Streaming users is fixed at 3.066. The maximum delay allowed for CBR and Video-Streaming users is fixed at 80 and 150 ms, respectively.

In each slot, the utility and power matrices along with the global and local power constraints are computed and fed as input to the algorithm. The output of the algorithm is the subcarrier assignment matrix S. The entire simulation sequence is run for 5000 ms, which is used to compute the performance metrics. We highlight for this section the performance results of our scheduling policy in terms of throughput, fairness, and delay. The results also highlight how the performance of the TA-Utility scheme can be enhanced with the help of the TA-Utility-based admission control scheme.

TABLE I
SIMULATION PARAMETERS

System Bandwidth	1.92MHz
Number of Subcarriers	64
Subcarrier Bandwidth	30KHz
BER Required	10^{-3}
Max. SAP Tx. Power	$1W$
Max. subcarrier Tx. Power	$0.05W$
Total Number of Users	15
Slot Length	10ms

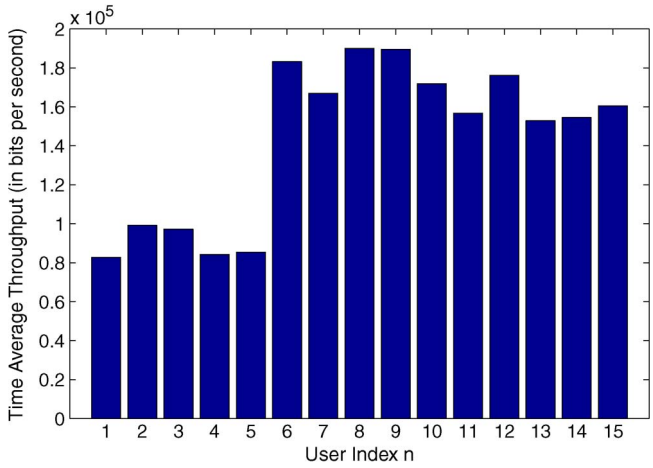

Fig. 3. Throughput performance under TA-Utility-based scheduling.

Fig. 3 shows the time average throughput achieved by the users. As shown in the figure, users 1 to 5 are the CBR users with a uniform arrival rate in the range of 75–125 kb/s. These users achieve a time-average throughput of approximately 100 kb/s. The priority index a_n of the users is set so that the utility of the SUs is higher than that of the EUs; hence, the SUs can achieve superior throughput. In the simulations, we have utilized $a_s = 2a_e = 0.2$ indicating that the probability of violation $\delta_s = 4\delta_e$, where indexes s and e correspond to the SU and the EU, respectively. The delay threshold T_n is set to the maximum delay values defined earlier based on the traffic type.

In addition to the throughput measurements, the mean of the queueing delay over the simulation time is computed for different arrival rates for each traffic class. The simulation parameters from Table I are retained. The mean delay of the worst case user of each of the traffic classes is plotted for different arrival rates. Two variants of the proposed schemes are implemented. Under Scheme 1, only the TA-Utility scheme is implemented, whereas under Scheme 2, the TA-Utility scheme is enhanced using the proposed admission control procedure. For Scheme 2, the simulation starts with 15 users, and then, new users of different traffic classes are randomly introduced in the interval [0, 2000] ms. User sessions are released based

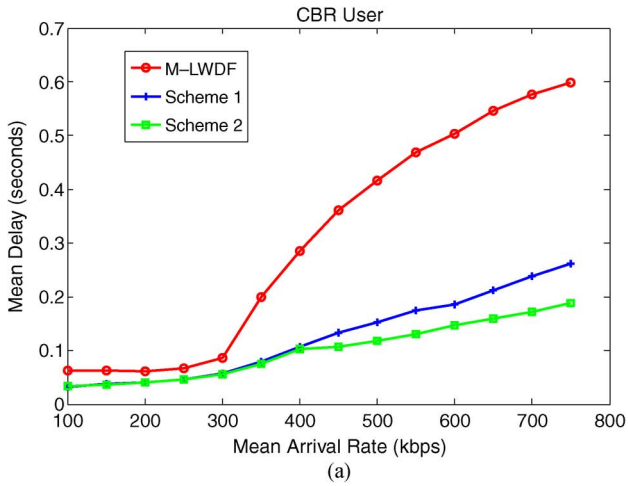

Fig. 4. Mean delay performance of different traffic classes. (a) Worst-case CBR user. (b) Worst-case Video user. (c) Worst-case BE user.

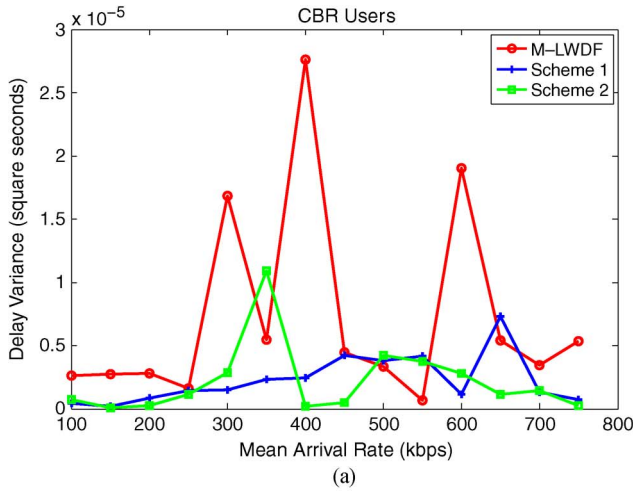

Fig. 5. Delay variance performance of different traffic classes. (a) CBR users. (b) Video users. (c) BE users.

on the admission control algorithm and for a given fixed utility threshold $U_{th} = 0.05$. Both these results obtained for the mean delay are compared with the baseline M-LWDF scheme, as shown in Fig. 4, for all the three traffic types.

As observed in Fig. 4(a), the mean delay of the worst case CBR user for both Scheme 1 and Scheme 2 remains within 50 ms for arrivals below 300 kb/s compared with a mean delay of 80 ms under the M-LWDF scheme. For arrivals beyond 300 kb/s, the M-LWDF scheme has the mean delay increasing almost linearly, whereas Scheme 1 and Scheme 2 result in a slowly increasing mean delay with Scheme 2 resulting in up to 50 ms less delay than Scheme 1 at arrival rates beyond 650 kb/s. For the case of video users shown in Fig. 4(b), Scheme 1 has comparable mean queueing delay to M-LWDF for all arrivals below 350 kb/s. Beyond 350 kb/s, the mean delay is increasing slower than the M-LWDF scheme. Scheme 2 results in further improvement from Scheme 1 with the maximum mean delay reaching up to 190 ms. Therefore, the TA-Utility scheduling

offers significant delay performance gains for delay-sensitive traffic classes. At the same time, the mean delay for BE users, as shown in Fig. 4(c), is much higher and reaches up to 1 s for both Scheme 1 and Scheme 2 when the arrival rate is 600 kb/s and above. This is still acceptable given the presence of CBR and $Video$ users with strict QoS requirements.

In addition to mean delay computation, it is also necessary to analyze the delay variance performance to determine the fairness of the proposed scheme for different traffic classes. The variance of the delay experienced by the users of a given traffic class is computed and plotted against the mean arrival rates of the users' traffic. The results obtained for Scheme 1 and Scheme 2 are compared with M-LWDF in Fig. 5. Fig. 5(a) shows the delay variance of CBR users for different arrival rates. Low values of delay variance for the proposed schemes compared with the M-LWDF scheme indicates that a higher degree of fairness is achieved for different CBR users. Similarly, Fig. 5(b) shows that the delay variance of both Scheme 1 and Scheme 2 is marginally lower compared with that of the M-LWDF scheme, showing good fairness performance. In Fig. 5(c), the delay variance is significantly higher for low data rates for the proposed schemes, particularly for Scheme 1; however, as the arrival rates of BE users increase, the delay variance converges toward that of the M-LWDF scheme.

V. Conclusion

In this paper, the problem of QoS provisioning in hybrid small cells with SUs and EUs has been considered. The need for an improved scheduling to provide stability and QoS performance in the presence of heterogeneous traffic has been explained, and a novel traffic-aware utility function aimed at this problem has been proposed. To this end, a traffic-aware utility maximization approach under power constraints has been proposed and is posed as an optimization objective. To obtain the optimal solution, a minimal algorithm that results in a minimal core of allocation vectors is presented. To further improve the proposed scheme, an admission control algorithm based on the traffic-aware utility function is proposed. In the end, the performance of the proposed scheme is illustrated using simulations.

Acknowledgment

The authors would like to thank Dr. I. F. Akyildiz, J. Miquel, and P. Wang of the Broadband Wireless Networking Laboratory for their constructive criticism, which has improved the quality of this paper.

References

[1] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, R. Vijaykumar, and P. Whiting, "CDMA Data QoS Scheduling on the Forward Link with Variable Channel Conditions," Bell Labs Tech. Memo., Murray Hill, NJ, USA, Tech. Rep., 2000.

[2] R. Balakrishnan, B. Canberk, and I. Akyildiz, "Traffic-aware utility based QoS provisioning in OFDMA hybrid small cells," in *Proc. IEEE ICC*, Jun. 2013.

[3] J. R. Barry, E. A. Lee, and D. G. Messerschmitt, *Digital Communication*, 3rd ed. Norwell, MA, USA: Kluwer, 2004.

[4] B. Canberk, I. Akyildiz, and S. Oktug, "A QoS-aware framework for available spectrum characterization and decision in cognitive radio networks," in *Proc. IEEE 21st Int. Symp. PIMRC*, Sep. 2010, pp. 1533–1538.

[5] V. Chandrasekhar, J. Andrews, and A. Gatherer, "Femtocell networks: A survey," *IEEE Commun. Mag.*, vol. 46, no. 9, pp. 59–67, Sep. 2008.

[6] D. Choi, P. Monajemi, S. Kang, and J. Villasenor, "Dealing with loud neighbors: The benefits and tradeoffs of adaptive femtocell access," in *Proc. IEEE GLOBECOM*, Dec. 2008, pp. 1–5.

[7] K. Dudzinski and S. Walukiewicz, "Exact methods for the knapsack problem and its generalizations," *Eur. J. Oper. Res.*, vol. 28, no. 1, pp. 3–21, Jan. 1987.

[8] M. Katoozian, K. Navaie, and H. Yanikomeroglu, "Utility-based adaptive radio resource allocation in OFDM wireless networks with traffic prioritization," *IEEE Trans. Wireless Commun.*, vol. 8, no. 1, pp. 66–71, Jan. 2009.

[9] M. Markakis, E. Modiano, and J. Tsitsiklis, "Scheduling policies for single-hop networks with heavy-tailed traffic," in *Proc. 47th Annu. Allerton Conf. Commun., Control, Comput.*, 2009, pp. 112–120.

[10] W.-H. Park, S. Cho, and S. Bahk, "Scheduler design for multiple traffic classes in OFDMA networks," in *Proc. IEEE ICC*, Jun. 2006, pp. 790–795.

[11] D. Pisinger, "A minimal algorithm for the multiple-choice knapsack problem," *Eur. J. Oper. Res.*, vol. 83, no. 2, pp. 394–410, Jun. 1995.

[12] S. Shakkottai and A. L. Stolyar, "Scheduling algorithms for a mixture of real-time and non-real-time data in HDR," in *Proc. ITC*, 2001, pp. 793–804.

[13] P. Sinha and A. A. Zoltners, "The multiple-choice knapsack problem," *Oper. Res.*, vol. 27, no. 3, pp. 503–515, May/Jun. 1979.

[14] G. Song and Y. Li, "Utility-based resource allocation and scheduling in OFDM-based wireless broadband networks," *IEEE Commun. Mag.*, vol. 43, no. 12, pp. 127–134, Dec. 2005.

[15] G. Song, Y. Li, and L. Cimini, "Joint channel- and queue-aware scheduling for multiuser diversity in wireless OFDMA networks," *IEEE Trans. Commun.*, vol. 57, no. 7, pp. 2109–2121, Jul. 2009.

[16] A. Valcarce, D. Lopez-Perez, G. de la Roche, and J. Zhang, "Limited access to OFDMA femtocells," in *Proc. IEEE PIMRC*, Sep. 2009, pp. 1–5.

[17] P. Wang and I. Akyildiz, "Network stability of cognitive radio networks in the presence of heavy tailed traffic," in *Proc. IEEE SECON*, Jun. 2012, pp. 165–173.

[18] P. Wang and I. Akyildiz, "On the origins of heavy-tailed delay in dynamic spectrum access networks," *IEEE Trans. Mobile Comput.*, vol. 11, no. 2, pp. 204–217, Feb. 2012.

[19] P. Xia, V. Chandrasekhar, and J. Andrews, "Femtocell access control in the TDMA/OFDMA uplink," in *Proc. IEEE GLOBECOM*, Dec. 2010, pp. 1–5.

Ravikumar Balakrishnan (S'12) received the B.E. degree in electrical and computer engineering from Anna University, Chennai, India, in 2009 and the M.S. degree in electrical and computer engineering from Georgia Tech, Atlanta, GA, USA, in 2011, where he is currently working toward the Ph.D. degree with the Broadband Wireless Networking Laboratory, School of Electrical and Computer Engineering.

In 2010 and 2013, he has held internship positions with Intel Corporation, USA and ORB Analytics, Atlanta, respectively. His areas of research include next-generation cellular networks and cognitive radio networks.

Berk Canberk (M'13) received the M.Sc. degree in digital communications engineering from Chalmers University of Technology, Göteborg, Sweden, in 2005 and the B.Sc. degree in electrical engineering and the Ph.D. degree in computer science from Istanbul Technical University, Istanbul, Turkey, in 2003 and 2011, respectively.

He is currently an Assistant Professor with the Department of Computer Engineering, Istanbul Technical University. He was a Visiting Scholar between August 2008 and May 2009 and a Visiting Postdoctoral Researcher between August 2011 and September 2011 with the Broadband Wireless Networking Laboratory, Georgia Tech, Atlanta, GA, USA. His main research areas include wireless communications, wireless network performance analysis and modeling, cognitive radio networks, and Long-Term Evolution small cells.

Dr. Canberk serves as an Area Editor for *Elsevier Computer Networks Journal*.