QoS and Energy Efficient Resource Allocation in Uplink SC-FDMA Systems

Dionysia Triantafyllopoulou, *Member, IEEE,* Konstantinos Kollias, and Klaus Moessner, *Member, IEEE*

***Abstract*—In this paper we present and evaluate the performance of a resource allocation algorithm to enhance the Quality of Service (QoS) provision and energy efficiency of uplink Long Term Evolution (LTE) systems. The proposed algorithm considers the main constraints in uplink LTE resource allocation, i.e., the allocation of contiguous sets of resource blocks of the localized Single Carrier Frequency Division Multiple Access (SC-FDMA) physical layer to each user, and the imperfect knowledge of the users' uplink buffer status and packet waiting time. The optimal resource allocation is formulated as a discrete connected cake-cutting problem, where different agents are allocated consecutive subsequences of a sequence of indivisible items. This problem is NP-hard, therefore a suboptimal algorithm is introduced, which performs resource allocation using information on the estimated uplink packet delay, the average delay and data rate of past allocations, as well as the required uplink power per resource block. Based on simulation results, the proposed algorithm achieves significant performance improvement in terms of packet timeout rate, goodput, and fairness. Moreover, the effect of poor QoS provision on energy efficiency is demonstrated through the evaluation of the performance in terms of energy consumption per successfully received bit.**

***Index Terms*—Delay, energy efficiency, Long Term Evolution (LTE), Quality of Service (QoS), resource allocation, uplink.**

I. Introduction

MODERN and next generation wireless communication systems are facing the challenge of major demand for increased capacity, resource utilization efficiency, advanced Quality of Service (QoS) provision, and optimal energy efficiency. This is a result of the fact that the network traffic growth is predicted to reach 1000-fold levels by 2020 [1]. In order to address this need for additional capacity, significant technological progress has been made. The respective approaches considered include network densification by the introduction of small cells and the creation of heterogeneous networks (HetNets), the employment of efficient spectrum sharing schemes, the use of new spectrum bands reaching even 90GHz, and the enhancement of cellular networks' efficiency [2], [3]. Therefore, in capacity demanding scenarios the role of resource allocation is of significant importance towards optimizing the resource management efficiency of future communication systems.

Manuscript received August 8, 2014; revised December 5, 2014; accepted January 28, 2015. This work was supported by the European Community's Seventh Framework Programme under Grant Agreement 318563 [CRS-i].

D. Triantafyllopoulou and K. Moessner are with the Institute for Communication Systems (ICS), University of Surrey, Guildford, Surrey, GU2 7XH, U.K. (e-mail: d.triantafyllopoulou@surrey.ac.uk, k.moessner@surrey.ac.uk).

K. Kollias is with the Department of Management Science and Engineering, Stanford University, Stanford, CA, USA. (e-mail: kkollias@stanford.edu).

Practical systems pose certain constraints in the resource allocation process, limiting the performance improvement potential that is indicated by analytical frameworks in the relevant literature, especially in the uplink direction. Specifically, in the case of Long Term Evolution (LTE) systems, the main challenges in uplink resource allocation are the following:

1) In the uplink direction, LTE systems operate on a Single Carrier Frequency Division Multiple Access (SC-FDMA) physical layer, which achieves considerably improved performance in terms of peak-to-average power ratio (PAPR) compared to Orthogonal Frequency Division Multiple Access (OFDMA) [4]. Two types of SC-FDMA are considered, i.e., localized SC-FDMA (LFDMA), where a set of adjacent subcarriers is allocated to each user, and distributed SC-FDMA, in which the subcarriers allocated to a user are distributed over the entire frequency band. One realization of distributed SC-FDMA is interleaved FDMA (IFDMA), where the allocated subcarriers are equidistant from each other. According to performance evaluation results, LFDMA, when combined with efficient channel-dependent scheduling, results in higher throughput than IFDMA, while their performance in terms of PAPR is similar when employing pulse shaping [4]. Therefore, in this paper, as in the vast majority of the relevant literature, we consider localized SC-FDMA transmission, according to which only groups of contiguous resource blocks can be allocated to each user. As a result, widely investigated and well-performing resource allocation algorithms that are designed for the LTE downlink, which allows the allocation of noncontiguous resource blocks to each user, cannot be applied on the uplink case.
2) The LTE eNodeB, which is the entity responsible for the resource allocation in both uplink and downlink directions, does not have accurate knowledge of the buffer status of the uplink User Equipment (UE) devices in terms of number and waiting time of pending packets. This becomes particularly challenging in the case of real-time applications, where the timely transmission of the user packets is of outmost importance. In order for the eNodeB to be informed on the UE devices' traffic demands, the LTE specifications describe in detail the procedures through which UE devices request for uplink scheduling grants and notify the eNodeB regarding their buffer status.

A. Review of the relevant literature

As a result of the imperfect knowledge of the users' buffer status and exact packet delay in the uplink direction, the majority of the proposals on uplink resource allocation found in the recent literature do not focus on Quality of Service (QoS) enhancement in terms of delay sensitivity of modern real-time applications or on the effect of poor QoS provision on energy efficiency [5]–[15]. In these proposals, the most common resource allocation objectives include, but are not limited to, the maximization of the throughput [5]–[8], the optimization of the spectral efficiency [9]–[11], the minimization of the energy consumption per transmitted bit [12], as well as the maximization of the resource allocation fairness [13], [14] and the minimization of the performance degradation with regards to the optimal solution [15].

In [5], [6], joint user pairing and resource allocation in the SC-FDMA LTE uplink is investigated. An optimal algorithm based on branch-and-bound search, aiming at weighted throughput maximization, is introduced as a benchmark. To reduce complexity, the original problem is divided into the subproblems of user pairing, and resource block allocation and suboptimal algorithms are developed. In [7], the authors formalize a general LTE uplink scheduling problem, which is suitable for various scheduling policies. This is proven to be MAX SNP-hard. Therefore, two approximation algorithms, i.e., a greedy one and an algorithm based on the local ratio technique, are designed. The proposed schemes perform resource allocation assuming knowledge of the users' queue size; however, the protocol through which this information is provided to the eNodeB and the effect of imperfect queue status information are not specified. The authors in [8] develop a joint optimization algorithm performing multiuser pairing and resource allocation with inter-cell interference avoidance. The main objective of this algorithm is the maximization of the weighted throughput of the network. Resource allocation is performed in the time, frequency and spatial domains. In order to address the problem of interference, multicell coordination is considered. In [9], three greedy algorithms are proposed, giving higher priority to the users with relatively poor channel quality for the purpose of fairness. The system is evaluated in terms of average spectral efficiency, Bit Error Rate (BER), and outage probability. The authors in [10] introduce a SC-FDMA resource allocation problem to determine the subchannel and power allocation with the aim to maximize the total user-weighted system capacity, subject to each user's total power and peak power constraint, while canonical duality theory is employed in [11] in order to perform joint power and subchannel allocation, and adaptive modulation. Energy efficiency is a performance metric considered in [12], where resource allocation is performed in the time and frequency domains. Moreover, the need for retransmissions in a system employing Hybrid Automatic Repeat reQuest (HARQ) is incorporated in the resource allocation through the use of a block scheduling interval specifically designed for synchronous HARQ. Two suboptimal approaches to minimize the average power allocation required are proposed. Proportional fairness is the main objective considered in [13] and [14]. More specifically, in [13] the well-known proportional fairness algorithm in the time domain is adapted to an uplink SC-FDMA framework. This problem is proved to be NP-hard, therefore a set of suboptimal algorithms considering frequency domain correlations and employing adaptive resource block grouping is also provided. Similarly, in [14], proportional fairness is the criterion based on which groups of resource blocks are allocated to the users.

Due to the fact that QoS provision in terms of delay sensitivity is not the main concern of the proposals in [5]–[15], the traffic models used are usually infinitely backlogged, or even unspecified. This assumption facilitates the required analytical modelling for the derivation of theoretical performance bounds. However, such traffic models do not reflect the variations of the user traffic demands in a realistic LTE system, especially in the case of real-time applications, and the limitation of imperfect user buffer status information on the uplink direction. Moreover, the packet delay as the result of traffic congestion in a cell, which poses the strictest constraints in the case of real-time applications, is not considered in the resource allocation problem formulations, therefore, disregarding the negative effect of packet losses due to excessive delays on the overall system performance.

On the other hand, QoS provision is the main objective of the proposals in [16]–[19]. In [16], the authors propose two resource allocation algorithms for multiclass services that consider the minimum throughput and the maximum allowed delay of each scheduling request. In order to guarantee fairness in the resource allocation, the algorithms dynamically adapt their operation to the number of requests in the system. However, the traffic model considered in this proposal is also infinitely backlogged, therefore not being able to accurately reflect the effect of varying traffic demands on the resource allocation performance. A Quality of Experience (QoE)-based approach for the joint optimization of the uplink transmission of both real-time as well as on-demand video is proposed in [17]. The optimization problem considers delay constraints in terms of both video request time and upload time. A three-stage uplink QoS-constraint resource allocation scheme is introduced in [18]. Firstly, a time domain scheduler prioritizes UE services according to their QoS requirements. Then, a frequency domain scheduler prioritizes users based on their channel quality. Finally, the modulation of the allocated resource blocks is determined to enhance system throughput. However, no power control is considered in [17] and [18] and the proposed solutions are not evaluated in terms of energy efficiency. In [19] the authors employ energy efficient resource allocation for uplink SC-FDMA systems under statistical QoS requirements using canonical duality theory. The proposed design is shown to enhance the energy efficiency while simultaneously satisfying the QoS requirements. However, the effect of varying traffic demands, especially in the case of real-time applications is not evaluated on the resource allocation performance.

B. Contributions and organization of the paper

Motivated by the review of the relevant literature, in this paper we propose a QoS-oriented and energy efficient resource

allocation algorithm for uplink LTE systems. Resource allocation is performed taking into consideration the estimated packet delays in the uplink direction, the average delay and data rate of allocations in the past, as well as the uplink power per resource block. More specifically, the main contributions of this paper with respect to the reviewed literature are summarized as follows:

1) Consideration of the constraints of uplink resource allocation in a realistic LTE system, which apart from the allocation of sets of contiguous resource blocks per user in localized SC-FDMA, include the specified LTE procedures of user requests for uplink transmission grants and buffer status reporting. The proposed algorithm uses the respective procedures in order to assess whether a user is in need for uplink resources and determine their required amount. Therefore, waste of resources that are allocated to users in excess of their actual needs is avoided, resulting in a more efficient resource management.
2) Consideration of the delay constraints of real-time applications in the resource allocation, through the estimation of the packet delays based on the received scheduling requests by the users. The incorporation of the estimated packet delay in the resource allocation allows the prioritization of users experiencing excessive delay, therefore reduces the probability of packet timeouts in real-time applications and results in improved QoS provision.
3) Formulation of the optimal uplink resource allocation as a discrete connected cake-cutting problem, where a sequence of indivisible items must be divided among different agents, with each agent being allocated a consecutive subsequence of these items. The original problem does not consider any upper bounds on the size of the allocated items to each agent and is more appropriate for systems with infinitely backlogged traffic, i.e., users that always have data to transmit. A modified discrete connected cake-cutting with pieces of bounded size is defined, which is more appropriate for a practical system with realistic traffic models. This problem is shown to be NP-hard as well. Therefore, a suboptimal resource allocation algorithm is proposed, which also considers realistic traffic patterns.
4) Consideration of the effect of QoS on energy efficiency through the evaluation of the system performance in terms of the total energy consumption per successfully received bit. It is shown that poor QoS provision also has a negative effect on the energy efficiency due to the wasted resources as a result of packet errors, caused by unfavorable wireless channel conditions, and timeouts, caused by excessive resource allocation delays.

This paper is organized as follows. Section II introduces the system model and provides a short overview of the subframe structure and the procedures for buffer status reporting and UE scheduling requests of LTE systems. Section III formulates LTE uplink resource allocation as a discrete connected cake-cutting problem. Section IV introduces and describes in detail the proposed suboptimal uplink resource allocation algorithm, whose performance is evaluated through simulations in section V. Finally, section VI contains conclusions and discusses on plans for future work.

TABLE I
DEFINITION OF SYSTEM MODEL PARAMETERS.

Parameter	Definition
$m_i^{UL}(t)$	Uplink resource allocation metric of user i
$m_{i,j}^{UL}(t)$	Uplink resource allocation metric of user i on scheduling block j
$d_i^{UL}(t)$	Estimated uplink queuing delay of user i (s)
$d_{th,i}$	Queuing delay threshold of user i (s)
$\overline{D}_i^{UL}(t)$	Average uplink delay of user i (s)
$\overline{R}_i^{UL}(t)$	Average uplink data rate of user i (b/s)
β	Average delay and data rate calculation factor
$r_i^{UL}(t)$	Instantaneous uplink data rate of user i (b/s)
$r_{i,j}^{UL}(t)$	Instantaneous uplink data rate of user i on scheduling block j (b/s)
$M_{i,j}$	Modulation of user i on scheduling block j (b/symbol)
L_{SB}^{UL}	Number of data carrying resource elements in an uplink scheduling block
$P_{1,i}$	Minimum uplink power per resource block of user i (dBm)
$P_{CMAX,C}$	Maximum uplink power (dBm)
$P_{0,PUSCH}$	Target received power (dBm)
α	Path loss compensation factor
PL_i	Path loss of user i (dB)
N_{RB}^{UL}	Total number of resource blocks per slot
N_{symb}^{UL}	Number of SC-FDMA symbols per uplink resource block
N_{SC}^{RB}	Number of subcarriers per resource block
$N_{RB,SB}$	Number of resource blocks per scheduling block
N_{SB}	Number of scheduling blocks per subframe
$BSR_i(t)$	Buffer Status Report of user i
$SR_i(t)$	Scheduling request indicator of user i
T_{sf}	Subframe length (s)
UE	Set of users
Φ	Set of available scheduling blocks
K_i	Set of scheduling blocks for which user i maximizes $m_{i,j}^{UL}(t)$
G_i	Set of allocated scheduling blocks to user i
G	Vector of allocated scheduling blocks
$\gamma_{i,j}$	Signal-to-Noise Ratio of user i on scheduling block j
$u_i(G_i)$	Utility of allocation G_i to user i
$U(G)$	Total cake-cutting utility
k_i	Upper bound of the piece allocated to user i

II. SYSTEM MODEL

The system model consists of a single LTE macro cell and a number of UE devices, randomly deployed in the macro cell coverage area. For the remainder of this document the terms user and UE are used interchangeably. Each user has an active real-time video connection on the uplink and the eNodeB is responsible to allocate the available resources in a fair, QoS and energy efficient manner, employing the proposed resource allocation algorithm. Table I summarizes the parameters used for the formulation and performance evaluation of the proposed algorithm.

A. Uplink resource allocation in LTE systems

In this subsection we briefly summarize the LTE protocol specification for the transmission of uplink scheduling requests (SRs) and the notification of the eNodeB regarding the buffer status of each UE.

Resources on the LTE uplink are allocated to the users in terms of uplink scheduling grants. A scheduling grant applies to a specific carrier of a UE, and is not limited to a specific

application class within the UE. A UE that requires uplink resources in order to transmit one or more of its pending data packets sends a SR to the uplink scheduler by raising a simple flag, which is transmitted on the Physical Uplink Control Channel (PUCCH) [20]. A SR can occur on a periodic manner, and its frequency is a UE-specific parameter provided by the higher layers [21], [22].

However, in order for the uplink scheduler to be able to determine the required amount of resources to be granted to each user, information on the amount of data available for transmission in the uplink UE buffers is also necessary. Therefore, as part of the uplink transmission through Medium Access Control (MAC) elements, information on the UE buffer situation is provided to the eNodeB in the form of Buffer Status Reports (BSRs). A BSR consists of a buffer size field, which contains information on the amount of data awaiting transmission across all logical channels in a logical channel group. The amount of data is indicated in number of bytes, and refers to all the data that are available for transmission in the Radio Link Control (RLC) and Packet Data Convergence Protocol (PDCP) layers. It has to be noted though that the size of the RLC and MAC headers are not considered in the buffer size computation [21].

B. LTE subframe structure

In the time domain, uplink LTE transmissions are organized into radio frames, each of which consists of two half-frames. A half-frame consists of five equally sized subframes of length T_{sf} each. Each subframe consists of two equally sized slots. Each slot consists of N_{symb}^{UL} SC-FDMA symbols, including cyclic prefix. The exact value of N_{symb}^{UL} depends on the cyclic prefix length, which is configured by the higher layers.

The resource grid describing the uplink transmitted signals in each slot consists of $N_{RB}^{UL} \times N_{SC}^{RB}$ subcarriers and N_{symb}^{UL} SC-FDMA symbols. The smallest physical resource in LTE is a resource element, consisting of one subcarrier during one SC-FDMA symbol. Resource elements are grouped into resource blocks, where each resource block consists of N_{SC}^{RB} consecutive subcarriers in the frequency domain and one slot consisting of N_{symb}^{UL} SC-FDMA symbols in the time domain [23]. A scheduling block consists of two consecutive resource blocks, spanning a subframe of length equal to T_{sf} and is the minimum amount of resources that can be allocated to a user in a subframe.

C. Resource allocation utility function

On the uplink direction of an LTE network, resource allocation is performed on a per subframe basis. In order to perform resource allocation in a fair, QoS and energy efficient manner and evaluate the utility of each scheduling block to a user, we introduce metric $m_{i,j}^{UL}(t)$ of user i, $i \in UE$, for scheduling block j, $j \in \{1, 2, \ldots, N_{SB}\}$, where N_{SB} is the number of scheduling blocks per subframe, as follows:

$$m_{i,j}^{UL}(t) = \frac{d_i^{UL}(t)}{d_{th,i}} \exp\left(\frac{\overline{D}_i^{UL}(t)}{\overline{R}_i^{UL}(t)}\right) \frac{r_{i,j}^{UL}(t)}{P_{1,i} \cdot N_{RB,SB}}. \quad (1)$$

$d_i^{UL}(t)$ is the time passed since the last uplink grant was allocated to user i or since a SR has been received from this user and $d_{th,i}$ is the delay threshold, beyond which a packet is no longer considered usable and is discarded by the user's buffer. Since the eNodeB does not have accurate information on the exact waiting time of the pending packets of each user, $d_i^{UL}(t)$ is used in order to allow a worst-case estimation of the packet delay, i.e., the case of a new packet entering the user's uplink buffer just after an uplink grant was allocated to the user or a SR was sent. Therefore, with the use of $d_i^{UL}(t)$, the prioritization of users who have waited for a higher amount of time since their last uplink grant or the latest SR, and are in higher risk of packet expiration, is achieved.

$\overline{D}_i^{UL}(t)$ and $\overline{R}_i^{UL}(t)$ are the average delay and data rate, respectively, experienced by user i in the past, and are calculated using a weighted moving average formula as follows:

$$\overline{D}_i^{UL}(t) = \beta d_i^{UL}(t) + (1-\beta)\overline{D}_i^{UL}(t-1), \quad (2)$$

$$\overline{R}_i^{UL}(t) = \beta r_i^{UL}(t) + (1-\beta)\overline{R}_i^{UL}(t-1), \quad (3)$$

where $r_i^{UL}(t)$ is the instantaneous uplink data rate of user i and $0 \leq \beta \leq 1$. The incorporation of $\overline{D}_i^{UL}(t)$ and $\overline{R}_i^{UL}(t)$ in $m_{i,j}^{UL}(t)$ allows the prioritization of users that were served with high average delay and low average data rate in the past, thus increasing the fairness of the proposed solution.

$P_{1,i}$ is the minimum uplink power per resource block of user i, which, based on the LTE uplink power control specification, is defined as follows:

$$P_{1,i} = \min\{P_{CMAX,C}, P_{0,PUSCH} + \alpha PL_i + 10\log_{10}(N_{RB}^{UL})\} - 10\log_{10} N_{RB}^{UL}. \quad (4)$$

$P_{1,i}$ is calculated based on the assumption that all the resource blocks of an uplink slot are allocated to user i. Of course, the actual uplink power per resource block will almost always be higher for the specific user, and will depend on the actual number of its allocated resource blocks, which, in principle, will be less than N_{RB}^{UL}. $P_{CMAX,C}$ is the configured UE transmit power, $P_{0,PUSCH}$ is the target received power per resource block, while PL_i is the user downlink path loss estimate calculated in the UE and α, $0 \leq \alpha \leq 1$, is a parameter for path loss compensation whose value is provided by the higher layers [22].

$r_{i,j}^{UL}(t)$ is the data rate achieved by user i on scheduling block j and is defined as follows:

$$r_{i,j}^{UL}(t) = \frac{\left(L_{SB}^{UL}\log_2 M_{i,j}\right)}{T_{sf}}, \quad (5)$$

where L_{SB}^{UL} is the number of data carrying resource elements per uplink scheduling block, which depends on the number of reference signals transmitted in a subframe, $M_{i,j}$ is the Modulation and Coding Scheme (MCS) of user i on scheduling block j and T_{sf} is the subframe length. In a generic SC-FDMA system that allows the selection of different MCS per scheduling block based on the perceived channel conditions, the value of $r_{i,j}^{UL}(t)$ is different for every scheduling block. However, since according to the LTE system specifications all

scheduling blocks allocated to the same user have the same MCS, the value of $r_{i,j}^{UL}(t)$ and, consequently, the value of $m_{i,j}^{UL}(t)$ will be the same across all scheduling blocks.

III. Optimal Uplink Resource Allocation (Cake-Cutting)

The problem of allocating a contiguous collection of scheduling blocks in a subframe to each user is strongly connected to the traditional *fair* division (or *cake-cutting*) problem from social choice theory [24]–[27]. In the traditional fair division problem there is a cake, represented as the $[0, 1]$ interval, and a set of agents with each one obtaining a given utility for each $[x, y]$ interval, with $0 \leq x \leq y \leq 1$. The cake must be divided among the agents and there are various objectives that one might wish to optimize or adhere to, e.g., some fairness criterion, maximizing social welfare, etc. The version of the fair division problem that is most closely related to our setting is *discrete connected cake-cutting*. In this case, the cake is a sequence of indivisible items, i.e., non-overlapping (x, y) intervals whose union equals $[0, 1]$, and each agent must be allocated a consecutive subsequence of these items. The agent utility functions are assumed to be additive, i.e., an agent's total utility upon receiving a subset of the items is equal to the sum of the individual utilities of each item.

There is a straightforward reduction from our setting and the problem of assigning uplink scheduling blocks to users seeking to maximize a total utility function, to the problem of allocating cake pieces to agents in discrete connected cake-cutting, seeking to maximize welfare, i.e., the sum of agents' utilities. More specifically, the users of the LTE system under consideration are mapped to agents in the cake-cutting setting, the uplink scheduling blocks of a subframe are mapped to the sequence of indivisible items that form the cake, and the users' $m_{i,j}^{UL}(t)$ metric functions are mapped to the agent utility functions, see Fig. 1. Therefore, if we define the set of allocated scheduling blocks to user i, $i \in UE$, as G_i, the total value that this user obtains from this allocation is referred to as $u_i(G_i)$. We define $u_i(G_i)$ as a complex function of the $m_{i,j}^{UL}(t)$ values, with the properties that (i) it is non-decreasing in all $m_{i,j}^{UL}(t)$'s, and (ii) there is a threshold τ, below which the Signal-to-Noise Ratio (SNR) of the scheduling block is very low, resulting in significantly increased BER, and making the scheduling block, and consequently all the allocated resources to the user, practically unusable.

The vector of allocated scheduling blocks G is defined as $G = \{G_i\}_{i \in UE}$. Let $U(G) = \sum_{i \in UE} u_i(G_i)$ be the total utility of allocation G. Therefore, the main objective of the cake-cutting algorithm is to identify the optimal allocation of sets of contiguous scheduling blocks to the different users in a manner that maximizes the total utility, i.e.,

$$G^* = \arg\max_{G} \{U(G)\}. \tag{6}$$

Results in [28] show that computing the allocation that maximizes welfare in discrete connected cake-cutting is NP-hard. Moreover, it is shown that it is not possible to achieve an arbitrary approximation of the optimal welfare unless P=NP.

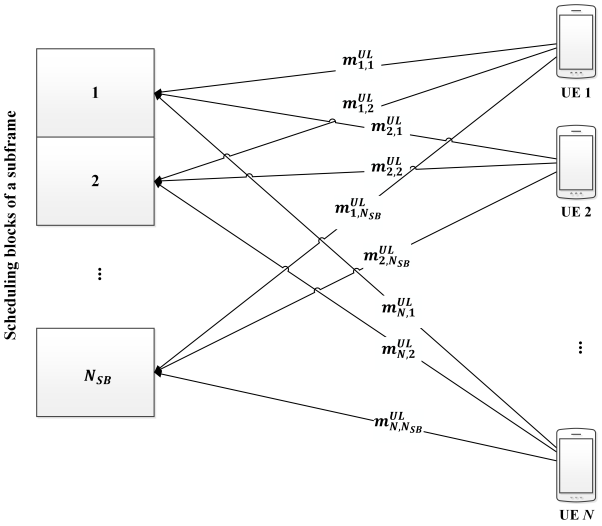

Fig. 1. Valuation of the scheduling block utilities by the users.

The best polynomial time approximation algorithm obtained in the same paper achieves an 8-approximation of the optimal welfare, which implies it is hard to obtain an algorithm that offers guarantees of practical importance. Our problem, however, is much more general and positive results (such as this approximation guarantee) do not carry over to our setting.

Concluding this section, we introduce the following modification, which is of interest in our setting. Consider the version of discrete connected cake-cutting, which includes an upper bound k_i on the cardinality of the set of contiguous resources G_i allocated to any user i. Each constant parameter k_i models the fact that agent i might be able to utilize at most k_i items. The established version of the problem, which does not consider parameters k_i, is appropriate for systems that assume infinitely backlogged traffic, i.e., users always having data to transmit, and always taking advantage of all their allocated scheduling blocks. However, in a realistic LTE system, the traffic models considered are not infinitely backlogged and a resource allocation algorithm needs to take into consideration the users' buffer status in order to avoid wasting resources by allocating them more scheduling blocks than actually needed. Therefore, we formally define the problem:

Definition 1. *Discrete Connected Cake-Cutting with Pieces of Bounded Size*

Suppose we are given a sequence of items $1, 2, \ldots, N_{SB}$, a set of players UE, and a utility $u_i(S)$, for every player $i \in UE$ and every contiguous subsequence of items S. Let $\mathcal{G}$ be the set of allocations, G, of items to players, such that G_i is a contiguous subsequence of items with $|G_i| \leq k_i$, for all $i \in UE$, and $|G_i \cap G_l| = 0$, for all $i \neq l, i, l \in UE$. We wish to find the optimal allocation of items to players that maximizes the total utility, i.e., $G^* = \arg\max_{G \in \mathcal{G}} \sum_{i \in UE} u_i(G_i)$.

The literature on maximizing welfare in discrete connected cake-cutting does not explicitly consider the modified version we defined above. However, we note that the reduction in [28] still applies, even if we restrict the number of indivisible items per player to a small constant. This is due to the fact that the

authors in [28] use a reduction from 3-dimensional matching to discrete connected cake-cutting, which results in instances such that any welfare maximizing allocation assigns at most 2 items to a player. Therefore, it can be concluded that the modified version of the cake-cutting problem defined in this section is also NP-hard.

IV. The Proposed Algorithm

Since, as discussed in the previous section, the optimal allocation of uplink scheduling blocks in a localized SC-FDMA LTE system is an NP-hard problem, in this section we introduce a suboptimal algorithm that takes into consideration the users' buffer status and real-time delay constraints, as well as the constraints of a realistic LTE system in order to perform uplink resource allocation in a QoS and energy efficient manner.

As a first step, the set of active users UE is sorted in descending order of $m_i^{UL}(t)$. This is a metric that aims to provide higher resource allocation priority to users with increased waiting time with respect to the delay threshold, high average delay and low average data rate of their allocations in the past, as well as low uplink power transmission requirements and high expected data rate per scheduling block. To this end, $m_i^{UL}(t)$ is defined as follows:

$$m_i^{UL}(t) = \frac{d_i^{UL}(t)}{d_{th,i}} \exp\left(\frac{\overline{D}_i^{UL}(t)}{\overline{R}_i^{UL}(t)}\right) \times \frac{1}{P_{1,i} \cdot N_{RB,SB}} E\left[r_{i,j}^{UL}(t)\right]. \quad (7)$$

The operation of the proposed resource allocation algorithm in each subframe of length equal to T_{sf} is formally described in Algorithm 1 and depicted in the flowcharts of Fig. 2 and Fig. 3. The algorithm iterates until either all the scheduling blocks of the subframe are allocated, i.e., the set Φ of available scheduling blocks is empty, or all users have received enough resources to accommodate their uplink transmission needs, i.e., the set UE of active users is empty. Therefore, for each user $i \in UE$, in descending order of $m_i^{UL}(t)$, the proposed uplink resource allocation algorithm performs the following steps:

1) Firstly, the user's need for an uplink transmission grant is assessed. This is based on whether a SR is received by the user, i.e., $SR_i(t) = 1$, or the value of the latest BSR verifies that the user buffer has uplink data waiting to be transmitted, i.e., $BSR_i(t) > 0$. If there is no need to allocate uplink resources in this subframe, the user is removed from UE and the algorithm proceeds to the next user.
2) If either $SR_i(t) = 1$ or $BSR_i(t) > 0$ the resource allocation algorithm determines the set K_i, which consists of the available scheduling blocks for which the user maximizes the value of $m_{i,j}^{UL}(t)$, i.e., $K_i = \left\{ j^{'} \in \Phi : i = \arg\max_{i' \in UE} \left(m_{i^{'},j^{'}}^{UL}(t) \right) \right\}$. It has to be noted that the scheduling blocks that comprise K_i are not necessarily contiguous.
3) If K_i is nonempty, the scheduling block j^* with the highest SNR $\gamma_{i,j}$ is determined, i.e., $j^* = \arg\max_{j \in K_i} (\gamma_{i,j})$ and, if its BER, i.e., BER_{i,j^*}, is lower than the threshold τ, it is the first scheduling block to be included in set G_i, i.e., the set of all scheduling blocks allocated to user i in this subframe.
4) The set G_i, which contains scheduling block j^*, as well as the maximum number of contiguous scheduling blocks neighboring j^* that can be allocated to user i is calculated. This depends on the user's buffer status, the availability of scheduling blocks that are neighbors to j^*, as well as on the value of $m_{i,j}^{UL}(t)$. Therefore, a scheduling block j is included in set G_i, if i) it is not already allocated to another user, i.e., $j \in \Phi$, ii) it maximizes the value of $m_{i,j}^{UL}(t)$, i.e., $i = \arg\max_{i' \in UE} \left(m_{i^{'},j}^{UL}(t) \right)$, iii) it is a neighbor to another scheduling block that is already included in G_i, therefore not violating the scheduling block contiguity constraint, i.e., $\exists j^{'} \in G_i : |j - j^{'}| = 1$, iv) its BER is lower than the threshold τ, and v) the number of scheduling blocks already included in G_i is not enough to accommodate all the traffic in the user's buffer, which is depicted as $L_{BSR}(BSR_i(t))$. The number of bytes that can be accommodated by scheduling block j depends on the user's MCS and is depicted as $L_{SB}(j)$. In order to determine the scheduling blocks that comprise G_i, the proposed algorithm uses j^*

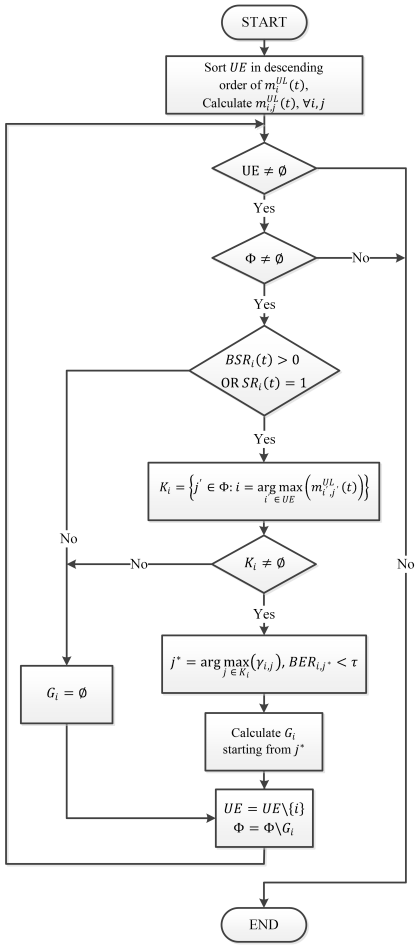

Fig. 2. Flowchart of the proposed uplink resource allocation algorithm in each Time Transmission Interval (TTI).

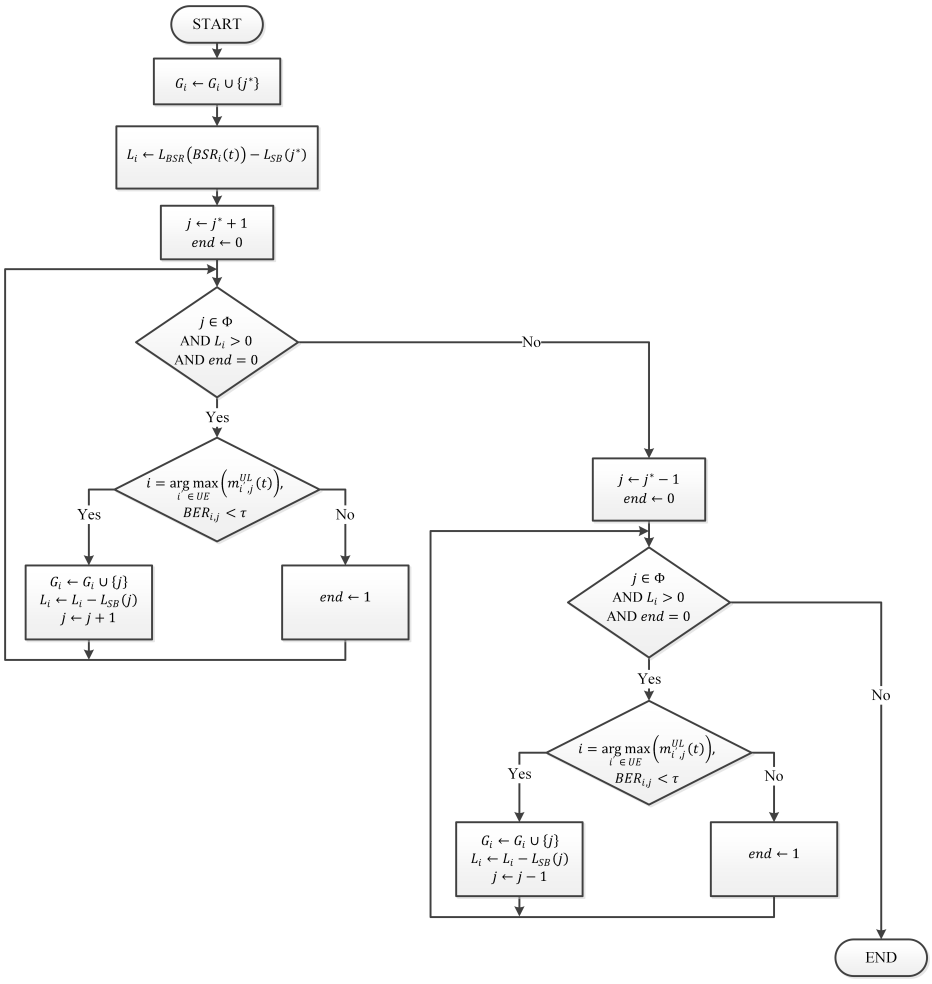

Fig. 3. Flowchart of the calculation of G_i.

as a starting point and attempts to expand the allocation towards both directions, i.e., scheduling blocks with $j < j^*$ and $j > j^*$. In each direction, the expansion terminates when a scheduling block that does not qualify one or more of the above five criteria for inclusion in G_i is met. The detailed steps of this process are described in Algorithm 1 and the flowchart of Fig. 3.

5) When the resource allocation for user i is finalized, the user is removed from UE and all its allocated scheduling blocks, i.e., belonging to G_i, are removed from the set Φ of available scheduling blocks.
6) If $\Phi \neq \emptyset$ and $UE \neq \emptyset$, the resource allocation algorithm proceeds to the next user, otherwise the resource allocation for this subframe is complete and the algorithm terminates.

A. Theoretical Analysis

Run-time analysis: In this setting, it is of practical importance that the scheduling block allocation algorithm is very fast. The proposed algorithm consists of a sequence of two main types of events: One event type is picking a starting point for a user and the other event type is a check of whether to allocate the scheduling block to the user. Let t^{sp} be the time required to find the best scheduling block for a user among a candidate set (i.e., complete an event of the first type) and let t^c be the time required to find whether the user is the best one for that scheduling block among a candidate set (i.e., complete an event of the second type). The running time of the algorithm is bounded by $|UE| \cdot t^{sp} + (|UE| + N_{SB}) \cdot t^c$. To see this, first note that events of the first type can happen at most $|UE|$ times, since we can have at most one for each user. Events of the second type can happen at most $|UE| + N_{SB}$ times since each time such an event occurs, either a user or a scheduling block is eliminated. Note that t^{sp} and t^c are, in the worst-case, linear in N_{SB} and $|UE|$ respectively. Hence the algorithm is $O(|UE| \cdot N_{SB})$, i.e., linear in the size of the input.

Performance Analysis: Note that the quality of a scheduling block j for a particular user i is given as $m^{UL}_{i,j}(t) = m^{UL}_i(t) \cdot r^{UL}_{i,j}(t)/E[r^{UL}_{i,j}(t)]$. Here $m^{UL}_i(t)$ can be considered as a user score, based on inherent properties of the user as well as its allocations in the past. On the other hand, $r^{UL}_{i,j}(t)$ depends on the value of $\gamma_{i,j}$, which is a Gamma distributed random variable. We analyze the algorithm in two distinct models with respect to the distributions of the $r^{UL}_{i,j}(t)$ variables. The first model assumes a hypothetical setting with very small perturbations on the $r^{UL}_{i,j}(t)$ variables, i.e., we assume the

Algorithm 1 Uplink Resource Allocation

Sort UE in descending order of $m_i^{UL}(t)$, $\forall i \in UE$
Calculate $m_{i,j}^{UL}(t)$, $\forall i \in UE$, $j \in \{1, 2, \ldots, N_{SB}\}$
for $i \in UE$ **do**
  if $\Phi \neq \emptyset$ **then**
    $G_i \leftarrow \emptyset$
    if $BSR_i(t) > 0$ **or** $SR_i(t) = 1$ **then**
      $K_i = \left\{ j^{'} \in \Phi : i = \arg\max_{i' \in UE} \left(m_{i^{'},j^{'}}^{UL}(t) \right) \right\}$
      if $K_i \neq \emptyset$ **then**
        $j^* \leftarrow \arg\max_{j \in K_i} (\gamma_{i,j})$, $BER_{i,j^*} < \tau$
        $G_i \leftarrow G_i \cup \{j^*\}$
        $L_i \leftarrow L_{BSR}(BSR_i(t)) - L_{SB}(j^*)$
        $j \leftarrow j^* + 1$, $end \leftarrow 0$
        while $j \in \Phi$ **and** $L_i > 0$ **and** $end = 0$ **do**
          if $i = \arg\max_{i' \in UE} \left(m_{i^{'},j}^{UL}(t) \right)$ **and** $BER_{i,j} < \tau$ **then**
            $G_i \leftarrow G_i \cup \{j\}$
            $L_i \leftarrow L_i - L_{SB}(j)$
            $j \leftarrow j + 1$
          else
            $end \leftarrow 1$
          end if
        end while
        $j \leftarrow j^* - 1$, $end \leftarrow 0$
        while $j \in \Phi$ **and** $L_i > 0$ **and** $end = 0$ **do**
          if $i = \arg\max_{i' \in UE} \left(m_{i^{'},j}^{UL}(t) \right)$ **and** $BER_{i,j} < \tau$ **then**
            $G_i \leftarrow G_i \cup \{j\}$
            $L_i \leftarrow L_i - L_{SB}(j)$
            $j \leftarrow j - 1$
          else
            $end \leftarrow 1$
          end if
        end while
      end if
    end if
    $UE \leftarrow UE \setminus \{i\}$
    $\Phi \leftarrow \Phi \setminus \{G_i\}$
  end if
end for

variance of the SNR $\gamma_{i,j}$ is very low, and consistently $r_{i,j}^{UL}(t)$ is arbitrarily close to $E[r_{i,j}^{UL}(t)]$. Moreover, in this model the algorithm is assumed to select the last scheduling block as a starting point in case all else is equal. This is a slight departure from realistic settings, however, we use this model to formally analyze the properties of the proposed algorithm and exhibit the intuition behind its design decisions. We call this the consistent model.

Theorem 1. *In the consistent model, the proposed algorithm achieves the optimal assignment.*

Proof. Note that, by definition, the distributions of $r_{i,j}^{UL}(t)$ for a fixed user i are the same. Then, in the consistent model, $E[r_{i,j}^{UL}(t)]$ is equal to some user-dependent $\overline{r}_i$. Then the utility of user i for every item j is equal to $m_i^{UL}(t)$. Given that the algorithm starts from the last available item as a tie-breaker, it always grants as many items as possible (up to k_i) to the user i who maximizes $m_i^{UL}(t)$ among the remaining users, hence achieving optimal welfare. □

We now analyze the performance of the proposed algorithm with respect to arbitrary distributions of the $r_{i,j}^{UL}(t)$ variables. We call this the arbitrary model. Theorem 2 provides a guarantee on the expected performance of our algorithm in the *worst case* instance. This means that, even if all $m_i^{UL}(t)$ values and the distributions of all $\gamma_{i,j}$ variables were picked by a malicious adversary, we would still achieve the guarantee of Theorem 2.

Theorem 2. *In the arbitrary model, the proposed algorithm, in expectation, achieves an $O(\log N_{SB})$ approximation of the optimal solution.*

Proof. Consider the optimal contiguous assignment of scheduling blocks to the users. Among all scheduling blocks j granted to user i, call j_i the one that maximizes $m_{i,j}^{UL}(t)$. We will first show the following statement:

For every user i, we can find a distinct block $j^{'}$ allocated by the proposed algorithm to some user $i^{'}$ with metric $m_{i^{'},j^{'}}^{UL}(t) \geq m_{i,j_i}^{UL}(t)$.

The proof is as follows. If j_i is assigned to i by the proposed algorithm, then we get our statement is true by setting $i^{'} = i$ and $j^{'} = j_i$. If i is not assigned j_i by the proposed algorithm, but is assigned at least one scheduling block j with higher $m_{i,j}^{UL}(t)$, then we similarly get that our statement holds with $i^{'} = i$ and $j^{'} = j$. On the other hand, if the algorithm does not assign j_i to i and every scheduling block assigned to i has lower metric than j_i, then this implies that another user i^* already received that scheduling block, resulting in a higher value of metric $m_{i^*,j_i}^{UL}(t)$. We do not yet map the i, j_i pair to the i^*, j_i pair, as it might have been already used in an argument of the type described above. Hence, we examine whether user i^* is receiving the corresponding j_{i^*} from the optimal algorithm. If this is true, or if i^* gets a better allocation, then we map this pair to i, j_i. If not, then again there is some other user that holds j_{i^*} with better metric than i^* would have, which means the exact same situation propagates to that user. This propagation can't keep happening forever, since the first user considered by the algorithm, i^1, by definition gets j_{i^1} or a better one. It remains to show that every user will get in expectation $O(\log N_{SB})$ scheduling blocks. The probability that some given scheduling block's quality is below the unacceptable error rate threshold τ is constant. Then, it is well-known that the expected maximum streak of heads of a biased coin in n trials is $O(\log n)$ [29], which completes the proof by mapping a heads coin toss to the event that a given resource is above the threshold for the user under consideration. □

V. Performance Evaluation

In order to evaluate the performance of the proposed uplink resource allocation algorithm, a simulation model was built in MATLAB. The performance of the system employing the proposed resource allocation algorithm is compared to a legacy

system that equally distributes the available uplink resources to the users, without estimating their delay constraints and buffer status, or taking into consideration their QoS and energy efficiency requirements, and the proportional fairness based "Riding Peaks" algorithm introduced in [13]. The simulation environment consists of a single LTE cell and a variable number of UE devices within the cell's coverage area. The maximum distance from the eNodeB is 330m.

The individual subsystems of the simulation model employed are as follows:

The *traffic generator* uses the Joint Scalable Video Model (JSVM) reference software [30] in order to generate variable-length video traffic frames for each UE, starting at a random instance within the first 33ms of a simulation run. The video sequence used is the well-known "Highway" video sequence [31], in a Quarter Common Intermediate Format (QCIF), i.e., an analysis of 176×144 pixels, with a rate of 30 frames per second (fps). The traffic generator provides the created video traffic frames to the resource allocator.

The *channel model* simulates the physical layer channel conditions by providing path loss, shadowing, and short-term fading. It produces bit errors randomly for each connection, based on the allocated scheduling blocks and the MCS per user. Path loss is $128.1 + 37.6 \log_{10} d$, where d is the distance from the eNodeB in km [32]. The shadowing is log-normal, with a standard deviation σ=8dB. Moreover, Rayleigh fading is assumed, with the instantaneous SNR per resource block being a Gamma distributed random variable with a probability density function (pdf) $p_\gamma(\gamma) = \frac{m^m \gamma^{m-1}}{\overline{\gamma}^m \Gamma(m)} \exp\left(-\frac{m\gamma}{\overline{\gamma}}\right)$. $\overline{\gamma}$ is the mean SNR value, as the result of path loss and shadowing, $\Gamma(m) = \int_0^\infty t^{m-1} e^{-t} dt$ is the Gamma function and m is the Nakagami fading parameter, which in the case of Rayleigh channel has a value m=1 [33]. The link budget considers transmitter and receiver antenna gain, cable loss, receiver Noise Floor (NF), Interference Margin (IM) and Control Channel overhead [34]. The values of these parameters considered in the simulation are summarized in Table II. The MCSs considered are QPSK $1/2$, 16-QAM $1/2$, and 64-QAM $3/4$. According to the LTE specifications, all the scheduling blocks allocated to a user in a subframe will have the same MCS. Moreover, perfect channel knowledge is assumed for the purposes of Adaptive Modulation and Coding (AMC).

The *resource allocator* is the entity that is responsible for allocating the uplink resources to the different UE devices following either the proposed algorithm, which takes into consideration parameters such as the connection delay constraints, user buffer status, QoS requirements and energy efficiency, or the equal distribution approach, which does not take into consideration any such information and allocates the uplink resources in a proportional manner, or the approach introduced in [13], which takes into consideration the instantaneous rate as well as the average data rate of past allocations.

The channel bandwidth is 10MHz, while the subframe length is 1ms. 2 Reference signal transmissions per uplink subframe are considered. Time Division Duplex (TDD) operation is assumed, following LTE TDD Configuration 1, resulting in a Downlink:Uplink ratio equal to 3:2. The maximum tolerable resource allocation delay $d_{th,i}$ for all users is 20ms. In the RLC layer the Unacknowledged Mode (UM) is considered, which supports segmentation/reassembly and in-sequence delivery, but not retransmissions. This is typical in the case of real-time applications since retransmissions increase the packet delay and, by the time a retransmitted packet segment is successfully received, the delay may have exceeded its upper threshold, resulting in the need to discard the whole packet. Robust Header Compression (ROHC) is considered for the Real-time Transport Protocol (RTP), the User Datagram Protocol (UDP) and the Internet Protocol (IP) layers, resulting in a RTP/UDP/IP header of 3 bytes. For the lower layers of the protocol stack, the header sizes are as follows: PDCP: 2 bytes, RLC: 3 bytes, MAC: 2 bytes, Cyclic Redundancy Check (CRC): 3 bytes.

The simulation scenario considers an increasing number of UE devices, each one with one uplink video connection. The total simulation time is 67s. The systems' performance is measured in terms of packet timeout rate, goodput, fairness, average delay, and energy efficiency of successfully received bits. All simulation model parameters are summarized in Table

TABLE II
PERFORMANCE EVALUATION PARAMETERS.

Parameter	Value
Physical layer parameters	Channel Bandwidth:10MHz, Subframe length T_{sf}: 1ms, Number of resource blocks (N_{RB}^{UL}): 50
Resource block format	Number of subcarriers per resource block (N_{SC}^{RB}): 12, Number of symbols per resource block (N_{symb}^{UL}): 7, Subcarrier spacing: 15kHz
Reference Signal transmissions	2 Reference Signal transmissions per subframe
TDD configuration	Configuration 1, DL:UL 3:2
Modulation and Coding Schemes	QPSK $1/2$, 16-QAM $1/2$, and 64-QAM $3/4$
Path loss model	$128.1 + 37.6 \log_{10} d$, d: distance from the eNodeB (km)
Transmitter antenna gain	0dBi
Receiver antenna gain	18dBi
Cable loss	0dB
Receiver Noise Floor	-116.4dBm
Interference Margin	1dB
Control Channel Overhead	0dB
Shadowing	Log normal, σ=8dB
Fading	Rayleigh
Maximum UE transmission power	23dBm
Target received power ($P_{0,PUSCH}$)	-57dBm
Uplink path loss compensation factor (α)	0.7
Maximum tolerable delay ($d_{th,i}$)	20ms
RLC mode	Unacknowledged mode (UM)
Traffic model	H264 video traffic QCIF 176×144
Protocol header sizes	RTP/UDP/IP with ROHC Compression: 3 bytes, PDCP: 2 bytes, RLC: 3 bytes, MAC: 2 bytes, CRC: 3 bytes
Moving average calculation factor (β)	0.2
Maximum distance from the eNodeB	330m
Simulation time	67s

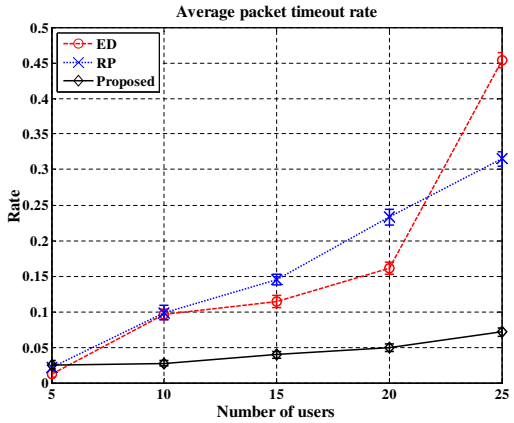

Fig. 4. Average packet timeout rate versus the number of users.

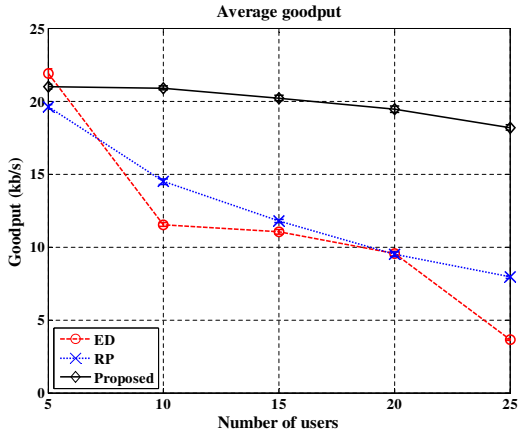

Fig. 5. Goodput versus the number of users.

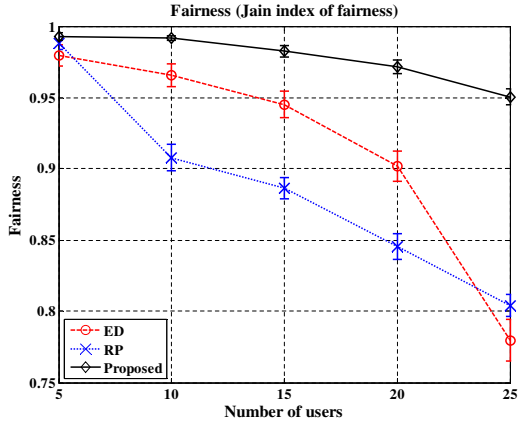

Fig. 6. Fairness (Jain index) versus the number of users.

II. In order to achieve statistical accuracy, 100 simulation runs were executed. In each case, the 95% Confidence Intervals (CI) are depicted in the form of error bars.

Fig. 4 depicts the average packet timeout rate versus an increasing number of users of the system that employs the proposed uplink resource allocation algorithm, the system that equally distributes the scheduling blocks in a QoS- and energy efficiency-agnostic manner, referred to as "ED", and the "Riding Peaks" algorithm of [13], referred to as "RP". The packet timeout rate is defined as the number of packets that expire in the unit of time, since in real-time applications excessive scheduling delay leads to discarding of expired packets. As it can be seen, the packet timeout rate of the ED and RP systems follows a sharp increase with the increase of the number of users due to the fact that the increased congestion results in excessive packet delays and packet expirations that cannot be avoided, since delay is not considered in these resource allocation processes. On the other hand, the system employing the proposed algorithm significantly outperforms the ED and RP systems in terms of packet timeout rate. This is a result of the prioritization of users based on an estimation of their packet delays with respect to their delay threshold, therefore significantly reducing the packet expirations.

Fig. 5 depicts the average goodput of all the systems under consideration. The goodput is defined as the throughput at the application layer, i.e., the rate of useful bits that reach the application layer in the unit of time. As it can be seen, in all three cases the goodput follows a declining course with the increase of the number of users, as a result of the increasing congestion, which leads to excessive packet delays and timeouts. However, the effect of increased congestion is more severe on the ED and RP systems that experience a rapid deterioration of the goodput with the increase of the number of users. On the contrary, the system employing the proposed algorithm achieves a significantly improved goodput, even in the cases of increased number of users.

Fig. 6 depicts the fairness of the three systems that employ the proposed, ED, and RP resource allocation algorithms, respectively. Fairness is evaluated using the Jain Index of Fairness, i.e., $FI = \left(\sum_{i \in UE} Th_i(t)\right)^2 / \left(|UE| \cdot \sum_{i \in UE} Th_i^2(t)\right)$ [35], where $Th_i(t)$ is the throughput of user i. The system that employs the proposed resource allocation algorithm achieves considerably improved fairness compared to the ED and RP systems. This is a result of the fact that the proposed algorithm takes into consideration the average packet delay $\overline{D}_i^{UL}(t)$ and the average data rate $\overline{R}_i^{UL}(t)$ in the user prioritization, therefore favoring users that have experienced high average delay and low average data rate in past allocations.

Fig. 7 depicts the average packet delay versus the number of users. As it can be seen, in the ED system the packet delay significantly increases with the increase of the number of users, as a result of the congestion and the inability of the resource allocation algorithm to prioritize users based on the expected expiration time of their packets. The average delay of the RP system follows a similar, though less sharp course. It is also shown that for small numbers of users the proposed system results in slightly higher delay, although significantly lower than the delay threshold, compared to the ED and RP systems, as a result of its need to accommodate larger queues, since users are efficiently prioritized and their packets are not dropped due to expiration.

In order to highlight the interdependency of the energy efficiency and QoS provision in resource allocation, Fig. 8 depicts the three systems' performance in terms of energy efficiency of successfully received bits, EE_s. This is defined as as the amount of data successfully

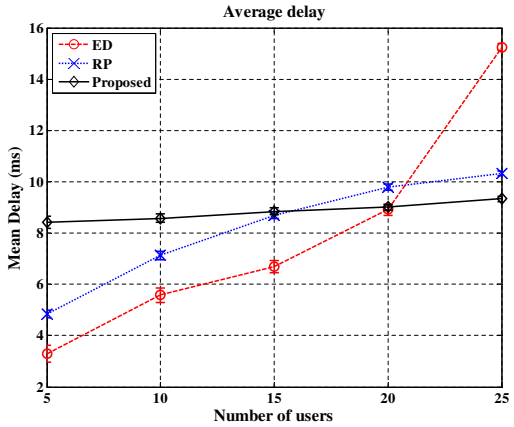

Fig. 7. Average delay versus the number of users.

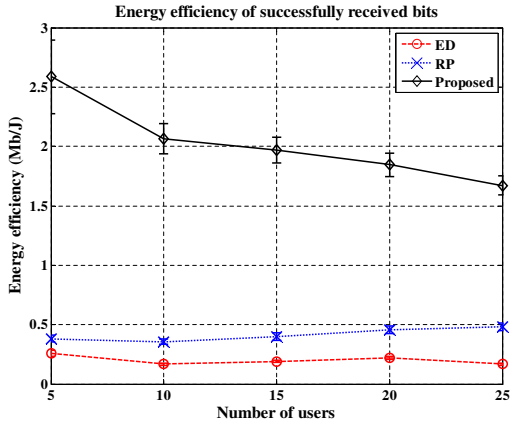

Fig. 8. Total energy efficiency of successfully received bits versus the number of users.

concatenated at the receiver's RLC layer (in Mb) for a given amount of energy (in J) and represents the average energy consumption per successfully received bit. EE_s can be formulated as follows: $EE_s = \left(\sum_{i \in UE} CR_i(1-PL_{t,i})(1-PL_{e,i})\right) / \left(\sum_{t=1}^{T} \sum_{i \in UE} \sum_{j \in \Phi} P_{i,j}(t)\right)$, where CR_i is the user's created data rate in b/s, $PL_{t,i}$ is the user's packet timeout rate, which depends on its delay, $PL_{e,i}$ is the user's packet error rate, which depends on its channel conditions, and $P_{i,j}(t)$ is the transmission power of user i on scheduling block j at time t, $t \in [1, \ldots, T]$. Therefore, the energy efficiency of successfully received bits highly depends on the QoS provision, as it is inverse proportional to the packet timeout rate $PL_{t,i}$ and the packet error rate $PL_{e,i}$ of the users. Therefore, the higher a user's packet timeout and packet error rate, the lower the energy efficiency of successfully received bits, given the fact that the created rate remains the same.

As it can be seen, in the proposed system the energy efficiency of received bits is more than 6-times improved compared to that of the ED system and almost 4-times improved compared to the RP system. This is a result of the fact that, due to packet segmentation performed at the RLC layer of LTE systems, a packet segment loss may be unrecoverable at the receiving side, therefore leading to waste of already received packet segments, whose transmission consumed energy. This could be partly mitigated by efficient ARQ schemes. However, these schemes are not appropriate for real-time applications, since the required retransmissions induce additional delays that may result in a packet having expired before being reassembled at the receiver side. This result highlights the effect that enhanced QoS provision has on energy efficiency, since the lower packet loss rate of the proposed system results in lower waste of already transmitted packet segments, and a larger amount of packets successfully being reassembled by the receiver RLC layer.

VI. Conclusion

In this paper we introduced an uplink resource allocation algorithm for LTE systems, which focuses on QoS provision in real-time applications and energy efficiency. We firstly formulated the problem of optimal uplink resource allocation as a discrete connected cake-cutting problem. However, this problem does not originally consider any upper bounds of the pieces allocated to each user, making it more appropriate for systems with infinitely backlogged traffic. To address this issue and adapt the problem to the traffic needs of a practical system, we defined a modified optimal cake-cutting problem that considers allocation of pieces of bounded size to each user, which similarly to the original problem, is NP-hard. Therefore, we also proposed a suboptimal algorithm, which complies with the constraints of a practical uplink localized SC-FDMA LTE system, i.e., lack of knowledge of the packet delays in the uplink direction, imperfect knowledge of the users' buffer status, and allocation of contiguous sets of resource blocks to each user. Focusing on addressing the delay sensitivity of real-time applications and the need for improved energy efficiency, the proposed algorithm prioritizes users based on their estimated packet delay, the average delay and data rate of past allocations, as well as the required transmission power per resource block. Extensive simulation results highlighted the considerable performance improvement achieved by the proposed algorithm compared to legacy systems in terms of packet timeout rate, goodput, and fairness. Moreover, in order to emphasize on the negative effect of poor QoS provision on energy efficiency, the system was also evaluated in terms of energy consumption per successfully received bit. Therefore, it was shown that poor QoS, as a result of increased packet losses, also results in poor energy efficiency, as the loss of packet segments leads to the inability of the system to perform packet reassembly at the receiver side, resulting in waste of already received packet segments whose transmission consumed energy. Our plans for future work include the extension of the proposed solution to a multicell scenario, also considering interference avoidance features.

Acknowledgment

We would like to acknowledge the support of the University of Surrey 5GIC (http://www.surrey.ac.uk/5gic) members for this work.

REFERENCES

[1] "Enhance mobile networks to deliver 1000 times more capacity by 2020," Nokia Solutions and Networks White Paper, Sept. 2013.

[2] B. Jalili, M. Dianati, B.G. Evans and K. Moessner, "Collaborative radio resource allocation for the downlink of multi-cell multi-carrier systems," *IET Commun.*, vol. 7, no. 5, pp. 430-438, March 2013.

[3] H. Lee, S. Vahid and K. Moessner, "A Survey of Radio Resource Management for Spectrum Aggregation in LTE-Advanced," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 2, pp. 745-760, Second Quarter 2014.

[4] H.G. Myung, J. Lim and D. Goodman, "Single carrier FDMA for uplink wireless transmission," *IEEE Veh. Technol. Mag.*, vol. 1, no. 3, pp. 30-38, Sept. 2006.

[5] J. Fan, Q. Yin, G.Y. Li, B. Peng and X. Zhu, "Joint User Pairing and Resource Allocation for Uplink SC-FDMA Systems," in *Proc. IEEE GLOBECOM*, Houston, TX, USA, 2011, pp. 1-5.

[6] J. Fan, G.Y. Li, Q. Yin, B. Peng and X. Zhu, "Joint User Pairing and Resource Allocation for LTE Uplink Transmission," *IEEE Trans. Wireless Commun.*, vol. 11, no. 8, pp. 2838-2847, Aug. 2012.

[7] F. Ren, Y. Xu, H. Yang, J. Zhang and C. Lin, "Frequency Domain Packet Scheduling with Stability Analysis for 3GPP LTE Uplink," *IEEE Trans. Mobile Comput.*, vol. 12, no. 12, pp. 2412-2426, Dec. 2013.

[8] J. Fan, G.Y. Li, Q. Yin and L. Li, "Multiuser pairing and resource allocation with interference avoidance for SC-FDMA cellular systems," in *Proc. IEEE GLOBECOM*, Anaheim, CA, USA, 2012, pp. 4993-4997.

[9] O. Nwamadi, X. Zhu and A.K. Nandi, "Dynamic physical resource block allocation algorithms for uplink long term evolution," *IET Commun.*, vol. 5, no. 7, pp. 1020-1027, May 2011.

[10] I.C. Wong, O. Oteri and W. Mccoy, "Optimal resource allocation in uplink SC-FDMA systems," *IEEE Trans. Wireless Commun.*, vol. 8, no. 5, pp. 2161-2165, May 2009.

[11] A. Ahmad, "Resource Allocation and Adaptive Modulation in Uplink SC-FDMA Systems," *Springer Wireless Personal Commun.*, vol. 75, no. 4, pp. 2217-2242, 2014.

[12] D.J. Dechene and A. Shami, "Energy-Aware Resource Allocation Strategies for LTE Uplink with Synchronous HARQ Constraints," *IEEE Trans. Mobile Comput.*, vol. 13, no. 2, pp. 422-433, Feb. 2014.

[13] S.B. Lee, I. Pefkianakis, A. Meyerson, S. Xu and S. Lu, "Proportional Fair Frequency-Domain Packet Scheduling for 3GPP LTE Uplink," in *Proc. IEEE INFOCOM*, Rio de Janeiro, Brazil, 2009, pp. 2611-2615.

[14] J. Kim, D. Kim and Y. Han, "Proportional fair scheduling algorithm for SC-FDMA in LTE uplink," in *Proc. IEEE GLOBECOM*, Anaheim, CA, USA, 2012, pp. 4816-4820.

[15] M. Assaad, W. Ben-Ameur and F. Hamid, "Resource Optimization of Non-Additive Utility Functions in Localized SC-FDMA Systems," *IEEE Trans. Signal Process.*, vol. 62, no. 18, pp. 4896-4910, Sept. 2014.

[16] O. Delgado and B. Jaumard, "Scheduling and resource allocation for multiclass services in LTE uplink systems," in *Proc. IEEE 6th Int. Conf. WiMob Comput. Netw. Commun.*, Niagara Falls, NY, USA, 2010, pp. 355-360.

[17] A. El Essaili, L. Zhou, D. Schroeder, E. Steinbach and W. Kellerer, "QoE-driven live and on-demand LTE uplink video transmission," in *Proc. IEEE 13th Int. Workshop MMSP*, Hangzhou, China, 2011, pp. 1-6.

[18] H.-C. Jang and Y.-J. Lee, "QoS-constrained resource allocation scheduling for LTE network," in *Proc. ISWPC*, Taipei, Taiwan, 2013, pp. 1-6.

[19] A. Aijaz, X. Chu and A.H. Aghvami, "Energy Efficient Design of SC-FDMA Based Uplink under QoS Constraints," *IEEE Wireless Commun. Lett.*, vol. 3, no. 2, pp. 149-152, April 2014.

[20] E. Dahlman, S. Parkvall and J. Skld, *4G LTE/LTE-Advanced for Mobile Broadband*. 1st ed., Kidlington, U.K.: Elsevier, 2011.

[21] 3GPP TS 36.321, V12.4.0 (2015-01): 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) protocol specification, Rel. 12.

[22] 3GPP TS 36.213, V12.4.0 (2015-01): 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures, Rel. 12.

[23] 3GPP TS 36.211, V12.4.0 (2015-01): 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation, Rel. 12.

[24] S. Brams, and A. Taylor, *Fair division: from cake cutting to dispute resolution*. Cambridge, U.K.: Cambridge Univ. Press, 1996.

[25] J.B. Barbanel, *The geometry of efficient fair division*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[26] H. Moulin, *Fair division and collective welfare*. The MIT Press, 2003.

[27] J.M. Robertson, and W.A. Webb, *Cake-cutting algorithms - be fair if you can*. Natick, MA: A K Peters, 1998.

[28] Y. Aumann, Y. Dombb, and A. Hassidim, "Computing socially efficient cake divisions," in *Proc. Int. Conf. AAMAS*, Saint Paul, MN, USA, 2013, pp. 343-350.

[29] M.F. Schilling, "The longest run of heads," *The College Mathematics Journal*, vol. 21, no. 3, pp. 196-207, May 1990.

[30] Joint Scalable Video Model (JSVM) reference software, http://www.hhi.fraunhofer.de/de/kompetenzfelder/image-processing/research-groups/image-video-coding/svc-extension-of-h264avc/jsvm-reference-software.html

[31] YUV Video Sequences, http://trace.eas.asu.edu/yuv/

[32] D. Triantafyllopoulou, T. Guo, and K. Moessner, "Energy Efficient ANDSF-assisted Network Discovery for non-3GPP Access Networks," in *Proc. IEEE Int. Workshop CAMAD Commun. Links Netw.*, Barcelona, Spain, 2012, pp. 297-301.

[33] Q. Liu, X. Wang, and G.B. Giannakis, "Cross-Layer Scheduler Design with QoS Support for Wireless Access Networks," *IEEE Trans. Wireless Commun.*, vol. 4, no. 3, pp. 1142-1153, May 2005.

[34] H. Holma, and A. Toskala, *LTE for UMTS: OFDMA and SC-FDMA based radio access*. Chichester, U.K.: John Wiley & Sons, 2009.

[35] R. Jain, D. Chiu, and W. Hawe, "A Quantitative Measure of Fairness and Discrimination for Resource Allocation in Shared Computer Systems," DEC Research Report TR-301, Sept. 1984.

Dionysia Triantafyllopoulou (S'06-M'09) received her B.Sc. in Computer Science in 2005 and her M.Sc. in Communication Systems and Networks in 2007 from the Dept. of Informatics and Telecommunications at the University of Athens, Athens, Greece. In 2009 she received her Ph.D. from the same Department. From 2005 to 2011, she worked as a researcher in the Dept. of Informatics and Telecommunications, University of Athens. Currently, she is a Research Fellow in the Institute for Communication Systems (formerly CCSR) of the University of Surrey, United Kingdom. Her research interests include radio resource management, spectrum sharing and mobility management in cognitive radio and heterogeneous networks.

Konstantinos Kollias is a PhD candidate in Information Science and Technology in the Department of Management Science and Engineering at Stanford University. He previously obtained a B.S. in Informatics and Telecommunications from the University of Athens and a M.S. in Operations Research from Stanford University.

Klaus Moessner is a Professor for Cognitive Networks and Services, in the Institute for Communication Systems (formerly CCSR) at the University of Surrey, UK. Klaus earned his Dipl-Ing (FH) at the University of Applied Sciences in Offenburg, Germany, an MSc from Brunel University and PhD from the University of Surrey (UK). His research interests include dynamic spectrum allocation, cognitive networks, reconfiguration management, service platforms and adaptability of multimodal user interfaces.