WIRELESS COMMUNICATIONS AND MOBILE COMPUTING
Wirel. Commun. Mob. Comput. yyyy; **00**:1–12
DOI: 10.1002/wcm

RESEARCH ARTICLE

Downlink Cross-layer Scheduling Strategies for LTE and LTE-Advanced Systems

Giulio Bartoli[1], Romano Fantacci[1], Dania Marabissi[1], Daniele Tarchi[2]* and Andrea Tassi[1]

[1]Department of Information Engineering, University of Florence, Firenze, Italy
[2]Department of Electrical, Electronic and Information Engineering, University of Bologna, Bologna, Italy

ABSTRACT

The most recent trend in the Information and Communication Technology (ICT) world is toward an ever growing demand of mobile heterogeneous services that imply the management of different Quality of Service (QoS) requirements and priorities among different type of users. The Long Term Evolution (LTE)/LTE-Advanced standards have been introduced aiming to cope with this challenge. In particular the resource allocation problem in downlink needs to be carefully considered. Herein, a solution is proposed by resorting to a modified Multidimensional Multiple-choice Knapsack Problem (MMKP) modeling, leading to an efficient solution. The proposed algorithm is able to manage different traffic flows taking into account users priority, queues delay and channel conditions achieving quasi-optimal performance results with a lower complexity. The numerical results show the effectiveness of the proposed solution with respect to other alternatives.

KEYWORDS

Radio Resource Allocation, OFDMA, LTE-A, Knapsack Problem, Multi-objective Optimization

***Correspondence**

Department of Electrical, Electronic and Information Engineering, University of Bologna, Viale Risorgimento 2, Bologna, Italy. E-mail: daniele.tarchi@unibo.it

1. INTRODUCTION

Nowadays there is an increasing number of mobile terminals (User Equipments, UEs) that handle several multimedia communications (such as audio/video calls) and are becoming effective computing nodes supporting new applications and services. For this reason future wireless broadband communications shall be able to provide a seamless interconnection of mobile users with multiple traffic flows and different QoS constraints. This is challenging in mobile wireless systems characterized by a rapidly changing scenario due to the user mobility, propagation effects and traffic burstiness.

The so-called 3G and beyond communication networks (e.g., UMTS, WiMAX, WCDMA, LTE, etc.) made possible building networks characterized by a multi-user scenario where the terminals can communicate by using multiple traffic classes even in mobility. Among others, the Long Term Evolution (LTE)/LTE-Advanced system, supported by the 3GPP consortium [1], is receiving great attention due to its flexibility and capabilities [2]. It is able to satisfy new communication requirements, thanks to its low latency and high spectral efficiency that guarantee high data rate and real time services.

The LTE physical layer is based on the Orthogonal Frequency Division Multiple Access (OFDMA) technique where the available bandwidth is divided into several

smaller bands, called subchannels, and disjunctive sets of subcarriers are allocated to different users thus providing a flexible multiuser access. An efficient OFDMA scheduler can exploit inherent multi-carrier nature of OFDMA and channel multiuser-diversity to allow link adaptation according to the behaviour of the narrow-band channels. In addition the correct amount of resources is assigned to each user and each traffic flow in order to respect the QoS constraints and the user priority.

The scheduling problem in an OFDMA system can be modeled as a joint subcarrier, rate and power allocation problem: at any scheduling instant, the resource allocation algorithm has to map each traffic flow into a given set of subcarriers with a suitable amount of power and rate taking into account priority, QoS constraints and channel state information. The resource allocation in OFDMA has received a great attention and different solutions can be found in the literature. Among them, in [3] a joint optimal subcarrier, power and rate allocation with the aim of sum-average-rate maximization is presented, while [4] proposes a resource allocation strategy that maximizes the instantaneous or the ergodic rate of the users optimizing the transmission power of each subcarrier. The resource allocation scheme presented in [5], aims to assign more resources to those users characterized by a better channel quality or having received in the past few downlink resources. In particular, this strategy represents an extension of the Proportional Fair (PF) scheduler [6, 7, 8]. In [9], the authors propose a set of subcarrier allocation techniques that, taking into account the channel state information, try to maximize the system capacity guaranteeing user fairness. Instead [10] proposes a scheduling technique that takes into account the users bandwidth requests and different traffic types in order to improve the system throughput. Similarly, [11] proposes a subcarrier allocation strategy that optimizes the overall system ergodic capacity modeling the scheduling problem as a combinatorial problem whose inputs are the total amount of downlink radio resource and the information about the propagation conditions experienced by each user.

Other different resource allocation approaches improve the user experience by imposing the minimization of the interference in a single and multi-cellular environment, as described in [12, 13]. We can note that all the resource allocation algorithms presented in [3]-[13] do not consider explicitly any QoS constraints that may characterize the different traffic flows directed to each UE.

There are several resource allocation strategies that maps the QoS requirements of a mobile user into a minimum sustained rate, however, this solution represents a severe limitation if we consider that a mobile terminal should be able to handle different traffic flows, characterized by different QoS constraints, at the same time.

To the best of our knowledge, the definition of a scheduling scheme able to explicitly take into account different QoS profiles in a multi-user environment is still an open issue. This is a difficult problem to be solved especially considering mobile users and optimal solutions result to be not affordable from a practical point of view.

For this reason different optimization approaches have been investigated. In [4] an optimization approach based on Lagrange multipliers is proposed. It is limited to the radio resource allocation without any constraint on the user requirements. In [14], the authors focus on a scenario similar to that considered in this paper, but the problem is modeled as a convex optimization problem and solved using a dual decomposition approach by exploiting the Hungarian Algorithm. The Hungarian Algorithm has been also considered in [15], while focusing on a video streaming scenario.

A different approach is to model the multiuser downlink resource allocation problem as a combinatorial optimization problem, with a particular attention to the Knapsack Problem (KP) approach, extensively used in operative research context. In the literature some proposal for exploiting the KP modeling in wireless resource allocation has been proposed. In [16] the authors propose to model the resource allocation problem in a WiMAX system as a KP.

Differently from the previous approaches we propose to use the Multidimensional Multiple-choice Knapsack Problem (MMKP) [17] where the optimization is based on the selection of multiple variables. Indeed, the MMKP differs from the classical KP, due to the presence in the model of sets of variables: the objective of the problem is to select the best variable in each set. The optimization problem we consider, where there are multiple users data flows grouped in different QoS classes can be more efficiently modeled as a MMKP rather than a KP.

Wirel. Commun. Mob. Comput. yyyy; **00**:1–12 © yyyy John Wiley & Sons, Ltd.
DOI: 10.1002/wcm
Prepared using wcmauth.cls

Even if the MMKP approach allows to have an optimal solution to the downlink resource allocation problem it suffers of a high complexity [18]. This paper proposes a novel heuristic, named EGRAS (Extended Greedy Resource Allocation Scheme) with the aim of reducing the computational complexity. The effectiveness of the proposed approach, in comparison with others widely adopted scheduling schemes, has been validated by resorting to computer simulations shown in Sec. 5. In this section the performance of the proposed heuristic scheduling strategy will be compared with the exact solution of the proposed MMKP-based scheduling model, showing that the EGRAS reaches a suitable compromise between optimal solution performance and complexity.

The organizations of this paper is the following. In Section 2 the main features of LTE systems and the considered system model are presented, while in Section 3 the focus is on the multi-class multi-user downlink scheduling problem. In Section 4 the proposed scheduling scheme is described by resorting to the MMKP optimization, in Section 5 performance comparisons are given as stated before, and finally conclusions are drawn in Section 6.

2. SYSTEM MODEL

In this section the system model is presented. We consider the downlink phase of a LTE system assuming TDD (Time Division Duplexing) mode of operation (also known as TD-LTE).

The radio resource in a TD-LTE system is organized in radio frames (10 ms long) each one composed by two half-frames, lasting 5 ms. One half-frame is composed by five sub-frames, 1 ms long. According to the TDD duplexing scheme, a sub-frame can carry both uplink or downlink traffic; in particular, in TD-LTE seven different radio frame configurations are defined [2]. This paper relies on the third downlink profile [19] for a TD-LTE radio frame characterized by three uplink sub-frames and five downlink sub-frames in the remaining part.

Since LTE relies on the OFDMA, a downlink sub-frame is a time-frequency grid divided in Resource Blocks (RBs). Each RB is composed by 7 or 6 OFDM symbols (depending on normal or extended cyclic prefix) and 12 contiguous subcarriers. The RB represents the basic resource allocation unit that a scheduling scheme can manage in an atomic manner. Different RBs within a sub-frame can use different modulation and coding schemes (MCS), see Tab. I. The selection of the best suitable MCS is performed assuming a target error rate as described in [20].

According to the LTE standard, different service types can be served; in particular, LTE supports communication flows belonging up to nine QoS profiles, four with Granted Bit Rate (GBR) and five not GBR (Non-GBR).

We consider a network topology composed by one eNodeB and a variable number of UEs randomly placed around the eNodeB. In order to detail more clearly the downlink resource allocation problem without loss of generality, we can imagine that the eNodeB can serve a set $\mathcal{U} = \{u_1, u_2, \ldots, u_M\}$ of M active UEs and is able to manage a set $\mathcal{Q} = \{q_1, q_2, \ldots, q_T\}$ of T downlink traffic flow types (characterized by different QoS constraints), mapped on T buffers (modeled as First Input First Output queues) for each served UE. Hereafter, we will refer to the above mentioned eNodeB buffers as *output queues*.

At each Transmission Time Interval (TTI), the downlink scheduler maps data on the RBs (belonging to the set $\mathcal{R} = \{r_1, r_2, \ldots, r_N\}$ of N elements) forming one downlink data traffic sub-frame.

The parameters that a downlink scheduler should take into account are:

- **Channel Quality Indicator (CQI)** - it represents the main source of information (at the eNodeB side) about the quality of the downlink channel between the eNodeB and a specific UE. Each UE sends to the eNodeB its *CQI report* computed by measuring the downlink reference signal;
- **Delay** - the eNodeB holds the downlink data traffic in one or more queues: a downlink scheduler has to reduce the delay suffered by the packets in each queue;
- **Priorities** - in a network there are users (for e.g., the UEs held by institutional users) and traffic flows (for e.g., VoIP communications), characterized by specific QoS constraints and priorities.

It is important to note that the 3GPP consortium does not provide any guideline for implementing scheduling schemes, demanding it to manufacturer.

DOI: 10.1002/wcm
Prepared using wcmauth.cls

3. MULTI-CLASS MULTI-USER DOWNLINK SCHEDULING

The aim of the scheduling is to maximize the overall throughput of the system by respecting specified QoS constraints. Thus, the problem can be modeled as a combinatorial problem where $M \times T$ variables need to be mapped into N resources, i.e., mapping all the possible QoS classes belonging to all the users into N resource blocks. Moreover, the resulting scheduling problem is modeled to take into account several aspects: the quality of the downlink channel between the eNodeB and each UE, the length of output queues and the data flow priority. Hence, the optimal resource allocation corresponds to solve the problem:

$$\max_{x_{i,j,l}} \left\{ \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{l=1}^{T} x_{i,j,l} v_{i,j,l} \right\} \tag{1a}$$

subject to

$$\sum_{j=1}^{M} \sum_{l=1}^{T} x_{i,j,l} \leq 1 \quad i = 1, \ldots, N \tag{1b}$$

$$\sum_{i=1}^{N} c_{i,j} x_{i,j,l} \leq w_{j,l} \quad j = 1, \ldots, M,\ l = 1, \ldots, T \tag{1c}$$

$$x_{i,j,l} \in \{0, 1\} \quad j = 1, \ldots, M,\ l = 1, \ldots, T \quad i = 1, \ldots, N$$

where $x_{i,j,l}$ is a binary variable equal to 1 if the i-th RB holds data traffic belonging to the l-th QoS class and directed to the j-th UE, or 0 otherwise.

By noticing (1a), the Resource Allocation Problem (RAP) relies on the maximization of an objective function where $v_{i,j,l}$ (called *mapping profit*) represents the profit that the eNodeB achieves if it maps the traffic belonging to the l-th QoS class and directed to the j-th UE in the i-th RB. The constraint (1b) states that one RB can hold at the same time (i.e., in the same TTI) only the traffic toward one UE and belonging to one QoS class (i.e., it corresponds to consider a RB holding data traffic belonging to one outgoing queue). The constraint (1c) states that from the l-th output queue related to the j-th user it is possible to send an amount of data less or equal to the total amount of data $w_{j,l}$ in the queue. The $c_{i,j}$ parameter (called RB capacity or, equivalently, mapping cost) represents the amount of bits that the i-th RB (directed to the j-th user) can hold according to the specific MCS used in the RB.

Table I. Available QI values.

MCS adopted in a RB	QI
QPSK 1/2 (BER $\geq 5 \cdot 10^{-4}$)	0
QPSK 1/2 (BER $< 5 \cdot 10^{-4}$)	2
16-QAM 1/2	4
64-QAM 1/2	6

Given a RB r_i, a user u_j and a QoS class q_l, the mapping profit can be defined as

$$v_{i,j,l} = \alpha g_{i,j} + \beta p_l w_{j,l} + \gamma d_{j,l} \tag{2}$$

where:

- $g_{i,j}$ represents the downlink channel Quality Index (QI) perceived by the j-th user and related to the i-th RB; it should be noted that the MCS cannot change within a RB. The values are reported in Tab. I*;
- p_l represents the relevance of the l-th QoS traffic class (hereafter called "priority"). In a system characterized by several traffic flows, belonging to several QoS classes and characterized by different priority indexes, the relevance of a QoS class is not necessarily function of its priority but can be derived from the statistical characterization (if it is available) of the QoS class itself;
- $d_{j,l}$, is the *transmission delay* and represents the time elapsed from the last transmission event related to the l-th QoS class of the j-th UE until the current scheduling instant.

The three terms of the linear combination are normalized respect to their maximum values in the profit expression. The parameters α, β and γ are non-negative real values representing the weights of the linear combination.

The downlink scheduler in the eNodeB can produce a feasible solution of the cross-layer resource allocation problem by solving the RAP at each TTI. Moreover, by using the RAP-based scheduling, the following goals can be achieved:

*In the case of QPSK, if an UE experiences a Bit-Error-Rate (BER) less than $5 \cdot 10^{-4}$ the QI is equal to 0, otherwise 2.

Wirel. Commun. Mob. Comput. yyyy; **00**:1–12 © yyyy John Wiley & Sons, Ltd.
DOI: 10.1002/wcm
Prepared using wcmauth.cls

- UEs characterized by a better downlink channel (i.e., higher QIs), longer output queues and higher transmission delays are preferred;
- by defining the mapping profit as a linear combination of physical (as the QIs) and MAC (as the $p_l w_{j,l}$) layer indexes, the eNodeB is able to manage scenarios where one or more UEs are characterized by low QIs and long output queue lengths, or scenarios where one or more UEs are characterized by an high QI but they are associated to almost empty output queues. In the first case, the RAP-based scheduler assigns a congruous amount of resources to these users avoiding the queue saturation and in the second one a waste of resources is prevented.
- the delay term introduces more fairness among the users by preventing the starvation of users with low QI and partially filled output queues.

The RAP problem is an integer optimization problem and cannot be solved in a computationally efficient way. In the literature there are several sub-optimal solution to the resource allocation problem. Our proposal is based on the modeling of the problem by using a combinatorial optimization approach, by exploiting the Knapsack Problem (KP) approach as explained in the following section.

4. THE MMKP-BASED DOWNLINK SCHEDULER

The KP approach is a widely known method to solve combinatorial problems. Due to the multidimensional nature of the RAP, we focus here on a variation of the KP, named MMKP. The MMKP is a combinatorial optimization problem, where it is supposed the presence of sets of variables, and the aim is to select the best variable in each set, subject to resource constraints, in order to maximizing the objective function. The MMKP problem can be formulated as [21]:

$$\max \sum_{i=1}^{N} \sum_{j=1}^{M} x_{i,j,l} v_{i,j} \tag{3a}$$

subject to

$$\sum_{j=1}^{M} x_{i,j} = 1 \quad i = 1, \ldots, N \tag{3b}$$

$$\sum_{i=1}^{N} \sum_{j=1}^{M} c_{l,i,j} x_{i,j} \leq w_l \quad l = 1, \ldots, T \tag{3c}$$

$$x_{i,j} \in \{0, 1\}$$

From (3), it is possible to note the similarity between the MMKP formulation and the RAP in (1).

The derivation of an exact solution for the MMKP is usually a complex task. The proposed algorithms, [22, 23, 18], to the best of our knowledge does not consider real-time requirements (for instance those concerning the resource scheduling in a multimedia system [22]), consuming an amount of computing time and resources actually not affordable. However, several heuristic strategies have been introduced addressing MMKP instances of practically interesting problems. The first heuristic method has been originally proposed to solve the 0-1 Knapsack Problem (0-1 KP) [17] and then extended to the MMKP family [17]. This is also the case of the heuristic procedure proposed by [24] and based on the Lagrange multiplier. In [22, 25, 23] the so-called Heuristic (HEU), Modified-HEU (M-HEU) and the Convex-HEU (C-HEU) algorithms have been proposed. Starting from a feasible solution, they update the components of a solution characterized by a low profit (in our case, the mapping profit) producing a new feasible solution characterized by a high profit. The updating process is performed until the maximum number of iterations is reached. In [26] a different heuristic approach is described dealing with a dimensional reduction of the admissible solution set.

Moreover, all these solutions cannot be used to solve the RAP. This is because the MMKP does not directly model the RAP due to the presence of the inequality in one constraint. MMKP approach can be used only to solve a particular case of the RAP optimization problem, where the constraint (1b) is substituted with the following relation:

$$\sum_{j=1}^{M} \sum_{l=1}^{T} x_{i,j,l} = 1 \quad i = 1, \ldots, N \tag{4}$$

This corresponds to consider that all the resources, i.e., the RBs, are always used and that the output queues have

DOI: 10.1002/wcm
Prepared using wcmauth.cls

Procedure 1 Extended Greedy Resource Allocation Scheme

$x_{i,j,l} = 0, \quad i = 1, \ldots, N, \quad j = 1, \ldots, M, \quad l = 1, \ldots, T$
for $i \leftarrow 1$ to N **do**
  $(u^*, q^*, \eta_{i,u^*,q^*}) \leftarrow$ findMaxEfficiency$(1, i)$
  if $u^* \neq$ undefined **and** $q^* \neq$ undefined **then**
    $x_{i,u^*,q^*} \leftarrow 1$
    $w_{u^*,q^*} \leftarrow w_{u^*,q^*} - c_{i,u^*}$
  end if
end for

Procedure 2 findMaxEfficiency(z, i)

if $z = MT$ **and** $w_{f_{i,z}[1],f_{i,z}[2]} - c_{i,f_{i,z}[1]} \geq 0$ **then**
  return $f_{i,z}$
else if $z = MT$ **then**
  return (undefined, undefined, 0)
end if
$f_{i,z+1} \leftarrow$ findMaxEfficiency$(z + 1, i)$
if $f_{i,z}[3] > f_{i,z+1}[3]$ **and** $w_{f_{i,z}[1],f_{i,z}[2]} - c_{i,f_{i,z}[1]} \geq 0$ **then**
  return $f^{i,z}$
else
  return $f_{i,z+1}$
end if

always an amount of resources at least equal to the RBs to be sent.

For this reason we are interested in the solution of a modified MMKP problem where the saturation hypothesis can be neglected leading to a more realistic model. The method proposed here (named EGRAS) resorts to a novel heuristic approach to solve an instance of RAP inspired to the modular dominance [27] widely adopted for the modified MMKP solution. Hence, the EGRAS heuristic allows to approach the modified MMKP modeling of the problem with the advantage of discarding the full buffer approximation needed for the MMKP approach.

4.1. The Extended Greedy Resource Allocation Scheme (EGRAS)

The *Extended Greedy Resource Allocation Scheme* (EGRAS), detailed in Proc. 1 allows to solve the RAP problem without any limitation in terms of output queues saturation.

Given the i-th RB we define a vector $\mathcal{F}_i = \{f_{i,1}, f_{i,2}, \ldots, f_{i,MT}\}$ of (MT) elements. The z-th element of $\mathcal{F}_i$ is a triplet $f_{i,z} = (j, l, \eta_{i,j,l})$ where $j \in \{1, \ldots, M\}$ and $l \in \{1, \ldots, T\}$ represent the user and class index, respectively. The $\eta_{i,j,l}$ value represents the *efficiency* of a triplet and can been defined as:

$$\eta_{i,j,l} = \frac{v_{i,j,l}}{1 + \hat{c} - c_{i,j}} \tag{5}$$

where $\hat{c}$ is the maximum capacity (expressed in bits) of a RB.

An important difference between EGRAS and other heuristic approaches based on the modular dominance is represented by the definition of the efficiency, since in EGRAS it is not a simple ratio between profits and costs. In the RAP modeling we introduced a strong correlation between mapping costs and profits: in particular, given a RB, we are interested in transmitting data traffic belonging to the output queue of the user with the higher mapping value and cost. For these reasons, in (5), the mapping profit is $1 + \hat{c} - c_{i,j}$ (for a given RB $r_i \in \mathcal{R}$ and user $u_j \in \mathcal{U}$).

The Proc. 1 computes for each RB r_i the triplet $(u^*, q^*, \eta_{i,u^*,q^*})$ such that the pair (u^*, q^*) guarantees the maximum efficiency evaluated in the Proc. 2. The procedures take explicitly into account the constraint (1c), while (1b) is implicitly verified.

It is possible to note that in a deployed TD-LTE network RBs and QoS classes (namely N and T) can be considered fixed; for these reasons the EGRAS scheme requires $O(M)$ comparisons, leading to computational complexity linearly increasing with respect to the number of users. It has to be noted that the complexity of EGRAS is lower respect to the MMKP, that is a NP-hard problem, while is comparable with the other heuristic approaches introduced before. However, it has to be noted that we consider a modified MMKP problem, where the inequality in (1b) is considered instead of the classical equality in (3b).

5. NUMERICAL RESULTS

The effectiveness of the proposed scheduling scheme will be validated in this section by resorting to computer simulations. The performance of the EGRAS method has been compared with two widely known downlink resource allocation schemes [2]: the max-C/I and the PF. These scheduling strategies can be summarized as follows:

- *max-C/I* - on each TTI it assigns more downlink resources (in terms of RBs) to the user characterized

Wirel. Commun. Mob. Comput. yyyy; **00**:1–12 © yyyy John Wiley & Sons, Ltd.
DOI: 10.1002/wcm
Prepared using wcmauth.cls

by the best instantaneous channel quality. This implies that the r-th downlink RB will be assigned to the $\hat{u}$-th user such that:

$$\hat{u} = \underset{u=1,\ldots,M}{\arg\max}\, \mathrm{CIR}(r,u)$$

where $\mathrm{CIR}(r,u)$ is the mean Carrier-to-Interference Ratio of the u-th user on the r-th RB. In order to be able to compare the performance with the EGRAS, the max-C/I proposed in [2] has been extended for scheduling downlink traffic belonging to a multi-user and multi-class scenario. Each user has a single data queue that is obtained by means of a suitable combination of all the output queues of that user (referred to data flows with different QoS): it is filled with traffic belonging to different output queues proportionally to the relevance of each QoS class.

- *PF* - it aims to reach the fairness among the users by scheduling the UE with the maximum relative advantage profit [2], defined as the ratio between the instantaneous data rate R_i of the i-th user and his average data rate $\overline{R_i}$. Moreover, for operating in a multi-class and multi-user environment, PF has been properly extended to be able to schedule, not only a particular user, but also a specific flow type. In particular, at a given TTI, the r-th downlink RB will hold data traffic directed to the $\hat{u}$-th UE and belonging to the $\hat{q}$-th traffic flow type such that:

$$(\hat{u},\hat{q}) = \underset{\substack{u=1,\ldots,M\\ q=1,\ldots,T}}{\arg\max}\, \frac{Q_{u,q,r}}{\overline{Q}_{u,q}}$$

where $Q_{u,q,r}$ is the amount of bits carried by the r-th RB holding data belonging to the q-th flow type and directed to the u-th user and $\overline{Q}_{u,q}$ is the total amount of bits transmitted to u-th user and belonging to the q-th flow type until the present scheduling instant.

As discussed in Sec. 3, the RAP model and, as a consequence, the EGRAS scheme are characterized by a set of parameters; all the numerical results discussed in this section have been obtained by setting α, β and γ respectively to 0.4, 0.2 and 0.4 after an extensive set of computer simulations aiming to optimize such parameters, and not reported here for space constraints.

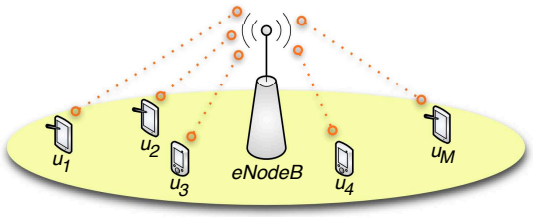


Figure 1. The multi-user simulation scenario.

In Sec. 3 the reference network scenario is detailed, and in Tab. II we report the main system parameters taken into account. Moreover, Tab. III shows the statistical distributions of the downlink data traffics related to the five considered service classes, listed in descending order. The traffic pattern of a service class is characterized by the statistical distributions[†] of the packet length, inter-arrival time and by the maximum sustained delay (defined as the maximum acceptable end-to-end delay for a packet belonging to a specific service class). Finally, Tab. III shows also the considered RAP priorities (Sec. 4).

In Fig. 1 a schematic representation of the considered scenario is reported, where it is possible to highlight the presence of multiple users randomly placed within a cell.

In order to highlight the performance of the different scheduling scheme for the QoS classes characterized by a higher relevance and, at the same time, to depict in a more synthetic way the QoS indexes of interest, we have defined (for a given user u) the *composite throughput* ($\widehat{\Gamma}_u$), the *composite average delay* ($\widehat{\Upsilon}_u$) and the *composite outage* ($\widehat{\Theta}_u$). They represent, respectively, the weighted average throughput, the average delay and the outage probability[‡] for each service class:

$$\widehat{\Gamma}_u = \frac{\sum_{t=1}^{T} \Gamma_{u,t}\, y_t}{\sum_{t=1}^{T} y_t} \tag{6}$$

$$\widehat{\Upsilon}_u = \frac{\sum_{t=1}^{T} \Upsilon_{u,t}\, y_t}{\sum_{t=1}^{T} y_t} \tag{7}$$

$$\widehat{\Theta}_u = \frac{\sum_{t=1}^{T} \Theta_{u,t}\, y_t}{\sum_{t=1}^{T} y_t} \tag{8}$$

[†] Looking at Tab. III: Par(a, b, c) stands for a bounded Pareto with a as a shape factor, b as the minimum and c as the maximum value, Exp(m) represents an exponential distribution with a mean value m and Const(c) refers to a constant value of c.

[‡] The outage probability of a QoS class is defined as the ratio between the number of data packets received within the maximum delay and the total number of generated packets.

DOI: 10.1002/wcm
Prepared using wcmauth.cls

Table II. System parameters.

Parameter	Value
LTE system duplexing type	TDD
radio frequency carrier	2.6 GHz
bandwidth	10 MHz
number of FFT point	1024
supported MCS	QPSK 1/2, 16-QAM 1/2, 64-QAM 1/2
RB size	12 subcarriers $\times$ 7 OFDM symbols
channel model	vehicular ITU-R A [28]

Table III. Traffic classes characterization.

Service Class	Interarrival times [ms]	Packet lengths [Byte]	Max. delay budget [ms]	RAP priorities
I	Par(1.2,0.10,0.50)	Par(1.2,20,125)	100	0.38
II	Exp(5.43)	Par(1.7,466,3000)	200	0.28
III	Exp(3.02)	Par(1.1,81.5,1500)	300	0.14
IV	Exp(40.00)	Const(2554)	400	0.11
V	Const(0.80)	Const(33)	400	0.07

where $\Gamma_{u,l}$, $\Upsilon_{u,l}$ and $\Theta_{u,l}$ are the throughput, the average delay and the outages of the traffic belonging to the l-th QoS class toward the u-th user, respectively. While y_n is the n-th weight computed as $y_n = 2^{T+1-n}$ and denotes the importance of the class (i.e., the relevance of the n-th QoS class).

The system performance has been investigated in a scenario characterized by a variable number of active UEs characterized by different average Signal-to-Noise Ratio (SNR) levels and downlink channel states. In particular, we have analyzed the performance of a reference UE by varying its SNR level, while the SNR levels of the other UEs have been set (and kept constant) to values belonging to an exponential distribution with mean value 10 dB.

First of all we aim to compare the exact solution of the RAP with that obtained by the EGRAS scheme. As presented in Sec. 4.1, the EGRAS scheme represents an heuristic method leading to an admissible solution to the RAP. To this aim, we considered 1000 instances of the RAP problem, for each of which we evaluated the objective function (1a), by comparing the exact solution and the solution provided by the EGRAS heuristic method. The exact solution of the RAP problem has been derived by exploiting the version 4.32 of the GLPK solver [29]. It is worth to notice that the exact solution refers to the solution of the RAP problem by using the MMKP modeling.

Tab. IV shows, for a variable number of UEs, the mean (μ_κ), variance (σ^2_κ) and maximum ($\max_i\{\kappa_i\}$) of the normalized gap defined for each instance i-th of the problem as:

Table IV. Mean, variance and maximum of the normalized gap as function of the number of UEs in the network.

Number of UEs	μ_κ [%]	σ^2_κ [%]	$\max_i\{\kappa_i\}$ [%]
4	3.85	0.04	15.48
6	1.13	0.02	14.12
8	1.03	0.01	17.18
10	0.94	0.01	12.95
12	2.20	0.01	8.21

$$\kappa_i = \frac{\chi_i - \psi_i}{\chi_i} \tag{9}$$

where χ_i and ψ_i are the value assumed by the objective function (1a) when the i-th instance of the RAP is solved by the optimum solver or by the EGRAS procedure, respectively. As we can note, the EGRAS performance is very close to the exact solution: the μ_κ value is not greater than 3.85% (with a variance of 0.04%). In that sense it is possible to conclude that, even if we can have a sub-optimal solution, the performance can be considered sufficiently close to the optimum solution, while the computational cost of the solution algorithm is drastically reduced. Moreover, the proposed solution allows to consider the original problem without any approximation in terms of saturated queues, as for the MMKP modeling.

DOI: 10.1002/wcm
Prepared using wcmauth.cls

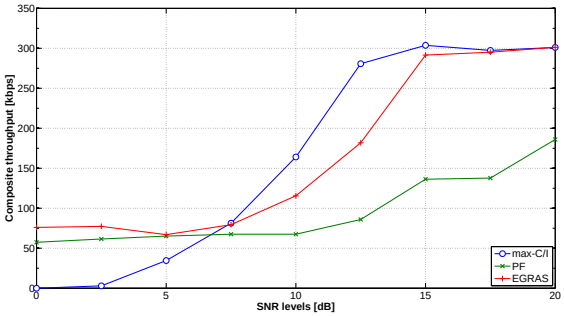


Figure 2. The composite throughput of the reference user as function of the SNR value.

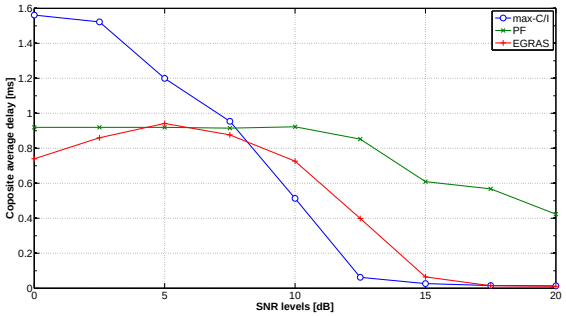


Figure 3. The composite average delay of the reference user as function of the SNR value.

Figs. 2 and 3 show, respectively, the composite throughput and average delay of the reference user, in a network scenario composed by 10 UEs, by considering all the introduced scheduling schemes: the EGRAS scheme achieves excellent results, only overcame by the max-C/I. However, differently from PF and EGRAS, max-C/I cannot take into account the fairness among the users, resulting in unbalanced downlink throughput among the UEs. This aspect is evident in Fig. 4 where the composite throughput of each user is shown. From this figure we can note that both EGRAS and PF ensure a fair distribution of the throughput, while max-C/I, as expected, assigns more resources to those users characterized by the highest SNR levels.

It is possible to note that while PF and max-C/I algorithms have a cross-point, the EGRAS algorithm allows to maximize the performance for all the considered SNR values. This is due to the fact that at low SNR values the other users have, in average, a higher SNR, so that the max-C/I algorithm is more prone to give them the resources. On the contrary the PF tries to give the same resources to all the users for increasing the fairness. At higher SNR, the reference user has a Channel State better than the other users, so that the max-C/I algorithm tends to give it more resources respect to the PF. The cross point at 7.5 dB is an equilibrium point where the reference user has an SNR almost corresponding to the average SNR of all the users, leading to similar performance for the considered algorithms. However, the EGRAS follows the PF algorithms for low SNR and the max-CI algorithm for high SNR maximizing the performance in terms of throughput and delay for all the SNR values.

The performance results in terms of fairness have been also reported in Fig. 5, where it is possible to note that

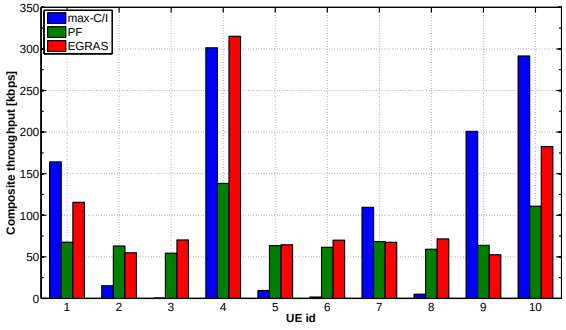


Figure 4. The composite throughput of each user.

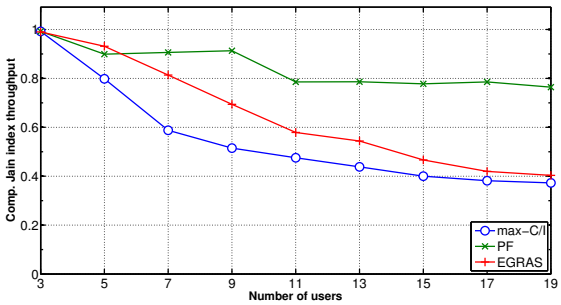
the fairness of the EGRAS method is always better respect to the max-C/I algorithm, while the PF algorithm has better fairness performance: this is expected because the PF algorithms works by maximizing the fairness.

Towards this end, we have resorted to the following formula [30]:

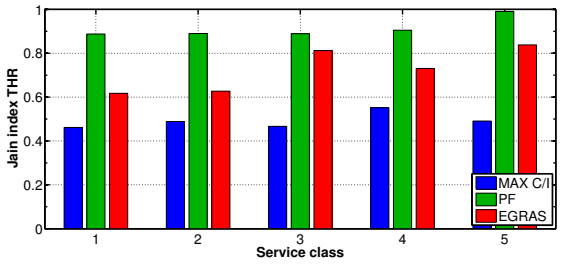
$$f(\mathbf{x}) = \frac{\left|\sum_{i=1}^{N} x_i\right|^2}{N \cdot \sum_{i=1}^{N} x_i^2} \qquad (10)$$

where x_i represents the composite throughput of the i-th user, i.e., $x_i = \Gamma_i$, $\mathbf{x}$ is a vector whose components are the x_i values, and N represents the number of users. The fairness index in (10) has a value between 0 and 1. Note that $f(\mathbf{x}) = 1$ denotes the best fairness performance and corresponds to a fair distribution of resource among all MSs.

However, by considering jointly the throughput and the fairness performance it is possible to note that the EGRAS algorithm maximizes the throughput performance with respect to the two other algorithms with fairness performance better than the max-C/I alternative.



(a) The composite fairness index for a variable number of users.



(b) The composite fairness index for the considered QoS classes.

Figure 5. The performance in terms of fairness for different number of users and classes.

The system performance has been also investigated with a variable number of UEs; in this case the SNR level of each user has been set (and kept constant) to 10 dB. In order to better represent the main QoS indexes of the downlink data traffic we introduce the *total composite throughput* ($\tilde{\Gamma}$) defined as:

$$\tilde{\Gamma} = \sum_{u=1}^{M} \widehat{\Gamma}_u \qquad (11)$$

Fig. 6 shows the total composite throughput; we can note that also here EGRAS and max-C/I have similar performance but, as shown in Fig. 7 and Fig. 8, EGRAS succeeds to manage the quality of service in a better way. EGRAS, in fact, is characterized by the lowest composite average delay; this leads to a reduced composite outage value among the other considered alternatives (in a network scenario composed by a number of UEs equal to or grater than 9).

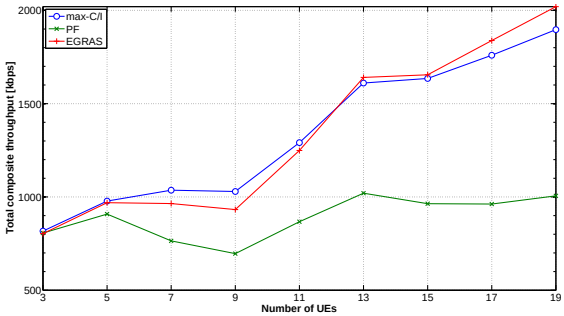


Figure 6. The total composite throughput as function of the number of UEs for a mean SNR value of 10 dB.

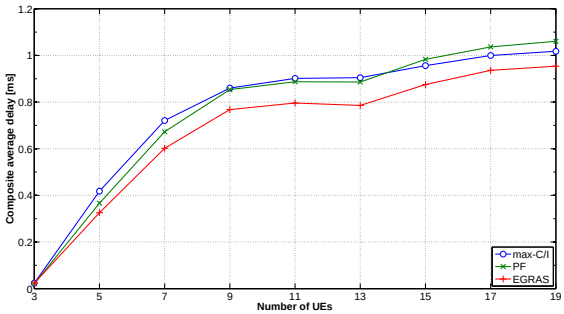


Figure 7. The average composite delay as function of the number of UEs for a mean SNR value of 10 dB.

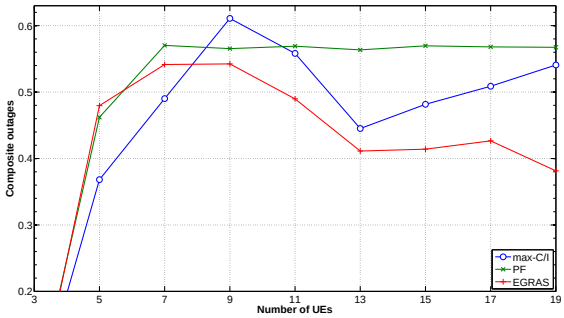


Figure 8. The average composite outage as function of the number of UEs for a mean SNR value of 10 dB.

6. CONCLUSION

One of the important characteristics of the modern broadband wireless communication systems is the support of multimedia communications among several users even in mobility. The LTE/LTE-A systems are designed to meet these requirements, however, when the number of users and application increases, each one with a specific QoS level, the complexity of the resource allocation problem increases. The aim of this paper is to present a possible solution to the resource allocation problem by exploiting

DOI: 10.1002/wcm
Prepared using wcmauth.cls

the MMKP modeling under specific assumptions. Due to the complexity of the MMKP model, we have focused on a suitable heuristic solution allowing to have performance close to the optimal solution of the MMKP problem whit a lower complexity. The effectiveness of the solution has been also proved by considering the performance results in a realistic scenario where multiple users with different multimedia traffic also in terms of QoS classes have been considered. The results have been compared to those obtained with the PF and max-C/I algorithms showing the effectiveness of the proposed solution.

A. PROOF OF THE LEMMA 1

Proof
Let

$$F : \{1, \ldots, (M\,T)\} \longrightarrow \{1, \ldots, M\} \times \{1, \ldots, T\}$$

be a bijective function, for $\hat{i} = 1, \ldots, N$, $\hat{j} = 1, \ldots, M$, $\hat{l} = 1, \ldots, T$, $\hat{h} = 1, \ldots, (M\,T)$ and $\hat{k} = 1, \ldots, (M\,T)$, without any loss of generality we have that:

- the $x_{\hat{i},\hat{j},\hat{l}}$ of RAP, for $F(\hat{h}) = (\hat{j}, \hat{l})$, can be rewritten as $x_{\hat{i},\hat{h}}$;
- the length of the output queue associated to the $\hat{j}$-th user, holding traffic belonging to the $\hat{l}$-th QoS class, for $F(\hat{k}) = (\hat{j}, \hat{l})$ can be also rewritten as $w_{\hat{k}}$;

Let the $r_{\hat{i},\hat{h},\hat{k}}$ parameter be defined as

$$r_{\hat{i},\hat{h},\hat{k}} = \begin{cases} c_{\hat{i},\hat{j}} & \text{if } F(\hat{h}) = F(\hat{k}) = (\hat{j}, \hat{l}) \\ 0 & \text{otherwise.} \end{cases} \quad (12)$$

the MRAP problem can be equivalently rewritten as follows:

$$\max_{x_{i,h}} \left\{ \sum_{i=1}^{N} \sum_{h=1}^{M\,T} x_{i,h} v_{i,h} \right\}$$

subject to

$$\sum_{i=1}^{N} \sum_{h=1}^{M\,T} x_{i,h} r_{i,h,k} \leq w_k \quad k = 1, \ldots, (M\,T) \quad (13)$$

$$\sum_{h=1}^{M\,T} x_{i,h} = 1 \quad i = 1, \ldots, N$$

$$x_{i,h} \in \{0, 1\} \quad i = 1, \ldots, N, h = 1, \ldots, (M\,T)$$

□

REFERENCES

1. 3GPP - The Mobile Broadband Standard. URL http://www.3gpp.org/.
2. Dahlman E, Parkvall S, Sköld J, Beming P. *3G Evolution: HSPA and LTE for Mobile Broadband*. 2nd edn., Elsevier Ltd: Oxford, United Kingdom, 2008.
3. Wang X, Giannakis GB. Resource allocation for wireless multiuser OFDM networks. *IEEE Transactions on Information Theory* Jul 2011; **57**(7):4359–4372.
4. Seong K, Mohseni M, Cioffi JM. Optimal resource allocation for OFDMA downlink systems. *Proc. of IEEE ISIT 2006*, Seattle, WA, USA, 2006; 1394–1398, doi:10.1109/ISIT.2006.262075.
5. Luo H, Ci S, Wu D, Wu J, Tang H. Quality-driven cross-layer optimized video delivery over LTE. *IEEE Communications Magazine* Feb 2010; **48**(2):102–109, doi:10.1109/MCOM.2010.5402671.
6. Kwan R, Leung C, Zhang J. Proportional Fair Multiuser Scheduling in LTE. *IEEE Signal Process. Lett.* Jun 2009; **16**(6):461–464, doi:10.1109/LSP.2009.2016449.
7. Kaneko M, Popovski P, Dahl J. Proportional fairness in multi-carrier system: upper bound and approximation algorithms. *IEEE Commun. Lett.* Jun 2006; **10**(6):462–464, doi:10.1109/LCOMM.2006.1638616.
8. Wang Y, Pedersen K, S andrensen T, Mogensen P. Carrier load balancing and packet scheduling for multi-carrier systems. *IEEE Trans. Wireless Commun.* May 2010; **9**(5):1780–1789, doi:10.1109/TWC.2010.05.091310.
9. Marabissi D, Tarchi D, Fantacci R, Biagioni A. Adaptive subcarrier allocation algorithms in wireless OFDMA systems. *Proc. of IEEE ICC'08*, Beijing, China, 2008.
10. Fantacci R, Marabissi D, Tarchi D. Adaptive scheduling algorithms for multimedia traffic in wireless OFDMA systems. *Physical Communication* 2009; **2**(3):228–234.

11. Larsson E. Optimal OFDMA Downlink Scheduling Under a Control Signaling Cost Constraint. *IEEE Trans. Commun.* Oct 2010; **58**(10):2776–2781, doi: 10.1109/TCOMM.2010.082010.090215.
12. Rahman M, Yanikomeroglu H. Enhancing cell-edge performance: a downlink dynamic interference avoidance scheme with inter-cell coordination. *IEEE Trans. Wireless Commun.* Apr 2010; **9**(4):1414–1425, doi:10.1109/TWC.2010.04.090256.
13. Ali S, Leung V. Dynamic frequency allocation in fractional frequency reused ofdma networks. *IEEE Trans. Wireless Commun.* Aug 2009; **8**(8):4286–4295, doi:10.1109/TWC.2009.081146.
14. Huang J, Subramanian VG, Agrawal R, Berry R. Joint scheduling and resource allocation in uplink OFDM systems for broadband wireless access networks. *IEEE Journal on Selected Areas in Communications* Feb 2009; **27**(2):226–234.
15. Lin WD, Hsieh HY. Joint optimization of resource allocation and modulation coding schemes for unicast video streaming in OFDMA networks. *Proc. of IEEE PIMRC 2012*, 2012; 571 –576.
16. Cicconetti C, Lenzini L, Lodi A, Martello S, Mingozzi E, Monaci M. A fast and efficient algorithm to exploit multi-user diversity in ieee 802.16 bandamc. *Computer Networks* 2011; **55**(16):3680–3693.
17. Kellerer H, Pferschy U, Pisinger D. *Knapsack Problems*. Springer-Verlag: Berlin, Germany, 2004.
18. Khan S. Quality adaptation in a multi-session adaptive multimedia system: Model and architecture. PhD Thesis, University of Victoria, Victoria, Canada 1998.
19. Sesia S, Toufik I, Baker M. *LTE, The UMTS long term evolution: From Theory to Practice*. 2nd edn., Wiley: Chichester, United Kingdom, 2011.
20. Marabissi D, Tarchi D, Fantacci R, Balleri F. Efficient adaptive modulation and coding techniques for WiMAX systems. *Proc. of IEEE ICC'08*, Beijing, China, 2008.
21. Parra-Hernandez R, Dimopoulos NJ. A new heuristic for solving the multichoice multidimensional knapsack problem. *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans* Sep 2005; **35**(5):708–717.
22. Khan S, Li KF, Manning EG, Akbar MM. Solving the knapsack problem for adaptive multimedia system. *Studia Informatica Universalis* 2003; **1**(1):157–178.
23. Akbar M, Manning E, Shoja G, Khan S. Heuristic solutions for the multiple-choice multi-dimension knapsack problem. *Computational Science - ICCS 2001*, *Lecture Notes in Computer Science*, vol. 2074, Alexandrov V, Dongarra J, Juliano B, Renner R, Tan C (eds.). Springer Berlin / Heidelberg, 2001; 659–668.
24. Moser M, Jokanovic D, Shiratori N. An algorithm for the multidimensional multiple-choice knapsack problem. *IEICE Trans. on Fundamentals of Electronics, Communications and Computer Sciences* Mar 1997; **80**(3):582–589.
25. Akbar MM, Rahman MS, Kaykobad M, Manning E, Shoja G. Solving the multidimensional multiple-choice knapsack problem by constructing convex hulls. *Computers & Operations Research* 2006; **33**(5):1259–1273, doi:10.1016/j.cor.2004.09.016.
26. Ibarra OH, Kim CE. Fast approximation algorithms for the knapsack and sum of subset problems. *Journal of the ACM* Oct 1975; **22**(4):463–468, doi:http://doi.acm.org/10.1145/321906.321909.
27. Martello S, Toth P. Algorithms for knapsack problems. *Surveys in Combinatorial Optimization*, Martello S, Laporte G, Minoux M, Ribeiro C (eds.). chap. 7, Elsevier Science Publishers: Amsterdam, Netherlands, 1987; 213–258.
28. *Guidelines for evaluation of radio transmission technologies for IMT-2000* Feb 1997. ITU-R Rec. M.1225.
29. GNU Linear Programming Kit. URL http://www.gnu.org/software/glpk.
30. Jain R, Chiu DM, Hawe W. A quantitative measure of fairness and discrimination for resource allocation in shared systems. *Technical Report TR-301*, Digital Equipment Corporation Sep 1984.

DOI: 10.1002/wcm
Prepared using wcmauth.cls