WIRELESS COMMUNICATIONS AND MOBILE COMPUTING
Wirel. Commun. Mob. Comput. (2014)
Published online in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/wcm.2526

RESEARCH ARTICLE

Maximization of user satisfaction in OFDMA systems using utility-based resource allocation

Emanuel B. Rodrigues*, Francisco R. M. Lima, Tarcisio F. Maciel and Francisco R. P. Cavalcanti

Wireless Telecommunications Research Group (GTEL), Department of Teleinformatics Engineering (DETI), Federal University of Ceará (UFC), Pici Campus, PO Box 6005, Building 722, Fortaleza, Brazil

ABSTRACT

In order to keep and/orexpand its share of the wireless communication market and decrease churn, it is important for network operators to keep their users (clients) satisfied. The problem to be solved is how to increase the number of satisfied non-real time (NRT) and real time (RT) users in the downlink of the radio access network of an orthogonal frequency division multiple access system. In this context, the present work proposes a method to solve the referred problem using a unified radio resource allocation (RRA) framework based on utility theory. This unified RRA framework is particularized into two RRA policies that use sigmoidal utility functions based on throughput or delay and are suitable for NRT and RT services, respectively. It is demonstrated by means of system-level simulations that a step-shaped sigmoidal utility function combined with a channel-aware opportunistic scheduling criterion is effective toward the objective of user satisfaction maximization. Copyright © 2014 John Wiley & Sons, Ltd.

KEYWORDS

resource allocation; utility theory; quality of service; 4G mobile communication

***Correspondence**

Emanuel B. Rodrigues, Wireless Telecommunications Research Group (GTEL), Department of Teleinformatics Engineering (DETI), Federal University of Ceará (UFC), Pici Campus, PO Box 6005, Building 722, Fortaleza, Brazil.
E-mail: emanuel@gtel.ufc.br

1. INTRODUCTION

The wireless shared channel in cellular networks is a medium over which many user equipments (UEs) compete for resources. In such a scenario, resource efficiency, quality of service (QoS)/user satisfaction, and user fairness are crucial aspects for resource allocation.

From a cellular operator's perspective, it is very important to use the limited radio resources efficiently in order to maximize revenues. From the users' point of view, it is more important to have a fair resource allocation so that they can meet their QoS requirements and maximize their satisfaction. The time-varying nature of the wireless environment, coupled with different channel conditions for different UEs, poses significant challenges to accomplish the goal of maximizing the number of satisfied users.

In order to keep and/or expand its share of the wireless communication market and decrease churn, it is important for network operators to keep their users (clients) satisfied. Therefore, it is desirable to have the maximum possible number of satisfied users. The problem to be solved is how to increase the number of satisfied non-real time (NRT) and real time (RT) users in the downlink of the radio access network of an orthogonal frequency division multiple access (OFDMA) system. In this context, the present work proposes a method to solve the referred problem using a dynamic radio resource allocation (RRA) framework based on utility theory, which runs distributedly in each base station (BS).

The rest of the paper is organized as follows. Section 2 presents the state-of-the-art and highlights the main contributions of this work. Section 3 presents the general theory behind utility-based optimization, while Section 4 shows how we can use it to propose a utility-based RRA framework for OFDMA systems. In Section 5, we consider a particular utility function that is suitable for the maximization of user satisfaction. Based on this choice, we describe two RRA techniques that maximize the number of satisfied NRT and RT users, respectively. Next, the system-level performance evaluation is presented in Section 6, and finally Section 7 draws the conclusions.

2. RELATED WORK

2.1. Non-real time services

The performance of RRA techniques in OFDMA-based systems with NRT services has been extensively evaluated; for example see [1,2]. However, as far as we are concerned, the specific problem of satisfaction maximization for NRT users using resource allocation techniques was the object of study of [3–5].

The satisfaction oriented resource allocation for NRT services (SORA-NRT) algorithm initially proposed in [3] was further evaluated in [4]. SORA-NRT is a heuristic downlink scheduling algorithm whose main goal is to guarantee that most of the users achieve the minimum QoS requirements in order to maximize the number of satisfied users. On the other hand, in [5] an adaptive RRA framework is proposed, which can be configured as different RRA policies. By means of the adaptation of a control parameter, this framework changes its configuration and can maintain user satisfaction at high levels for different system loads. In particular, previous algorithms proposed in the literature are either computationally complex [3,4] or not easily tunable or scalable [5].

Other works try to maximize some metric under minimum rate constraints [6–8]. The approach used in [6] is the formulation of an optimization problem based on a long-term objective, where the system rate is to be maximized in an expected sense, and the guarantee of a long-term constraint, where the constraint is an instantaneous rate requirement to be guaranteed in expectation. The optimal solution presented in [6] requires the calculation of Karush–Kuhn–Tucker (KKT) multipliers in each time slot using a stochastic approximation algorithm, which is computationally complex. Moreover, their optimal solution is limited because it does not give any information about the system feasibility.

The algorithm proposed in [7] was designed to maximize the frequency diversity gain, i.e., allow the largest amount of users to obtain non-outage sub-carriers. In contrast to the allocation based on the multi-user diversity, which allocates the sub-carriers to the users with best channel conditions, this algorithm allocates non-outage (may not be the best) sub-carriers to every user using a matching method based on the random bipartite graph theory. The assumptions and definitions in [7] are simple and not realistic, but their main objective was to provide insights for understanding OFDMA systems, and also set up a basic theoretical framework.

Finally, the scheme described in [8] first performs a priority-based resource allocation algorithm to fulfill the rate requirement of the users with highest priority thresholds. Next, it performs a channel state information (CSI)-based resource allocation algorithm for the remaining users to enhance system throughput.

2.2. Real time services

In order to increase the percentage of satisfied users in a scenario with RT services, RRA techniques should take into account efficiency in the resource usage and QoS guarantees (delay bounds). We have classified some works that dealt with both factors in three main approaches: heuristics [4,9], cross-layer packet scheduling (PSC) [10–14], and utility theory [15–17].

A heuristic algorithm called satisfaction oriented resource allocation for RT services (SORA-RT), which is the counterpart of the SORA-NRT algorithm in [3], was initially proposed in [9] and further evaluated in [4]. It is a downlink scheduling algorithm initially designed for the voice over IP (VoIP) service and whose main objective is to maximize the number of satisfied RT users in the system. To the best of our knowledge, this is the only work in the literature so far that deals with the specific problem of user satisfaction maximization in RT service scenarios.

The opportunistic PSC algorithms suitable for RT services found in the literature have priority functions that use an efficiency indicator, such as the instantaneous transmission rate (rate maximization policy) [10–12] or the ratio between the instantaneous transmission rate and throughput (proportional fairness policy) [12–14], and a QoS indicator based on delay. The idea behind these algorithms is not only using the resources in the most efficient way but also giving priority to users with poorer QoS (higher delays).

The utility-based PSC algorithms adopted a similar but more general procedure. The difference is that the QoS indicator used in the priority functions is now a marginal utility function based on delay. For example, [15] and [16] used z-shaped utility functions while [17] used particularly designed utility functions suitable to the services investigated therein. Since the utility functions can be freely designed to provide the desired result, the utility-based approach is more general than classic PSC priority functions.

2.3. Contributions

The main contributions of the present work are:

(1) **Proposal of a generalized utility-based RRA framework**: Being a flexible and powerful tool, utility theory is used here to formulate a utility-based optimization problem, which is presented in Section 3, that allows us to develop a novel and generalized RRA framework able to solve our user satisfaction maximization problem, as it will be shown in Section 4. The proposed framework utilizes a suitable combination between the efficiency and QoS indicators, which is governed by the chosen utility function. In contrast to other RRA techniques based on heuristics [3,4,9], or common sense [5,10–15], our framework uses an RRA policy derived from a mathematical formulation.

Furthermore, our framework can be used in both NRT and RT scenarios, which is an advantage compared with other solutions that are suitable for either NRT or RT services.

(2) **Use of sigmoidal utility functions for satisfaction maximization**: We figured out that sigmoidal utility functions are especially suited for satisfaction maximization problems, both for NRT and RT services, as explained in Section 5.1.

(3) **Proposal of two low-complexity RRA techniques**: Choosing specific QoS metrics and sigmoidal shapes that are appropriate to NRT or RT services, we were able to develop two RRA techniques to be used in each of these services scenarios. These techniques are described in Sections 5.2 and 5.3. Notice that they were particularized from the general RRA framework described in Section 4. Finally, the derived RRA techniques are tunable, scalable, and have low computational complexity. These are important advantages compared with other techniques that do not present a simple configurability or a good scalability, such as [5,7], or those that demand a high computational cost, e.g. [4,6].

3. UTILITY-BASED OPTIMIZATION FORMULATION

3.1. General formulation

Utility theory can be used in communication networks to quantify the benefit of usage of certain resources, e.g., bandwidth and/or power; or evaluate the degree to which a network satisfies service requirements of users' applications, e.g., in terms of throughput and delay.

The general utility-based optimization problem considered in this work is formulated as:

$$\max_{\mathcal{S}_j,\mathbf{p}} \sum_{j=1}^{J} U(x_j) \tag{1a}$$

$$\text{subject to} \bigcup_{j=1}^{J} \mathcal{S}_j \subseteq \mathcal{S} \tag{1b}$$

$$\mathcal{S}_i \bigcap \mathcal{S}_j = \varnothing, \quad i \neq j, \quad \forall i,j \in \{1,2,\ldots,J\} \tag{1c}$$

$$\sum_{k=1}^{K} p_k \leq P_{\mathrm{t}} \tag{1d}$$

$$p_k \geq 0, \quad \forall k \in \{1,2,\ldots,K\} \tag{1e}$$

where J is the total number of UEs in a cell, K is the total number of resources in the system (sub-carriers, codes, or the like) to be assigned to the users, $\mathcal{S}$ is the set of all resources in the system, $\mathcal{S}_j$ is the subset of resources assigned to the j^{th} UE, $\mathbf{p} = \begin{bmatrix} p_1 & p_2 & \ldots & p_K \end{bmatrix}$ is the vector of powers for all resources, p_k is the power of the k^{th} resource, P_{t} is total transmit power of the BS, and $U(x_j)$ is a utility function based on a generic variable x_j that can represent a resource usage or QoS metric of user j. Constraints (1b) and (1c) state that the union of all subsets of resources assigned to different users must be contained in the total set of resources available in the system, and that these subsets must be disjoint, i.e., the same resource cannot be shared by two or more users in the same (TTI). On the other hand, constraints (1d) and (1e) require that the total sum of the powers over all resources must not surpass the total transmit power of the BS, and that these powers must be positive.

The optimum solution for the joint optimization problem (1) is very difficult to be found [18]. Most of the suboptimum solutions proposed in the literature are based on the problem-splitting technique, which splits the problem in two stages: first, dynamic resource assignment (DRA) with fixed power allocation, and next, adaptive power allocation (APA) with fixed resource assignment. In the present work, we also use the problem-splitting technique. It has been shown for OFDMA-based systems that APA provides limited gains in comparison with Equal Power Allocation (EPA) with much more complexity [18]. Therefore, we also consider EPA among the resources.

Depending on the utility function and the variable x, several RRA policies can be designed. In this study, we are interested at formulating general RRA techniques suitable for maximizing the satisfaction of NRT or RT services. Therefore, we consider the variable x to be either the users' throughput (average data rates) or the users' head-of-line (HOL) packet delay, which are QoS parameters suitable for NRT and RT services, respectively. The particular optimization formulations suitable for each of these classes of service are described in the following.

3.2. Non-real time services

Regarding NRT services, the considered optimization problem is the maximization of the total utility with respect to the throughput so that the objective function (1a) becomes

$$\max_{\mathcal{S}_j} \sum_{j=1}^{J} U(T_j[n]) \tag{2}$$

where $U(T_j[n])$ is an increasing utility function based on the current throughput $T_j[n]$ of the j^{th} UE at the n^{th} TTI. Notice that the power vector is not an optimization variable anymore because EPA was chosen in the second stage of the problem-splitting approach, as explained previously. Therefore, the optimization problem considering the objective function (2) and constraints (1b)-(1e) is a DRA problem.

It is demonstrated in Appendix A that we are able to derive a new simplified DRA problem that is equivalent to our original DRA problem (2) regarding NRT services. The objective function of the new simplified DRA problem

is linear in terms of the instantaneous user's data rate and given by

$$\max_{\mathcal{S}_j} \sum_{j=1}^{J} U'(T_j[n-1]) \cdot R_j[n] \tag{3}$$

where $R_j[n]$ is the instantaneous data rate of the j^{th} UE and $U'(T_j[n-1]) = \left.\frac{\partial U}{\partial T_j}\right|_{T_j=T_j[n-1]}$ is the marginal utility of the j^{th} UE with respect to its throughput in the previous TTI.

According to the logical assumptions and mathematical simplifications described in appendix A, we claim that the instantaneous optimization that maximizes (3) leads to a long-term optimization that maximizes (2).

The simplified objective function (3) characterizes a weighted sum rate maximization problem [19], whose weights are adaptively controlled by the marginal utilities. For simplicity of notation, we represent the marginal utility corresponding to the j^{th} NRT UE as the weight

$$w_j^{\text{nrt}} = U'(T_j[n-1]) \tag{4}$$

which will be later used in the proposed RRA framework.

3.3. Real time services

For the case of RT services, the considered optimization problem is the maximization of the total utility with respect to the users' HOL packet delays. We have that the objective function (1a) becomes

$$\max_{\mathcal{S}_j} \sum_{j=1}^{J} U\left(d_j^{\text{hol}}[n]\right) \tag{5}$$

where $U\left(d_j^{\text{hol}}[n]\right)$ is a decreasing utility function based on the current HOL delay $d_j^{\text{hol}}[n]$ of the j^{th} UE. The HOL delay is the time that the oldest packet in the user's buffer has to wait before gaining access to the wireless channel.

It is also possible to derive a new simplified DRA problem that is equivalent to our original DRA problem (5) regarding RT services. According to Appendix B, the objective function of the new simplified DRA problem is also linear in terms of the instantaneous user's data rate and given by

$$\max_{\mathcal{S}_j} \sum_{j=1}^{J} \left|U'\left(d_j^{\text{hol}}[n]\right)\right| \cdot R_j[n] \tag{6}$$

where $U'\left(d_j^{\text{hol}}[n]\right) = \left.\frac{\partial U\left(d_j^{\text{hol}}\right)}{\partial d_j^{\text{hol}}}\right|_{d_j^{\text{hol}}=d_j^{\text{hol}}[n]}$ is the marginal utility of the j^{th} UE with respect to its current HOL delay.

Appendix B describes some mathematical simplifications that allows us to assume that the instantaneous optimization that maximizes (6) leads to a long-term optimization that maximizes (5). The objective function (6) also characterizes a weighted sum rate maximization [19], where the weights are given by the absolute value of the marginal utility with respect to the current HOL delay. For simplicity sake, let us also define an RT user-specific weight w_j^{rt} given by

$$w_j^{\text{rt}} = \left|U'\left(d_j^{\text{hol}}[n]\right)\right| \tag{7}$$

This RT utility-based weight also plays an important role on the RRA framework proposed in the following.

4. UTILITY-BASED RESOURCE ALLOCATION FRAMEWORK FOR OFDMA SYSTEMS

The optimization formulation described in Section 3 is general and can be applied to any modern cellular system. Since Fourth Generation (4G) cellular systems, such as 3GPP long term evolution (LTE) and IEEE Worldwide Interoperability for Microwave Access (WiMAX), are based on OFDMA, we propose on this work a resource allocation framework suitable for this particular air interface.

The weighted sum rate maximization problems given by (3) and (6) have linear objective functions with respect to $R_j[n]$, whose DRA solutions have a closed form and are simpler to obtain [20,21]. The UE with index $m(k,n)$ is chosen to transmit on the k^{th} resource in the n^{th} TTI if it satisfies the condition given by (8):

$$m(k,n) = \arg\max_{j} \left\{w_j \cdot c_{j,k}[n]\right\} \tag{8}$$

where $c_{j,k}[n]$ denotes the instantaneous achievable transmission efficiency of the k^{th} resource with respect to the j^{th} UE (Shannon capacity), and w_j is the utility-based weight factor of the j^{th} UE. One one hand, if the UE uses an NRT service, we have that $w_j = w_j^{\text{nrt}} = U'(T_j[n-1])$, according to (4). On the other hand, for RT services we have $w_j = w_j^{\text{rt}} = \left|U'\left(d_j^{\text{hol}}[n]\right)\right|$, according to (7).

Figure 1 explains how the utility-based DRA algorithm proposed above works. Consider a scenario in which two NRT users i and j compete for 7 resources, where the former user has better channel conditions than the latter in all resources. The channel qualities $\gamma_{i,k}^*$ and $\gamma_{j,k}^*$ plotted in the figure are utility-scaled versions of their original channel qualities $\gamma_{i,k}$ and $\gamma_{j,k}$, respectively, i.e., $\gamma_{i,k}^* = w_i^{\text{nrt}} \cdot \gamma_{i,k}$ and $\gamma_{j,k}^* = w_j^{\text{nrt}} \cdot \gamma_{j,k}$. According to (8), resources $k = 1, \ldots, 3$ are assigned to user i and resources $k = 4, \ldots, 7$ are assigned to user j. Notice that if the utility-based weights w_i^{nrt} and w_j^{nrt} were not used, all resources would have been assigned to user i, who originally had better channel conditions ($\gamma_{i,k} > \gamma_{j,k}$). Thus, the utility-based weights provided a QoS-based resource allocation. The same reasoning is valid for the case of RT services, where the weight w_j^{rt} should be used.

Even though it was not explicitly derived here, the complexity of the proposed DRA algorithm involves the

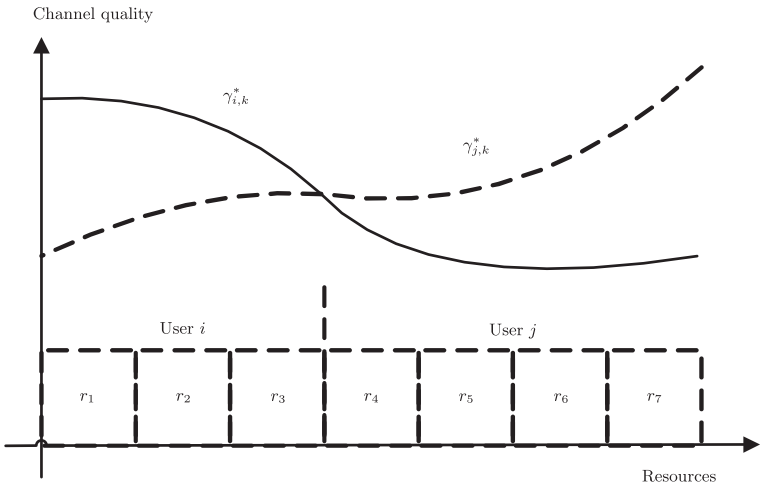


Figure 1. Utility-based dynamic resource assignment.

computation of the following: calculation of the marginal utility per user, calculation of the product between the marginal utility and the transmission efficiency per user and resource, and sorting of the users by this product value on each resource. Notice that all these computations are rather simple operations.

5. MAXIMIZATION OF USER SATISFACTION USING SIGMOIDAL UTILITY FUNCTIONS

The present work claims that it is possible to provide high user satisfaction for NRT or RT users with low complexity if we consider a step-shaped function, such as the sigmoidal function, as the utility function in the optimization problem formulated in Section 3. This utility function should be based on a particular QoS parameter suitable for either NRT or RT services.

In this work, we propose two RRA policies able to maximize the number of satisfied users in the system. They are the throughput-based satisfaction maximization (TSM) policy, whose formulation is based on the users' throughput and is suitable for NRT services, and the delay-based satisfaction maximization (DSM) policy, whose formulation is based on the users' HOL delay and is suitable for RT services.

Since these policies share the same RRA framework described in Section 4, they are similar and have many characteristics in common. These similarities are discussed considering a joint formulation, which is described in Section 5.1. Furthermore, their differences are highlighted in Sections 5.2 and 5.3.

5.1. General formulation

In order to achieve high user satisfaction levels, we propose to use a sigmoidal utility function based on a generic QoS metric x, as indicated below:

$$U(x) = \frac{1}{1 + e^{\mu \cdot \sigma (x - x_{\text{req}})}} \tag{9}$$

Notice that this function is monotonic, continuous and differentiable, which makes it mathematically tractable. The function is parameterized by the use of three parameters: μ is a constant (-1 or 1) that determines if the sigmoidal function is an increasing or decreasing function, respectively; σ is a non-negative parameter that determines the slope of the function; and x_{req} is the QoS requirement of a given service and determines the abscissa shift of the function. The abscissa is the QoS metric x.

The marginal utility given by the utility-based weight plays an important role in the DRA algorithm described in Section 4. The higher the weight, the higher the priority of the UE to get a resource. The utility-based weight based on a generic QoS metric x_j of the j^{th} UE is given by

$$\begin{aligned} w_j &= \frac{\partial U(x_j[n])}{\partial x_j[n]} = \sigma \cdot U(x_j[n]) \cdot (1 - U(x_j[n])) \\ &= \frac{\sigma \cdot e^{\mu \cdot \sigma (x_j[n] - x_{\text{req}})}}{\left(1 + e^{\mu \cdot \sigma (x_j[n] - x_{\text{req}})}\right)^2} \end{aligned} \tag{10}$$

The corresponding DRA algorithm, which is given by (8), must use the particular expression of w_j presented in (10). Note that expression (10) is easy to calculate, which has a positive impact on the computational complexity of the utility-based RRA framework.

The higher the value of σ, the closer to a step-shaped function the utility function will be. Otherwise, considering lower values of σ, the utility function becomes more linear. In order to have a desired step-shaped sigmoidal function no matter the value of the QoS requirement, the fixed σ parameter must be a function of the QoS requirement x_{req}. The sigmoidal function should be equal to a given value δ

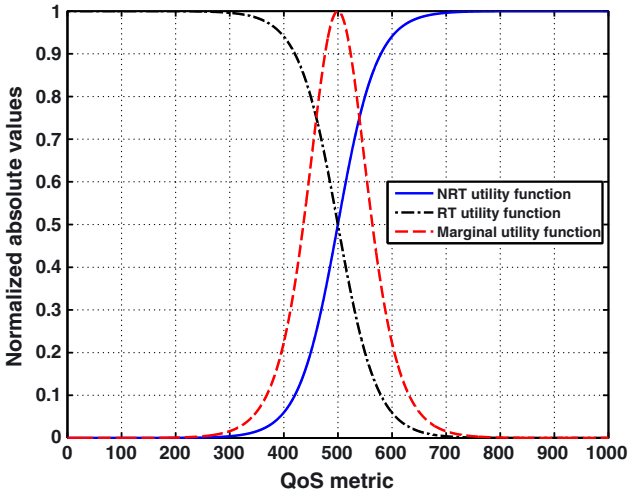


Figure 2. Examples of sigmoidal utility functions and the absolute value of a bell-shaped marginal utility function ($x_{\text{req}} = 500$).

when the QoS metric x achieves a given proportion ρ of the QoS requirement x_{req}. Therefore, we have that

$$\sigma = \frac{\log \frac{1-\delta}{\delta}}{\rho \cdot x_{\text{req}}} \tag{11}$$

For illustration purposes, Figure 2 presents examples of normalized sigmoidal utility functions suitable for NRT and RT services. These utility functions follow the expression given by (9). For the case of NRT services, we assume that the QoS metric is the users' throughput and that $\mu = -1$, which yields an increasing sigmoidal function. On the other hand, $\mu = 1$ and a decreasing sigmoidal function based on the users' HOL delay are related to RT services. Considering the absolute value of the derivatives of the sigmoidal utility functions, both NRT and RT marginal utility functions have a bell shape, as can be seen in Figure 2.

Regarding the NRT utility function in Figure 2, we have $\mu = -1, \delta = 0.01, \rho = 0.5$, and $x_{\text{req}} = 500$ (see (11)). It means that the NRT function starts to increase noticeably, i.e. $U(x) = \delta = 0.01$, when x is half of the QoS requirement, i.e. $x = \rho \cdot x_{\text{req}} = 250$. A similar behavior is observed with the decreasing RT utility function, but in that case we have $\mu = 1$. This means that the RT function starts to decrease noticeably, i.e. $U(x) = 1 - \delta = 0.99$, when x is half of the QoS requirement, i.e. $x = \rho \cdot x_{\text{req}} = 250$.

In the following, we give more details about how the general formulation described above can be configured as each of the proposed utility-based policies proposed in this work, namely TSM and DSM. It will also become clear why the shape of the sigmoidal utility function is very appropriate to characterize the satisfaction of NRT and RT services.

5.2. Throughput-based satisfaction maximization for NRT services

The TSM policy uses an increasing sigmoidal utility function, like the NRT utility function depicted in Figure 2. The function is based on the users' throughput and is centered on a throughput requirement T_{req}. An increasing utility function means that the higher the throughput, the higher the users' utility derived from the network (higher satisfaction). In order to have this increasing sigmoidal function, we use $\mu = -1$ in (9). A step-shaped utility function means that a given user becomes satisfied rapidly if the throughput approaches and exceeds the throughput requirement. The opposite occurs when the user throughput decreases to values lower than the requirement. This behavior is in accordance with the definition of satisfaction for NRT services widely used in the literature.

Figure 2 also depicts the marginal utility as a bell-shaped function. It means that UE experiencing throughput levels close to the requirement will have higher priority in the resource allocation process. Therefore, one can conclude that the UEs in the edge of becoming unsatisfied or satisfied are benefited. Moreover, the TSM technique has the property of avoiding the users to become unsatisfied by giving priority to users with QoS levels just above the requirement.

5.3. Delay-based satisfaction maximization for RT services

The DSM policy considers the users' HOL delay as the QoS metric. Thus, the RT utility function should be decreasing (see Figure 2), which means that the higher the delay, the lower the users' utility derived from the network (lower satisfaction). In that sense, we have $\mu = 1$ in (9). This utility function is also centered on a QoS requirement,

which is called d_{req} and must be equal to or lower than the RT delay budget.

Notice in Figure 2 that the marginal utility is a symmetric function around the QoS requirement (RT delay budget). Notice that depending on the values of the DSM delay requirement (central value of the sigmoidal function) and the RT delay budget, there could be some portion of the marginal utility function (abscissa values higher than the RT delay budget) that will be neglected. This is due to the fact that we assume a packet discard procedure where a HOL packet is discarded at the transmitter if its delay is already higher than the RT delay budget, since this packet would be considered lost at the receiver anyway.

Generally, the definition of satisfaction for RT services used in the literature is based on frame erasure rate (FER) [4,9,22]. If the user's FER is lower than or equal to a requirement, the user is considered satisfied; otherwise it is assumed unsatisfied. Besides the packet losses due to channel errors, we have packet losses due to unbearable delays. As commented before, if a packet from an RT user arrives at the receiver later than the RT delay budget, this packet is considered lost. Taking this into account, we consider that a utility function based on HOL delay is suitable for an RRA policy that intends to provide high levels of user satisfaction, even if this satisfaction is measured based on FER. Therefore, a step-shaped utility function based on HOL delay means that a given user becomes unsatisfied rapidly if the HOL delay approaches and exceeds the delay requirement. The opposite occurs when the user delay decreases to values lower than the requirement.

Looking at Figure 2, which depicts the marginal utility as a bell-shaped function, we can see that the UEs who will have higher priority in the resource allocation process are the ones who experience HOL delays close to the requirement. If this requirement is set to be equal or close to the RT delay budget, one can conclude that the UEs in the edge of becoming unsatisfied are benefited.

6. PERFORMANCE EVALUATION

In this section, we investigate the performance of the TSM and DSM policies proposed in this work. Firstly, the classical algorithms found in the literature that are used as references for comparison are described in Section 6.1. Next, the main simulation models and parameters are presented in Section 6.2. Finally, the performance of the TSM and DSM policies are evaluated by means of system-level simulations in Sections 6.3.1 and 6.3.2, respectively.

6.1. Classical algorithms

Each of the classical algorithms considered in this work uses a different DRA criterion. In order to have a fair comparison with the proposed TSM and DSM policies, all classical algorithms use EPA among the resources. Table I shows which RRA algorithms will be compared in each of the scenarios considered in this study. Moreover, we present the priority function (argument of the DRA expression, see (8)) for each of the algorithms, when applicable.

6.2. Simulation scenario

The simulations took into account the main characteristics of an OFDMA system. The general simulation parameters are depicted in Table II. All simulation results in this work are presented with the 95% bootstrap confidence interval of the mean of the samples.

In order to evaluate the RRA techniques in terms of fairness, we use the well-known Jain's fairness index [23]. The Jain's fairness function is independent of the allocation metric being used. Furthermore, this allocation metric must be directly proportional to the utility derived from the network. In this work, we use the throughput and the inverse of the HOL packet delay of the UEs as the allocation metrics to calculate the respective fairness indexes for NRT and RT services, respectively. In the case of NRT services, if few users have high throughput and the others have low throughput, the allocation metrics of the former will be higher than the latter. This means that the users with high throughput received more resources and so the fairness index is low. In the case of RT services, if few users have low packet delay and the others have high packet delay, the allocation metrics of the former will be higher than the latter (notice that the allocation metric is the inverse of the delay). This means that the users with low packet delay received more resources and so the fairness index is also low. In both cases, if the users' allocation metrics are similar, the fairness index is high.

The definition of satisfaction depends on the type of service that the UE uses. A NRT user is considered satisfied if its session throughput is higher than or equal to a threshold $\left(T_j[n] \geq T_{\text{req}}\right)$. The session duration depends on the time span of each independent simulation snapshot. An RT user is considered satisfied if its FER is lower than or equal to a threshold. In our simulation model, we assume that a frame is lost if a packet arrives at the UE receiver later than the delay budget of the RT service.

6.3. Simulation results

6.3.1. Non-real time services.

In this section, we compare the performance of the TSM technique with other classical RRA techniques found in the literature, namely SORA-NRT [3,4], Rate Maximization (RM) [24] and Proportional Fairness (PF) [25].

Figure 3 depicts the system capacity in terms of total cell throughput. As expected, the RM policy provides the best results. It is able to achieve the maximum allowed system capacity for all system loads by assigning each physical resource block (PRB) to the UEs that can transmit at the highest Coding Scheme (MCS). The PF policy also presents good performance with an almost flat behavior for different traffic loads. Since SORA-NRT and TSM give more importance at achieving high satisfaction levels, they show the worst performance in terms of efficiency in the

Table I. RRA algorithms that are compared in the NRT and RT scenarios.

Algorithms	NRT scenario	RT scenario	Priority function
RM [24]	X	X	$c_{j,k}[n]$
PF [25]	X	X	$\frac{c_{j,k}[n]}{T_j[n-1]}$
SORA-NRT [3,4]	X		Heuristic
TSM	X		$\frac{\sigma \cdot e^{-\sigma\left(T_j[n]-T_{\text{req}}\right)}}{\left(1+e^{-\sigma\left(T_j[n]-T_{\text{req}}\right)}\right)^2} \cdot c_{j,k}[n]$
MLWDF [13]		X	$d_j^{\text{hol}}[n] \cdot \frac{c_{j,k}[n]}{T_j[n-1]}$
UEPS [15]		X	$\frac{\sigma \cdot e^{\sigma\left(d_j^{hol}[n]-d_{\text{req}}\right)}}{\left(1+e^{\sigma\left(d_j^{hol}[n]-d_{\text{req}}\right)}\right)^2} \cdot \frac{c_{j,k}[n]}{T_j[n-1]}$
SORA-RT [4,9]		X	Heuristic
ADS [12]		X	$\left(1+\frac{1}{d_{\text{req}}-d_j^{\text{hol}}[n]}\right) \cdot \frac{c_{j,k}[n]}{T_j[n-1]}$
DSM		X	$\frac{\sigma \cdot e^{\sigma\left(d_j^{hol}[n]-d_{\text{req}}\right)}}{\left(1+e^{\sigma\left(d_j^{hol}[n]-d_{\text{req}}\right)}\right)^2} \cdot c_{j,k}[n]$

Table II. Simulation parameters.

Parameter	Value
Number of cells	1
Maximum BS transmit power	1 W
Cell radius	500 m
UE speed	3 km/h
Carrier frequency	2 GHz
System bandwidth	5 MHz
Number of sub-carriers	512
Sub-carrier bandwidth	15 kHz
Number of PRBs	25
Path loss[a]	$L = 128.1 + 37.6 \log_{10} d$
Log-normal shadowing standard deviation	8 dB
Small-scale fading	3GPP typical urban (TU) [26]
AWGN power per sub-carrier	−123.24 dBm
Noise figure	9 dB
Link adaptation	Link level curves from [27]
SNR threshold of MCS 1 [27]	−6.9 dB
TTI duration	1 ms
NRT traffic model	Full buffer
RT packet size (S_p)	256 bits
RT packet interarrival time ($1/L$)	2 ms
FER threshold	2%
Throughput filtering time constant (f_{thru})	1/1000
Throughput requirement (T_{req})	512 kbps
HOL delay requirement (d_{req})	40 ms
RT delay budget	40 ms
Parameter μ for TSM	−1
Parameter μ for DSM	1
Parameter σ for TSM	$2.441 \cdot 10^{-5}$
Parameter σ for DSM	345.338
Parameter δ	0.01
Parameter ρ	0.5
Simulation time span	30 s
Number of simulation runs	30

[a] d is the distance to the BS in km.

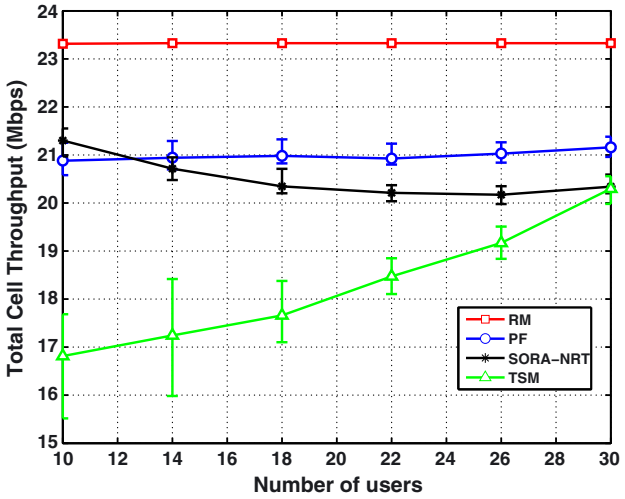


Figure 3. Total cell throughput as a function of the number of NRT users.

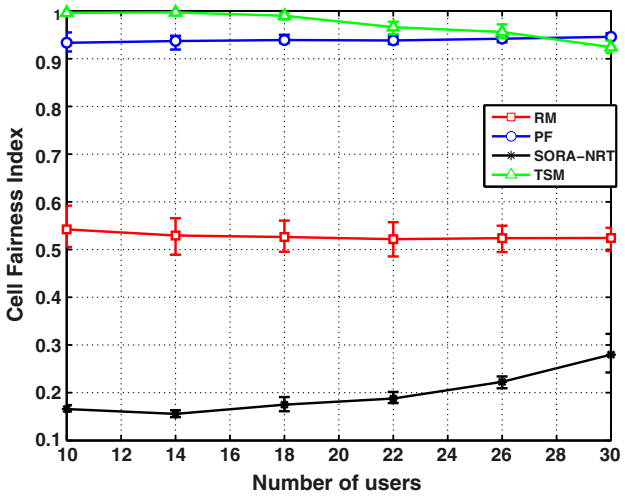


Figure 4. Mean cell fairness index as a function of the number of NRT users.

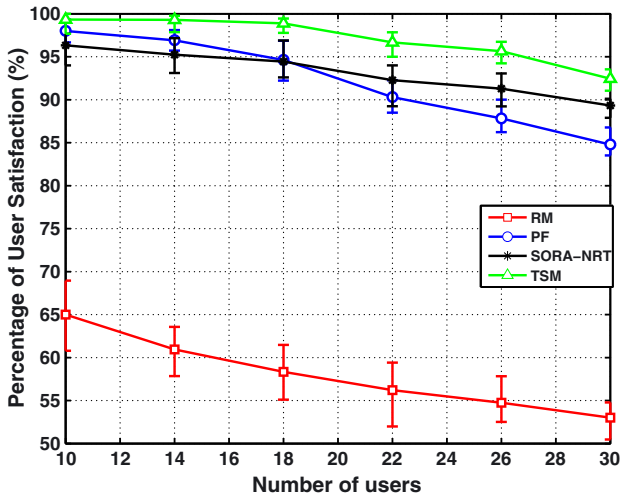


Figure 5. User satisfaction as a function of the number of NRT users.

resource usage. It is interesting to notice that SORA-NRT outperforms TSM for low system loads, when there are sufficient resources for all UEs[†]. When all users are satisfied, SORA-NRT allocates the remaining resources to the users with the best channel conditions, which explains the higher system throughput. As it will be seen later on, TSM is a fairer policy and tries to keep the throughput of the satisfied users as close to each other as possible, which is not so efficient in terms of system capacity. When the traffic load increases, more users become unsatisfied and SORA-NRT does not have a pool of extra resources to improve capacity anymore. In this situation, both SORA-NRT and TSM present similar behavior. Notice that TSM is clearly able to exploit multiuser diversity in order to achieve higher system capacity when the number of users increases.

The throughput-based cell fairness index averaged over all snapshots is shown in Figure 4. Analyzing Figure 3 and Figure 4, one can notice the intrinsic trade-off between system capacity and user fairness. The RM policy is able to use the resources very efficiently but is very unfair in the resource and QoS distribution. TSM is not so good as RM in terms of system capacity but compensates by providing very high fairness among the users. PF turns out to be a very good trade-off with good spectral efficiency and high fairness. Since SORA-NRT did not present so high system capacity, we would expect higher fairness levels, but it is not the case, as can be seen in Figure 4.

Figure 5 depicts the percentage of satisfied users as a function of the system load. Looking also at Figure 3 and Figure 4, one can notice that user satisfaction and fairness are positively correlated while user satisfaction and system capacity are negatively correlated. The latter comparison also express clearly the fundamental trade-off between resource efficiency and user satisfaction.

The TSM technique shows the best satisfaction results for all considered system loads, which demonstrates the advantage of using sigmoidal utility functions when satisfaction maximization is desired. Regarding TSM, the percentage of satisfied users is highly correlated with the Jain's fairness index. This is due to the way that TSM allocates resources and a property of the Jain's fairness index adopted in this study. Let us assume that we have J users in the cell. TSM tends to share the resources equally among Q users that can be satisfied, while the remaining $J - Q$ users do not receive any resource. In this situation, both the satisfaction percentage and the fairness index will be Q/J (see [23] for more details). RM provides an overall degraded QoS because it leaves many users in outage situations. PF and SORA-NRT also present good satisfaction results. It is important to highlight the satisfaction gain provided by the TSM technique compared with SORA-NRT, since the latter was specially designed to provide maximum satisfaction levels.

Another advantage of the TSM technique is the low computational complexity. Its complexity is in the same order of PF and SORA-NRT. The computational complexity of TSM depends only on the calculation of the users' utility based weights (mathematical function of the users' throughputs) with dimension J, and K sorting operations of a scheduling metric with dimension J.

6.3.2. Real time services.

In this section, the performance of the proposed DSM policy is compared with other classical algorithms, such as Modified Largest Weighted Delay First (MLWDF) [13], urgency and efficiency-based packet scheduling (UEPS) [15], SORA-RT [4,9], asymptotic delay scheduler (ADS) [12], PF [25] and RM [24].

Figure 6 depicts the total cell throughput (system capacity) as a function of the number of RT users for a delay budget of 40 ms. In general, the delay-aware policies have better performance in terms of capacity because they are more successful at avoiding packet losses due to unbearable delays. If more packets are successfully transmitted, the system capacity is higher. At first sight, one could expect that the pure opportunistic policy RM would present the highest system capacity. But in the scenario we are evaluating, this is not the case, as it can be seen in Figure 6. The reason for that behavior is because RM chooses few users with best channel quality to transmit, but the buffers of these users do not have so much data to transmit because of the nature of the RT traffic model considered in this work. Therefore, the PRBs, which have a huge transmission capability, will not be efficiently used due to lack of data. That is why the system capacity provided by RM in this scenario is poor.

However, if we combine the opportunistic characteristic of RM with a proper delay-based component, just like the DSM policy does, we have a remarkable improvement in system capacity. The DSM policy, together with MLWDF and UEPS, show the best results. They are followed closely by the ADS scheduler. It is interesting to notice that the SORA-RT algorithm initially shows a good performance, but suddenly starts to lose capacity when the offered load achieves 120 RT users. This is an indication that the SORA-RT algorithm fails at avoiding system congestion when the offered load is high.

The mean cell fairness index based on HOL delay is shown in Figure 7. As expected, the RM algorithm provides the lowest levels of fairness because it leaves many users unattended due to bad channel quality. On the other extreme, we have the proposed DSM algorithm, which is able to provide both the highest system capacity and fairness in a remarkable way. Other delay-aware algorithms, such as UEPS, ADS and MLWDF have also good performance in terms of fairness, where the latter is the worst among them.

The SORA-RT algorithm does not show good fairness results for high offered loads. When the system load increases, the algorithm heuristics is not good enough to avoid users from becoming unsatisfied, and so many users are neglected and fairness decreases. The fairness decrease in this case is associated with an increasing number of

[†]We are considering a system bandwidth of 5 Mhz, which accounts for 25 PRBs.

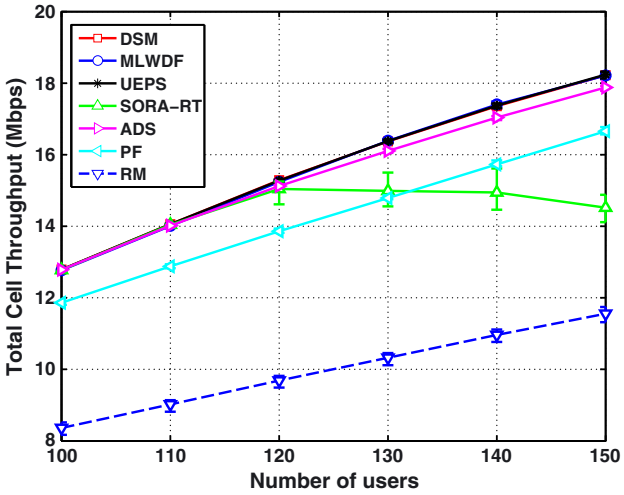


Figure 6. Total cell throughput as a function of the number of RT users.

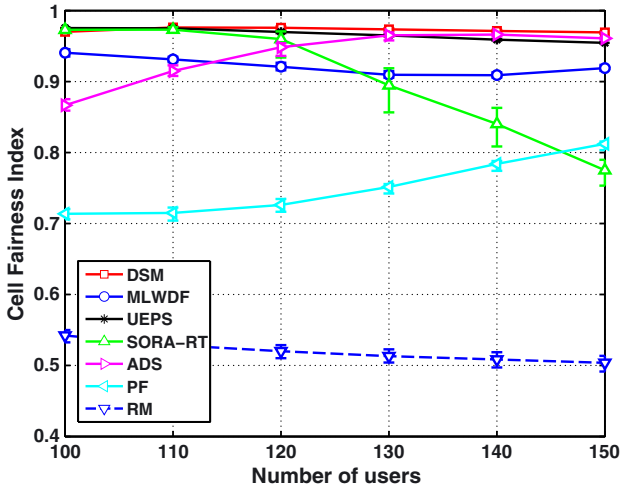


Figure 7. Mean cell fairness index as a function of the number of RT users.

users that do not have chance to transmit. The same behavior can be observed with the RM algorithm. On the other hand, there could be the case of fairness increase when the offered load also increases, like in the case of PF and ADS. Generally this is an indication that the system is becoming congested, i.e., the users' buffers are becoming overloaded and the HOL delays of the users present similar values close to the delay budget. Therefore, the best pattern to expect is an almost constant cell fairness index no matter the system load, which is the case of DSM, UEPS and MLWDF.

Figure 8 depicts the percentage of satisfied RT users. The algorithms that take into account the HOL delay in their formulations are those ones that provide the highest user satisfaction. The resource allocation criteria of these algorithms are based on the combination of two indicators: a QoS indicator that is a function of the HOL delay, and an efficient indicator that can be the achievable transmission efficiency (DSM) or the ratio between the transmission efficiency and the user throughput (MLWDF, UEPS and ADS). Comparing DSM and UEPS, which have the same QoS indicator (bell-shaped marginal utility function), it can be concluded that the achievable transmission efficiency is a better efficiency indicator for the maximization of user satisfaction, since DSM outperforms UEPS. Furthermore, MLWDF, UEPS and ADS, which have the same efficiency indicator, have different delay-based functions as the QoS indicator: linear, bell-shaped and exponential, respectively (see Table I). Comparing these three algorithms in Figure 8, it can be concluded that the linear function is better than the bell-shaped function, which is also better than the exponential function.

Special attention must be given to the proposed DSM policy, which achieved its objective of maximizing user

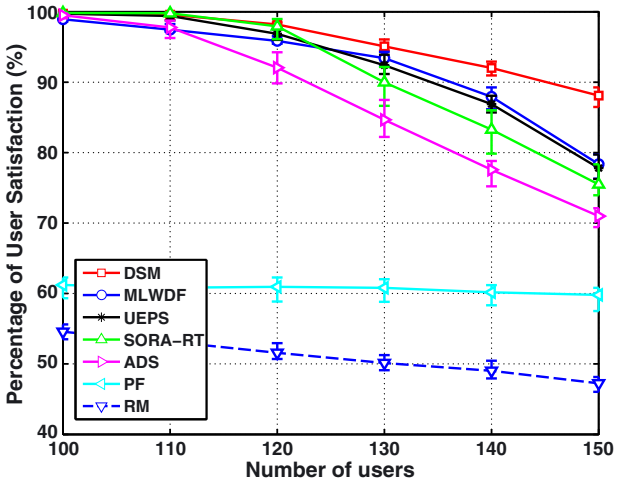


Figure 8. User satisfaction as a function of the number of RT users.

satisfaction. The combination of the bell-shaped delay-based indicator and the achievable transmission efficiency indicator proved to be the best option. Furthermore, its computational complexity is in the same order of the classical algorithms, except RM which presents the lowest complexity.

SORA-RT, which was especially designed to provide high user satisfaction levels, provides reasonable percentage of user satisfaction, but not the sufficient to surpass DSM, MLWDF or UEPS.

7. CONCLUSIONS

In this work we used Utility Theory to propose two novel RRA policies called TSM and DSM, whose main objectives are to maximize satisfaction among NRT or RT users in a cellular network, respectively. These policies share the same RRA framework composed of a utility-based DRA algorithm and an equal power allocation among frequency resources. Both policies use a sigmoidal utility function. On one hand, TSM uses an increasing sigmoidal function based on throughput with inflection point in the users' throughput requirement. On the other hand, DSM uses a decreasing sigmoidal function based on HOL delay with inflection point in the users' HOL delay requirement, which is usually equal to the RT delay budget of the system. It was demonstrated by means of system-level simulations that a step-shaped sigmoidal utility function combined with a channel-aware opportunistic criterion is effective toward the objective of user satisfaction maximization.

When compared to other RRA techniques found in the literature, namely SORA-NRT, PF and RM, the TSM technique showed the highest satisfaction, very high fairness and increasing system capacity with the number of users due to multiuser diversity. Moreover, TSM presents low computational complexity in the same order of PF and SORA-NRT.

When the DSM policy is compared with classical algorithms, such as MLWDF, UEPS, ADS, PF, RM and SORA-RT, it was observed that the former presented simultaneously the highest satisfaction, fairness and system capacity. Although RM presents the lowest computation complexity among all policies, the complexity of the DSM policy is also low and approximately the same of the other algorithms.

Notice that NRT and RT traffic are optimized separately with different objective functions in our proposed utility-based RRA framework. In order to extend the applicability of the framework in more realistic scenarios, we intend to address the user maximization problem in a mixed traffic scenario, where both NRT and RT traffic would be optimized at the same time. Other perspectives for future work are the investigation of the dynamic adaptation of the slope of the sigmoidal function according to the users' states and/or network operator's interests (capacity improvement, QoS guarantees, etc), and the evaluation of how adaptive power allocation algorithms and multiple antenna schemes can help the proposed techniques to achieve even higher user satisfaction.

APPENDIX A: UTILITY-BASED OPTIMIZATION FORMULATION FOR NRT SERVICES

As explained in Section 3.2, the considered optimization problem for NRT services is the maximization of the total utility with respect to the throughput. Thus, the objective function is

$$\max_{\mathcal{S}_j} \sum_{j=1}^{J} U(T_j[n]) \tag{A.1}$$

The throughput of the j^{th} UE is calculated using an exponential smoothing filtering, as indicated below:

$$T_j[n] = (1 - f_{\text{thru}}) \cdot T_j[n-1] + f_{\text{thru}} \cdot R_j[n] \tag{A.2}$$

where $R_j[n]$ is the instantaneous data rate of the j^{th} UE and f_{thru} is a filtering constant.

Evaluating the objective function in (A.1) and the throughput expression in (A.2), the derivative of $U(T_j)$ with respect to the transmission rate R_j is given by:

$$\frac{\partial U}{\partial R_j} = \frac{\partial U}{\partial T_j} \cdot \frac{\partial T_j}{\partial R_j} = f_{\text{thru}} \cdot \left.\frac{\partial U}{\partial T_j}\right|_{T_j = (1-f_{\text{thru}}) \cdot T_j[n-1] + f_{\text{thru}} \cdot R_j[n]}$$

In the case that f_{thru} is sufficiently small, the expression above can be simplified as follows [20]:

$$\frac{\partial U(T_j[n])}{\partial R_j[n]} \approx f_{\text{thru}} \cdot \left.\frac{\partial U}{\partial T_j}\right|_{T_j = T_j[n-1]} \tag{A.3}$$

where the previous resource allocation totally determines the current values of the marginal utilities. Using the one-order Taylor formula for the utility function [20,28] and considering (A.3), we have

$$\begin{aligned} \sum_{j=1}^{J} U(T_j[n]) \approx & \sum_{j=1}^{J} U(T_j[n-1]) \\ & + \sum_{j=1}^{J} \left.\frac{\partial U}{\partial T_j}\right|_{T_j = T_j[n-1]} \\ & \cdot \left(f_{\text{thru}} \cdot R_j[n] - f_{\text{thru}} \cdot T_j[n-1]\right) \end{aligned} \tag{A.4}$$

Let us consider the maximization of (A.4). Notice that the maximization of the left side of expression (A.4) is our original DRA problem given by (A.1). The maximization of the right side of expression (A.4) is the new simplified DRA problem. Since f_{thru} is a constant and $T_j[n-1]$ is known and fixed before the resource allocation at the current TTI n, the new simplified DRA problem becomes linear in terms of the instantaneous user's data rate, and is given by

$$\max_{\mathcal{S}_j} \sum_{j=1}^{J} U'(T_j[n-1]) \cdot R_j[n] \tag{A.5}$$

Notice that we started with an optimization formulation based on throughput given by (A.1), made some logical assumptions and mathematical simplifications, and ended up with a linear optimization formulation based on instantaneous rates given by (A.5). According to these arguments, we claim that the instantaneous optimization maximizing (A.5) leads to a long-term optimization that maximizes (A.1).

APPENDIX B: UTILITY-BASED OPTIMIZATION FORMULATION FOR RT SERVICES

According to Section 3.3, the considered optimization problem for RT services is the maximization of the total utility with respect to the users' HOL packet delays. The objective function is given by

$$\max_{\mathcal{S}_j} \sum_{j=1}^{J} U\left(d_j^{\text{hol}}[n]\right) \tag{B.1}$$

In order to understand the model used in this work for the calculation of the HOL delays, Figure B.1 is presented. This figure illustrates a packet queue for a given RT user. As it can be seen in the figure, the traffic model for RT services used in this work assumes a packet arrival rate of L packets per second, i.e., a new packet of S_{p} bits (fixed size) arrives in the user buffer every $1/L$ seconds.

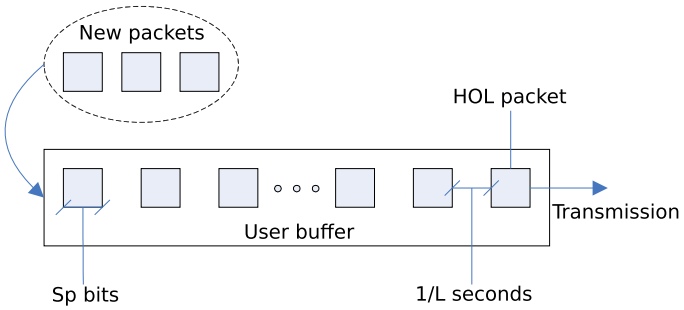
Taking into account Figure B.1 and considering a generic UE j, we propose in this work a recursive model for calculating an approximate value of the HOL delay. The recursive equation is

$$d_j^{\text{hol}}[n+1] = d_j^{\text{hol}}[n] + t_{\text{tti}} - \frac{1}{L} \cdot \left(\frac{R_j[n] \cdot t_{\text{tti}}}{S_{\text{p}}}\right) \tag{B.2}$$

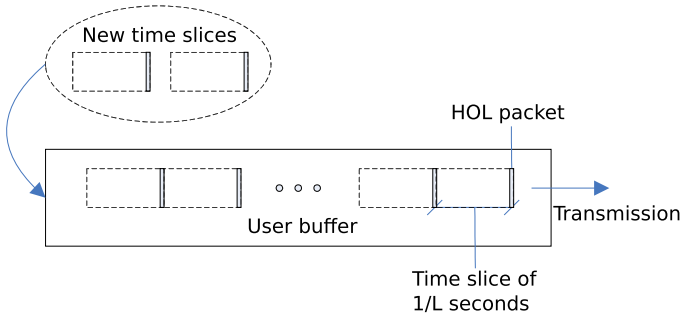
where t_{tti} is the duration of the TTI in seconds, L is the packet arrival rate, S_{p} is the packet size, and $R_j[n]$ is the instantaneous achievable transmission rate on TTI n. In this queue model, we assume that the packet size S_{p} is sufficiently small, so that the queue can be represented ideally by a sequence of time slices with duration $1/L$ seconds each (see Figure B.1). Notice that this assumption does not invalidate the mathematical and conceptual RRA framework, and makes the optimization model much more tractable.

Looking at (B.2), firstly it can be seen that the HOL delay is always incremented by at least the duration of one TTI, no matter how many bits were transmitted in the current transmission interval. This represents the passing of time in the system, which means that all packets in the queue will be one TTI older. Secondly, the decrement of the HOL delay depends on the number of time slices (duration of $1/L$ seconds each) that is decremented due to the transmission in the current TTI. If the j^{th} UE has not been served by any resource in the n^{th} TTI, $R_j[n]$ is equal to zero and no time slices are decremented. If the instantaneous transmission rate is such that the HOL packet is totally transmitted in the current TTI, it means that one time slice with duration of $1/L$ seconds should be decremented in the HOL delay. If the instantaneous transmission rate is sufficiently high so that many packets in the queue can be transmitted, the corresponding number of time slices should be decremented in the HOL delay.

Assessing the objective function in (B.1) and the HOL delay expression in (B.2), we can see that the derivative



(a) User buffer as a sequence of packets



(b) User buffer as a sequence of time slices

Figure B.1. Modeling of an RT user buffer.

of $U\left(d_j^{\text{hol}}\right)$ with respect to the transmission rate R_j can be expressed as

$$\frac{\partial U}{\partial R_j} = \frac{\partial U}{\partial d_j^{\text{hol}}} \cdot \frac{\partial d_j^{\text{hol}}}{\partial R_j} = \frac{\partial U}{\partial d_j^{\text{hol}}} \cdot \left(-\frac{t_{\text{tti}}}{L \cdot S_{\text{p}}}\right)$$

Using the result above and assuming that the TTI duration is sufficiently small, the Lagrange theorem of the mean can be used [16,28], which says that

$$\begin{aligned}\sum_{j=1}^{J} U\left(d_j^{\text{hol}}[n+1]\right) &\approx \sum_{j=1}^{J} U\left(d_j^{\text{hol}}[n]\right) \\ &\quad + \sum_{j=1}^{J} \left.\frac{\partial U}{\partial R_j}\right|_{R_j = R_j[n-1]} \\ &\quad \cdot \left(R_j[n] - R_j[n-1]\right) \\ &= \sum_{j=1}^{J} - \left.\frac{\partial U}{\partial d_j^{\text{hol}}}\right|_{d_j^{\text{hol}} = d_j^{\text{hol}}[n]} \cdot \frac{t_{\text{tti}}}{L \cdot S_{\text{p}}} \\ &\quad \cdot \left(R_j[n] - R_j[n-1]\right) \\ &= \sum_{j=1}^{J} \left.\left|\frac{\partial U}{\partial d_j^{\text{hol}}}\right|\right|_{d_j^{\text{hol}} = d_j^{\text{hol}}[n]} \cdot \frac{t_{\text{tti}}}{L \cdot S_{\text{p}}} \\ &\quad \cdot \left(R_j[n] - R_j[n-1]\right)\end{aligned} \tag{B.3}$$

The absolute value operator is used in (B.3) because the utility function is assumed to be decreasing, which yields negative marginal utilities and cancels the negative sign in (B.3).

On one hand, the maximization of the left side of expression (B.3) is our original DRA problem given by (B.1). On the other hand, the maximization of the right side of expression (B.3) is the new simplified DRA problem. We have that t_{tti}, L and S_{p} are constants, and that $d_j^{\text{hol}}[n]$ and $R_j[n-1]$ are known and fixed before the resource allocation at the n^{th} TTI. Therefore, the new simplified DRA problem becomes linear in terms of the instantaneous user's data rate, and is given by

$$\max_{\mathcal{S}_j} \sum_{j=1}^{J} \left|U'\left(d_j^{\text{hol}}[n]\right)\right| \cdot R_j[n] \tag{B.4}$$

Taking into account (B.3), we are able to assume that the instantaneous optimization maximizing (B.4) leads to a long-term optimization that maximizes (B.1).

ACKNOWLEDGEMENTS

This work was supported (in part) by Ericsson Research, Wireless Access Network Department, Luleå, Sweden, and by the Ericsson Innovation Center, Indaiatuba, Brazil. Emanuel B. Rodrigues would like to thank CAPES/CNPq/FUNCAP/PNPD for financial support.

REFERENCES

1. Tarchi D, Fantacci R, Bonciani E. Analysis and comparison of scheduling techniques for a BWA OFDMA mobile system. *Wiley Wireless Communications and Mobile Computing* July 2010; **10**(7): 888–898.
2. Pitic R, Serrelli F, Redana S, Capone A. Performance evaluation of utility-based scheduling schemes with QoS guarantees in IEEE 802.16/WiMAX systems. *Wiley Wireless Communications and Mobile Computing* July 2010; **10**(7): 912–931.
3. Santos RB, Lima FRM, Freitas W, Cavalcanti FRP. QoS Based Radio Resource Allocation and Scheduling with Different User Data Rate Requirements for OFDMA Systems. In *IEEE 18th International Symposium on Personal, Indoor and Mobile Radio Communications, 2007*, Athens, Greece, 2007; 1–5.
4. Lima FRM, Wänstedt S, Cavalcanti FRP, Freitas WC. Scheduling for Improving System Capacity in Multiservice 3GPP LTE. *Journal of Electrical and Computer Engineering* 2010; **2010**(819729): 1–16.
5. Rodrigues EB, Casadevall F. Adaptive Radio Resource Allocation Framework for Multi-User OFDM. In *Proceedings of IEEE 69th Vehicular Technology Conference - VTC Spring*, Barcelona, Spain, 2009; 1–6.
6. Kulkarni S, Rosenberg C. Opportunistic scheduling: generalizations to include multiple constraints, multiple interfaces, and short term fairness. *Wireless Networks* May 2005; **11**(5): 557–569.
7. Bai B, Chen W, Cao Z, Letaief KB. Max-matching diversity in OFDMA systems. *IEEE Transactions on Communications* April 2010; **58**(4): 1161–1171.
8. Chung YH, Chang CJ. A balanced resource scheduling scheme with adaptive priority thresholds for OFDMA downlink systems. *IEEE Transactions on Vehicular Technology* March 2012; **61**(3): 1276–1286.
9. Lima FRM, Freitas WC, Cavalcanti FRP. Scheduling Algorithm for Improved System Capacity of Real-Time services in 3GPP LTE. In *XXVII Telecommunications Brazilian Symposium - SBrT*, Blumenau, Brazil, October 2009; 1–6.
10. Ho QD, Ashour M, Le-Ngoc T. Opportunistic delay-margin-based resource allocation for next-generation wireless networks. *Wiley Wireless Communications and Mobile Computing* September 2011; **11**(9): 1254–1265.
11. Gueguen C, Baey S. Scheduling in OFDM Wireless Networks without Tradeoff between Fairness and Throughput. In *Proceedings of IEEE 68th Vehicular Technology Conference - VTC-Fall*, Calgary, Canada, September 2008; 1–5, DOI: 10.1109/VETECF.2008.397.
12. Braga AR, Rodrigues EB, Cavalcanti FRP. Packet Scheduling for VoIP over HSDPA in Mixed Traffic Scenarios. In *Proceedings IEEE 17th International Symposium on Personal, Indoor and Mobile Radio Communications - PIMRC*, Helsinki, Finland, September 2006; 1–5, DOI: 10.1109/PIMRC.2006.254119.
13. Andrews M, Kumaran K, Ramanan K, Stolyar A, Whiting P, Vijayakumar R. Providing Quality of Service over a Shared Wireless Link. *IEEE Communications Magazine* June 2001; **32**(2): 150–154.
14. Shakkottai S, Stolyar AL. Scheduling Algorithms for a Mixture of Real-Time and Non-Real-Time Data in HDR. In *Proceedings of 17th International Teletraffic Congress (ITC)*, Salvador, Brazil, 2001; 793–804.
15. Ryu S, Ryu B, Seo H, Shin M. Urgency and Efficiency based Wireless Downlink Packet Scheduling Algorithm in OFDMA System. In *Proceedings of IEEE 61st Vehicular Technology Conference - VTC Spring,* vol. 3, Stockholm, Sweden, 2005; 1456–1462.
16. Lei H, Zhang L, Zhang X, Yang D. A Packet Scheduling Algorithm Using Utility Function for Mixed Services in the Downlink of OFDMA Systems. In *Proceedings of IEEE 66th Vehicular Technology Conference - VTC Fall*, Baltimore, USA, 2007; 1664–1668.
17. Song G, Li YG. Utility-Based Resource Allocation and Scheduling in OFDM-Based Wireless Broadband Networks. *IEEE Communications Magazine* December 2005; **43**(12): 127–134.
18. Gross J, Bohge M. Dynamic Mechanisms in OFDM Wireless Systems: A Survey on Mathematical and System Engineering Contributions. *Technical Report TKN-06-001*, Telecommunication Networks Group (TKN), Technical University Berlin, 2006.
19. Hoo LMC, Halder B, Tellado J, Cioffi JM. Multiuser Transmit Optimisation for Multicarrier Broadcast Channels: Asymptotic FDMA Capacity Region and Algorithms. *IEEE Transactions on Communications* June 2004; **52**(6): 922–930.
20. Song G, Li YG. Cross-Layer Optimization for OFDM Wireless Networks - Part II: Algorithm development. *IEEE Transactions on Wireless Communications* March 2005; **4**(2): 625–634.
21. Hosein PA. QoS Control for WCDMA High Speed Packet Data. In *Proceedings 4th International Workshop on Mobile and Wireless Communications Network*, Stockholm, Sweden, 2002; 169–173, DOI: 10.1109/MWCN.2002.1045716.

22. Rodrigues EB, Cavalcanti FRP, Wänstedt S. QoS-Driven Adaptive Congestion Control for Voice over IP in Multiservice Wireless Cellular Networks. *IEEE Communications Magazine* 2008; **46**(1): 100–107.
23. Jain R, Chiu D, Hawe W. A Quantitative Measure of Fairness and Discrimination for Resource Allocation in Shared Computer Systems. *Technical Report TR-301*, DEC Research, September 1984.
24. Jang J, Lee KB. Transmit Power Adaptation for Multiuser OFDM Systems. *IEEE Journal on Selected Areas in Communications* January 2003; **21**(2): 171–178.
25. Kelly F. Charging and Rate Control for Elastic Traffic. *European Transactions on Telecommunications* 1997; **8**: 33–37.
26. 3GPP. Deployment Aspects. *Technical Report TR 25.943 V9.0.0*, Third Generation Partnership Project, December 2009.
27. Mehlführer C, Wrulich M, Ikuno JC, Bosanska D, Rupp M. Simulating the Long Term Evolution Physical Layer. In *Proceedings of the 17th European Signal Processing Conference (EUSIPCO 2009)*, Glasgow, Scotland, 2009; 1471–1478.
28. Rodrigues EB. Adaptive Radio Resource Management for OFDMA-Based Macro- and Femtocell Networks. *Ph.D. Thesis*, Universitat Politècnica de Catalunya, Barcelona, Spain, 2011.

AUTHORS' BIOGRAPHIES

Emanuel B. Rodrigues received his BSc and MSc degrees in Electrical Engineering from the Federal University of Ceará (UFC), Fortaleza, Brazil, in 2001 and 2004, respectively. He also received his PhD degree with honors in Signal Theory and Communications from the Universitat Politècnica de Catalunya (UPC/BarcelonaTech), Barcelona, Spain, in 2011. In 2014, he joined UFC, where he is an associate professor in the Computer Science Department. He has been working in the Wireless Telecom Research Group (GTEL) since 2001 and has actively participated in several projects in a technical and scientific cooperation between GTEL and Ericsson Research. Within this cooperation, he has been in an internship at Ericsson Research at Linköping, Sweden, in 2004, where he studied admission control algorithms for HSDPA systems. During the last 12 years, he has published several conference papers, journal/magazine articles, and book chapters and has been a reviewer of important international conferences and IEEE journals and magazines. His main research interests are radio resource management and QoS control for macrocell and femtocell networks.

Francisco R. M. Lima received his BS degree with honors in Electrical Engineering in 2005, and MSc and DSc degrees in Telecommunications Engineering from the Federal University of Ceará, Fortaleza, Brazil, in 2008 and 2012, respectively. In 2008, he has been in an internship at Ericsson Research in Luleä, Sweden, where he studied scheduling algorithms for LTE system. Since 2010, he has been a professor of Computer Engineering Department of Federal University of Ceará, Sobral, Brazil. Prof. Lima is also a researcher at the Wireless Telecom Research Group (GTEL), Fortaleza, Brazil where he works in projects in cooperation with Ericsson Research. He has published several conference and journal articles as well as patents in the wireless telecommunications field. His research interests include radio resource allocation algorithms for QoS guarantees in scenarios with multiple services, resources, antennas, and users.

Tarcisio F. Maciel received his BSc and MSc degrees in Electrical Engineering from the Federal University of Ceará (UFC), Fortaleza, Brazil, in 2002 and 2004, respectively, and the Dr-Ing degree in Electrical Engineering from the Technische Universität Darmstadt (TUD), Darmstadt, Germany, in 2008. From 2001 to 2004, he was a researcher with the Wireless Telecom Research Group (GTEL) of the UFC. From 2005 to 2008, he was a Research Assistant with the Communications Engineering Laboratory of the TUD. In 2009, he was a professor of the computer engineering course with UFC. Since 2008, he has been a researcher with GTEL and a member of the Post-Graduation Program in Teleinformatics Enginnering (PPGETI) of the UFC. Since 2010, he has been a professor with the Center of Technology, UFC. His research interests include radio resource management, numerical optimization, and multiuser/multiantenna communications.

Francisco R. P. Cavalcanti received his BSc and MSc degrees in Electrical Engineering from Federal University of Ceará, Fortaleza, Brazil, in 1994 and 1996, respectively, and DSc degree in Electrical Engineering from the State University of Campinas, São Paulo, Brazil, in 1999. Upon graduation, he joined the Federal University of Ceará, where he is currently an Associate Professor and holds the Wireless Communications Chair with the Department of Teleinformatics Engineering. In 2000, he founded and, since then has directed, the Wireless Telecom Research Group (GTEL), which is a research laboratory

based on Fortaleza, which focuses on the advancement of wireless telecommunications technologies. At GTEL, he manages a program of research projects in wireless communications sponsored by the Ericsson Innovation Center in Brazil and Ericsson Research in Sweden. Prof. Cavalcanti has produced a varied body of work including one book, conference and journal papers, international patents, and computer software dealing with subjects such as radio resource allocation, cross-layer algorithms, service quality provisioning, transceiver architectures, signal processing, and project management. Prof. Cavalcanti is a distinguished researcher of the Brazilian Scientific and Technological Development Council for his technology development and innovation record. He also holds a Leadership and Management professional certificate from the Massachusetts Institute of Technology, Cambridge.