

Chapter 2

Resource Allocation for Improved User Satisfaction with Applications to LTE

Francisco R.M. Lima, Emanuel B. Rodrigues, Tarcisio F. Maciel
and Mats Nordberg

2.1 Introduction

Cellular networks have experienced an incredible development in the past decades since the commercial launch of 2nd Generation (2G)'s Global System for Mobile Communications (GSM) in the beginning of 1990s to the specification of the 4th Generation (4G)'s Long-Term Evolution-Advanced (LTE-A) by 3rd Generation Partnership Project (3GPP). Several technology improvements have been introduced that enable higher data rates in both downlink and uplink, low packet latencies, and support to new multimedia services. Even with these technological improvements, the cellular networks face nowadays an important challenge that is the steep increase of mobile traffic expected for the next years. According to Ericsson [4], global mobile data traffic will increase 12-fold and the number of mobile subscriptions will be 9.3 billion by the end of 2018. Even with this increased data traffic, cellular operators should be still able to guarantee the user Quality of Service (QoS) for all provided services.

Cellular operators should guarantee the satisfactory provision of the services in order to maintain a high number of subscribers, decrease churn, and attract new subscribers. From the user's point of view, acceptable (QoS) is one of the most

F. R. M. Lima (✉) · E. B. Rodrigues · T. F. Maciel
Wireless Telecommunications Research Group (GTel), Federal University of Ceará,
Caixa Postal 6005, Fortaleza 60440-900, Brazil
e-mail: rafaelm@gtel.ufc.br

E. B. Rodrigues
e-mail: emmanuel@gtel.ufc.br

T. F. Maciel
e-mail: maciel@gtel.ufc.br

M. Nordberg
Ericsson Research, Luleå, Sweden
e-mail: mats.nordberg@ericsson.com

important aspects for guaranteeing user satisfaction and loyalty to the cellular operator. According to this, a reasonable objective to be pursued in cellular networks is the maximization of the number of satisfied users.

User satisfaction is a broad concept that depends on many aspects, e.g., technical parameters such as service type, throughput, and delay, as well as economic issues such as subscription fees. When the service type is concerned, we can identify two groups based on delivery requirements: Real Time (RT) and Non-Real Time (NRT) services. RT services usually relate to applications whose data packets should be delivered with a short and bounded delay in order to be useful to the receivers. As examples of this class of service we can cite online games and Voice over IP (VoIP). On the other hand, packet delay requirements for NRT services are not so strict as for RT services. The important aspect to be guaranteed in this class of service is the information integrity and the performance is usually measured in terms of average data rate (throughput). As examples of applications associated with this class of service we can mention web browsing and File Transfer Protocol (FTP).

In order to face the challenges of cellular operators and increase the number of satisfied users for different service types, Radio Resource Allocation (RRA) is of utmost importance. RRA is responsible for managing and distributing the available scarce resources of the radio interface to the active connections. Among the resources to be dealt by RRA we have frequency resources, transmit power and time slots, among others. In order to design RRA strategies, different directions can be followed. In this chapter, we present two approaches for designing RRA solutions, namely: heuristic and utility-based approaches. As it will be presented in the following, while the heuristic design provides simple and quick solutions to the RRA problems, the utility-based approach is a flexible and general tool for RRA design. In this chapter, we consider that RRA is applied in the context of multiple antennas at the transmitter and/or receiver that is an effective technology to obtain multiplexing and diversity gains.

The remainder of this chapter is organized as follows. First, some related works are discussed in Sect. 2.2. Next, we present in Sects. 2.3 and 2.4 the general heuristic and utility-based RRA frameworks, respectively, that are suitable for increasing the number of satisfied users for RT and NRT services. In Sect. 2.5, we present the performance evaluation by means of computer simulations of algorithms following the considered RRA frameworks. Finally, in Sect. 2.6 we summarize the main conclusions achieved in this chapter.

2.2 Background and Related Work

In this section, we provide a short review of RRA and scheduling solutions for NRT and RT services in Sects. 2.2.1 and 2.2.2.

2.2.1 Satisfaction Maximization for Non-Real Time Services

Studies in scheduling strategies for NRT services for Orthogonal Frequency-Division Multiple Access (OFDMA) systems have begun with the generalization of basic time-domain packet schedulers [7] such as maximum rate, fair throughput, and proportional fair schedulers to the frequency domain [13]. The main performance objectives studied for NRT services were spectral efficiency maximization and improved fairness. The works [13, 20, 23] show that dividing the packet scheduler into a time domain and a frequency domain component and utilizing different algorithms in both domains, the throughput fairness between users can be effectively controlled. However, in modern networks the fairness performance criterion is not able to capture whether the minimum QoS requirements expected by the connected users is fulfilled or not. We propose that user satisfaction, which consists in the ratio between the number of users that have the minimum QoS fulfilled and the total number of users, is a more suitable performance objective in this context.

A literature review was performed and we found that the specific topic of satisfaction maximization for NRT users using resource allocation techniques was the object of study of Ref. [26]. This work proposes an adaptive RRA framework that 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. However, this framework is not easily tunable or scalable, because not only it needs to know the load regions where each RRA policy achieves the highest satisfaction, but also the satisfaction thresholds that trigger the shift among the policies.

2.2.2 Satisfaction Maximization for 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 into two main approaches: opportunistic (PS) [2, 3, 6, 31, 36], and utility theory [15, 29, 33].

The opportunistic PS 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) [3, 6, 36] or the ratio between the instantaneous transmission rate and throughput (proportional fairness policy) [2, 3, 31], and a delay-based QoS indicator. 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 PS 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, Refs. [15] and [29] used z -shaped

utility functions while [33] 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 classical PS priority functions.

To the best of our knowledge, the techniques proposed in the present work are the first ones to deal explicitly with the problem of user satisfaction maximization in RT service scenarios.

2.3 Heuristic Radio Resource Allocation

Heuristic solutions comprise methods to find satisfactory answers to the studied problems based on experience and common sense. These solutions are especially suitable for the cases where the best possible solution is hard or impossible to obtain. In these cases, heuristic methods accelerate the problem-solving process and provide us accessible and simple solutions, which usually are more suited to real-life implementation in the systems.

In this section, we present a heuristic RRA framework to the problem of maximizing user satisfaction. The basic ideas of the RRA framework are to estimate the required number of resources that each user demands to be satisfied and sort the connected users according to a specific priority based on the current satisfaction status, QoS requirements, and traffic state. Then, the users with high priorities get resources in an opportunistic way. This heuristic framework is able to perform resource allocation for either NRT or RT services. In Sect. 2.3.1, the problem to be solved is formulated mathematically. The heuristic RRA framework is particularized for OFDMA systems in Sect. 2.3.2. Finally, in Sects. 2.3.3 and 2.3.4 we present two algorithms based on the proposed heuristic framework that are able to improve user satisfaction. The contributions presented in this section were first shown in the seminal works [16–18, 30].

2.3.1 Problem Formulation

As commented before, the problem to be addressed here is the user satisfaction maximization for NRT and RT services. One of the most important aspects for NRT services is the information integrity, i.e., information loss is not tolerable. Furthermore, this class of services does not impose strict delay requirements although too high packet delays are unacceptable. According to this, a meaningful performance metric for NRT services is the average data rate given by

$$\bar{R}_j[n] = \frac{\varphi_j[n]}{(t_j[n] \cdot t_{\text{tti}})}, \quad (2.1)$$

where $t_j[n]$ is the total active time of user j at Transmission Time Interval (TTI) n since the session beginning; $\varphi_j[n]$ is the number of correctly transmitted bits from user j at TTI n since the session beginning and t^{tti} is the time duration of one TTI. We assume here that an NRT service is satisfied when the average data rate at the end of the data session is higher than or equal to the average data rate requirement, \bar{R}_j^{req} .

RT services are characterized by the short time response between the communicating parts which leads to strict requirements regarding packet delay and jitter. In order to measure the performance of RT services we consider the Frame Erasure Rate (FER) that is directly related to packet delay and loss and is given by

$$\text{FER}_j[n] = \frac{\eta_j^{\text{lost}}[n]}{\eta_j^{\text{lost}}[n] + \eta_j^{\text{succ}}[n]}, \quad (2.2)$$

where $\eta_j^{\text{succ}}[n]$ and $\eta_j^{\text{lost}}[n]$ are the number of successfully transmitted and lost packets (frames) of user j at TTI n since the session beginning, respectively. An RT user is satisfied if the FER is lower than or equal to the required FER denoted by $\text{FER}_j^{\text{req}}$.

According to these definitions, the problem of user satisfaction maximization in NRT or RT traffic scenario can be presented mathematically in the following optimization form

$$\begin{aligned} & \max_{\mathcal{K}_j \forall j \in \mathcal{J}} \sum_{j=1}^J u(\bar{R}_j - \bar{R}_j^{\text{req}}), \text{ in NRT scenario} \\ & \text{or} \end{aligned} \quad (2.3a)$$

$$\max_{\mathcal{K}_j \forall j \in \mathcal{J}} \sum_{j=1}^J u(\text{FER}_j^{\text{req}} - \text{FER}_j), \text{ in RT scenario}$$

$$\text{subject to } \bigcup_{j=1}^J \mathcal{K}_j \subseteq \mathcal{K}, \quad (2.3b)$$

$$\mathcal{K}_i \cap \mathcal{K}_j = \emptyset, \quad i \neq j, \quad \forall i, j \in \{1, 2, \dots, J\}, \quad (2.3c)$$

where J is the total number of users in a cell, \mathcal{J} is the set of all users, \mathcal{K} is the set of all resources in the system, \mathcal{K}_j is the subset of resources assigned to the user j , and $u(\cdot)$ is a step function that assumes the value 1 when its argument is greater than or equal to zero, and assumes the value 0 otherwise. Note that K is defined as the number of available radio resources¹ in the system.

The optimization problem (2.3) is composed of an objective function presented in (2.3a) and constraint functions presented in (2.3b) and (2.3c). The objective of the optimization problem (2.3) is to maximize the number of satisfied users. Note

¹ In the context of OFDMA systems, a radio resource is represented by a subcarrier or a set of them in the frequency domain and a sequence of Orthogonal Frequency-Division Multiplexing (OFDM) symbols in the time domain.

that the objective can be written into two forms depending on the service scenario. When we have that all users are from an NRT service the objective is to maximize the number of users with average data rate not lower than the required average data rate, i.e., $\bar{R}_j \geq \bar{R}_j^{\text{req}}$. On the other hand, when all users are from an RT service, the objective is to maximize the number of users with FER not higher than the required FER, i.e., $\text{FER}_j \leq \text{FER}_j^{\text{req}}$. Constraints (2.3b) and (2.3c) state that the union of all subsets of resources assigned to different users must be contained in the 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 at the same TTI.

Problem (2.3) belongs to the class of combinatorial problems that in general are hard to solve optimally. Furthermore, the relationship between the resource assignment variable and the average data rate ($\bar{R}_j[n]$) or the FER ($\text{FER}_j[n]$) is given by a nonlinear function which further increases the complexity of the problem. Motivated by the computational complexity of problem (2.3) we present in Sect. 2.3.2 an alternative solution to the problem.

2.3.2 Heuristic Resource Allocation Framework for OFDMA Systems

In order to provide a possibly suboptimal but simple solution to problem (2.3) we followed a heuristic framework based on the “divide to conquer idea”. Basically, in order to define which resource should be assigned to which user, we tackle two intermediate problems:

1. Which users will get resources?
2. Which resources will be assigned to the selected users?

The solution to the first problem is called Resource Allocation part whereas the solution to the second problem is named Resource Assignment. Therefore, in the Resource Allocation part we select the users that should get resources and in the Resource Assignment part we perform the proper association between the selected users in the Resource Allocation part and the available resources. In Fig. 2.1 we present the main building blocks of the proposed heuristic framework.

In the Resource Allocation, we perform three main tasks: (1) the definition of the data rate that should be transmitted at the current TTI to each user; (2) the definition of the number of resources demanded at the current TTI to satisfy each user, and (3) the building of a priority list.

The data rate that each user j should transmit at the current TTI ($\Delta R_j[n]$) is calculated based on the traffic conditions and QoS requirements. Note that this variable is different depending on the service class, i.e., NRT or RT. Based on the calculation of $\Delta R_j[n]$ we are able to estimate the number $\kappa_j[n]$ of resources demanded by each user j . The higher is $\Delta R_j[n]$, the higher is $\kappa_j[n]$. The priority list is built by sorting in the descending order the users according to the user priority. We define p_j as the priority of user j that should be calculated according to the satisfaction status of each

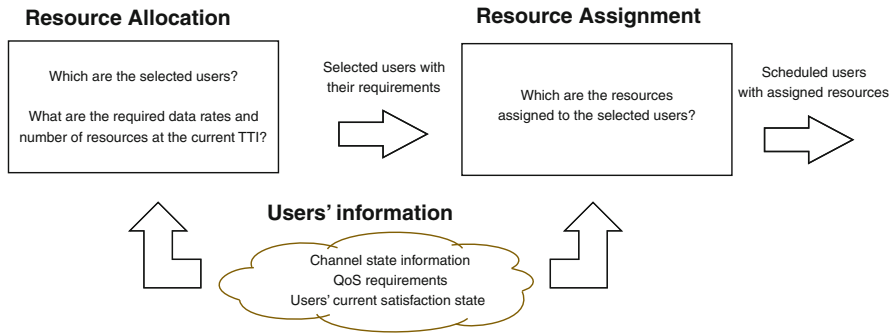


Fig. 2.1 Heuristic framework for maximization of user satisfaction

user. Further details on this will be shown in the Sects. 2.3.3 and 2.3.4, i.e., how this priority can be calculated for NRT and RT users.

Finally, the Resource Allocation part selects the first J' users from the priority list to get resources in the Resource Assignment part. J' is the maximum integer lower than or equal to J so that the sum of the estimated required number of resources does not surpass the total number of available resources K .

In the Resource Assignment part the selected users of the Resource Allocation part get resources in an opportunistic manner. More specifically, the resources are assigned to the users in an opportunistic round-robin fashion, i.e., at each round one user chooses one resource. The first user to choose a resource is the one that has the resource with the best channel condition among all others. Then, the user with the resource with second best channel condition chooses its resource, and so on. At each round, the users get resources until achieving $\kappa_j[n]$ resources, that is, each user gets the estimated number of resources required to achieve the data rate of $\Delta R_j[n]$. Once a given user gets the estimated number of resources, this user is taken out of the Resource Assignment process. If all selected users get the estimated number of resources and there are still unassigned resources, these remaining resources are equally distributed among the users since they are already satisfied.²

In the following, we provide more details on how the framework can be particularized for NRT and RT services in the form of two heuristic RRA techniques called Satisfaction-Oriented Resource Allocation for Non-Real Time Services (SORA-NRT) and Satisfaction-Oriented Resource Allocation for Real-time Services (SORA-RT), which are described in Sects. 2.3.3 and 2.3.4, respectively.

² A slight modification of the algorithm could be made by suppressing this latter step (distribution of unassigned resources) and avoiding oversatisfaction to users. In this case some frequency resources would not be used avoiding extra interference in other cells as well as transmit power would be saved motivated by energy efficiency concerns.

2.3.3 Application of the Heuristic Framework for NRT Services

This section describes in detail how the SORA-NRT technique works. Based on the reasoning explained above, we have that the data rate that each user j should transmit at the current TTI is given by

$$\Delta R_j[n] = \bar{R}_j^{\text{req}} \cdot (t_j[n] + 1) - \bar{R}_j[n-1] \cdot t_j[n-1]. \quad (2.4)$$

The transmit data rate calculated in Eq. (2.4) is the one that should be allocated to an unsatisfied user at the current TTI in order to this user become and stay satisfied even if this user does not have transmit opportunity at the next TTI.

The estimated number $\kappa_j[n]$ of resources required at the current TTI is given by

$$\kappa_j[n] = \frac{\Delta R_j[n]}{\left(\frac{\sum_{k \in \mathcal{K}} r_{j,k}}{K} \right)}, \quad (2.5)$$

where $r_{j,k}$ is the transmit data rate to user j on resource k . Note that, the denominator of Eq. (2.5) consists in the average data rate of user j assuming all resources. Therefore, κ_j in this equation means the estimated number of resources in order to user achieve a transmit data rate of $\Delta R_j[n]$.

The main idea in the prioritization process is to give precedence in transmission to the NRT users that are unsatisfied over the satisfied ones. According to this, we give opportunity to the unsatisfied users become satisfied and avoid resource over provision for satisfied users. Therefore, the unsatisfied users have absolute priority over the satisfied ones. In order to prioritize the users inside the groups of unsatisfied and satisfied users we define the priority of user j as

$$p_j = \frac{1}{|\kappa_j[n]|} \quad (2.6)$$

where $|\cdot|$ represents the absolute value of a scalar. Therefore, within the group of unsatisfied users we give priority to the users that need fewer resources to become satisfied, i.e., lower value of $\kappa_j[n]$. In the group of satisfied users we prioritize the satisfied users that are near to unsatisfaction. In this way we avoid that new users become unsatisfied. In summary, the priority list for NRT users is illustrated in Fig. 2.2.

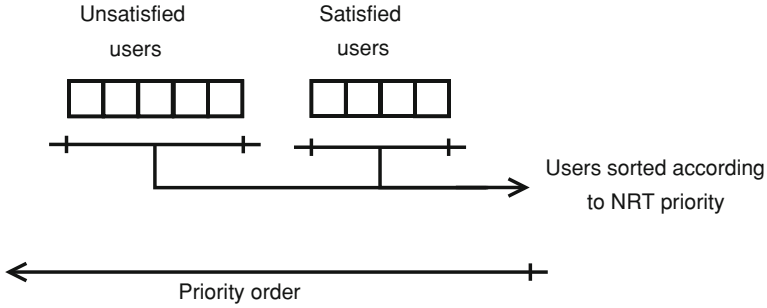


Fig. 2.2 Priority list for NRT users

2.3.4 Application of the Heuristic Framework for RT Services

This section describes in detail how the SORA-RT technique works. The transmit data rate that should be allocated to an RT user at the current TTI is given by

$$\Delta R_j[n] = \frac{b_j^{\text{hol}}[n]}{t_{\text{tti}}}, \quad (2.7)$$

where $b_j^{\text{hol}}[n]$ comprises the number of bits of the oldest packet in the transmit buffer of the Base Station (BS) at TTI n corresponding to user j . As for RT services, such as VoIP, the upper layers fragment the information data into multiple small packets so that it is completely feasible transmitting the whole packet in a single TTI.

As in the NRT case, the estimated number of resources of user j at the current TTI is given by Eq. (2.5).

When the prioritization for RT users is concerned, we have some differences compared to the NRT case. Basically, in the RT traffic scenario absolute priority is given to the satisfied users instead of the unsatisfied users as it is done for the NRT case. The reason for this difference is that RT users do not tolerate fluctuations in the provided QoS differently of NRT users. In other words, it is easier to satisfy an unsatisfied NRT user than an unsatisfied RT user due to its strict delay requirements. Therefore, it is important to keep the maximum number of RT users satisfied in order to get high user satisfaction.

Before defining the priority of each user within the group of satisfied and unsatisfied users we define the following variable

$$\omega_j[n] = \begin{cases} \left\lceil \frac{(\eta_j^{\text{succ}}[n] + \eta_j^{\text{lost}}[n]) \cdot \text{FER}_j^{\text{req}} - \eta_j^{\text{lost}}[n]}{1 - \text{FER}_j^{\text{req}}} \right\rceil, & \text{if } \text{FER}_j[n] \leq \text{FER}_j^{\text{req}} \\ \left\lceil \frac{\eta_j^{\text{lost}}[n] - (\eta_j^{\text{succ}}[n] + \eta_j^{\text{lost}}[n]) \cdot \text{FER}_j^{\text{req}}}{\text{FER}_j^{\text{req}}} \right\rceil, & \text{otherwise,} \end{cases} \quad (2.8)$$

where the operators $\lceil \cdot \rceil$ and $\lfloor \cdot \rfloor$ return the first integer greater than or equal to and the first integer lower than or equal to a real number, respectively. The variable $\omega_j[n]$ has two possible meanings depending on the satisfaction status of user j . If user j is satisfied, ω_j means the maximum number of packets that user j can successively lose and still be satisfied. On the other hand, for an unsatisfied user this variable means the number of consecutive packets that user j should transmit in order to become satisfied. Note that high values of $\omega_j[n]$ mean that the FER of user j is much lower or much higher than the required FER depending on user j is satisfied or unsatisfied, respectively.

The priority of user j is given by

$$p_j = \frac{1}{(d_j^{\text{req}} - d_j^{\text{hol}}) \cdot (\omega_j[n] + 1)}, \quad (2.9)$$

where d_j^{hol} and d_j^{req} are the current delay of the oldest packet in the transmit buffer for user j , i.e., Head Of Line (HOL) packet delay of user j , and the packet delay requirement, respectively. According to Eq. (2.9), users with high HOL packet delays (close to the deadline) and that have low value for $\omega_j[n]$ have higher priority than the other users. In Fig. 2.3 we illustrate the structure of the priority list for RT users.

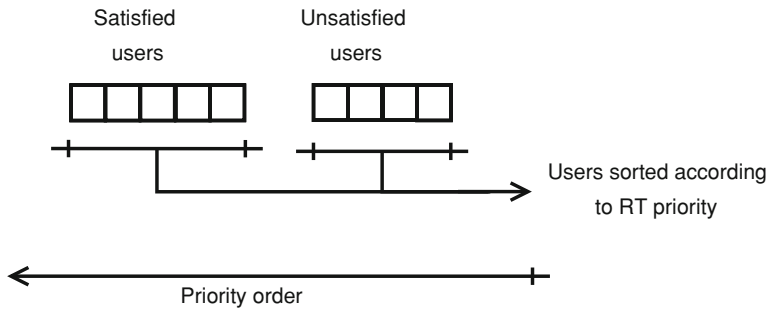


Fig. 2.3 Priority list for RT users

2.4 Utility-Based Radio Resource Allocation

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

Utility theory is a flexible tool that is employed in this work to design a general RRA framework that is able to improve user satisfaction in next generation cellular networks. Section 2.4.1 presents the general theory behind utility-based optimization, while Sect. 2.4.2 shows how we can use it to propose a utility-based RRA framework for OFDMA systems. We consider a particular utility function that is suitable for improving user satisfaction. Based on this choice, we describe two RRA techniques in Sects. 2.4.3 and 2.4.4 that maximize the number of satisfied NRT and RT users, respectively.

2.4.1 Problem Formulation

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

$$\max_{\mathcal{K}_j} \sum_{j=1}^J U(x_j) \quad (2.10a)$$

$$\text{subject to } \bigcup_{j=1}^J \mathcal{K}_j \subseteq \mathcal{K}, \quad (2.10b)$$

$$\mathcal{K}_i \cap \mathcal{K}_j = \emptyset, \quad i \neq j, \quad \forall i, j \in \{1, 2, \dots, J\}, \quad (2.10c)$$

where J is the total number of users 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{K} is the set of all resources in the system, \mathcal{K}_j is the subset of resources assigned to user j , 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 the user j . Constraints (2.10b) and (2.10c) 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.

The optimization problem (2.10a, 2.10b, and 2.10c) could be formulated considering the power allocated to the resources as another optimization variable. However, the optimum solution for this joint optimization problem is very difficult to be found [5]. Most of the sub-optimum solutions proposed in the literature split the problem into two stages: first, dynamic resource assignment with fixed power allocation, and next, adaptive power allocation with fixed resource assignment. Furthermore,

it has been shown for OFDMA-based systems that equal power allocation provides almost the same gains in comparison with adaptive power allocation with much less complexity [5]. Therefore, we consider the simplified optimization problem (2.10a, 2.10b and 2.10c), which can be solved by a suitable dynamic resource assignment with equal power allocation among the resources.

Depending on the utility function and the variable x_j , several RRA policies can be designed. In this study, we are interested at formulating general RRA techniques suitable for NRT or RT services. Therefore, we consider the variable x to be either the users' throughput (average data rates) or the users' HOL packet delay, which are QoS parameters suitable for NRT and RT services, respectively.

It is demonstrated in appendices 1 and 2 that we are able to derive simplified optimization problems that are equivalent to our original problem regarding NRT and RT services. According to appendices 1 and 2, the objective function of our simplified problem is linear in terms of the instantaneous user's data rate and given by

$$\max_{\mathcal{X}_j} \sum_{j=1}^J U'(x_j) \cdot R_j[n], \quad (2.11)$$

where $R_j[n]$ is the instantaneous data rate of user j (see Sect. 2.5.2.4) and $U'(x_j)$ is the marginal utility (derivative of the utility function) of the user j with respect to its QoS metric. The objective function (2.11) characterizes a weighted sum rate maximization problem [8], whose weights are adaptively controlled by the marginal utilities. Based on appendices 1 and 2, we represent the marginal utility corresponding to user j as the weight

$$w_j^{\text{nrt}} = U'(T_j[n-1]), \quad (2.12)$$

if the user j has an NRT service, or as the weight

$$w_j^{\text{rt}} = \left| U'(d_j^{\text{hol}}[n]) \right|, \quad (2.13)$$

if the user j has an RT service. We have that $T_j[n-1]$ is the average throughput of user j calculated up to the previous TTI, and $d_j^{\text{hol}}[n]$ is the HOL packet delay of user j at the current TTI n .

These utility-based weights play an important role on the RRA framework proposed in the following.

2.4.2 Utility-Based Resource Allocation Framework for OFDMA Systems

The general optimization formulation described in Sect. 2.4.1 can be applied to any modern cellular system. In this work, we focus on 4G cellular systems such as 3GPP Long-Term Evolution (LTE) which is based on OFDMA.

We call the optimization problem (2.10) with subcarriers or Physical Resource Blocks (PRBs) as the resources and considering equal power allocation, as the Dynamic Resource Assignment (DRA) problem. This problem has a closed form solution when the objective function is linear with respect to $R_j[n]$. The solution of the problem when the objective function is given by (2.39) can be found in Ref. [32], while the solution for the objective function (2.43) is described in Ref. [9]. Based on that, we have that the user with index j^* is chosen to transmit on the resource k at the TTI n if the condition below is satisfied:

$$j^* = \arg \max_j \{w_j \cdot r_{j,k}[n]\}, \quad (2.14)$$

where w_j is the utility-based weight factor of user j , and $r_{j,k}[n]$ denotes the instantaneous achievable transmission rate of the resource k with respect to the user j . On the one hand, if the user has an NRT service, we have that $w_j = w_j^{\text{nrt}} = U'_j(T_j[n-1])$, according to (2.12). 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 (2.13).

Figure 2.4 explains how the utility-based DRA algorithm proposed above works. Consider a scenario in which two NRT users i and j compete for seven resources, where the former user has better channel conditions than the latter in all resources. This is represented in Fig. 2.4a, where the Signal-to-Noise Ratios (SNRs) of users i and j are given by $\gamma_{i,k}$ and $\gamma_{j,k}$, respectively, and we have $\gamma_{i,k} > \gamma_{j,k}, \forall k$. In this case, all resources would be assigned to user i in accordance with (2.14). On the other hand, the SNRs $\gamma_{i,k}^*$ and $\gamma_{j,k}^*$ plotted in Fig. 2.4b are utility-scaled versions of their original SNRs $\gamma_{i,k}$ 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 (2.14), resources $k = 1, \dots, 3$ would be assigned to user i and resources $k = 4, \dots, 7$ would be assigned to user j . 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.

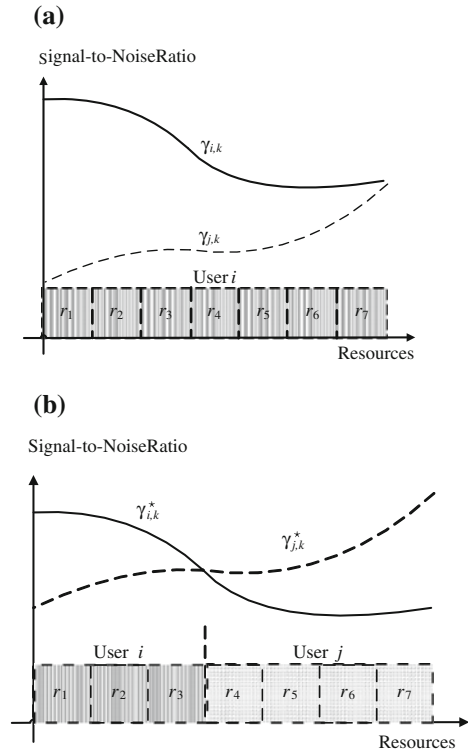
If we consider a step-shaped function, such as the sigmoidal function, as the utility function in the optimization problem formulated in Sect. 2.4.1, it is possible to achieve high user satisfaction for NRT or RT users with low complexity, as will be seen later. This utility function should be based on a particular QoS parameter suitable for each of these services.

Two utility-based RRA policies able to maximize the number of satisfied users in the system are proposed in this work. The first one is the Throughput-based Satisfaction Maximization (TSM) policy, whose formulation is based on the users' throughput and is suitable for NRT services, and the second one is the Delay-based Satisfaction Maximization (DSM) policy, whose formulation is based on the users' HOL packet delay and is suitable for RT services.

The similarities among these techniques are discussed considering a joint formulation, which is described in the following.

We propose to use a sigmoidal utility function based on a generic QoS metric $x_j[n]$ of the user j , as indicated below:

Fig. 2.4 Utility-based dynamic resource assignment (DRA). **a** Resource assignment without utility-based weights. **b** Resource assignment with utility-based weights



$$U(x_j[n]) = \frac{1}{1 + e^{\mu \cdot \sigma (x_j[n] - x_j^{\text{req}})}}, \quad (2.15)$$

where $x_j[n]$ and x_j^{req} are the current QoS metric and the QoS requirement of the user j , respectively; σ is a nonnegative parameter that determines the shape of the sigmoidal function; and μ is a constant (-1 or 1) that determines if the sigmoid is an increasing or decreasing function.

As explained in Sect. 2.4.2, the utility-based weight plays an important role in the DRA algorithm. The higher the weight, the higher the priority of the user to get a resource. The utility-based weight based on a generic QoS metric x_j of the user j is given by the marginal utility, which is the derivative of the utility function $U(x_j[n])$ with respect to the QoS metric $x_j[n]$, i.e., $w_j = \frac{\partial U(x_j[n])}{\partial x_j[n]}$. Therefore, we have that

$$w_j = \frac{\sigma \cdot e^{\mu \cdot \sigma (x_j[n] - x_j^{\text{req}})}}{\left(1 + e^{\mu \cdot \sigma (x_j[n] - x_j^{\text{req}})}\right)^2}. \quad (2.16)$$

The particular expression of w_j presented in (2.36) must be used in the corresponding DRA algorithm given by (2.14).

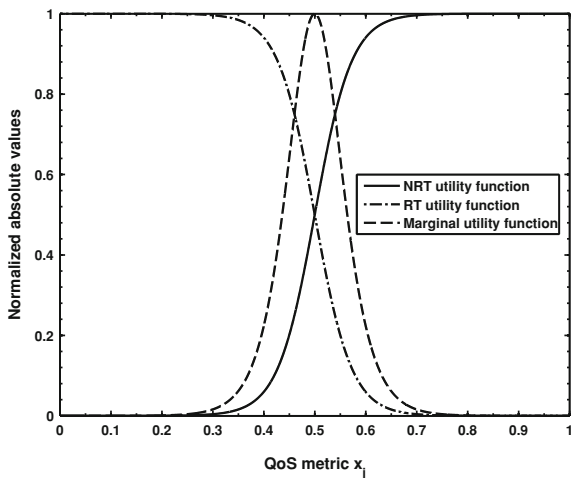
In the following, we give more details about how the general formulation described above can be configured as each of the proposed utility-based policies proposed in this work, namely TSM and DSM (described in Sects. 2.4.3 and 2.4.4, respectively).

2.4.3 Application of the Utility-Based Framework for NRT Services

The utility function used by the TSM policy is an increasing step-shaped sigmoid, which is suitable for NRT services, as illustrated in Fig. 2.5. The function is based on the users' throughput $T_j[n]$ and is centered on a throughput requirement T_j^{req} . An increasing utility function means that the higher the throughput, the higher the users' utility derived from the network. This increasing sigmoid is achieved when we set $\mu = -1$ in (2.15). A step-like utility function means that when the throughput approaches exceeds the throughput requirement, a given user becomes satisfied rapidly. 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 [27].

The marginal utility (utility-based weight) is illustrated in Fig. 2.5 as a bell-shaped function, which is the derivative of the sigmoidal utility function. It means that the users who have higher priority in the resource allocation process are the ones experiencing throughput levels close to the requirement. Therefore, one can conclude that the users most benefited are those in the imminence of becoming unsatisfied or satisfied. Moreover, the TSM technique has an interesting property, which is to avoid

Fig. 2.5 Examples of step-shaped sigmoidal utility functions and the absolute value of a bell-shaped marginal utility function ($x_j^{\text{req}} = 0.5$)



the users from becoming unsatisfied by giving priority to those users with QoS levels just above the requirement.

The higher the value of the parameter σ , the steeper the sigmoid. For the case of NRT services, we have achieved satisfactory results with $\sigma = 2.441 \times 10^{-5}$ in (2.15) and (2.16), which is suitable for the case of $T_j^{\text{req}} = 512$ kbps (see Table 2.4).

Notice that a different value of the QoS requirement will require an adjustment of the σ parameter accordingly, so that we have a desired step-shaped sigmoid no matter the value of the QoS requirement. In order to do that, the fixed σ parameter must be a function of the QoS requirement T_j^{req} . A possible way to do that is to force the sigmoid to be equal to a given value δ when the QoS metric $T_j[n]$ achieves a given proportion ρ of the QoS requirement T_j^{req} . Therefore, we have that

$$\sigma = \frac{\log \frac{1-\delta}{\delta}}{\rho \cdot T_j^{\text{req}}}. \quad (2.17)$$

Regarding the NRT utility function in Fig. 2.5, we have $\mu = -1$, $\delta = 0.01$, $\rho = 0.5$, and $T_j^{\text{req}} = 0.5$. It means that the NRT function starts to increase noticeably, i.e., $U_j(T_j) = \delta = 0.01$, when T_j is half of the QoS requirement, i.e., $T_j = \rho \cdot T_j^{\text{req}} = 0.25$.

2.4.4 Application of the Utility-Based Framework for RT Services

The DSM policy considers the users' HOL packet delay $d_j^{\text{hol}}[n]$ as the QoS metric. Since the users' utility derived from the network is lower when the delay is higher, we have that the RT utility function should be decreasing (see Fig. 2.5). In that sense, we have $\mu = 1$ in (2.15). This utility function is also centered on a QoS requirement, which is called d_j^{req} and must be equal to or lower than the RT delay budget.

The absolute value of the resulting marginal utility is a bell-shaped function (see Fig. 2.5), which is a symmetric function around the QoS requirement (RT delay budget). In our model, a packet discard procedure is used, where the 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. In this way, depending on the values of the RT delay budget and the DSM delay requirement (central value of the sigmoid), there could be some portion of the marginal utility function (abscissa values higher than the RT delay budget) that will be neglected.

Most of the works found in the literature define the satisfaction of RT services based on FER, e.g., Ref. [28]. If the user's FER is higher than a requirement, the user is considered unsatisfied; otherwise it is assumed satisfied. Besides the packet losses due to channel errors, we also have packet losses due to unbearable delays. Taking this into account, we consider that even if the satisfaction is measured in terms of FER, a utility function based on HOL packet delay is suitable for an RRA policy that intends to provide high levels of user satisfaction. Therefore, a decreasing step-like utility

function based on HOL packet delay means that a given user becomes unsatisfied rapidly if the HOL packet delay approaches and exceeds the delay requirement. The opposite occurs when the user delay decreases to values below the requirement.

As it can be seen in Fig. 2.5, the marginal utility is a bell-shaped function, which means that the users who experience HOL packet delays close to the requirement are the ones who will have higher priority in the resource allocation process. If this requirement is set to be equal or close to the RT delay budget, one can conclude that the users in the imminence of becoming unsatisfied are benefited.

For the case of RT services, we have achieved satisfactory results with $\sigma = 138.135$ in (2.15) and (2.16), which is suitable for the case of $d_j^{\text{req}} = 100$ ms (see Table 2.5).

Notice again that a different value of the QoS requirement will require an adjustment of the σ parameter accordingly. For the case of RT services, the fixed σ parameter must be a function of the QoS requirement d_j^{req} . The same expression (2.17) can be used to calculate the value of σ suitable for RT services, substituting T_j^{req} by d_j^{req} and using suitable values for the other parameters. Taking Fig. 2.5 as an example, suppose that we have $\mu = 1$, $\delta = 0.01$, $\rho = 0.5$, and $d_j^{\text{req}} = 0.5$. The RT function starts to decrease noticeably, i.e., $U_j(d_j) = 1 - \delta = 0.99$, when d_j is half of the QoS requirement, i.e., $d_j = \rho \cdot d_j^{\text{req}} = 0.25$.

2.5 Performance Evaluation

A model of an LTE-like system was implemented in a computational simulator. In this section, we present simulation results concerning the presented solutions (heuristic and utility-based) and analyze the relative performance and existing trade-offs between them and some prior art algorithms found in the literature.

2.5.1 Prior-art Algorithms

Each of the prior-art algorithms considered in this work uses a different DRA criterion. However, in order to have a fair comparison with the proposed SORA-NRT, SORA-RT, TSM, and DSM techniques, all prior-art algorithms use equal power allocation among the resources.

The SORA-NRT, SORA-RT, TSM, and DSM techniques were already described in Sects. 2.3.3, 2.3.4, 2.4.3, and 2.4.4. The following sections describe in more details each of the other studied algorithms.

2.5.1.1 Rate Maximization

The Rate Maximization (RM) RRA for OFDMA-based systems was first studied in Ref. [12]. The objective was to maximize the sum of data rates of the users subject to a maximum transmission power constraint. The solution is to assign each resource to the user that has the highest channel gain on it and next apply the waterfilling power allocation, which characterizes a pure opportunistic policy.

The mathematical formulation of the RM policy is presented in (2.18). The user with index j^* is chosen to transmit on resource k in TTI n if it satisfies the condition given by

$$j^* = \arg \max_j \{r_{j,k} [n]\}. \quad (2.18)$$

Although the original RM policy proposed in Ref. [12] consider waterfilling as the solution of the power allocation, we consider Equal Power Allocation (EPA) in order to have a fair comparison with the SORA-NRT, SORA-RT, TSM, and DSM policies proposed in this work, as explained before.

2.5.1.2 Proportional Fairness

The Proportional Fair (PF) algorithm intends to serve users with favorable radio conditions in order to provide a high instantaneous throughput relative to their average throughput [14, 35]. The user with index j^* is chosen to transmit on resource k in TTI n if the condition below is satisfied:

$$j^* = \arg \max_j \left\{ \frac{r_{j,k} [n]}{T_j [n-1]} \right\}, \quad (2.19)$$

The throughput of user j is averaged using a Simple Exponential Smoothing (SES) filtering according to (2.36).

2.5.1.3 Modified Largest Weighted Delay First

The Modified Largest Weighted Delay First (MLWDF) criterion was originally proposed in Ref. [2] to be used as a PS algorithm in single-carrier systems. We consider in this work a generalized version of this algorithm suitable for OFDMA systems. The user with index j^* is chosen to transmit on resource k in TTI n if it satisfies the condition given by

$$j^* = \arg \max_j \left\{ d_j^{\text{hol}} [n] \cdot \frac{r_{j,k} [n]}{T_j [n-1]} \right\}, \quad (2.20)$$

Since MLWDF considers the HOL packet delay in its formulation, it is especially suitable for RT services.

The authors in Ref. [2] also consider a weight for each user in the priority function that is dependent on the maximum due delay time and the maximum allowed probability of the packet delay exceeding this due time, which provides a QoS differentiation among users. However, in this work we assume that all users have the same characteristics and there is no need to use this kind of QoS differentiation. This policy was proved to be throughput-optimal, i.e., it makes the queues stable if it is feasible to do so with any other scheduling rule [2].

2.5.1.4 Urgency and Efficiency-Based Packet Scheduling

The Urgency and Efficiency-based Packet Scheduling (UEPS) is a utility-based PS algorithm, which uses the time-utility function as a scheduling urgency factor and the relative status of the current channel to the average one as an efficiency indicator of radio resource usage [29]. Its design goal is to maximize throughput of NRT traffics while satisfying QoS requirements of RT traffics. In this work, we evaluate the performance of the UEPS algorithm in the RT traffic scenario.

The utility function used by the UEPS algorithm is a sigmoid represented by

$$U_j \left(d_j^{\text{hol}}[n] \right) = \frac{e^{-\sigma \left(d_j^{\text{hol}}[n] - d_j^{\text{req}} \right)}}{1 + e^{-\sigma \left(d_j^{\text{hol}}[n] - d_j^{\text{req}} \right)}}. \quad (2.21)$$

According to the UEPS criterion, the user with index j^* is chosen to transmit on resource k in TTI n if the condition below is satisfied:

$$j^* = \arg \max_j \left\{ \left| U'_j \left(d_j^{\text{hol}}[n] \right) \right| \cdot \frac{r_{j,k}[n]}{T_j[n-1]} \right\}, \quad (2.22)$$

where $\left| U'_j \left(d_j^{\text{hol}}[n] \right) \right|$ is the absolute value of the derivative of the utility function given by (2.21).

Tables 2.1 and 2.2 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 (2.14)] for each of the algorithms, when applicable.

It is important to highlight here the difference between the UEPS technique [29] and the new proposed DSM technique. According to Table 2.2, the difference in the priority functions is a factor related to the opportunistic use of the resources. While UEPS uses the ratio between the instantaneous and average data rates, DSM uses only the instantaneous data rate. Notice that the expression for the priority function of the DSM technique was found by following a mathematical development based on utility theory (see Sect. 2.4 and Appendix 2), while the priority function of the UEPS technique was chosen empirically without a mathematical foundation (see Ref. [29] for more details).

Table 2.1 RRA algorithms that are compared in the NRT traffic scenario

Algorithms	Priority function
RM [12]	$r_{j,k}[n]$
PF [14]	$\frac{r_{j,k}[n]}{T_j[n-1]}$
SORA-NRT	Heuristic
TSM ^a	$\frac{\sigma \cdot e^{-\sigma(T_j[n]-T_j^{\text{req}})}}{\left(1 + e^{-\sigma(T_j[n]-T_j^{\text{req}})}\right)^2} \cdot r_{j,k}[n]$

^a $\sigma = 2.441 \times 10^{-5}$ **Table 2.2** RRA algorithms that are compared in the RT traffic scenario

Algorithms	Priority function
RM [12]	$r_{j,k}[n]$
MLWDF [2]	$d_j^{\text{hol}}[n] \cdot \frac{r_{j,k}[n]}{T_j[n-1]}$
UEPS ^a [29]	$\frac{\sigma \cdot e^{\sigma(d_j^{\text{hol}}[n]-d_j^{\text{req}})}}{\left(1 + e^{\sigma(d_j^{\text{hol}}[n]-d_j^{\text{req}})}\right)^2} \cdot \frac{r_{j,k}[n]}{T_j[n-1]}$
SORA-RT	Heuristic
DSM ^a	$\frac{\sigma \cdot e^{\sigma(d_j^{\text{hol}}[n]-d_j^{\text{req}})}}{\left(1 + e^{\sigma(d_j^{\text{hol}}[n]-d_j^{\text{req}})}\right)^2} \cdot r_{j,k}[n]$

^a $\sigma = 138.135$

2.5.2 Scenario Characterization and Simulation Modeling

The simulations took into account the main characteristics of an LTE-like system. The main simulation models are described in the following sections, and a summary with the general simulation parameters are depicted in Table 2.3.

2.5.2.1 General Assumptions

Some general simulation assumptions considered in this work are listed below.

- We consider frequency-selective Rayleigh fading and each PRB experiences flat fading. In this way, we assume that the channel gains are constant over a TTI, but vary from one TTI to another.
- The BS has perfect knowledge of the Channel State Information (CSI) of all users in all PRBs.

Table 2.3 Simulation parameters

Parameter	Value
Number of cells	1
Maximum BS transmission power	1 W
Cell radius	500 m
UE speed	3 km/h
Carrier frequency	2 GHz
System bandwidth	5 MHz
Total number of sub-carriers	512
Total number of useful sub-carriers	300
Sub-carrier bandwidth	15 kHz
Number of PRBs	25
Path loss	Using (2.23)
Log-normal shadowing standard dev.	8 dB
Small-scale fading ^a	3GPP Typical Urban (TU) [1, 11]
AWGN power per sub-carrier	−123.24 dBm
Noise figure	9 dB
Link adaptation	Using link level curves from [19]
SNR threshold of MCS 1 [19]	−6.9 dB
Transmission time interval (TTI)	1 ms
Multiple antenna configurations ^a	SISO, MISO ^b 1×2 , SIMO ^c 2×1 , SU-MIMO ^d 2×2
Simulation time span	30 s

^a We are considering the 3GPP TU channel profile, which is symmetric between the downlink and uplink when there is no antenna correlation. Therefore, the performance of the RRA algorithms for the MISO 2×1 and SIMO 1×2 configurations are the same

^b Maximum ratio transmission (MRT) Precoding

^c Maximum ratio combining (MRC) Precoding

^d Zero forcing (ZF) Precoding

- The resource allocation information (PRB assignment, modulation, and coding schemes, etc.) is sent to each user in a separate control channel, so that the users can decode the data in their own PRBs.
- The users are static, i.e., there is no mobility. However, in each simulation scenario several independent snapshots with different user distributions are simulated, which captures the system performance in different coverage situations. Each snapshot has a given duration, and although long-term fading is kept constant, fast fading is correlated over time considering a 3 km/h user speed (see Table 2.3).
- The downlink transmission scheme is based on OFDM using a normal cyclic prefix length and considering 14 OFDMA symbols per TTI. The total subcarrier bandwidth is 15 kHz, which accounts for both data and pilot symbols.
- The minimum resource block considered in the simulations is a time-frequency chunk, which is called PRB and is formed by a time slot of 1 ms (TTI) and 12 subcarriers.

2.5.2.2 Propagation

The path loss follows the model proposed in Ref. [34] for a test scenario in urban and suburban areas. Considering a 2 GHz carrier frequency, and a mean BS antenna height of 15 m, the equation of the path loss L_j^{path} in dB as a function of the distance d between the BS and user j in km is presented as follows:

$$L_j^{\text{path}} = 128.1 + 37.6 \log_{10} d. \quad (2.23)$$

The modeling of the large-scale fading used in this work is the well-known zero-mean lognormal shadowing model characterized by a given standard deviation σ_{sh} [24].

In this work, we assume that the small-scale fading (fast fading) follows a Rayleigh distribution. One of the most popular approaches to generate the Rayleigh fading suitable for simulation purposes is the Jakes' model [11], which is the approach considered in this work. We consider the power-delay profile according to the Typical Urban (TU) model proposed by the 3GPP [1].

In order to calculate a proper value for the BS transmit power, we assume the parameters indicated in Table 2.3. We have found that 1 W of BS transmit power is sufficient to provide 99 % of coverage probability considering the minimum Modulation and Coding Scheme (MCS) taken from the link adaptation curves presented in Ref. [19].

2.5.2.3 Spatial Filtering with Multiple Antennas

We consider a Multiple-Input-Multiple-Output (MIMO) channel with M^{tx} transmit antennas and M^{rx} receive antennas and \mathbf{H} is an $M^{\text{rx}} \times M^{\text{tx}}$ matrix whose elements $h_{a,b}$ consist in the channel transfer function between the receive antenna a and transmit antenna b . Before transmission, the signals are filtered by a transmit matrix \mathbf{M} with dimension $M^{\text{tx}} \times q$ and, at the receiver the signals are filtered by a receiver filter \mathbf{D} with dimension $q \times M^{\text{rx}}$, where q is the number of transmitted signals, $q \leq \min(M^{\text{tx}}, M^{\text{rx}}, \nu)$, and ν is the rank of the channel matrix \mathbf{H} . Therefore, the input-output relation for the MIMO channel corresponding to user j is given by

$$\tilde{\mathbf{y}}_j = \mathbf{D}_j \mathbf{y}_j = \mathbf{D}_j \mathbf{H}_j \mathbf{M}_j \mathbf{x}_j + \mathbf{D}_j \mathbf{n}_j \quad (2.24)$$

where \mathbf{y}_j and $\tilde{\mathbf{y}}_j$ are the prior-filtering received signal vector and the postfiltering received signal vector with dimension $q \times 1$, \mathbf{x}_j is the transmit signal vector with dimension $q \times 1$ and \mathbf{n}_j is the $M^{\text{rx}} \times 1$ white Zero Mean Circularly Symmetric Complex Gaussian (ZMCSCG) noise vector. We assume that the channel is perfectly known at the transmitter and receiver.

In this work, we consider single-user MIMO (SU-MIMO) so that there is no spatial resource sharing among different users. Each user transmission can be divided into

many streams (depending on the channel and antenna configuration) and sent through different spatial subchannels.

The following spatial filtering schemes were used in this work for Single-Input Multiple-Output (SIMO), Multiple-Input-Single-Output (MISO), and SU-MIMO configurations.

Maximum Ratio Transmission Precoding

The Maximum Ratio Transmission (MRT) precoding is designed to maximize the transmitter SNR for MISO scenarios [22]. The precoding matrix \mathbf{M}_j and the decoding matrix \mathbf{D}_j for the user j are defined, respectively, as

$$\mathbf{M}_j = \frac{\mathbf{H}_j}{\|\mathbf{H}_j\|_2}, \quad \text{and} \quad \mathbf{D}_j = 1, \quad (2.25)$$

where \mathbf{H}_j is the channel matrix and $\|\cdot\|_2$ is the Euclidian norm of a matrix.

Maximum Ratio Combining Precoding

The Maximum Ratio Combining (MRC) precoding is used to maximize the SNR for SIMO scenarios [22]. The precoding matrix \mathbf{M}_j and the decoding matrix \mathbf{D}_j for the user j are defined, respectively, as

$$\mathbf{M}_j = 1, \quad \text{and} \quad \mathbf{D}_j = \frac{\mathbf{H}_j^H}{\|\mathbf{H}_j\|_2}, \quad (2.26)$$

where \mathbf{H}_j is the channel matrix.

Zero Forcing Precoding

Zero-Forcing (ZF) precoding is conceived to totally decorrelate the transmit signals so that the signal at every receiver output is free of interference [21]. When $M^{\text{rx}} \leq M^{\text{tx}}$, the columns of the precoding matrix \mathbf{M}_j for the user j are achieved as the normalized columns of the matrix $\tilde{\mathbf{M}}_j$ defined as

$$\tilde{\mathbf{M}}_j = \mathbf{H}_j^H (\mathbf{H}_j \mathbf{H}_j^H)^{-1} \quad (2.27)$$

while the decoding matrix \mathbf{D}_j is defined as

$$\mathbf{D}_j = \mathbf{I}_{M^{\text{rx}}} \quad (2.28)$$

where $\mathbf{I}_{M^{\text{rx}}}$ denotes an $M^{\text{rx}} \times M^{\text{rx}}$ identity matrix, $(\cdot)^{-1}$ is the inverse of a matrix and $(\cdot)^H$ is the conjugate transpose operation in a matrix.

2.5.2.4 Link Adaptation

Depending on the channel condition, an appropriate number of bits is transmitted on each PRB. This is accomplished by the link adaptation procedure. The link adaptation curves used in this work were taken from Ref. [19], which characterize a 3GPP LTE system.

Using the precoding schemes described in Sect. 2.5.2.3 and considering a link adaptation scheme that allows a user to transmit at different data rates according to the SNRs, we have that the possible transmit rate of user j at the PRB k is

$$r_{j,k} = \sum_{l=1}^{\nu} f(\gamma_{j,k,l}) \quad (2.29)$$

where $f(\cdot)$ maps the SNR $\gamma_{j,k,l}$ of each spatial dimension l (stream) of user j on PRB k to the possible data rate. The SNR is given by

$$\gamma_{j,k,l} = \frac{p_{k,l} \cdot \sigma_l^2}{\eta} \quad (2.30)$$

where $p_{k,l}$ is the available transmit power per resource k and spatial dimension l , η is the noise power, and σ_l is the l th singular value of the channel matrix \mathbf{H} .

Once we have the achievable transmission rate of each PRB taking into account all available streams, the downlink data transmission rate for each user can be calculated. In the resource assignment process, we assume that each PRB can only be assigned to one single user. Assuming that a PRB set \mathcal{K}_j is assigned to user j , its transmission rate is calculated as

$$R_j = \sum_{k \in \mathcal{K}_j} r_{j,k} \quad (2.31)$$

where $r_{j,k}$ is given by (2.29). The total rate of the system is the sum of R_j among all users, as indicated below:

$$R_{\text{cell}} = \sum_{j=1}^J R_j. \quad (2.32)$$

2.5.2.5 Traffic Model

NRT services, such as World Wide Web (WWW) and FTP, are not delay-sensitive and require an overall high throughput. The traffic model used for NRT services

is the full-buffer model. It assumes that the buffers of the users located in the BS always have data to be transmitted. It is an assumption widely used in many works in the literature that evaluate RRA techniques for OFDMA-based systems. The idea behind this model is that some NRT multimedia services to be provided by next-generation mobile broadband systems require the transfer of large amounts of data, for example high definition images, music, and video. Furthermore, the full-buffer model characterizes a worst-case scenario regarding system load. Since all RRA techniques studied in this work consider the same model, the relative performance comparison remains valid.

We consider a simple traffic model for RT services, which consists on the regular generation of packets of b_j^{hol} bits into the buffer of user j every $1/\lambda$ ms. The delay of each packet is accounted and it must respect the RT delay budget of the radio access network. If the packet arrives at the receiver later than this delay budget, it is discarded.

2.5.2.6 Performance Metrics

Fairness

In order to evaluate the RRA techniques in terms of fairness, we use the well-known Jain's fairness index, which is a quantitative fairness measure originally proposed by Jain et al. in Ref. [10]. The general Jain's fairness function is independent of the allocation metric being used. Considering a generic allocation metric $\mathbf{x} = [x_1, \dots, x_j, \dots, x_J]$, the Jain's fairness function can be interpreted in terms of the variance and expected value of \mathbf{x} , as follows:

$$F(\mathbf{x}) = \frac{(E(\mathbf{x}))^2}{E(\mathbf{x}^2)} = \frac{1}{1 + \frac{\text{Var}(\mathbf{x})}{(E(\mathbf{x}))^2}} = \frac{\left(\sum_{j=1}^J x_j\right)^2}{J \cdot \sum_{j=1}^J x_j^2}, \quad (2.33)$$

where $E(\cdot)$ and $\text{Var}(\cdot)$ represent the expectation operator and variance of their arguments, respectively.

The Jain's fairness index has some interesting properties [10]:

- The fairness is bounded between 0 and 1 (or 0 and 100 %). A totally fair allocation (with all x_j 's equal) has a fairness of 1, while a totally unfair allocation (with all resources given to only one user) has a fairness of $1/J$, which is 0 in the limit as $J \rightarrow \infty$.
- The fairness is independent of scale, i.e., unit of measurement does not matter.
- The fairness is a continuous function. Any slight change in allocation is reflected into fairness.

- If only Q of J users share the resources equally with the remaining $J - Q$ users not receiving any resource, then the fairness is Q/J .

The 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 users 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.

Satisfaction

The definition of user satisfaction depends on the type of service that the user has, i.e., NRT or RT service. An NRT user is considered satisfied if its session throughput is higher or equal to a threshold ($T_j[n] \geq T_j^{\text{req}}$). 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 receiver later than the delay budget of the RT service.

The percentage of satisfied users is calculated as

$$\psi^{\text{cell}}[n] = \frac{J^{\text{sat}}[n]}{J}, \quad (2.34)$$

where $J^{\text{sat}}[n]$ is the number of satisfied users in the cell. The metric given by (2.34) is the satisfaction metric adopted for both scenarios with NRT and RT services.

2.5.3 Simulation Results

Simulation results are presented for two case studies. The first one considers a scenario where we have an NRT service (Sect. 2.5.3.1), and the second one assumes that we have an RT service (Sect. 2.5.3.2).

Table 2.4 Simulation parameters for the evaluation of satisfaction maximization for NRT services

Parameter	Value
NRT traffic model	Full buffer
Throughput filtering time constant (f_{thru})	1/1,000
User throughput requirement (T_j^{req})	512 kbps
Parameter μ	-1
Parameter σ	2.441×10^{-5}
Number of independent simulation runs	30

2.5.3.1 Non-real Time Services

In this section, we compare the performance of the proposed SORA-NRT and TSM techniques with other classical techniques found in the literature, namely RM and PF. The algorithms are assessed assuming different multiantenna configurations, such as Single-Input-Single-Output (SISO), MISO 1×2 , SIMO 2×1 , and SU-MIMO 2×2 , where $M^{\text{rx}} \times M^{\text{tx}}$ represents the number of antennas in the receiver and transmitter, respectively. The particular simulation parameters used in this analysis are depicted in Table 2.4.

The total cell throughput (system capacity) as a function of the number of NRT users considering different antenna configurations is depicted in Fig. 2.6. As expected, the RM policy provides the best results for all multiantenna scenarios. Since it assigns each spatial stream to the users that can transmit at the highest MCS, it is able to achieve the maximum allowed system capacity for all system loads. The SORA-NRT policy also presents good performance, but the resulting system capacity decreases when the number of NRT users increases. For low system loads, when there are sufficient resources to satisfy all users,³ SORA-NRT allocates the remaining resources to the users with the best channel conditions, which explains the higher system throughput. When the traffic load increases, more users become unsatisfied and SORA-NRT does not have a pool of extra resources to improve capacity anymore. This explains the capacity degradation for high traffic loads. As will be seen later on, TSM is a fair policy and tries to keep the throughput of the satisfied users as close as possible, which is not so efficient in terms of system capacity. Notice that TSM is able to exploit multiuser diversity in order to achieve higher system capacity when the number of users increases. Finally, the PF policy presents an almost flat behavior for different traffic loads.

It can also be noticed in Fig. 2.6 that the system capacity increases when more antennas are used in either the BS or the user, as expected. On one hand, MISO provides slight higher cell throughput than SISO due to transmit diversity, which increases the SNR in the receiver. On the other hand, SU-MIMO presents huge gains because it takes advantage of the higher number of spatial streams in order to boost capacity.

³ We are considering a system bandwidth of 5 Mhz, which accounts for 25 PRBs.

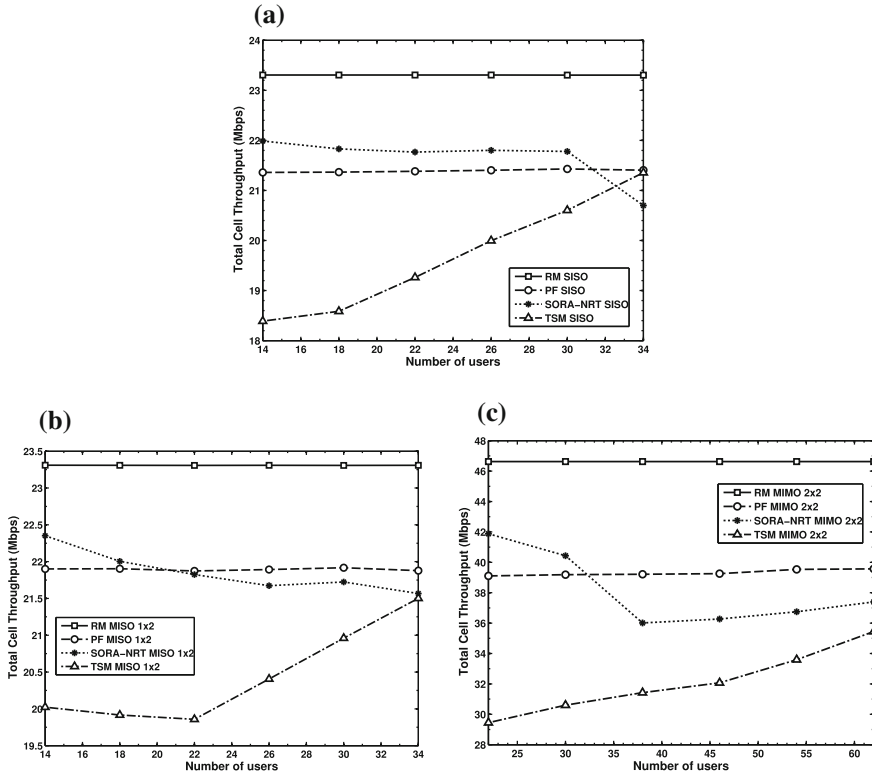


Fig. 2.6 Total cell throughput as a function of the number of NRT users considering different antenna configurations **a** SISO **b** MISO 1×2 **c** SU-MIMO 2×2

Figure 2.7 shows the throughput-based cell fairness index for different multi-antenna schemes. It is calculated by (2.33) and averaged over all snapshots. The intrinsic trade-off between system capacity and user fairness is identified when we analyze Figs. 2.6 and 2.7. More details about this trade-off can be found in Chap. 4 of this book. The RM policy is able to use the resources very efficiently but is very unfair in the resource and QoS distribution. PF turns out to be a good trade-off with reasonable spectral efficiency and high fairness. On the one hand, SORA-NRT presents a reasonable system capacity, but its performance in terms of throughput-based fairness is not good. On the other hand, TSM is not so good in terms of system capacity but compensates by providing very high fairness among the users.

Regarding the multiantenna configurations, it can be observed that the addition of an extra antenna in the BS (MISO) does not make a big difference in terms of throughput-based fairness. However, extra antennas in both the BS and user (SU-MIMO) causes a fairness decrease for all RRA techniques, except TSM. This happens because extra spatial streams provide more diversity, which is exploited by the opportunistic policies and may cause unfair resource distribution more frequently.

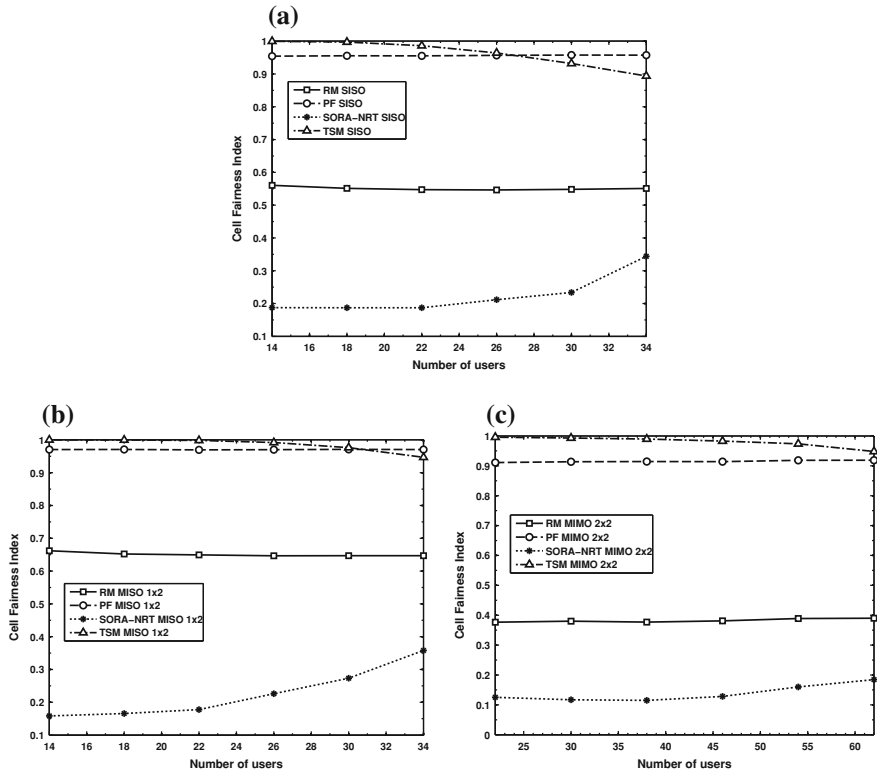


Fig. 2.7 Mean cell fairness index as a function of the number of NRT users considering different antenna configurations **a** SISO **b** MISO 1×2 **c** SU-MIMO 2×2

Since TSM achieves a good balance between channel access opportunism and QoS awareness, it is able to maintain high fairness levels even in the SU-MIMO scenario.

Figure 2.8 presents the most important result of the study regarding NRT services: the percentage of satisfied users as a function of the system load for distinct antenna schemes.

Looking also at Figs. 2.6 and 2.7, one can notice that user satisfaction and system capacity are negatively correlated, while user satisfaction and fairness are positively correlated. The former comparison also express clearly the fundamental trade-off between resource efficiency and user satisfaction. Some of these trade-offs are studied in detail in Chap. 4 of this book.

The TSM and SORA-NRT techniques show better satisfaction results than the classical techniques for all considered system loads, except for SORA-NRT that presents slightly lower satisfaction levels than PF for a reduced number of users. Among all, TSM provides the best results, which demonstrates the advantage of using the flexible utility-based RRA framework when satisfaction maximization is desired. Looking at the TSM results, one can notice that the percentage of satisfied

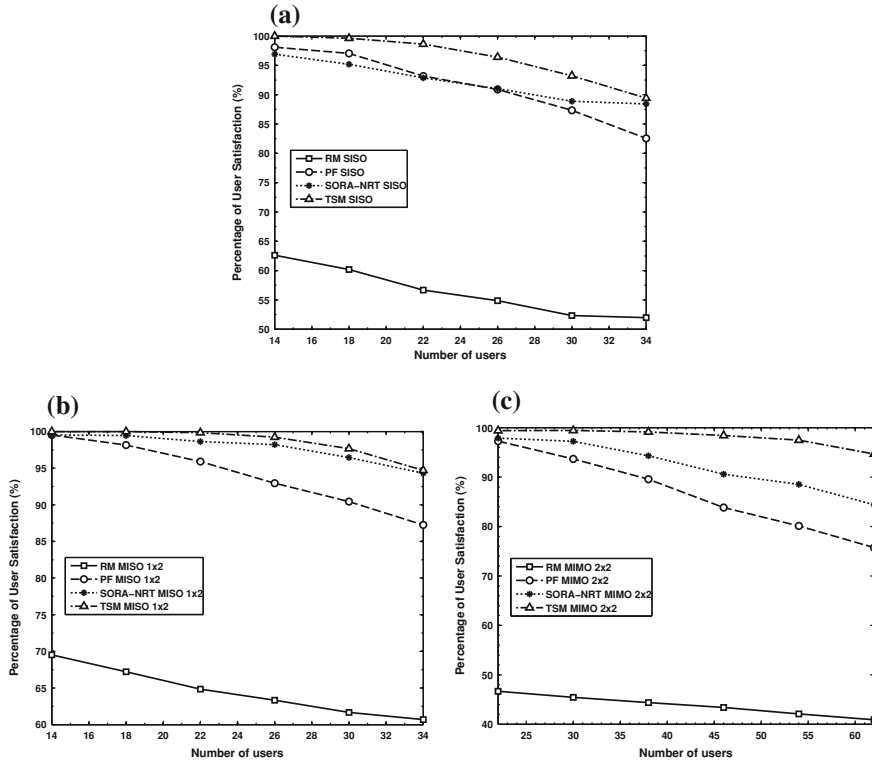


Fig. 2.8 User satisfaction as a function of the number of NRT users considering different antenna configurations **a** SISO **b** MISO 1×2 **c** SU-MIMO 2×2

users is highly correlated with the Jain's fairness index (see Figs. 2.7 and 2.8). This is due to a property of the Jain's fairness index adopted in this study and the way that TSM allocates resources. 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 , what explains the highly positive correlation (see Ref. [10] for more details). RM provides an overall degraded QoS because it leaves many users in outage situations. The classical PF technique also presents good satisfaction results. Another good characteristic of the proposed SORA-NRT and TSM techniques is the low computational complexity.

Notice that the use of more antennas increases the percentage of satisfied users, except for the RM technique in the SU-MIMO scenario. In the RM case, the increased diversity due to more spatial streams tends to concentrate the resources in the hands of even fewer users, which decreases the overall percentage of user satisfaction in the system.

The result presented in Fig. 2.8 confirms that our objective of proposing efficient and low complexity RRA techniques is able to improve user satisfaction in a scenario with NRT services was accomplished.

2.5.3.2 Real Time Services

In this section, the performance of the proposed SORA-RT and DSM techniques is compared with other classical algorithms, such as MLWDF, UEPS, and RM. The algorithms are assessed assuming different multiantenna configurations, such as SISO, MISO 1×2 , SIMO 2×1 , and SU-MIMO 2×2 . Table 2.5 presents the simulation parameters used in this analysis.

Figure 2.9 depicts the total cell throughput (system capacity) as a function of the number of RT users for different multiantenna configurations. In general, the delay-aware policies show higher capacity because they are more successful at avoiding unbearable delays and preventing packets from being lost. The higher the number of successfully transmitted packets, the higher the system capacity. At first sight, one could expect that the pure opportunistic policy RM would present the highest system capacity. In the scenario we are evaluating, this is not true 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 the lack of data.

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 technique, together with MLWDF and UEPS, shows the best results. 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, 130, and 190 RT users in the SISO, MISO, and SU-MIMO scenarios, respectively. This happens because the SORA-RT heuristics does not give priority to users who need many resources to become satisfied. Since we have a small amount of data to be transmitted in the system with a restrictive delay requirement, this unfair policy has a negative impact on system capacity.

Table 2.5 Simulation parameters for the evaluation of satisfaction maximization for RT services

Parameter	Value
RT packet size (b_j^{hol})	256 bits
RT packet interarrival time ($1/\lambda$)	2 ms
FER threshold	2 %
HOL packet delay requirement (d_j^{req})	100 ms
RT delay budget	100 ms
Parameter μ	1
Parameter σ	138.135
Number of independent simulation runs	10

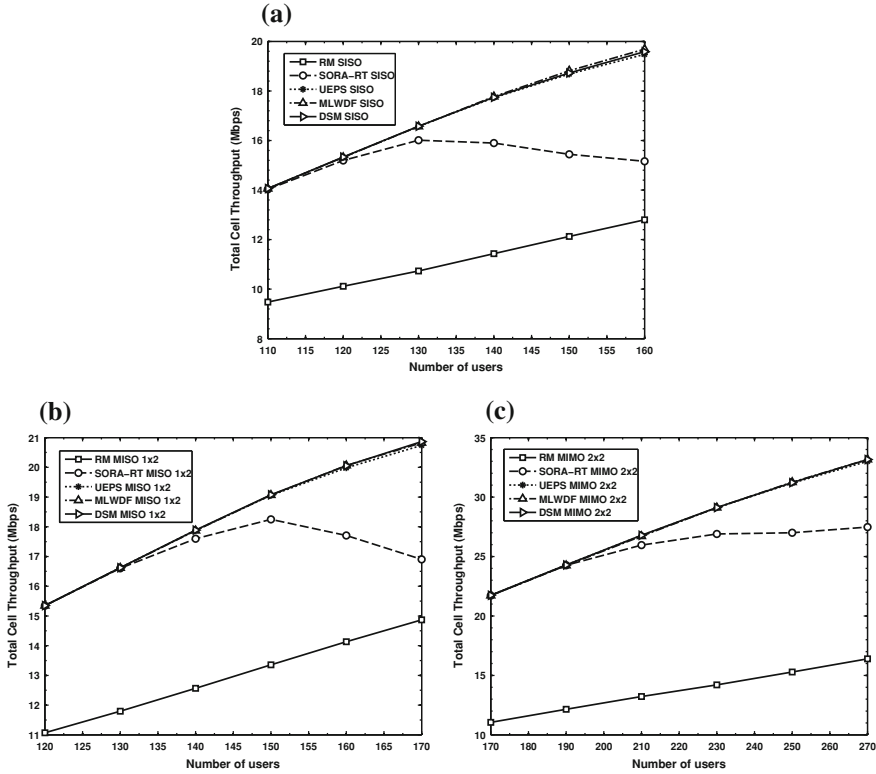


Fig. 2.9 Total cell throughput as a function of the number of RT users considering different antenna configurations **a** SISO **b** MISO 1×2 **c** SU-MIMO 2×2

Comparing the multiantenna schemes in Fig. 2.9, one can observe the same behavior among the RRA algorithms, and a capacity increase with the addition of more antennas in the transmitter and receiver. As expected, the capacity gain is small from SISO to MISO, which is a consequence of higher SNR due to transmit diversity, but the gain is higher from MISO to SU-MIMO, which is caused by the exploitation of an extra spatial stream. Comparing the capacity results of the RT and NRT traffic scenarios (see Figs. 2.6 and 2.9), it can be noticed that the cell throughputs on the RT traffic scenario is lower than the NRT traffic scenario, which can be explained by the difference in the traffic models of the service classes (see Sect. 2.5.2.5).

The mean cell fairness index based on HOL packet delay is shown in Fig. 2.10. 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 and MLWDF have also good performance in terms of fairness, where the latter is worse than the former. Finally, the SORA-RT algorithm presents worse fairness

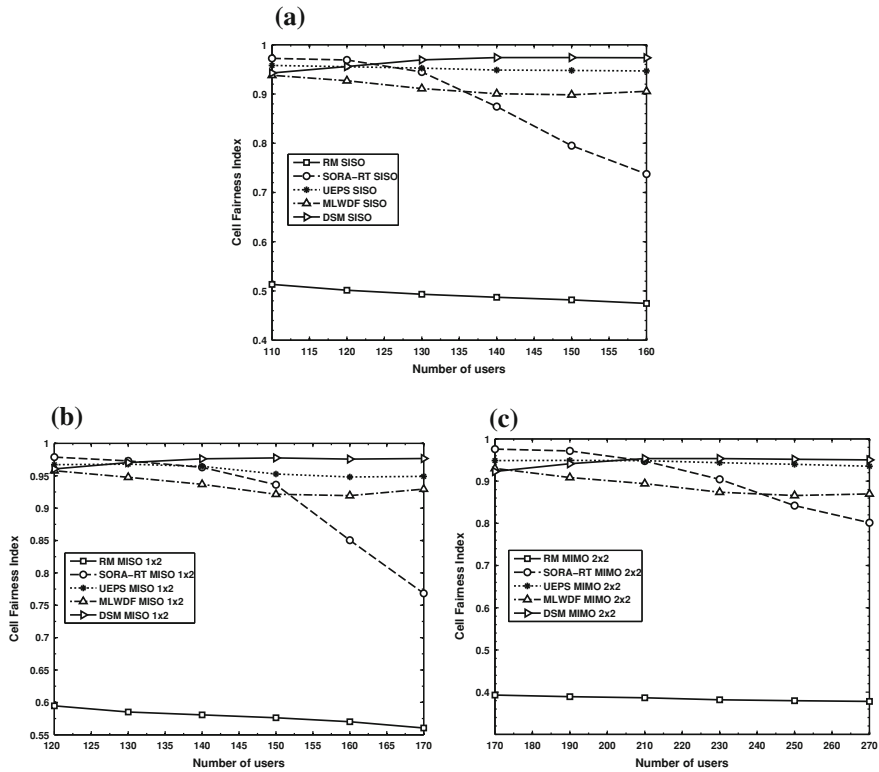


Fig. 2.10 Mean cell fairness index as a function of the number of RT users considering different antenna configurations **a** SISO **b** MISO 1×2 **c** SU-MIMO 2×2

results for high offered loads. As explained before, the reason for that is because the worst users with highest delays would need many resources to become satisfied in this scenario, so they are neglected by the SORA-RT heuristics, which decreases fairness. Comparing the multiantenna configurations in Fig. 2.10, one can observe the same relative behavior among the RRA algorithms.

Finally, Fig. 2.11 presents the most important result of the study regarding RT services: the percentage of satisfied RT users for various multiantenna schemes. The algorithms that take into account the HOL packet delay in their formulations are those ones that provide the highest user satisfaction. The resource allocation criteria of the nonheuristic algorithms (DSM, MLWDF and UEPS) are based on the combination of two indicators: a QoS indicator that is a function of the HOL packet delay, and an efficient indicator that can be the transmission rate (DSM) or the ratio between the transmission rate and the user throughput (MLWDF and UEPS). Comparing DSM and UEPS, which have the same QoS indicator (bell-shaped marginal utility function), it can be concluded that the achievable transmission rate is a better efficiency indicator for the maximization of user satisfaction, since DSM outper-

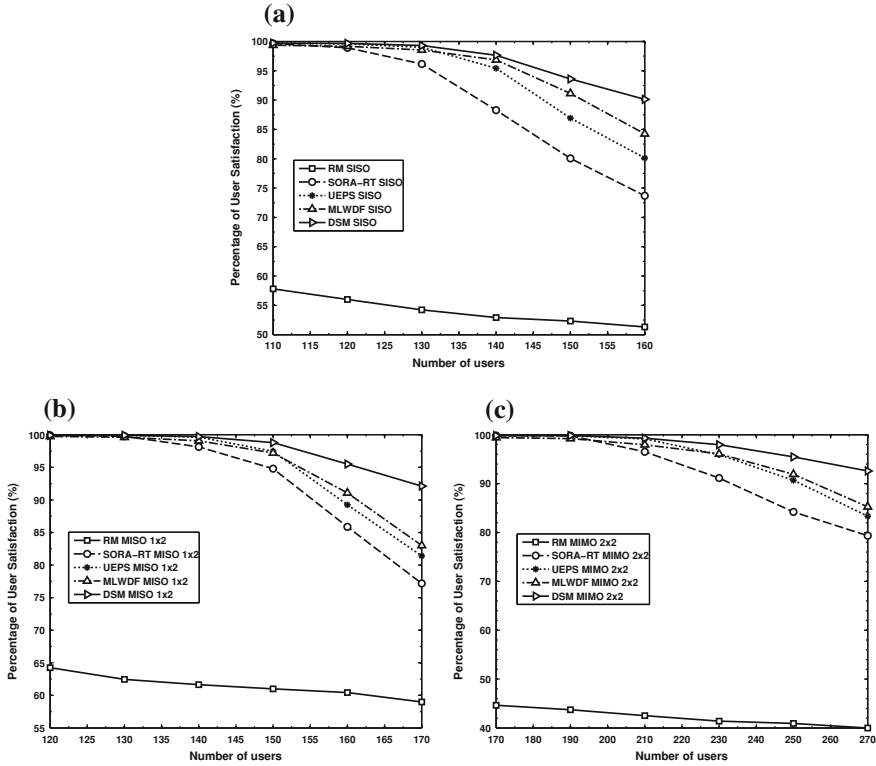


Fig. 2.11 User satisfaction as a function of the number of RT users considering different antenna configurations **a** SISO **b** MISO 1×2 **c** SU-MIMO 2×2

forms UEPS. Furthermore, MLWDF and UEPS, which have the same efficiency indicator, have different delay-based functions as the QoS indicator: linear and bell-shaped, respectively (see Table 2.2). Comparing these two algorithms in Fig. 2.11, it can be concluded that the linear function is better than the bell-shaped function. The proposed SORA-RT provides good results for low offered loads, but reasonable percentage of user satisfaction for high offered loads. Comparing Figs. 2.10 and 2.11, one can notice that the fairness and satisfaction results of SORA-RT are highly correlated, which strengthens our understanding that the policy of neglecting the worst RT users carried out by SORA-RT has a negative impact on capacity, fairness, and also satisfaction.

Special attention must be given to the proposed DSM technique, which achieved its objective of maximizing user satisfaction for all considered multiantenna schemes. The combination of the bell-shaped delay-based indicator and the transmission rate 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.

2.6 Conclusions

In this chapter, we dealt with the RRA problem of maximizing the number of satisfied users in cellular networks considering NRT or RT services. In order to solve this problem we presented two approaches, namely: heuristic and utility-based RRA design.

We first presented a general heuristic framework for maximizing the number of satisfied users based on two steps: resource allocation and resource assignment. In the resource allocation, we define which users will get resources whereas in the resource assignment part we define the proper association between the selected users and the resources. Furthermore, the general framework is particularized into two novel RRA techniques called SORA-NRT and SORA-RT, which are suitable for NRT and RT services, respectively.

After that, we presented a utility-based framework for solving the RRA problem. This framework uses a sigmoidal utility function, and is composed of a utility-based DRA algorithm, which takes into account QoS-based prioritization and channel opportunism, and an equal power allocation among frequency resources. Two novel RRA techniques are derived from the utility-based framework: TSM and DSM. On the one hand, TSM uses an increasing sigmoidal function based on throughput with inflection point in the users' throughput requirement, and its main objective is to improve satisfaction among NRT users in a cellular network. On the other hand, DSM uses a decreasing sigmoidal function based on HOL packet delay with inflection point in the users' HOL packet delay requirement, which is usually equal to the RT delay budget of the system. Its main objective is to improve satisfaction among RT users.

According to system-level simulation results in a scenario with an NRT service, the proposed TSM and SORA-NRT show the best satisfaction results, when compared to other classical RRA techniques, such as PF and RM. Furthermore, the proposed techniques have reduced computational complexity.

In a scenario with an RT service, the proposed utility-based DSM technique outperforms the other techniques, namely MLWDF, UEPS, and also the proposed heuristic-based SORA-RT. DSM presents simultaneously the highest satisfaction, fairness, and system capacity. Moreover, the proposed techniques also shows reduced computational complexity.

All techniques were evaluated in different multiple antenna scenarios, namely SISO, MISO 2×1 , SIMO 1×2 , and SU-MIMO 2×2 . It was observed that the addition of more antennas in the transmitter and/or receiver helped the proposed RRA techniques to achieve higher cell throughput, improve fairness, and increase the percentage of satisfied users.

Based on the simulation results, it can be concluded that our objective of proposing efficient and low complexity RRA techniques able to improve user satisfaction in a scenario with NRT or RT services was accomplished. However, the heuristic policy of the SORA-RT technique still needs to be optimized in order to provide even better satisfaction results.

To sum up, we conclude that RRA design is an effective tool for satisfying operators' and users' needs in cellular networks. Heuristic and utility-based frameworks are very valuable tools to design efficient frameworks to solve a variety of RRA problems, in particular the user satisfaction maximization problem. Comparing the heuristic and utility-based approaches, some advantages of the latter were observed: higher flexibility, solid mathematical formulation, lower computational complexity, and better user satisfaction results in the scenarios evaluated in this study.

Appendix 1: Utility-Based Optimization Formulation for NRT Services

Let us consider a utility-based optimization problem in a scenario with NRT services formulated as:

$$\max_{\mathcal{K}_j} \sum_{j=1}^J U(T_j[n]) \quad (2.35a)$$

$$\text{subject to } \bigcup_{j=1}^J \mathcal{K}_j \subseteq \mathcal{K}, \quad (2.35b)$$

$$\mathcal{K}_i \cap \mathcal{K}_j = \emptyset, \quad i \neq j, \quad \forall i, j \in \{1, 2, \dots, J\}, \quad (2.35c)$$

where J is the total number of users 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{K} is the set of all resources in the system, \mathcal{K}_j is the subset of resources assigned to user j , and $U(T_j[n])$ is an increasing utility function based on the current throughput $T_j[n]$ of the user j in TTI n . Constraints (2.35b) and (2.35c) 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.

The throughput of user j is calculated using an exponential smoothing filtering, as indicated below:

$$T_j[n] = (1 - f^{\text{thru}}) \cdot T_j[n-1] + f^{\text{thru}} \cdot R_j[n], \quad (2.36)$$

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

Evaluating the objective function in (2.35a) and the throughput expression in (2.36), 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 [32]:

$$\frac{\partial U(T_j[n])}{\partial R_j[n]} \approx f^{\text{thru}} \cdot \left. \frac{\partial U}{\partial T_j} \right|_{T_j=T_j[n-1]}, \quad (2.37)$$

where the previous resource allocation totally determines the current values of the marginal utilities. Using the one-order Taylor formula for the utility function [25, 32] and considering (2.37), we have

$$\begin{aligned} \sum_{j=1}^J U(T_j[n]) &\approx \sum_{j=1}^J U(T_j[n-1]) + \\ &\sum_{j=1}^J \left. \frac{\partial U}{\partial T_j} \right|_{T_j=T_j[n-1]} \cdot (f^{\text{thru}} \cdot R_j[n] - f^{\text{thru}} \cdot T_j[n-1]). \end{aligned} \quad (2.38)$$

Notice that maximizing (2.38) leads to the maximization of the original objective function (2.35a). 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 objective function of our simplified optimization problem becomes linear in terms of the instantaneous user's data rate, and is given by

$$\max_{\mathcal{X}_j} \sum_{j=1}^J U'(T_j[n-1]) \cdot R_j[n], \quad (2.39)$$

where $U'(T_j[n-1]) = \left. \frac{\partial U}{\partial T_j} \right|_{T_j=T_j[n-1]}$ is the marginal utility (derivative of the utility function) of user j with respect to its throughput in the previous TTI. The objective function (2.39) characterizes a weighted sum rate maximization problem [8], whose weights are adaptively controlled by the marginal utilities.

Notice that we started with an optimization formulation based on throughput given by (2.35a), made some logical assumptions and mathematical simplifications, and ended up with a linear optimization formulation based on instantaneous rates given by (2.39). According to these arguments, we claim that the instantaneous optimization maximizing (2.39) leads to a long-term optimization that maximizes (2.35a).

Appendix 2: Utility-Based Optimization Formulation for RT Services

Let us consider a utility-based optimization problem in a scenario with RT services formulated as:

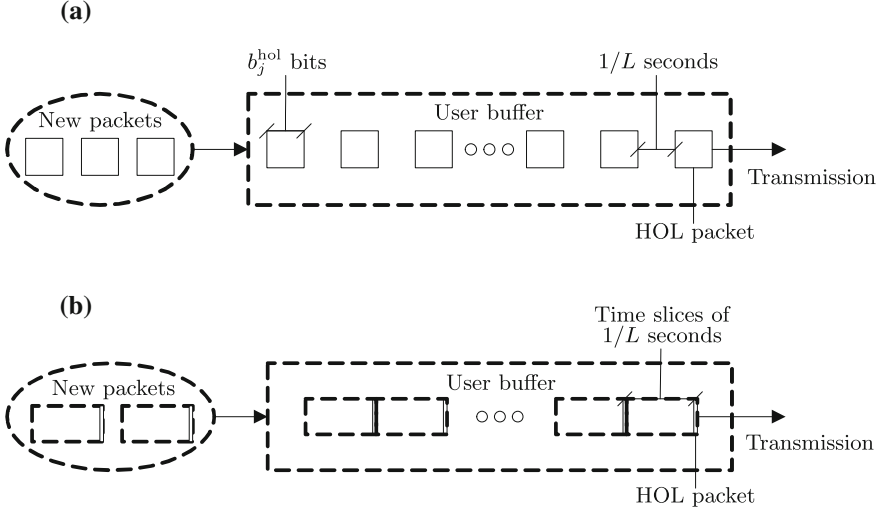


Fig. 2.12 Modeling of a RT user buffer **a** User buffer as a sequence of packets **b** User buffer as a sequence of time slices

$$\max_{\mathcal{K}_j} \sum_{j=1}^J U \left(d_j^{\text{hol}} [n] \right) \quad (2.40a)$$

$$\text{subject to } \bigcup_{j=1}^J \mathcal{K}_j \subseteq \mathcal{K}, \quad (2.40b)$$

$$\mathcal{K}_i \cap \mathcal{K}_j = \emptyset, \quad i \neq j, \quad \forall i, j \in \{1, 2, \dots, J\}, \quad (2.40c)$$

where J is the total number of users 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{K} is the set of all resources in the system, \mathcal{K}_j is the subset of resources assigned to user j , and $U \left(d_j^{\text{hol}} [n] \right)$ is a decreasing utility function based on the current HOL delay $d_j^{\text{hol}} [n]$ of user j at TTI n . Constraints (2.40b) and (2.40c) 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.

In order to understand the model used in this work for the calculation of the HOL delays, Fig. 2.12a 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 b_j^{hol} bits (fixed size) arrives in the buffer of user j every $1/L$ s.

Taking into account Fig. 2.12a and considering a generic user j , we propose in this work a recursive model for calculating an approximate value of the HOL delay. The recursive equation is

$$d_j^{\text{hol}}[n+1] = d_j^{\text{hol}}[n] + t^{\text{tti}} - \frac{1}{L} \cdot \left(\frac{R_j[n] \cdot t^{\text{tti}}}{b_j^{\text{hol}}} \right), \quad (2.41)$$

where t^{tti} is the duration of the TTI in seconds, L is the packet arrival rate, b_j^{hol} is the packet size in bits, and $R_j[n]$ is the instantaneous achievable transmission rate on TTI n . In this queue model, we assume that the packet size b_j^{hol} is sufficiently small, so that the queue can be represented ideally by a sequence of time slices with duration $1/L$ s each (see Fig. 2.12b). Notice that this assumption does not invalidate the mathematical and conceptual RRA framework, and makes the optimization model much more tractable.

Looking at (2.41), first 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. Second, 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 user j has not been served by any resource in TTI n , $R_j[n]$ 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 (2.40a) and the HOL delay expression in (2.41), we can see that the derivative of $U(d_j^{\text{hol}})$ with respect to the transmission rate R_j can be expressed as

$$\frac{\partial U}{\partial R_j} = \frac{\partial U}{\partial d_j^{\text{hol}}} \cdot \frac{\partial d_j^{\text{hol}}}{\partial R_j} = \frac{\partial U}{\partial d_j^{\text{hol}}} \cdot \left(-\frac{t^{\text{tti}}}{L \cdot b_j^{\text{hol}}} \right).$$

Using the result above and assuming that the TTI duration is sufficiently small, the Lagrange theorem of the mean can be used [15, 25], which says that

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

$$= \sum_{j=1}^J \left| \frac{\partial U}{\partial d_j^{\text{hol}}} \right|_{d_j^{\text{hol}}=d_j^{\text{hol}}[n]} \cdot \frac{t^{\text{tti}}}{L \cdot b_j^{\text{hol}}} \cdot (R_j[n] - R_j[n-1]). \quad (2.42)$$

The absolute value operator was used in (2.42) because the utility function was assumed to be decreasing, which yields a negative marginal utility (derivative of the utility function) and cancels the negative sign in (2.42). Notice that the maximization of (2.42) leads to the maximization of (2.40a). Taking into account (2.42), we have that t^{tti} , L , and b_j^{hol} are constants, and that $d_j^{\text{hol}}[n]$ and $R_j[n-1]$ are known and fixed before the resource allocation at TTI n . Therefore, our simplified optimization objective function is given by

$$\max_{\mathcal{K}_j} \sum_{j=1}^J \left| U' \left(d_j^{\text{hol}}[n] \right) \right| \cdot R_j[n], \quad (2.43)$$

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 user j with respect to its current HOL delay. The objective function (2.43) is a weighted sum rate maximization [8], where the weights are given by the absolute value of the marginal utility with respect to the current HOL delay.

Notice that after some logical assumptions and mathematical simplifications made upon (2.40a), a linear optimization formulation based on instantaneous rates given by (2.43) was achieved. Therefore, we claim that the instantaneous optimization maximizing (2.43) leads to a long-term optimization that maximizes (2.40a).

References

1. 3GPP: Deployment aspects. Technical Report TR 25.943 V9.0.0, Third Generation Partnership Project (2009)
2. Andrews, M., Kumaran, K., Ramanan, K., Stolyar, A., Whiting, P., Vijayakumar, R.: Providing quality of service over a shared wireless link. *IEEE Commun. Mag.* **32**(2), 150–154 (2001)
3. Braga, A.R., Rodrigues, E.B., Cavalcanti, F.R.P.: Packet scheduling for VoIP over HSDPA in mixed traffic scenarios. In: *IEEE International Symposium on Personal, Indoor and Mobile Radio, Communications*, pp. 1–5 (2006)
4. Ericsson: Ericsson mobility report: on the pulse of the networked society. Whitepaper (2012)
5. 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 of Berlin, Berlin (2006)
6. Gueguen, C., Baey, S.: Scheduling in OFDM wireless networks without tradeoff between fairness and throughput. In: *IEEE Vehicular Technology Conference*, pp. 1–5 (2008)
7. Holma, H., Toskala, A. (eds.): *WCDMA for UMTS: radio access for third generation mobile communications*, 3rd edn. Wiley, New York (2004)

8. Hoo, L.M.C., Halder, B., Tellado, J., Cioffi, J.M.: Multiuser transmit optimisation for multi-carrier broadcast channels: asymptotic FDMA capacity region and algorithms. *IEEE Trans. Commun.* **52**(6), 922–930 (2004)
9. Hosein, P.A.: QoS control for WCDMA high speed packet data. In: *International Workshop on Mobile and Wireless Communications, Network*, pp. 169–173 (2002)
10. 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 (1984)
11. Jakes, W.C.: *Microwave mobile communications*. Wiley / The Institute of Electrical and Electronics Engineers (IEEE) (1994)
12. Jang, J., Lee, K.B.: Transmit power adaptation for multiuser OFDM systems. *IEEE J. Sel. Areas Commun.* **21**(2), 171–178 (2003)
13. Kela, P., Puttonen, J., Kolehmainen, N., Ristaniemi, T., Henttonen, T., Moisio, M.: Dynamic packet scheduling performance in UTRA long term evolution downlink. In: *International Symposium on Wireless, Pervasive Computing*, pp. 308–313 (2008)
14. Kelly, F.: Charging and rate control for elastic traffic. *Eur. Trans. Commun.* **8**, 33–37 (1997)
15. 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: *IEEE Vehicular Technology Conference*, pp. 1664–1668 (2007)
16. Lima, F.R.M.: Satisfaction oriented resource allocation for wireless OFDMA systems, Master's thesis, Federal University of Ceará, Fortaleza, Brazil (2008)
17. Lima, F.R.M., dos Santos, R.B., Cavalcanti, F.R.P., Freitas, W.C.: Radio resource allocation for maximization of user satisfaction. In: *IEEE Workshop on Signal Processing Advances in Wireless Communications*, pp. 565–569 (2008)
18. Lima, F.R.M., Wänstedt, S., Cavalcanti, F.R.P., Freitas, W.C.: Scheduling for improving system capacity in multiservice 3GPP LTE. *J. Electr. Comput. Eng.* 21–36 (2010) <http://www.hindawi.com/journals/jece/2010/819729/cta/>
19. Mehlführer, C., Wrulich, M., Ikuno, J.C., Bosanska, D., Rupp, M.: Simulating the long term evolution physical layer. In: *European Signal Processing Conference*. Glasgow, Scotland (2009)
20. Mongha, G., Pedersen, K.I., Kovacs, I.Z., Mogensen, P.E.: QoS oriented time and frequency domain packet schedulers for the UTRAN long term evolution. In: *IEEE Vehicular Technology Conference*, pp. 2532–2536 (2008)
21. Paulraj, A., Biglieri, E., Goldsmith, A.: *MIMO wireless communications*, 1st edn. Cambridge University Press, New York (2007)
22. Paulraj, A., Nabar, R., Gore, D.: *Introduction to space-time wireless communications*, 1st edn. Cambridge University Press (2003)
23. Pokhariyal, A., Pedersen, K.I., Monghal, G., Kovacs, I.Z., Rosa, C., Kolding, T.E., Mogensen, P.E.: HARQ aware frequency domain packet scheduler with different degrees of fairness for the UTRAN long term evolution. In: *IEEE Vehicular Technology Conference*, pp. 2761–2765 (2007)
24. Rappaport, T.S. (ed.): *Wireless communications: principles and practice*, 2nd edn. Prentice Hall, Upper Saddle River, USA (2002)
25. Rodrigues, E.B.: Adaptive radio resource management for OFDMA-based macro- and femto-cell networks. Ph.D. thesis, Universitat Politècnica de Catalunya, Barcelona, Spain (2011)
26. Rodrigues, E.B., Casadevall, F.: Adaptive radio resource allocation framework for multi-user OFDM. In: *IEEE Vehicular Technology Conference*, pp. 1–6 (2009)
27. Rodrigues, E.B., Casadevall, F.: Control of the trade-off between resource efficiency and user fairness in wireless networks using utility-based adaptive resource allocation. *IEEE Commun. Mag.* **49**(9), 90–98 (2011)
28. Rodrigues, E.B., Cavalcanti, F.R.P., Wänstedt, S.: QoS-driven adaptive congestion control for voice over IP in multiservice wireless cellular networks. *IEEE Commun. Mag.* **46**(1), 100–107 (2008)
29. Ryu, S., Ryu, B., Seo, H., Shin, M.: Urgency and efficiency based wireless downlink packet scheduling algorithm in OFDMA system. In: *IEEE Vehicular Technology Conference*, vol. 3, pp. 1456–1462 (2005)

30. dos Santos, R.B., Lima, F.R.M., Freitas, W.C., Cavalcanti, F.R.P.: Qos based radio resource allocation and scheduling with different user data rate requirements for OFDMA systems. In: International Symposium on Personal, Indoor and Mobile Radio, Communications, pp. 1–5 (2007)
31. Shakkottai, S., Stolyar, A.L.: Scheduling algorithms for a mixture of real-time and non-real-time data in HDR. In: International Teletraffic Congress, pp. 793–804 (2001)
32. Song, G., Li, Y.G.: Cross-layer optimization for OFDM wireless networks - part II: algorithm development. *IEEE Trans. Wirel. Commun.* **4**(2), 625–634 (2005)
33. Song, G., Li, Y.G.: Utility-based resource allocation and scheduling in OFDM-based wireless broadband networks. *IEEE Commun. Mag.* **43**(12), 127–134 (2005)
34. UMTS: Selection procedures for the choice of radio transmission technologies of the umts. Tech. Rep. TR 101 112 V3.2.0 - UMTS 30.03, Universal Mobile Telecommunications System (UMTS), Sophia Antipolis, France (1998)
35. Viswanath, P., Tse, D.N.C., Laroia, R.: Opportunistic beamforming using dumb antennas. *IEEE Trans. Inf. Theor.* **48**(6), 1277–1294 (2002)
36. Zhang, Y., Liew, S.C.: Link-adaptive largest-weighted-throughput packet scheduling for real-time traffics in wireless OFDM networks. In: IEEE Global Telecommunications Conference, vol. 5, pp. 2490–2494 (2005)

Resource Allocation and MIMO for 4G and Beyond

Porto, R. (Ed.)

2014, XXXVIII, 527 p. 217 illus., 36 illus. in color.,

Hardcover

ISBN: 978-1-4614-8056-3