Chapter 4
Capacity, Fairness, and QoS Trade-Offs in Wireless Networks with Applications to LTE

Emanuel B. Rodrigues, Francisco R. M. Lima, Ferran Casadevall and Francisco Rodrigo Porto Cavalcanti

4.1 Introduction

The design of wireless mobile networks is driven by a multitude of objectives. As an example, among the requirements for 4th Generation(4G) given by International Mobile Telecommunications (IMT)-Advanced we can highlight maximum average cell spectral efficiency and cell border spectral efficiency in bits/s/Hz, maximum packet latency and minimum number of supported Voice over IP (VoIP) users [15]. In this case we can identify as design objectives spectral efficiency (average cell spectral efficiency), cell coverage (cell border spectral efficiency), Quality of Service (QoS) (packet latency), and user satisfaction or user capacity (number of supported VoIP users). Other objectives could be present in network design such as energy efficiency and fairness.

One of the most important tools for optimizing wireless mobile networks is Radio Resource Allocation (RRA). RRA is responsible for managing the available resources in the radio access interface such as frequency chunks, transmit power, and time slots. When the system bottleneck is in the radio access instead of the core network, efficient RRA can dictate the performance of the overall system.

E. B. Rodrigues (✉) · F. R. M. Lima · Francisco Rodrigo Porto Cavalcanti
Wireless Telecommunications Research Group (GTEL), Federal University of Ceará,
Caixa Postal 6005, Fortaleza 60440-900, Brazil
e-mail: emanuel@gtel.ufc.br

F. R. M. Lima
e-mail: rafaelm@gtel.ufc.br

Francisco Rodrigo Porto Cavalcanti
e-mail: rodrigo@gtel.ufc.br

F. Casadevall
Universitat Politècnica de Catalunya, Barcelona, Spain
e-mail: ferranc@tsc.upc.edu

Francisco Rodrigo Porto Cavalcanti (ed.), *Resource Allocation and MIMO for 4G and Beyond*, DOI: 10.1007/978-1-4614-8057-0_4,

However, in general all these network design objectives cannot be achieved at the same time by RRA strategies. A well-known case that illustrates this issue is the RRA strategies that aim at maximizing spectral efficiency. In order to maximize spectral efficiency, the system resources should be assigned to the users that can use them in the most efficient way in terms of b/s/Hz. These users are the ones that have better channel quality states. However, this RRA solution in general leads to reduced fairness and poor QoS provision to the other users that do not have the best channel quality states. Clearly, spectral efficiency is a contradicting objective with regard to both fairness and QoS provision.

Different RRA strategies can be designed to maximize one objective in detriment of another as well as to balance them. In this context, RRA strategies can be static or adaptive. By adaptive RRA strategies we mean solutions that can be configured to achieve different points in the trade-off between opposing objectives, whereas static strategies are able to achieve only a fixed point in the trade-off between the system objectives.

In order to conceive RRA solutions for the existing design objectives, many strategies can be followed. We highlight in this chapter the heuristic and utility-based approaches. As will be shown in the following sections, the heuristic design provides simple and quick solutions to the RRA problems, while the utility-based approach is a flexible and general tool for RRA design.

In this chapter we study important trade-offs in the downlink of modern wireless mobile networks and show adaptive RRA solutions based on the heuristic and utility-based approaches. The focus in this chapter is on Non-Real Time (NRT) services that have as main QoS metric throughput or average data rate. The remainder of this chapter is organized as follows. In Sect. 4.2 we review important objectives and trade-offs in wireless networks, whereas in Sects. 4.3 and 4.4 we present the heuristic and utility-based frameworks for conceiving RRA solutions, respectively. Then, in Sects. 4.5 and 4.6 we propose adaptive RRA strategies for the capacity versus fairness and capacity versus QoS trade-offs, respectively. Finally, in Sect. 4.7 we summarize this chapter with the main conclusions of the presented study.

4.2 Trade-Offs in Wireless Networks

Resource allocation for wireless mobile communications systems can have different objectives, such as the maximization of system capacity, cell coverage, user QoS (user satisfaction), fairness in the resource distribution, etc. Unfortunately, in general all these objectives cannot be achieved at the same time. Below we list some fundamental compromises that appear in wireless cellular networks:

- **Coverage Versus QoS**: Due to propagation losses, the QoS of the users located in the cell edge is usually worse than the one perceived by the users that are close to the base station. A procedure used in the planning and dimensioning of cellular systems is to determine the cell radius depending on the required percentage of the users that should use the minimum allowed Modulation and Coding Scheme (MCS). The trade-off is also evident in this dimensioning procedure, because the higher the minimum QoS requirement, the smaller the cell coverage will be.

- **Capacity Versus Coverage**: Excessive capacity can have a negative impact on the coverage of interference-limited systems. This is the case of Third Generation (3G) systems based on Code Division Multiple Access (CDMA), where the cells shrink when they become heavily loaded (cell breathing phenomenon) [10]. Another aspect is that base stations with high power provide good coverage, but also generate excessive interference to the neighbor cells, which can decrease the overall system capacity.
- **Fairness Versus Coverage**: The random user location in the coverage area and the wireless channel variability cause differences in the channel quality perceived by the users. This quality variability is directly proportional to the cell coverage: the larger the cell size, the higher the variability. Normally, resource allocation algorithms take into account the Channel State Information(CSI) of the users. So, the higher the variability of the users' CSI, the lower the fairness of the corresponding resource allocation.
- **Fairness Versus QoS**: Since the wireless resources are limited, the QoS of the users cannot be improved indefinitely. If the QoS of few users is maximized, the others will feel the lack of resources. This imbalance is translated into a fairness decrease. On the other hand, if a high fairness is assured and consequently the users have more or less the same QoS, the maximum achievable QoS in this situation is lower.
- **Capacity Versus Fairness**: This compromise is also known as the *efficiency versus fairness* trade-off. In order to maximize system capacity, the wireless resources must be allocated in the most efficient way possible. This is accomplished by using opportunistic resource allocation algorithms, which assign the resources to the users who have the best channel conditions with respect to these resources. As commented before, mobile cellular systems present a high variability on the channel quality experienced by the users. The use of opportunistic RRA in order to maximize capacity will inevitably concentrate the resources among the users in good propagation conditions, while the ones in worse channel conditions would starve. This situation is characterized by low fairness. On the other hand, if a high fairness is required, the system is forced to cope with the bad channel conditions of the worst users and allocate resources to them. Since this allocation is not efficient in the resources' point-of-view, the overall system capacity will be degraded.
- **Capacity Versus QoS**: This is also known as the *capacity versus satisfaction* trade-off. A clear compromise between system capacity and user QoS is the fact that the existence of more users in the system decreases the QoS per capita, because less resources would be available for each of the users. Furthermore, the use of opportunistic resource allocation in order to maximize capacity can degrade the QoS of the worst users, which decreases the total percentage of satisfied users in the system. On the other hand, system capacity is decreased if the users with bad channel conditions are contemplated in order to increase total user satisfaction.

Notice that the compromises described above are fundamental trade-offs found in mobile cellular systems and most of them are technology-independent. System design, the deployment of specific technologies and the use of suitable RRA techniques can help the network operators to decrease the gap between these opposing factors. If these compromises cannot be solved in a "win-win" approach, adaptive RRA strategies are still very useful at finding an appropriate trade-off between these objectives.

In this chapter, we are interested in studying and evaluating two of the aforementioned trade-offs: capacity versus fairness and capacity versus QoS.

4.3 Heuristic-Based Resource Allocation Framework

In general, the RRA problems that address the capacity versus fairness and capacity versus QoS trade-offs can be represented in a mathematical form as optimization problems. Basically, optimization problems are composed of an objective, constraints and variables to be optimized. The variable to be optimized in the RRA problems are the resources in mobile networks such as frequency chunks and transmit power. The objective of an optimization problem consists in the aspect of mobile networks that should be improved. Common objectives in RRA are the maximization of transmit data rate and minimization of transmit power. In this chapter we focus on the former objective. Finally, the constraints in optimization problems are restrictions imposed by mobile systems and users. Constraints are able to limit the search space of all possible solutions, i.e., a given RRA solution that leads to an improved objective is infeasible when it does not comply with the problem constraints. We call optimal solution the solution that best improves the objective of the optimization problem and obeys the problem constraints.

The RRA problems studied in this chapter assume that the variable to be optimized is the frequency resource assignment. As the frequency resources are discrete, the optimization problems to be solved belong to the class of combinatorial or integer optimization problems. Furthermore, the mathematical expressions that appear in the objective and constraints of the RRA optimization problems studied here are nonlinear functions of the optimization variable. The combination of combinatorial problems with nonlinear objective and constraints in general turns the task of finding the optimal solution or best RRA solution impractical. Often, the optimal solution can be found only by exhaustive search that enumerates all possible solutions and tests the attained objective in order to find the best one, or other techniques that are able to discard part of the search space but still have exponential-order worst case complexity [29].

In order to find good-enough RRA solutions with reduced computational effort we can use heuristic solutions. Heuristic solutions are simple solutions found by methods, techniques, or algorithms that are conceived based on experience and common sense. In general, the outputs of heuristic methods are suboptimal but acceptable solutions for practical deployments. These solutions are especially suitable for

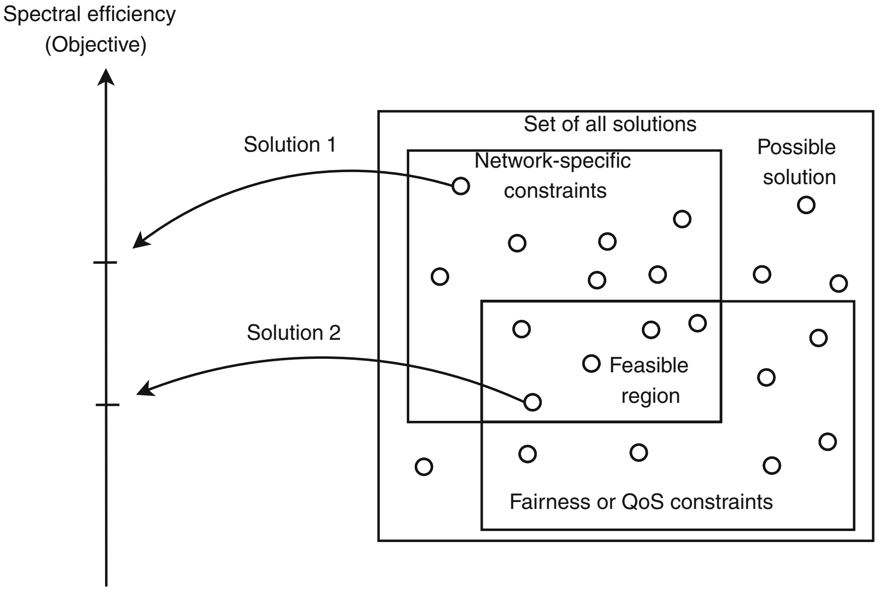

Fig. 4.1 Illustration of the problems to be solved in the capacity versus fairness and capacity versus QoS trade-offs

cases where the optimal solution is hard or impossible to obtain. In these cases, heuristic methods accelerate the problem-solving process and provide us accessible and simple solutions.

The problems to be solved in both capacity versus fairness and capacity versus QoS trade-offs have a common structure and are illustrated in Fig. 4.1. Each possible resource assignment or solution to the problem is represented by circles in this figure. Also, the objective to be pursued is to maximize the total data rate or spectral efficiency that is shown on the left-hand side of Fig. 4.1. Note that each possible solution has a different value for the spectral efficiency. Regarding the problem constraints, we have network specific constraints such as the multiple access constraints, and the QoS or fairness constraints that are directly related to the data service provided to the users by the network. In this figure we illustrate the space of all solutions and two inner spaces that represent the network specific and fairness or QoS constraints. Note that we are interested in the solutions that obey both set of constraints located in the intersection region (feasible region). Therefore, although "solution 1" is able to achieve higher spectral efficiency than "solution 2" in Fig. 4.1, we are interested in the latter solution since it is in accordance with both sets of constraints.

The heuristic framework to solve the presented problem consists in two parts: **Unconstrained Maximization** and **Resource Reallocation**. In the Unconstrained Maximization part we relax the fairness or QoS constraints and solve the problem in order to find the solution that obeys the network-specific constraints that leads to the maximum spectral efficiency. In general, due to the propagation properties of the

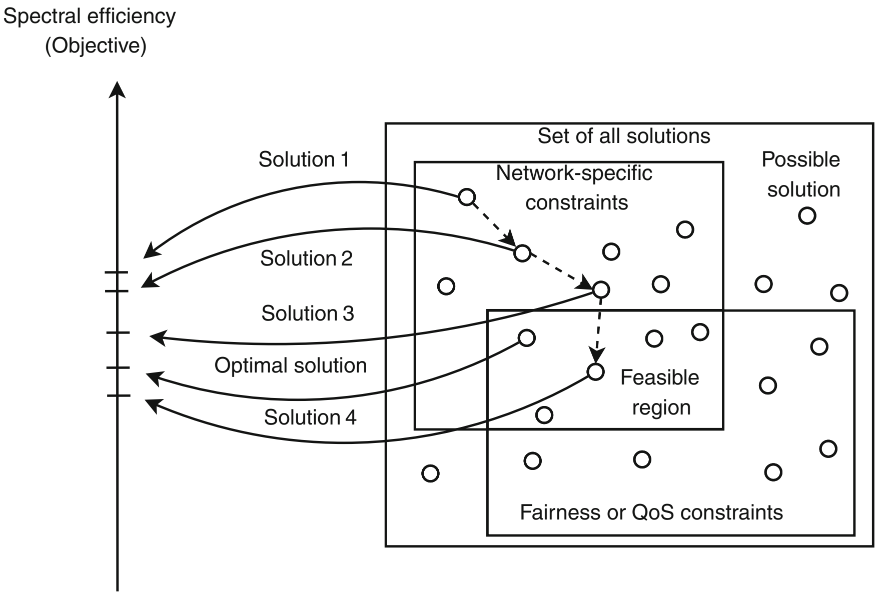

Fig. 4.2 Illustration of the general heuristic framework to solve the problems in the capacity versus fairness and capacity versus QoS trade-offs

wireless medium, only few users that are next to the transmit antennas get most of the system resources. Therefore, we expect that the fairness or QoS constraints are not met in this initial solution. This initial solution is illustrated by "Solution 1" in Fig. 4.2. In the Resource Reallocation part of the proposed framework, we have an iterative phase where the system resources assigned in the Unconstrained Maximization part are exchanged between the users in order to meet either fairness or QoS constraints. In Fig. 4.2, we assume that the final solution, "Solution 4", is found after three iterations or resource reallocations. The reallocations in each iteration are represented by the dashed-line arrows where the intermediate solutions "Solution 2" and "Solution 3" are obtained after the first and second iterations. Note that the main idea in the reallocation procedure is that the loss in spectral efficiency after each iteration should be kept as minimum as possible. At the end of the proposed framework, we expect that the solution found by the proposed heuristic method achieves a spectral efficiency as close as possible to the optimal solution represented in Fig. 4.2 by "Optimal solution".

4.4 Utility-Based Resource Allocation Framework

In communication networks, the benefit of the usage of certain resources, e.g., bandwidth and/or power, can be quantified by using utility theory. This theoretical tool can also be used to evaluate the degree to which a network satisfies service requirements of users' applications, e.g., in terms of throughput and delay.

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

$$\max_{\mathscr{K}_j} \sum_{j=1}^{J} U\left(T_j\left[n\right]\right) \tag{4.1a}$$

$$\text{subject to} \quad \bigcup_{j=1}^{J} \mathscr{K}_j \subseteq \mathscr{K}, \tag{4.1b}$$

$$\mathscr{K}_i \bigcap \mathscr{K}_j = \emptyset, \quad i \neq j, \quad \forall i, j \in \{1, 2, \ldots, J\}, \tag{4.1c}$$

where J is the total number of users in a cell, $\mathscr{K}$ is the set of all resources in the system, $\mathscr{K}_j$ is the subset of resources assigned to user j, K is the total number of resources in the system (subcarriers, codes, etc) to be assigned to the users, and $U\left(T_j\left[n\right]\right)$ is a monotonically increasing utility function based on the current throughput $T_j\left[n\right]$ of the user j in Transmission Time Interval (TTI) n. Constraints (4.1b) and (4.1c) state that the union of all subsets of resources assigned to different users must be contained in the total set of resources available in the system, and that the same resource cannot be shared by two or more users in the same TTI, i.e., these subsets must be disjoint.

The power allocated to the resources could be considered as another optimization variable in the optimization problem (4.1a)–(4.1c). However, this joint optimization problem is very difficult to be solved optimally [8]. Revising the literature, we can find out that most of the sub-optimum solutions 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 Orthogonal Frequency-Division Multiple Access (OFDMA)-based systems that adaptive power allocation provides limited gains in comparison with equal power allocation with much more complexity [8]. Therefore, we consider the simplified optimization problem (4.1a)–(4.1c), which can be solved by a suitable dynamic resource assignment with equal power allocation.

Many RRA policies can be proposed if different utility functions are used. In this study, we are interested in formulating general RRA techniques suitable for controlling the capacity versus fairness and capacity versus QoS trade-offs in a scenario with NRT services.

It is demonstrated in the Appendix that we are able to derive a simplified optimization problem that is equivalent to our original problem. According to the Appendix, the objective function of our simplified problem is linear in terms of the instantaneous user's data rate and given by

$$\max_{\mathscr{K}_j} \sum_{j=1}^{J} U^{'}\left(T_j\left[n-1\right]\right) R_j\left[n\right], \tag{4.2}$$

where $R_j\,[n]$ is the instantaneous data rate of user j and $U^{'}\left(T_j\,[n-1]\right) = \left.\frac{\partial U}{\partial T_j}\right|_{T_j=T_j[n-1]}$ is the marginal utility of user j with respect to its throughput in the previous TTI. The objective function (4.2) characterizes a weighted sum rate maximization problem [11], whose weights are adaptively controlled by the marginal utilities.

The weighted sum rate maximization problem given by (4.2) has a linear objective function with respect to $R_j\,[n]$, whose solution is simple to obtain. Particularly, the Dynamic Resource Assignment (DRA) problem in OFDMA systems, which is the optimization problem (4.1) with subcarriers or physical resource blocks (PRB) as the resources and considering equal power allocation, has a closed form solution when the objective function is given by (4.2) [12, 41]. The user with index $j^{\star}$ is chosen to transmit on resource k in TTI n if it satisfies the condition given by

$$j^{\star} = \arg\max_{j} \left\{U^{'}\left(T_j\,[n-1]\right) r_{j,k}\,[n]\right\}, \tag{4.3}$$

where $r_{j,k}\,[n]$ denotes the instantaneous achievable transmission rate of resource k with respect to user j. Notice that this utility-based resource allocation performs a balance between the QoS-dependent factor $U^{'}\left(T_j\,[n-1]\right)$ and the efficiency-dependent factor $r_{j,k}\,[n]$.

The chosen utility function must be parameterized, for example by a parameter δ, i.e. $U\left(\cdot\right) = U\left(T_j[n], \delta\right)$, in order to allow the control of trade-offs between two objectives. The parameter δ is limited by $\delta^{\min} \leq \delta \leq \delta^{\max}$. On the one hand, $\delta^{\min}$ is associated to the maximization of one objective, for example system capacity. On the other hand, $\delta^{\max}$ is associated to the maximization of the other objective, for example, system fairness or user satisfaction. The adaptation of δ according to a suitable metric and a desired target allows the control of trade-offs.

4.5 Capacity Versus Fairness Trade-Off

In this section, we study the trade-off between capacity and fairness. First, a general definition of the trade-off is presented in Sect. 4.5.1. Next, two RRA techniques are proposed, namely Fairness-based Sum Rate Maximization (FSRM) and Adaptive Throughput-based Efficiency-Fairness Trade-off (ATEF), which are described and evaluated in Sects. 4.5.2 and 4.5.3, respectively. The former is based on the heuristic-based RRA framework described in Sect. 4.3, while the latter is based on the utility-based RRA framework presented in Sect. 4.4. Finally, the conclusions about the study of the capacity versus fairness trade-off are shown in Sect. 4.5.4.

4.5.1 General Definition

It is well known that the scarcity of radio resources is one of the most important characteristics of wireless communications, which demands a very efficient usage of the available resources. Different criteria can be used for resource allocation; for instance, the users that present the best channel quality can be chosen to use the resources. In this case, the efficiency indicator of the resources is the channel quality. The efficiency in the resource usage can be maximized if opportunistic RRA algorithms are used [26]. The opportunism comes from the fact that the resources are dynamically allocated to the users that present the highest efficiency indicator with regard to the radio resources. When the resources have different efficiency indicators to different users (multi-user diversity), the trade-off between efficiency (capacity) and fairness appears. The use of opportunistic resource allocation to exploit these diversities causes unfair situations in the resource distribution.

From a network operator perspective, it is very important to use the channel efficiently because the available radio resources are scarce and the revenue must be maximized. From the users' point of view, it is more important to have a fair resource allocation in a way that they are not on a starvation/outage situation and their QoS requirements are guaranteed.[1] Then the question is: how can the network operator manage this trade-off? In this section we try to answer this question and highlight important clues toward this goal.

In order to better understand the aforementioned trade-off, it is indispensable to define what fairness means. There are two main fairness definitions: resource- or QoS-based [31]. In the former, fairness is related to the equality of opportunity to use network resources, for example, the number of frequency resources a user is allowed to use or the amount of time during which a user is permitted to transmit. In the latter, fairness is associated with the equality of utility derived from the network, e.g., flow throughput. Resource and QoS-based fairness are related to the notion of how equal is the number of resources allocated or how similar is the service quality experienced by the users, respectively. If all users in a given instant approximately have the same number of allocated resources, or perceive more or less the same QoS level, we can say that the system provides a high fairness. On the contrary, if the resources are concentrated among few users, or few of them experience a very good QoS while the others are unsatisfied, the resource allocation can be considered unfair.

Focusing on QoS-based fairness, it is well known that the characteristics and transmission requirements of NRT traffic differ from those of Real Time (RT) data traffics. NRT services, such as World Wide Web (WWW) and File Transfer Protocol

[1] Mobile operators are becoming increasingly more concerned about fairness issues in their networks. It has been observed that most of the Internet traffic is coming from a few end-users, thereby congesting the network for the rest of the users. A small number of customers use their broadband service inappropriately, for example, when sending or downloading very large files, or using 'peer to peer' and file sharing software. In order to solve this problem, network operators are implementing 'Fair Use Policies' in order to manage inappropriate use and make sure the service can be used fairly by everyone.

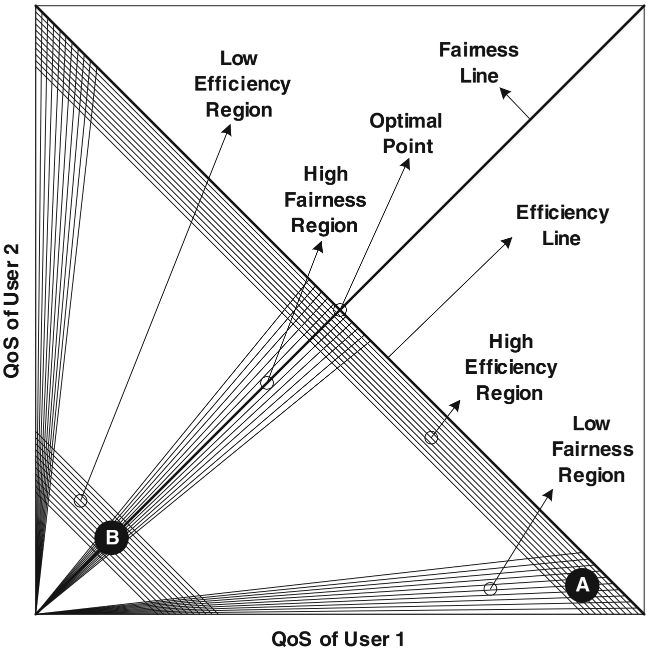

Fig. 4.3 Illustration of different operation points of the trade-off between capacity and fairness in a wireless network with two users

(FTP), are not delay-sensitive but require an overall high throughput. Therefore, rate or throughput can be used as fairness indicators in a scenario with NRT services.

Let us consider a simplified scenario of two users in a wireless system. Figure 4.3 depicts a conceptual view of the trade-off between system capacity and QoS-based user fairness in such a scenario. This conceptual analysis is also valid for the case of resource-based fairness.

The QoS experienced by the two users after the resource allocation is represented by the axes on the figure. One can notice that there are two main lines on the figure: efficiency and fairness. Since the radio resources in the wireless system are limited, the efficiency line delimits a capacity region. The fairness line indicates that the QoS of the users are the same in any point along this line, i.e., the fairness is maximum. The crossing between these lines is the optimal network operation point, which characterizes a resource allocation with maximum efficiency and fairness. In the figure, one can see regions of low and high efficiency and fairness. Wired networks can effectively work near the optimal point due to the implementation of congestion control techniques, such as Transmission Control Protocol (TCP) [6]. However, the frequency and time-varying wireless channel poses significant challenges to the solution of this problem, and the optimal RRA technique that always provides maximum efficiency and fairness in wireless networks is still an open problem.

Referring again to Fig. 4.3, let us assume that user 1 has better channel conditions than user 2. If an opportunistic RRA policy that gives importance only to the efficiency in the resource usage were used, we would have the region marked as “A”. In this case, the majority of the resources were allocated to user 1, which would cause an unfair situation. On the other hand, the region marked as “B” characterizes an RRA policy that provides absolute fairness but causes a significant loss in efficiency since it has to deal with the bad channel conditions of user 2. Therefore, one can observe that in most of the times the optimal point of maximum efficiency and fairness may be unfeasible due to the channel quality of the users.

4.5.2 *Fairness-Based Sum Rate Maximization*

The Fairness-based Sum Rate Maximization (FSRM) technique is based on the heuristic-based RRA framework described in Sect. 4.3 and tries to solve the problem of controlling the trade-off between capacity and fairness. It was first proposed in [33, 35].

This section is organized as follows. Section 4.5.2.1 revises the state of the art about the topic, while the RRA problem to be solved by FSRM is formulated in Sect. 4.5.2.2. The details of the FSRM technique are presented in Sect. 4.5.2.3, and finally, simulation results in Sect. 4.5.2.4 show the comparison between FSRM and other classical RRA techniques.

4.5.2.1 Background

In general, heuristic-based RRA strategies are derived from combinatorial optimization formulations. The optimization-based RRA strategies for OFDMA systems found in the literature typically follow two approaches: *margin adaptive* and *rate adaptive*. The former formulates the dynamic resource allocation with the goal of minimizing the transmitted power with a rate constraint for each user [22, 44]. The latter aims at maximizing the instantaneous data rate with a power constraint [18, 32, 38]. Since the capacity versus fairness trade-off is an explicit consequence of the use of opportunistic rate adaptive RRA algorithms, this latter approach is the one studied in this section.

There are three main classical approaches to cope with the rate adaptive optimization problem: Max–Min Rate (MMR) [21, 32], Sum Rate Maximization (SRM) [18], and Sum Rate Maximization with Proportional Rate Constraints (SRM-P) [38, 45].

The rate adaptive approach was first proposed in [32], where the objective was to maximize the minimum rate of the users. A sub-optimum heuristic solution comprising subcarrier assignment and equal power allocation was proposed. After the resource allocation the users have almost the same rate, which results in the fairest policy in terms of data rate distribution. The MMR optimization problem was reformulated in [21] in order to be solved by Integer Programming techniques. Notice that such a policy is able to maximize fairness at the expense of degraded system capacity (see region “B” in Fig. 4.3).

Reference [18] presented the solution of the SRM problem, which is the classical opportunistic rate adaptive policy. SRM maximizes the system capacity regardless of the QoS of the individual users. The subcarriers are assigned to the users who have the highest channel quality, and next the power is allocated among the subcarriers following the waterfilling procedure [30]. This resource allocation ignores the users with bad channel conditions, who may not receive any resources, and benefits the users close to the base station. According to Fig. 4.3, this policy would be located in region “A”.

The SRM-P optimization problem attempts to be a trade-off solution between system capacity and user fairness [38]. The same objective function of the problem described in [18] was considered and a new optimization constraint of rate proportionality for each user was added. This constraint aims to rule the rate distribution in the system. This new optimization problem is suitable for a scenario where there are different service classes with different proportional rate requirements. The solution was divided into two steps: a sub-optimum subcarrier assignment based on [32] and an optimal power allocation. The SRM-P problem was further addressed by [45], which linearized the power allocation problem avoiding the solution of a set of nonlinear equations that was required by the solution proposed in [38].

In this section, a new proposed fairness/rate adaptive policy called FSRM is described. It is a generalization of a classical rate adaptive policy SRM found in the literature [18].

4.5.2.2 Problem Formulation

The generalization of the classical SRM policy takes into account a new way to control the trade-off between system capacity and fairness. This control is applied on a cell fairness index and is formulated as a new constraint in the optimization problem.

The considered RRA optimization problem is formulated as follows:

$$\max_{\mathbf{X}} \sum_{j=1}^{J} \sum_{k=1}^{K} r_{j,k} x_{j,k} \tag{4.4}$$

$$\text{subject to } x_{j,k} = \{0, 1\}, \quad \forall j \in \mathscr{J} \text{ and } \forall k \in \mathscr{K}, \tag{4.5}$$

$$\sum_{j=1}^{J} x_{j,k} = 1, \quad \forall k \in \mathscr{K}, \tag{4.6}$$

$$\Phi^{\text{cell}} = \Phi^{\text{target}}, \tag{4.7}$$

where J and K are the total number of active users and available frequency resources, respectively; $\mathscr{J}$ and $\mathscr{K}$ are the sets of users and resources, respectively; $\mathbf{X}$ is a

$J \times K$ assignment matrix whose elements $x_{j,k}$ assume the value 1 if the resource k is assigned to the user j and 0 otherwise; Φ^{cell} is the instantaneous Cell Fairness Index (CFI); and Φ^{target} is the Cell Fairness Target (CFT), i.e., the desired target value of the CFI.

Constraints (4.5) and (4.6) say that each frequency resource must be assigned to only one user at any instant of time. A new fairness control mechanism is explicitly introduced into the optimization problem of the fairness/rate adaptive policy by means of the fairness constraint (4.7). A short-term (instantaneous) fairness control can be achieved, because this constraint requires that the instantaneous CFI Φ^{cell} must be equal to the CFT Φ^{target} at each TTI.

The fairness/rate adaptive optimization (4.4)–(4.7) is a nonlinear combinatorial optimization problem, because it involves an integer variable $x_{j,k}$ and a nonlinear constraint (4.7), as will be explained in the following. This problem is not convex because the integer constraint (4.5) makes the feasible set nonconvex.

Constraint (4.7) is the main novelty in comparison with the classical SRM rate adaptive policy. It has a deep impact on the design of the RRA technique used to solve the optimization problem (4.4)–(4.7), as will be shown in Sect. 4.5.2.3. In order to better comprehend the importance of this constraint, let us further elaborate on the concept of the fairness index.

It is assumed that each user has a rate requirement R_j^{req} that will indicate whether this user is satisfied or not. In order to evaluate how close the user's transmission rate is from its rate requirement, the UFI is defined as

$$\phi_j = \frac{R_j}{R_j^{\text{req}}}, \tag{4.8}$$

where R_j is the instantaneous transmission rate of user j.

In order to measure the fairness in the rate distribution among all users in the cell, the CFI is calculated by

$$\Phi^{\text{cell}} = \frac{\left(\sum_{j=1}^{J} \phi_j\right)^2}{J \sum_{j=1}^{J} \left(\phi_j\right)^2}, \tag{4.9}$$

where J is the number of users in the cell and ϕ_j is the UFI of user j given by (4.8). This proposed CFI is a particularization of the well-known Jain's fairness index proposed by Jain et al. in [16]. Notice that $1/J \leq \Phi^{\text{cell}} \leq 1$. On one hand, the worst allocation occurs when $\Phi^{\text{cell}} = 1/J$, which means that all resources were allocated to only one user. On the other hand, a perfect fair allocation is achieved when $\Phi^{\text{cell}} = 1$, which means that the instantaneous transmission rates allocated to all users are equally proportional to their requirements R_j^{req} (all UFIs are equal).

The objective function (4.4) is the same of the classical rate adaptive SRM policy [18]. The constraint (4.7) does not exist in the original SRM problem. Therefore, SRM is a pure channel-based opportunistic policy, where the resources are allocated to the users with better channel conditions, which maximizes the cell throughput.

However, such a solution is extremely unfair because the other users with worse channel conditions are neglected.

Although the objective function of the proposed FSRM policy seeks the maximization of capacity, the fairness constraint (4.7) acts as a counterpoint, provoking the explicit appearance of a trade-off. FSRM tries to answer the following question: *How can a given fairness level be achieved while keeping the system capacity as high as possible?* Guided by this criterion, the FSRM policy can achieve different fairness levels and draw a complete capacity-fairness curve. We will answer in the next section this design question.

4.5.2.3 Algorithm Description

The underlying concept behind the FSRM policy is that resource allocation can be based on two possible approaches:

- ***Resource-centric/efficiency-oriented***: the RRA policy allows the resource to "choose" who is the best user to use it;
- ***User-centric/fairness-oriented***: the RRA policy allows the user to choose which is the most adequate resource to him/her.

Whether the RRA policy uses the former, the latter, or both approaches, will determine its ability to control the intrinsic trade-off between resource efficiency and user fairness found in wireless networks.

Three "actors" play an important role in the proposed technique: the "richest" user (the one with the maximum proportional rate), the "poorest" user (the one with the minimum proportional rate), and the resource.

The FSRM policy is able to increase the fairness in the system. This process is illustrated in Fig. 4.4. In this hypothetical example, we have the distribution of the QoS among 20 users. The user IDs are ordered in such a way that the users with best QoS are given IDs around 10, and the users with worst service quality are given the extreme IDs (close to 1 or 20). A QoS distribution depicted by the dashed curve shows an unfair resource usage. If fairness is to be increased from that point, the resources, and consequently the QoS, should be divided more equally among the users. This is accomplished removing resources from the rich and giving them to the poor. The solid curve is an example of a fair QoS distribution.

The fairness/rate adaptive problem formulated in (4.4)–(4.7) is a nonconvex optimization problem, which makes it very difficult to find the optimum solution. This work proposes an RRA technique able to solve the proposed fairness/rate adaptive problem in a sub-optimum way. Based on the heuristic-based framework described in Sect. 4.3, the FSRM policy is implemented by a sequence of two RRA algorithms, as explained in the following.

1. ***Unconstrained Maximization***: An initial fairness level (CFI) is achieved after the execution of the DRA algorithm of the classical SRM policy. Therefore, an initial positioning on the capacity-fairness plane is determined.

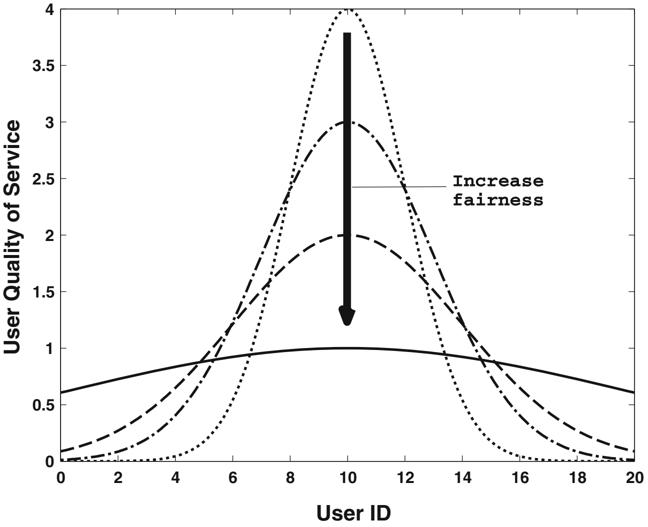

Fig. 4.4 Relation between QoS distribution and fairness adaptation

2. ***Resource Reallocation***: The initial CFI is, in general, low because SRM is an unfair policy. Thus, in order to meet the desired CFT, fairness must be increased by means of resource reallocations among users. Fairness variation is only possible if resources are moved between different users. The first step is to decide from which user a resource will be removed. Next, a small amount of resource (resource with worst channel quality) is removed from this user. Finally, this resource is given to the user that can take the most benefit of it, or in other words, this resource is given to the user that can use it in the most efficient way. This means to assign the removed resource to the user that has the highest channel gain on it. Hopefully, after this procedure an accurate approximation of the CFT is achieved.

After the Unconstrained Maximization and Resource Reallocation parts, we perform Equal Power Allocation (EPA), i.e., the power is divided equally among the resources.

In the Unconstrained Maximization part, a resource should be assigned to only one user who has the best channel gain for that resource, as indicated in Algorithm 7.

Algorithm 7 Unconstrained Maximization part of the FSRM technique

```
Initialization
$\mathscr{J} \leftarrow \{1, 2, 3, \cdots J\}$ {Users set}
$\mathscr{K} \leftarrow \{1, 2, 3, \cdots K\}$ {Resources set}
for all $j \in \mathscr{J}$ and $k \in \mathscr{K}$ do
    $x_{j,k} \leftarrow 0$ {Reset connection matrix}
    $\mathscr{K}_j \leftarrow \varnothing$ {Reset user's resources subset}
end for
Resource assignment
for all $k \in \mathscr{K}$ do
    $j^* \leftarrow \arg\max_j\{\gamma_{j,k}\}$ {Find user with maximum SNR on resource $k$}
    $x_{j^*,k} \leftarrow 1$ {Set the connection}
    $\mathscr{K}_{j^*} \leftarrow \mathscr{K}_{j^*} \bigcup \{k\}$ {Update user's resources subset}
end for
```

The detailed pseudo-code of the Resource Reallocation part of FSRM is presented in Algorithm 8 while its flowchart is depicted in Fig. 4.5.

Algorithm 8 is an iterative heuristic algorithm that adapts the CFI by means of a resource reallocation procedure. Initially, the CFI according to (4.9) is calculated. As previously mentioned, the initial DRA procedure is performed by the classical SRM technique, which in general provides low levels of CFI. Therefore, it is most likely that the initial CFI provided by SRM is lower than the desired CFT value Φ^{target}. Based on that, the fairness-based DRA algorithm of the FSRM technique must increase the fairness until a value close to Φ^{target}. This is accomplished by an iterative procedure that stops when the CFT is achieved. Details are given below.

1. Select a user j^* from the set of available users in such a way that fairness can be increased if a resource is removed from this user. This can be accomplished by taking resources from the user with maximum proportional rate (richest user) and give them to other users.
2. From the subset of resources assigned to user j^*, select the one with the minimum Signal-to-Noise Ratio (SNR) with respect to this user (resource k^*).
3. Find the user j^{**} (different of user j^*) who can be most benefited from the resource reallocation. This is the user with maximum SNR on resource k^*.
4. Remove resource k^* from user j^* and give it to user j^{**} (resource reallocation). The rates and subsets of assigned resources of users j^* and j^{**} must be updated.
5. Re-calculate the new value of CFI and repeat the process until the CFT is achieved.

Algorithm 8 Resource Reallocation part of the FSRM technique

```
Initialization
$\mathscr{J} \leftarrow \{1, 2, 3, \ldots, J\}$; $\mathscr{K} \leftarrow \{1, 2, 3, \ldots, K\}$; $\mathscr{B} \leftarrow \varnothing$ {Initialize users set, resources set and blocked resources subset}
for all $j \in \mathscr{J}$ and $k \in \mathscr{K}$ do
  $x_{j,k} \leftarrow 0$ {Reset connection matrix}
  $\mathscr{K}_j \leftarrow \varnothing$ {Reset resources subset of each user}
  $\mathscr{Q}_k \leftarrow \varnothing$ {Reset blocked users subset of each resource}
end for
Resource reallocation to increase fairness
Calculate $\Phi^{\text{cell}}$ according to (4.9)
if $\Phi^{\text{cell}} < \Phi^{\text{target}}$ then {Increase fairness}
  while $\Phi^{\text{cell}} < \Phi^{\text{target}}$ do
    $j^* \leftarrow \arg\max_j \{R_j / R_j^{\text{req}}\}, \forall j \in \mathscr{J}$ {Find user with maximum proportional rate}
    $k^* \leftarrow \arg\min_k \{\gamma_{j^*,k}\}, \forall k \in \mathscr{K}_{j^*}$ and $\forall k \notin \mathscr{B}$ {Find available resource assigned to user $j^*$ with minimum SNR}
    $\mathscr{Q}_{k^*} = \mathscr{Q}_{k^*} + \{j^*\}$ {Update subset of blocked users for resource $k^*$}
    $j^{**} \leftarrow \arg\max_j \{\gamma_{j,k^*}\}, \forall j \in \mathscr{J}$ and $\forall j \notin \mathscr{Q}_{k^*}$ {Find available user with maximum SNR on resource $k^*$}
    if $j^{**}$ exists then
      Remove resource $k^*$ from user $j^*$ and give it to $j^{**}$; update $R_{j^*}$, $R_{j^{**}}$, $\mathscr{K}_{j^*}$ and $\mathscr{K}_{j^{**}}$
    else
      $\mathscr{B} = \mathscr{B} + \{k^*\}$ {Update set of blocked resources}
    end if
    Re-calculate $\Phi^{\text{cell}}$ according to (4.9)
  end while
end if
```

During the fairness increase procedure, the resources have more freedom to move between the users. In order to avoid ping-pong effects, the resource k^* cannot return to its original owner (user j^*) in subsequent iterations of the algorithm. Due to this restriction, after some iterations, the resource k^* may not have any user eligible to receive it. In this case, this resource is removed from the set of available resources.

As can be noticed, the way the resources are reallocated in the FSRM policy guarantees that a desired CFT is met while maximum capacity is achieved.

4.5.2.4 Simulation Results

In this section, we compare the performance of the proposed FSRM technique with three classical rate adaptive techniques, namely SRM [18], SRM-P [45] and MMR [32]. Table 4.1 shows the parameters considered in the system-level simulations,

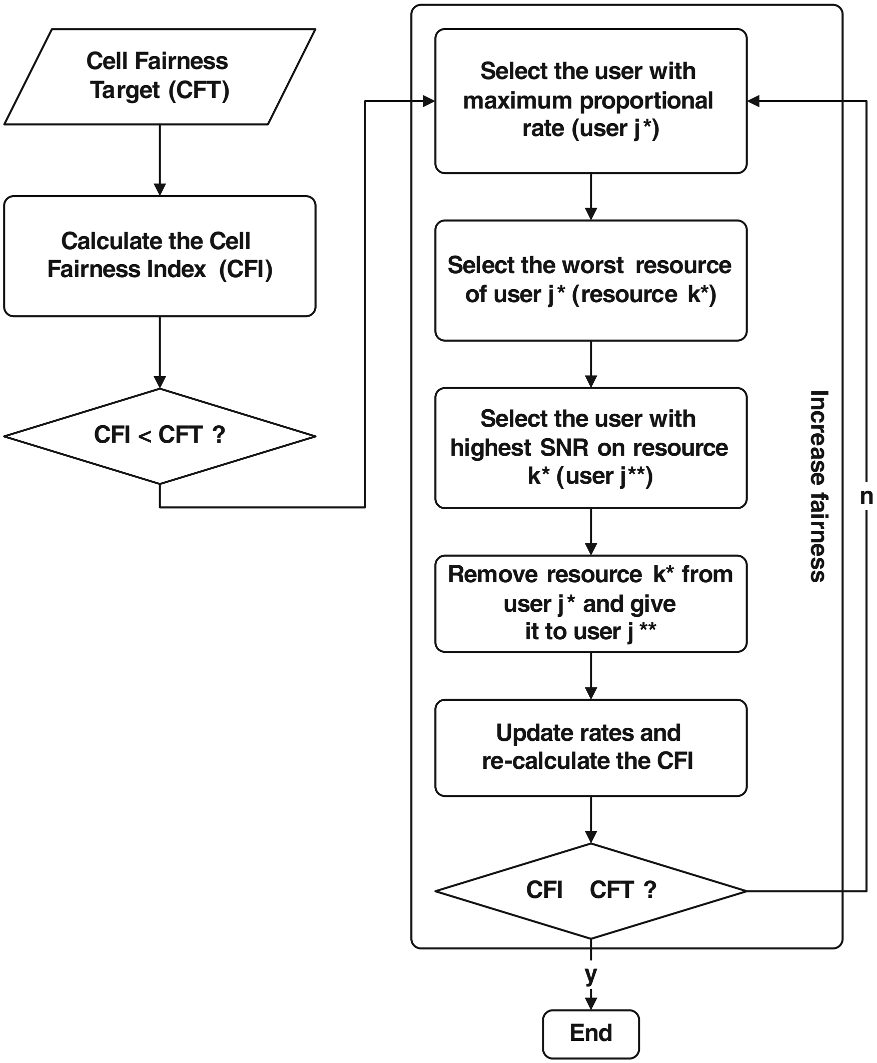
Cell Fairness Target (CFT)
Calculate the Cell Fairness Index (CFI)
CFI < CFT ?
Select the user with maximum proportional rate (user j*)
Select the worst resource of user j* (resource k*)
Select the user with highest SNR on resource k* (user j**)
Remove resource k* from user j* and give it to user j **
Update rates and re-calculate the CFI
CFI CFT ?
Increase fairness
n
y
End

Fig. 4.5 Flowchart of the Resource Reallocation part of the FSRM technique

where the main characteristics of a single-cell[2] Long Term Evolution (LTE)-based system were modeled.

[2] There is a trend in next generation mobile communication networks that RRA techniques should be executed in the base stations, not in the radio network controllers anymore, as was the case for 3G systems. Moreover, all the information needed by the RRA techniques proposed in this chapter is available in each base station locally. The reasons explained above support our decision of evaluating the RRA techniques in a single-cell scenario. Finally, we expect that the performance evaluation on a multi-cell scenario would present only a performance degradation for all studied

Table 4.1 Simulation parameters for the evaluation of the FSRM technique

Parameter	Value
Maximum BS transmission power	1 W
Cell radius	500 m
MT speed	Static
Carrier frequency	2 GHz
Number of subcarriers	192
Effective subcarrier bandwidth[a]	14 kHz
Path loss[b]	$L = 128.1 + 37.6 \log_{10} d$
Log-normal shadowing standard deviation	8 dB
Small-scale fading	Typical urban (TU)
AWGN power per subcarrier	-123.24 dBm
BER requirement	10^{-6}
Link adaptation	Shannon capacity with SNR gap [40]
Transmission time interval (TTI)	0.5 ms
NRT traffic model	Full buffer
User satisfaction requirement (R_j^{req})	512 kbps
Proportional rate requirements[c]	$1/J$
Target CFI (Φ^{target})	Variable
Number of independent snapshots	10,000

[a] The effective subcarrier bandwidth takes into account the signaling overhead
[b] Distance d in km
[c] In the SRM-P technique, we considered that all users had the same proportional rate requirements, which is given by $1/J$, where J is the number of users

Figure 4.6 depicts the mean CFI averaged over all snapshots as a function of the number of users for all classical rate adaptive algorithms and the fairness/rate adaptive technique proposed in this work. It can be observed that the SRM technique, which uses a pure opportunistic policy that allocates the resources only to the best users, is the one that presents the highest rates. However, this benefit comes at the expense of a very unfair distribution of the QoS among the users, since many of them do not have the opportunity to transmit due to the lack of resources. Notice that the higher the number of users, the lower the fairness provided by SRM. This is due to the multi-user diversity which is fully exploited by the opportunistic resource allocation of the SRM technique. At the other extreme we have the SRM-P and MMR techniques, where the transmission rates of the users are more equalized, and therefore the fairness in the system is higher. However, the transmission rates of the users are also lower, which characterizes a capacity loss. Notice that the users' rates provided by SRM-P are slight higher than the ones achieved with MMR, because the former takes extra actions that allow a better utilization of the resources [45].

(Footnote 2 continued)
techniques due to inter-cell interference, which would not change the conclusions taken from their relative comparison.

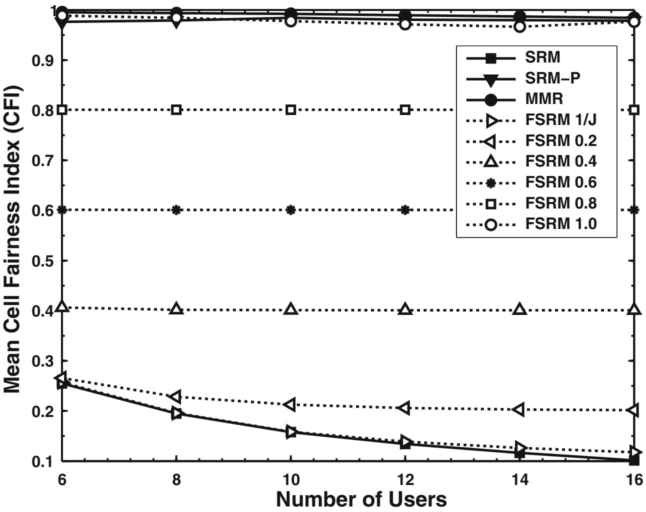

Fig. 4.6 Mean cell fairness index as a function of the number of users for the classic (*solid lines*) and the FSRM technique (*dashed line*)

Figure 4.6 also shows that the FSRM technique is successful at guaranteeing the fairness targets, which were $[1/J, 0.2, 0.4, 0.6, 0.8, 1.0]$, where J is the total number of NRT users in the cell. It can be observed that for lower system loads that FSRM is not able to exactly meet very low CFTs (see for instance 6 or 8 users and CFT=$1/J$ in Fig. 4.6). This happens due to two interrelated factors: (1) the performance of the FSRM technique is lower-bounded by the classic SRM policy; and (2) the multi-user diversity is not sufficient with a low number of users. As explained in Sect. 4.5.2.3, the initial resource assignment performed by the classic SRM is the first step of the FSRM technique. If the CFT is larger than the initial CFI, fairness should be increased, and resource reallocations are done in the reallocation part of the heuristic-based framework. This explains the lower bound given by SRM. On the other hand, the performance of the FSRM strategy converges to the performance of the classic MMR for extremely high values of CFT, since the latter presents the highest values of CFI.

Figure 4.7 compares the performance of the proposed FSRM strategy with the classical rate adaptive techniques in terms of total cell throughput, which is the efficiency indicator that we use in this analysis. As a consequence of the trade-off, we have that the total cell throughput is inversely proportional to the CFT. As can be seen in Fig. 4.7, the higher the CFT, the lower the total cell throughput. Regarding the classical strategies, SRM provides much better results in terms of system capacity than SRM-P and MMR. SRM-P also shows slightly better results than MMR due to its resource assignment algorithm that seeks the maximization of the capacity whenever possible. One can see the inverse proportion between capacity and CFT by the performance of the FSRM technique.

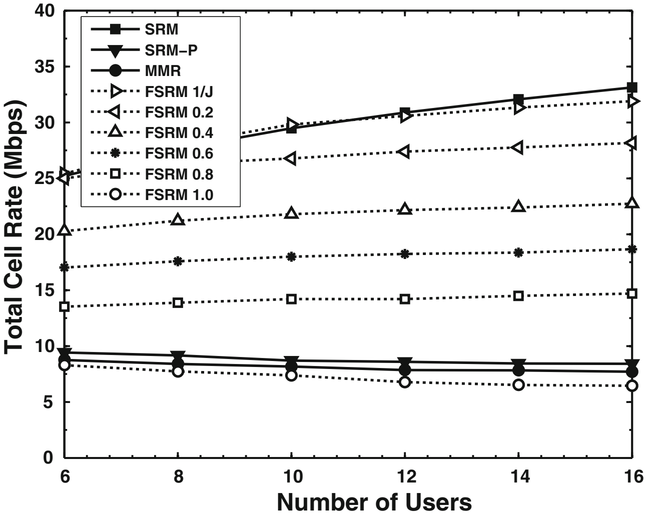

Fig. 4.7 Total cell rate as a function of the number of users for the classic (*solid lines*) and the FSRM technique (*dashed line*)

The best way to evaluate the trade-off between resource efficiency and user fairness is plotting the 2D capacity-fairness plane. The chosen efficiency and fairness indicators are the total cell data rate (capacity) and cell fairness index, respectively. Figure 4.8 summarizes the most relevant aspects discussed so far. It compares the performance of the classical rate adaptive techniques (SRM, SRM-P and MMR), which are indicated as single markers, and the generalized fairness/rate adaptive strategy (FSRM), which is indicated as solid line. In order to plot the capacity-fairness plane, the number of users must be fixed, which in this case is 16.

The classical rate adaptive techniques are represented as single points in the capacity-fairness plane because they represent static policies, i.e., each policy provides only one trade-off operation point. SRM provides maximum capacity at the expense of very poor fairness among users, while SRM-P and MMR are very fair in the rate distribution (CFI close to one) but as a consequence they achieve much lower system capacity.

On the other hand, the FSRM technique is able to achieve a desired cell fairness target thanks to a new fairness constraint in the optimization problem. It is able to cover the whole path between extreme points in the capacity-fairness plane (classical rate adaptive points), drawing a complete curve. One can observe that the performance of the proposed fairness/rate adaptive strategy converges to the results of the classical rate adaptive techniques in both extremes of the CFI, which are $1/J$ and 1.

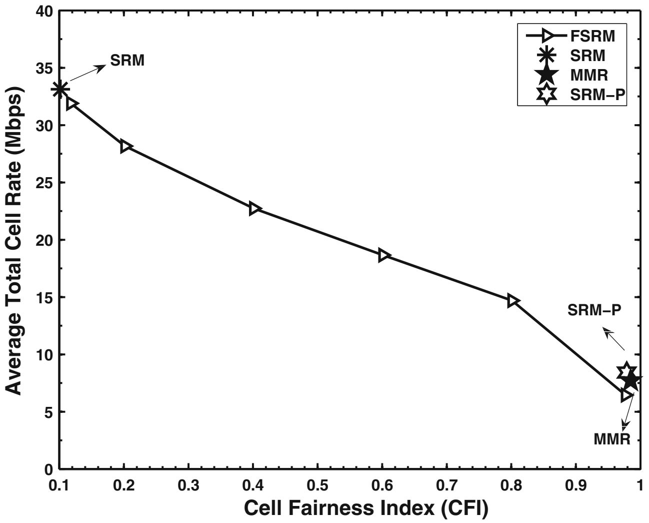

Fig. 4.8 Capacity-Fairness plane for the classical and FSRM techniques

4.5.3 Adaptive Throughput-Based Efficiency-Fairness Trade-Off

The Adaptive Throughput-based Efficiency-Fairness Trade-off (ATEF) technique is based on the utility-based RRA framework described in Sect. 4.4 and tries to solve the problem of controlling the trade-off between capacity and fairness. It was first proposed in the seminal works [33, 34].

This section is organized as follows. Section 4.5.3.1 presents some works related to the topic, while the RRA problem to be solved is formulated in Sect. 4.5.3.2. The proposed technique is described in Sect. 4.5.3.3, while Sect. 4.5.3.4 shows the performance evaluation of ATEF and other classical RRA techniques.

4.5.3.1 Background

Most of the works that proposed packet scheduling (PS) algorithms to effect a compromise between efficiency and fairness among NRT flows [4, 5, 13, 46] are based on the Proportional Fair (PF) PS algorithm proposed in [43] for High Data Rate (HDR) CDMA systems. However, there are some works [7, 23] that used different approaches. The former introduced a PS algorithm with a fairness controlling parameter that accounts for any intermediate policy between the instantaneous fairness and the opportunistic policies, while the latter evaluated a scheduling algorithm whose priority function is a linear combination between instantaneous channel capacity and the average throughput. As a generalization of the PF criterion, we can highlight the weighted α-proportional fairness PS algorithm, which is also known as the alpha-rule and was initially proposed by [28] and later used in [20]. The idea behind this algorithm is to embody a number of fairness concepts, such as rate maximization, proportional fairness and max–min fairness, by varying the values of the parameter α and the weight parameter.

A more general class of RRA algorithms is based on utility fairness. Utility fairness is defined with a utility function that composes the optimization problem, where the objective is to find a feasible resource allocation that maximizes the utility function specific to the fairness concept used. Some examples of utility functions can be found in [9, 19, 39]. There is a general family of utility functions that were presented and/or evaluated in [36, 37, 42] that includes the weighted α-proportional fairness algorithm as a special case. Some works followed a similar approach, but using different utility functions, e.g., [3, 40, 41].

The utility fairness concept is used in this section to propose the utility-based alpha-rule, which is a generalized parametric RRA framework suitable for NRT services that can balance efficiency and fairness in wireless systems according to the network operator's interest. This framework is composed of dynamic resource assignment algorithm and can be designed to work as any of well-known classical RRA policies by adjusting only one parameter in their corresponding parametric structures.

4.5.3.2 Problem Formulation

We consider a family of utility functions based on throughput of the form presented in (4.10) below [37].

$$U\left(T_j\left[n\right]\right) = \frac{T_j\left[n\right]^{1-\alpha}}{1-\alpha} \tag{4.10}$$

where $\alpha \in [0, \infty)$ is a nonnegative parameter that determines the degree of fairness.

Figure 4.9 depicts, for different values of α, the utility and marginal utility functions. A family of concave and increasing utility functions is shown, which represents that the satisfaction of the users increases when their throughput increases. The marginal utilities play an important role in the DRA algorithm, as explained in Sect. 4.4. Let us consider a utility-based weight of user j as its marginal utility, i.e., $w_j = U^{'}\left(T_j\left[n-1\right]\right)$. The higher the weight, the higher the priority of the user to get a resource. The marginal utility functions also show that users experiencing poor QoS (low throughput) will have higher priority in the resource allocation process. And such priority is higher when α increases. Therefore, one can conclude that when α increases, the users with poorest QoS are benefited, and so the fairness in the system becomes stricter.

Taking into account (4.10), the expression of the weight w_j becomes

$$w_j = U^{'}\left(T_j\left[n-1\right]\right) = \frac{1}{T_j\left[n-1\right]^{\alpha}}. \tag{4.11}$$

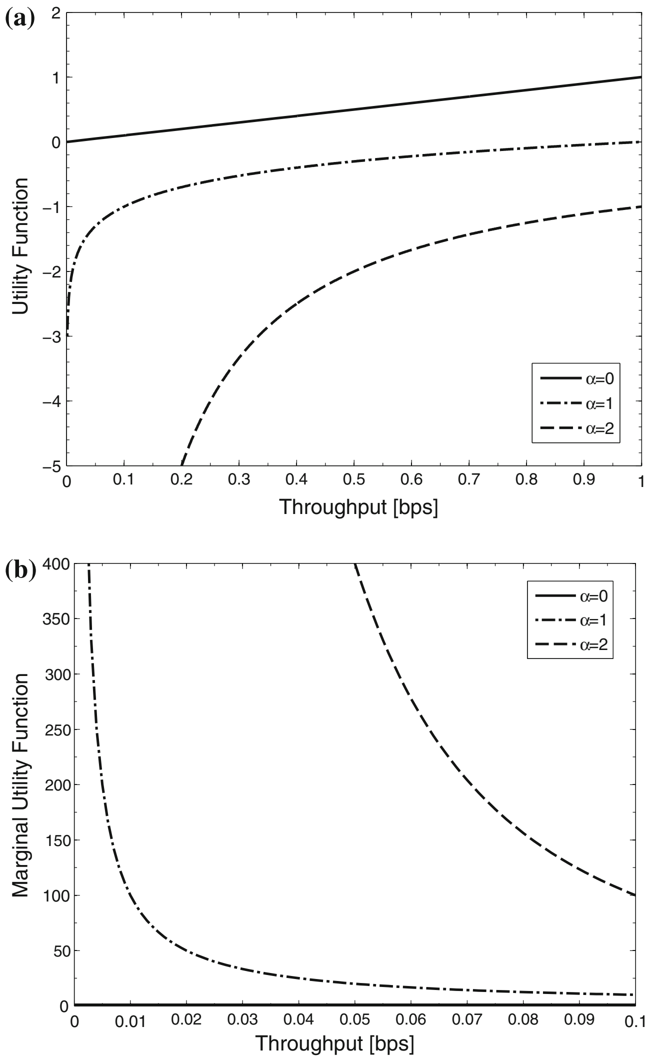

Fig. 4.9 Family of utility functions used in the utility-based alpha-rule framework. **a** Utility functions. **b** Marginal utility functions

The corresponding DRA algorithm, which is given by (4.3), must use the particular expression of w_j presented in (4.11).

Depending on the value of the fairness controlling parameter α, the alpha-rule framework presented above can be designed to work as different RRA policies, achieving different performances in terms of resource efficiency and throughput-based fairness. The main characteristics of the alpha-rule framework and the four particular RRA policies contemplated by this framework are presented in Table 4.2.

Table 4.2 Features of the utility-based alpha-rule framework: $U\left(T_j\left[n\right]\right) = \frac{T_j[n]^{1-\alpha}}{1-\alpha}$

Policies	Parameter α	Weight w_j	Characteristics
RM	0	1	High resource efficiency and low throughput-based fairness
PF	1	$\dfrac{1}{T_j\left[n-1\right]}$	Static trade-off between resource efficiency and throughput-based fairness
MMF	$\alpha \to \infty$	$\lim\limits_{\alpha\to\infty} \dfrac{1}{T_j\left[n-1\right]^{\alpha}}$	Low resource efficiency and high throughput-based fairness
ATEF	Adaptive	$\dfrac{1}{T_j\left[n-1\right]^{\alpha}}$	Dynamic trade-off between resource efficiency and throughput-based fairness

The first three RRA policies are well-known classical policies, namely Rate Maximization (RM) [18] (also known as SRM), Max–Min Fairness (MMF) [37] and Proportional Fair (PF) [19]. The novel adaptive policy ATEF is described in detail in the following.

4.5.3.3 Algorithm Description

The ATEF policy is an adaptive version of the utility-based alpha-rule. It aims to achieve an efficient trade-off between resource efficiency and throughput-based fairness planned by the network operator in a scenario with NRT services. This is done by means of the adaptation of the fairness controlling parameter α in the utility function presented in (4.10). The user priority in the resource allocation is very sensitive to the value of α, as can be seen in Fig. 4.9. So small values are sufficient to provide the desired fairness degrees on the ATEF DRA algorithm.

The ATEF policy is based on the definition of a user fairness index (UFI) ϕ_j, which is based on throughput and calculated for each user in the cell. The instantaneous UFI is defined as

$$\phi_j\left[n\right] = \frac{T_j\left[n-1\right]}{T_j^{\text{req}}}, \tag{4.12}$$

where T_j^{req} is the throughput requirement of user j.

Next, we define a cell fairness index considering all users connected to it as follows:

$$\Phi^{\text{cell}}\left[n\right] = \frac{\left(\sum_{j=1}^{J}\phi_j\left[n\right]\right)^2}{J\sum_{j=1}^{J}\left(\phi_j\left[n\right]\right)^2}, \tag{4.13}$$

where J is the number of users in the cell. This proposed CFI is based on the well-known Jain's fairness index [16]. This fairness index was also used by the heuristic-based FSRM technique, whose formulation is presented in Sect. 4.5.2.2.

The objective of the ATEF policy is to assure that the instantaneous CFI $\Phi^{\text{cell}}[n]$ is kept around a planned value Φ^{target}, i.e., a strict throughput-based fairness distribution among the users is achieved. Therefore, the ATEF policy adapts the parameter α in the utility-based alpha-rule framework in order to achieve the desired operation point. Therefore, the new value of the parameter α is calculated using a feedback control loop of the form:

$$\alpha[n] = \alpha[n-1] - \eta\left(\Phi^{\text{filt}}[n] - \Phi^{\text{target}}\right) \tag{4.14}$$

where the parameter η is a step size that controls the adaptation speed of the parameter α; $\Phi^{\text{filt}}[n]$ is a filtered version of the CFI $\Phi^{\text{cell}}[n]$ using an exponential smoothing filtering, which is used to smooth time series with slowly varying trends and suppress short-run fluctuations; and Φ^{target} is the CFT, i.e. the desired value for the CFI.

The ATEF technique is an iterative and sequential process. At each TTI, the steps indicated in Fig. 4.10 are executed. This process is executed indefinitely. After some iterations (TTIs), the ATEF technique reaches a stable convergence of the fairness pattern defined by the target CFI. The simplicity of the ATEF policy makes it a robust and reliable way to control the trade-off between capacity and fairness. By keeping the cell fairness around a planned target value, the network operator can have a stricter control of the network QoS and also have a good prediction about the performance in terms of system capacity.

4.5.3.4 Simulation Results

In this section, the performance of the utility-based alpha-rule is evaluated by means of system-level simulations. The performance of the ATEF policy is compared to the three classic RRA policies (MMF, PF and RM). In this simulation scenario, several CFTs were considered for the ATEF policy, namely $\Phi^{\text{target}} = [1/J, 0.2, 0.4, 0.6, 0.8, 1.0]$. The simulations took into account the main characteristics of an LTE-based cellular system. The general simulation parameters are the same as used for the evaluation of the FSRM technique in Sect. 4.5.2.4 (see Table 4.1). Table 4.3 shows the specific simulation parameters used in the performance evaluation of the utility-based alpha-rule framework.

The throughput-based CFI calculated by (4.13) averaged over all simulation snapshots is depicted in Fig. 4.11 for various system loads. It can be observed that ATEF is successful at achieving its main objective, which is to guarantee a strict fairness distribution among the users. This is achieved due to the feedback control loop that dynamically adapts the parameter α of the alpha-rule framework.

Notice that the structure of the utility-based alpha-rule framework bounds the performance of the ATEF policy between the performances of the RM and MMF

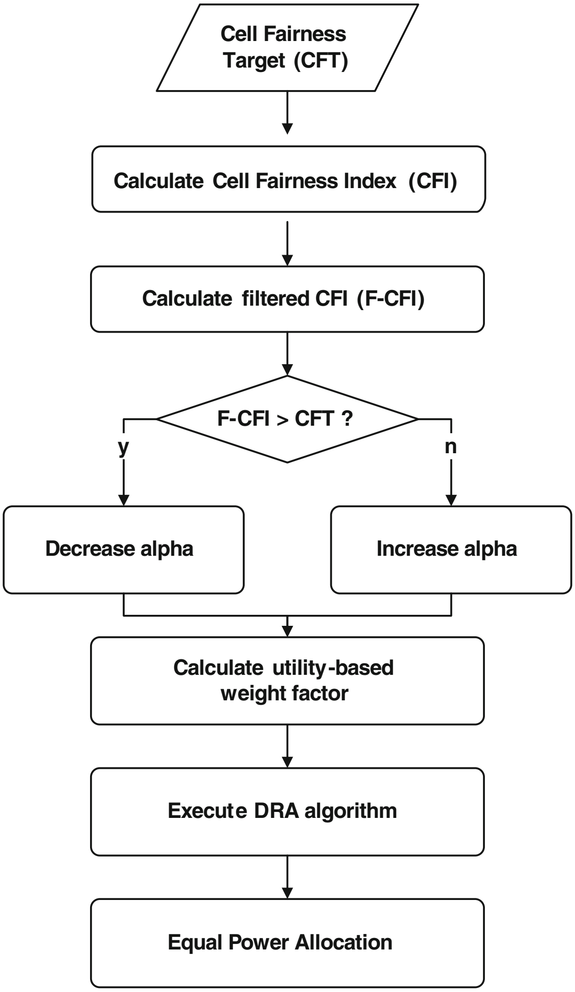
Cell Fairness Target (CFT)
Calculate Cell Fairness Index (CFI)
Calculate filtered CFI (F-CFI)
F-CFI > CFT ?
y
n
Decrease alpha
Increase alpha
Calculate utility-based weight factor
Execute DRA algorithm
Equal Power Allocation

Fig. 4.10 Flowchart of the ATEF technique

policies. According to Table 4.2, the extreme values of the parameter α are 0 and ∞ (in practice a very large number), which correspond to RM and MMF policies, respectively. We considered in the simulations a range of values from 0 to 10 for the adaptation of the parameter α by the ATEF policy. Notice that this upper limit of $\alpha = 10$ was sufficient for the ATEF policy configured with $\Phi_{\text{target}} = 1.0$ to be very close to the performance of the MMF policy. On the other extreme, it is clear that RM works as a lower bound for ATEF configured with $\Phi_{\text{target}} = 1/J$.

Regarding the classic RRA policies, as expected, MMF provided the highest fairness, very close to the maximum value of 1, while RM was the unfairest strategy with a high variance on the fairness distribution for high cell loads. PF presented a good intermediate fairness distribution. From this fairness analysis, it can be concluded

Table 4.3 Specific simulation parameters for the evaluation of the utility-based alpha-rule framework

Parameter	Value
Throughput filtering constant (f^{thru})	1/1,000
Minimum α value	0
Maximum α value	10
ATEF control time window	0.5 ms
ATEF fairness target (Φ^{target})	Variable
ATEF step size (η)	0.1
ATEF filtering time constant	10
User throughput requirement (T_j^{req})	512 kbps
Simulation time span	5 s
Number of independent simulation runs	70

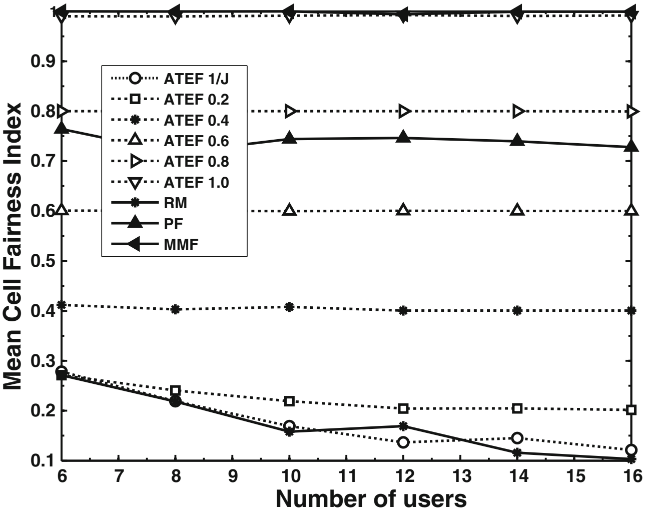

Fig. 4.11 Mean cell fairness index as a function of the number of users for the utility-based alpha-rule framework

that the advantage of the ATEF policy compared with the classic RRA strategies is that the former can be designed to provide any required fairness distribution, while the latter are static and do not have the freedom to adapt themselves and guarantee a specific performance result.

We consider the total cell throughput (cell capacity) as the efficiency indicator, which is presented in Fig. 4.12 as a function of the number of users.

As expected, RM was able to maximize the system capacity, while MMF presented the lowest cell throughput, since it is not able to exploit efficiently the available resources. PF is a trade-off between RM and MMF, so its performance is laid between them. The ATEF policy is able to achieve several cell throughput performances

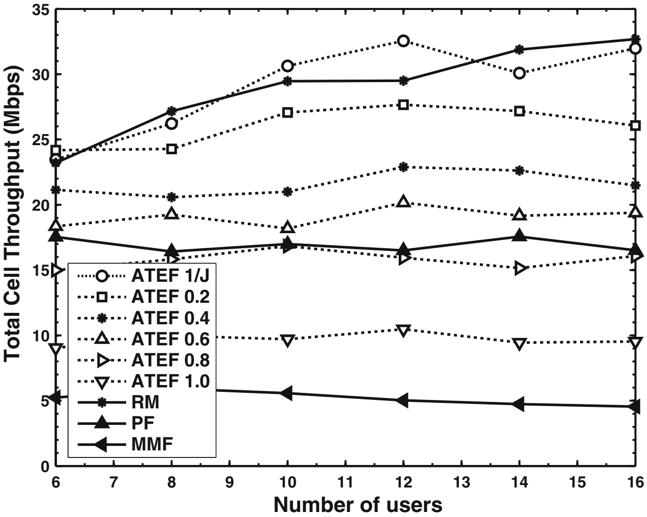

Fig. 4.12 Total cell throughput as a function of the number of users for the utility-based alpha-rule framework

depending on the value of the chosen CFT. In this way, we realize that ATEF is able to work as a hybrid policy between any classic RRA strategy contemplated in the framework.

Looking at Figs. 4.11 and 4.12, one can clearly see the conflicting objectives of capacity and fairness maximization, and how RM and MMF are able to achieve one objective in detriment of the other. PF and ATEF were able to achieve a static and a dynamic trade-off, respectively. A didactic way to explicitly evaluate the trade-off between resource efficiency and user fairness is to combine Figs. 4.11 and 4.12 and plot a 2D plane between total cell throughput (capacity) and the cell fairness index. Figure 4.13 presents the plane built from the simulations of all studied RRA policies on a scenario with 16 active NRT flows.

In Fig. 4.13, the classic RRA policies are indicated as single markers, and the adaptive policy ATEF is indicated as a solid line. The classic policies show a static behavior on the capacity-fairness plane. RM is the most efficient on the resource usage but provides an unfair throughput distribution among users, while MMF is able to provide maximum throughput-based fairness at the expense of low system capacity. The PF policy appears as a fixed trade-off between MMF and RM, with intermediate system capacity and throughput-based fairness.

In order to achieve a desired cell fairness target, the ATEF policy controls the parameter α adaptively according to (4.14). In this way, it is able to cover the whole path between the classic policies in the capacity-fairness plane. Notice in the ATEF curve that the fairness targets set in the simulations (0.2, 0.4, 0.6, 0.8, and 1.0) are successfully met. As expected, the performance of the ATEF policy for very low fairness range converges to the performance of the RM policy. Therefore, it can be

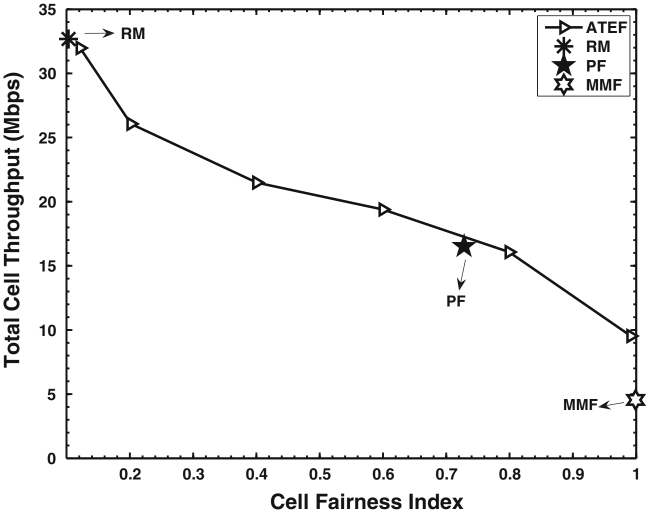

Fig. 4.13 Capacity-Fairness plane for the utility-based alpha-rule framework

concluded that the ATEF policy can adaptively adjust the utility-based alpha-rule framework presented in Table 4.2 in order to provide a dynamic trade-off between resource efficiency and throughput-based fairness.

4.5.4 Conclusions

Two adaptive RRA techniques for the control of the capacity versus fairness trade-off are proposed: FSRM and ATEF. We propose to manage this trade-off by means of fairness control. FSRM and ATEF use two different ways to control the fairness in the system: instantaneousor average fairness control, respectively.

FSRM is able to cover the whole path between the extreme points in the capacity-fairness plane, drawing a complete capacity-fairness curve. One can observe that the performance of FSRM converges to the results of the classical rate adaptive strategies in both extremes of the cell fairness index, which are $1/J$ and 1. These classical techniques are SRM and MMR, respectively. The performance of FSRM is constrained by SRM and MMR because FSRM plays with the competition of two paradigms: efficiency-oriented (resource-centric) and fairness-oriented (user-centric). SRM is the maximum exponent of the former paradigm, while MMR is the best representative of the latter.

The fairness control performed by ATEF is bounded by the structure of the alpha-rule framework, i.e., the minimum and maximum fairness performance depends on the allowed range of values for the parameter α. In the alpha-rule framework, minimum and maximum α correspond to the classical RM and MMF policies, respectively. The ATEF technique dynamically adapts the fairness-controlling parameter α of the alpha-rule framework using a feedback control loop, in order to achieve a desired fairness distribution in terms of throughput (average data rate).

ATEF is able to provide equal or better cell capacity than the respective classical policies for the same cell fairness indexes. Furthermore, it is also able to provide dynamic trade-offs covering the capacity-fairness plane. This is a remarkable strategic advantage to the network operators, because they can now control the aforementioned trade-off and decide in which point on the plane they want to operate.

4.6 Capacity Versus QoS Trade-Off

In this section, we study the trade-off between capacity and QoS. First, a general definition of the trade-off is presented in Sect. 4.6.1. Next, two RRA techniques are proposed: Constrained Rate Maximization (CRM) and Adaptive Throughput-based Efficiency-Satisfaction Trade-off (ATES). The former is based on the heuristic-based RRA framework described in Sect. 4.3, while the latter is based on the utility-based RRA framework presented in Sect. 4.4. The CRM and ATES techniques are described and evaluated in Sects. 4.6.2 and 4.6.3, respectively. Finally, the conclusions about the study of the capacity versus QoS trade-off are shown in Sect. 4.6.4.

4.6.1 General Definition

Capacity and QoS are two contradicting objectives in wireless networks. Without loss of generality, let us consider the case of opportunistic RRA that take into account the channel quality of the users. As it was previously mentioned, the objective of such opportunistic RRA is to allocate more resources to the users with better channel conditions, which leads to a higher resource utilization and system capacity. However, this strategy benefits the users closer to the Base Station (BS), i.e., the ones with highest SNR, and can cause starvation to the users with worse channel conditions. This can severely degrade some users' experience as a result of unfair resource allocation and increased variability in the scheduled rate and delay. Moreover, long delays in the scheduling of packets coming from bad channels can cause severe degradation in the overall performance of the system for higher layer protocols, such as TCP.

On the other hand, schemes that aim to maximize the overall satisfaction have to fulfill QoS requirements and guarantee specific targets of throughput, packet delays, among others. Sometimes, system resources should be assigned to users independently of channel quality state in order to take into account users with degraded QoS, which penalizes users with better channel conditions and reduces system efficiency. Therefore, in general maximizing the system capacity leads to poor QoS provision and vice versa.

The compromise between efficiency and fairness has been widely studied in the literature, as explained in Sect. 4.5. However, to the best of our knowledge, the explicit evaluation of the capacity versus QoS trade-off has not been covered in the literature.

4.6.2 Constrained Rate Maximization

The capacity versus QoS tradeoff will be characterized, in Sect. 4.6.2.1, by the optimization problem of maximizing the system capacity under minimum satisfaction constraints. Then we present the optimal and the heuristic solutions to this problem in Sects. 4.6.2.2 and 4.6.2.3, respectively. Finally, simulation results for performance evaluation are presented in Sect. 4.6.2.4. The contributions presented in this section were first shown in the seminal works [24, 25].

4.6.2.1 Problem Formulation

We consider that in a given TTI, J active users compete for K available resources. We define $\mathscr{J}$ and $\mathscr{K}$ as the set of active users and available resources, respectively. As we are dealing with a multiservice scenario we assume that the number of services provided by the system operator is S and that $\mathscr{S}$ is the set of all services. We consider that the set of users from service $s \in \mathscr{S}$ is $\mathscr{J}_s$ and that $|\mathscr{J}_s| = J_s$, where $|\cdot|$ denotes the cardinality of a set. Note that $\bigcup\limits_{s\in\mathscr{S}} \mathscr{J}_s = \mathscr{J}$ and $\sum\limits_{s\in\mathscr{S}} J_s = J$. We define $\mathbf{X}$ as a $J \times K$ assignment matrix with elements $x_{j,k}$ that assume the value 1 if the resource $k \in \mathscr{K}$ is assigned to the user $j \in \mathscr{J}$ and 0 otherwise. According to the link adaptation functionality, the BS can transmit at different data rates according to the channel state, allocated power, and perceived noise/interference. We consider that user $j \in \mathscr{J}$ can transmit using resource $k \in \mathscr{K}$ with the data rate $r_{j,k}$. The transmit power is uniformly distributed among the available resources. A user j is satisfied if its transmit data rate is higher than or equal to its data rate requirement R_j^{req} after resource allocation. Furthermore, the system operator requires that κ_s users of service s should be satisfied after resource allocation.

The problem of maximizing capacity under minimum satisfaction constraints is formulated as

$$\max_{\mathbf{X}} \left(\sum_{j=1}^{J} \sum_{k=1}^{K} r_{j,k} x_{j,k} \right), \tag{4.15a}$$

$$\text{subject to} \quad \sum_{j=1}^{J} x_{j,k} = 1, \ \forall k \in \mathscr{K}, \tag{4.15b}$$

$$x_{j,k} \in \{0, 1\}, \ \forall j \in \mathscr{J} \text{ and } \forall k \in \mathscr{K}, \tag{4.15c}$$

$$\sum_{j \in \mathscr{J}_s} u\left(\sum_{k=1}^{K} r_{j,k} x_{j,k}, \; R_j^{\text{req}} \right) \geq \kappa_s, \;\; \forall s \in \mathscr{S}, \tag{4.15d}$$

where $u(x, b)$ is a step function that assumes the value 1 if $x \geq b$ and 0 otherwise, where b is a constant. The first part of this optimization problem is the objective function in (4.15a). The objective of this problem is to maximize the total downlink data rate transmitted by the BS to the connected users. When the problem constraints are concerned, we can see that constraints (4.15b) and (4.15c) assure that the each resource k should be allocated exclusively to a given user, i.e., a given resource cannot be shared by multiple users. Another consequence of these constraints is that within a cell covered by a given BS there is no intra-cell interference. The last constraint (4.15d) addresses QoS and user satisfaction issues. In this constraint, for each provided service s in the system, a minimum number of users should be satisfied (κ_s). This is equivalent to satisfy a certain percentage of the connected users for each service in the system.

4.6.2.2 Method for Obtaining the Optimal Solution

Note that problem (4.15) has a binary optimization variable $x_{j,k}$. Therefore, this problem belongs to the class of combinatorial optimization problems. Moreover, constraint (4.15d) is a nonlinear function of the optimization variable $x_{j,k}$. Therefore, problem (4.15) is a nonlinear combinatorial problem that is hard to solve optimally depending on the problem dimensions [47].

A well-known method to solve problem (4.15) consists in the brute force method that consists in numerating all possible solutions, testing whether they obey the constraints (4.15b)–(4.15d), and evaluating the achieved total data rate. The optimal solution is the one that presents the highest total data rate. The total number of possible solutions that can be enumerated is J^K. Therefore, this method only works for small J and K, which is not the case in cellular networks.

Fortunately, problem (4.15) can be simplified by modifying constraint (4.15d). Consider a binary selection variable ρ_j that assumes the value 1 if user j is selected to be satisfied and 0 otherwise. According to this, the problem (4.15) can be restated as

$$\max_{\mathbf{X}, \boldsymbol{\rho}} \left(\sum_{j=1}^{J} \sum_{k=1}^{K} r_{j,k} x_{j,k} \right), \tag{4.16a}$$

$$\text{subject to} \quad \sum_{j=1}^{J} x_{j,k} = 1, \quad \forall k \in \mathscr{K}, \tag{4.16b}$$

$$x_{j,k} \in \{0, 1\}, \quad \forall j \in \mathscr{J} \text{ and } \forall k \in \mathscr{K}, \tag{4.16c}$$

$$\sum_{k=1}^{K} r_{j,k} x_{j,k} \geq R_j^{\text{req}} \rho_j, \quad \forall j \in \mathscr{J}, \tag{4.16d}$$

$$\rho_j \in \{0, 1\}, \quad \forall j \in \mathscr{J}, \tag{4.16e}$$

$$\sum_{j \in \mathscr{J}_s} \rho_j \geq \kappa_s, \quad \forall s \in \mathscr{S}. \tag{4.16f}$$

As can be seen, the constraint (4.15d) of problem (4.15) was replaced by constraints (4.16d), (4.16e) and (4.16f) in problem (4.16). Now, the optimization variables are $x_{j,k}$ and ρ_j, and all problem constraints and objective function are linear. Therefore, we managed to convert problem (4.15) to an Integer Linear Problem (ILP). This special class of optimization problems can be solved by standard numerical solvers based on the Branch and Bound (BB) algorithm. The main idea of the BB algorithm is to decrease the search space by solving a relaxed version of the original optimization problem [29].

Although, the optimal solution of problem (4.16) can be obtained with much less processing time with BB-based solvers compared to the brute force method, the worst-case complexity of the BB-based solvers is exponential with the number of variables and problem constraints [47]. In problem (4.16), we have $J \times K + J$ variables and $J + K + S$ constraints. Consequently, obtaining the optimal solution to the studied problem is not feasible for the short time basis of cellular networks even for moderated number of users, resources and services.

4.6.2.3 Algorithm Description

In this section we present an algorithm to solve the problem presented in Sect. 4.6.2.1 following the heuristic framework presented in Sect. 4.3. As it was shown previously, the first part of the solution consists in solving the studied problem without the minimum satisfaction constraints. In other words, we are interested in finding the solution that maximizes the spectral efficiency. The implementation of this first part, called **Unconstrained Maximization**, is presented in Fig. 4.14.

In step (1) of the Unconstrained Maximization part we define two temporary user sets: auxiliary user set represented by $\mathscr{B}$ and the available user set denoted by $\mathscr{A}$. The auxiliary user set contains the users that can be disregarded without violating the minimum satisfaction constraints per service. The available user set contains the users that were not disregarded along the Unconstrained Maximization part and will get resources in the Reallocation part. Both user sets are initialized with the set of all users $\mathscr{J}$.

In step (2) we solve the relaxed version of problem (4.16), i.e., without the minimum satisfaction constraints, with the users of the available user set. The optimal solution to the relaxed problem is simple; basically the resources should be assigned to the users with best channel quality on them [18]. According to the RRA performed

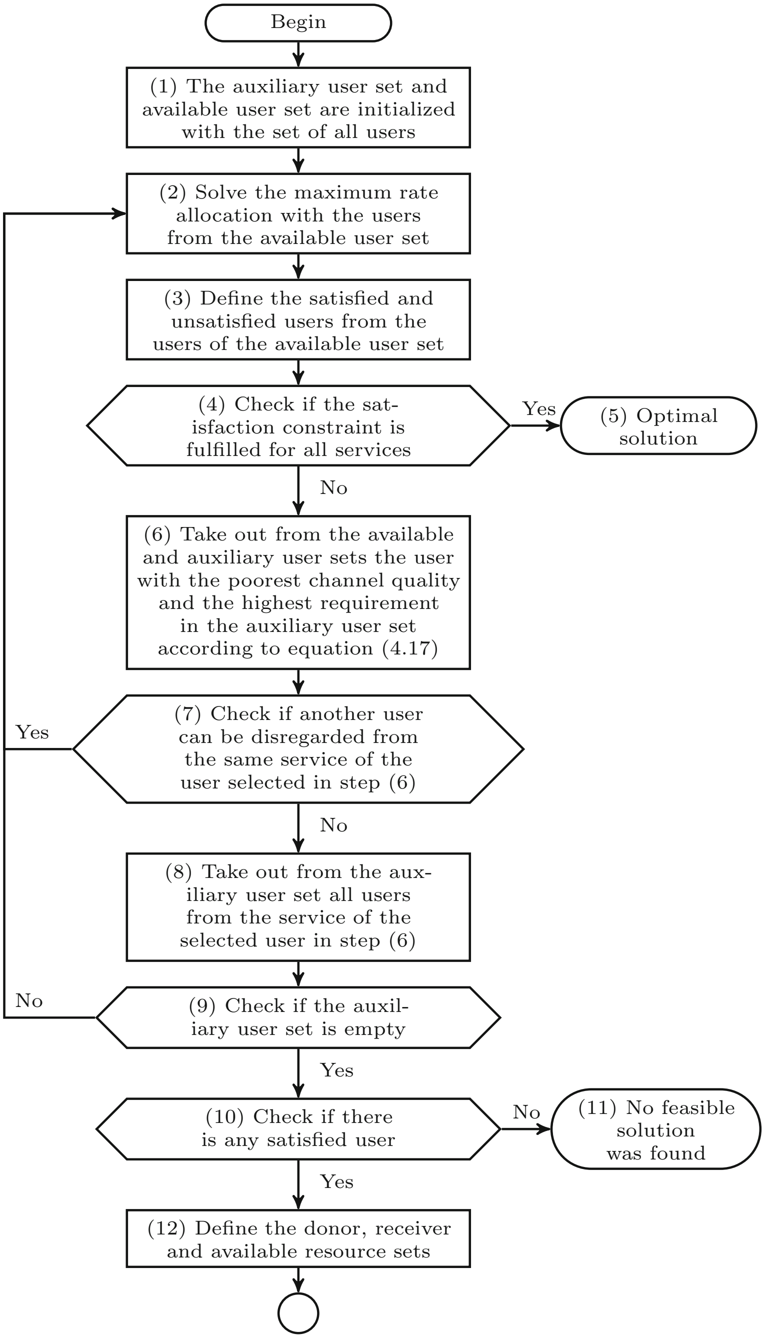
Begin
(1) The auxiliary user set and available user set are initialized with the set of all users
(2) Solve the maximum rate allocation with the users from the available user set
(3) Define the satisfied and unsatisfied users from the users of the available user set
(4) Check if the satisfaction constraint is fulfilled for all services
Yes
(5) Optimal solution
No
(6) Take out from the available and auxiliary user sets the user with the poorest channel quality and the highest requirement in the auxiliary user set according to equation (4.17)
Yes
(7) Check if another user can be disregarded from the same service of the user selected in step (6)
No
(8) Take out from the auxiliary user set all users from the service of the selected user in step (6)
No
(9) Check if the auxiliary user set is empty
Yes
(10) Check if there is any satisfied user
No
(11) No feasible solution was found
Yes
(12) Define the donor, receiver and available resource sets

Fig. 4.14 Flowchart of the Unconstrained Maximization part of the CRM technique

in step (2), some of the users would get an allocated data rate higher than or equal to the required data rate, R_j^{req}, whereas other users would get an allocated data rate lower than the data rate requirement. Therefore, in step (3) we define the former users as the satisfied users while the latter are the unsatisfied users.

In step (4) we evaluate if the minimum number of users that should be satisfied per service, κ_s, is fulfilled with the RRA performed in step (2). Basically, in step (4) we evaluate if the set of constraints (4.16d), (4.16e) and (4.16f) of problem (4.16) are fulfilled. If so, the RRA solution in step (2) is the optimal solution of the studied problem as presented in step (5). Note that this is an uncommon situation because of the wireless propagation characteristics where few users present the best channel qualities in most of the resources. In this way, only few users would get satisfied with the solution in step (2).

In step (6), a user is taken out of the RRA process. The main idea here is to take out of the RRA process the user that demands more resources to be satisfied. According to this, the selected user is chosen according to the following equation:

$$j^* = \arg \max_{j \in \mathscr{B}} \frac{R_j^{\text{req}}}{\frac{1}{K} \sum\limits_{k=1}^{K} r_{j,k}}. \tag{4.17}$$

As it can be seen in (4.17), the denominator of the fraction in the argument of the arg max $(\cdot)$ function consists in the estimated average transmit data rate of user j per resource, whereas the numerator is the required data rate of user j. Therefore, the ratio between these two quantities consists in the estimated number of resources that user j needs to be satisfied. The objective is to disregard the user that needs more resources. As it will be shown later, there is a limit in the number of users that can be disregarded that depends on the minimum satisfaction constraints of the studied problem.

Note that if there are initially J_s users from service s and a minimum of κ_s users should be satisfied, the maximum number of users that can be disregarded is $J_s - \kappa_s$ in order to be still possible guaranteeing the minimum satisfaction constraint for service s. In step (7), we check whether the service of the user selected in step (6) (represented here by s^*) can have another user disregarded without violating the minimum satisfaction constraints. If so, the algorithm returns to step (2) where the relaxed version of problem (4.16) is solved with the users of the available user set. Otherwise, all users from service s^* will be taken out of the auxiliary user set in step (8), i.e., these users could not be disregarded in the Unconstrained Maximization part.

In step (9) we check if the auxiliary user set is empty which means that we could not disregard any user without violating the minimum satisfaction constraints. If so, we check in step (10) if at least one user is satisfied. If the output of step (10) is positive, we define from the available user set and the resource set three new sets: the donor ($\mathscr{D}$) and receiver ($\mathscr{R}$) user sets, and the available resource set ($\mathscr{K}$). The donor user set $\mathscr{D}$ is composed of the satisfied users in the available user set $\mathscr{A}$ and can donate resources to unsatisfied users. The receiver user set $\mathscr{R}$ is composed of the unsatisfied users from the available user set $\mathscr{A}$ that need to receive resources from the donors to have their data rate requirements fulfilled. Finally, the available resource set $\mathscr{K}$ is composed of all the resources from the users in the donor user set, i.e., the resources that can be donated to the unsatisfied users (receiver users).

Note that if the auxiliary user set is not empty in step (9), step (2) is executed again with the users of the available user set. Also, if there is no satisfied user in step (10) the algorithm is not able to find a feasible solution. A satisfied user or donor user is necessary in the second part of the proposed algorithm in order to donate resources to the unsatisfied users or receiver users.

In Fig. 4.15 we present the flowchart of the second part of the proposed solution named as **Resource Reallocation**. In step (1) of the Reallocation part of the proposed solution, the user from the receiver user set with the worst channel condition is chosen to receive resources. The main motivation for choosing the user with worst channel condition is to increase the probability that this user will get resources in good channel conditions, and therefore, need few resources to become satisfied. Then, in step (2) a resource previously assigned to a donor user is reassigned to the receiver user selected in step (1). The criterion to select the resource k^* is presented in the following:

$$k^* = \arg\max_{k \in \mathscr{K}} \frac{r_{j^*,k}}{r_{j^+,k}}, \tag{4.18}$$

where j^* is the selected receiver user in step (1) and j^+ is the user from the donor user set $\mathscr{D}$ that has got assigned the resource k in the first part of the proposed solution (Unconstrained Maximization). The numerator of the fraction in the argument of the arg max $(\cdot)$ function represents the transmit data rate of the selected user j^* on resource k whereas the denominator comprises the transmit data rate of user j^+ (donor user) on resource k. Therefore, the chosen resource k^* is the one belonging to user j^* that presents the lowest loss in transmit data rate compared to the previous allocation.

The selected resource in step (2) is reassigned to the receiver user only if the donor user does not become unsatisfied with the resource reallocation. This test is performed in step (3). If the donor cannot donate resources without becoming unsatisfied, the selected resource in step (2) is taken out of the available resource set. Otherwise, the resource is reallocated in step (4) and the data rates of the receiver and donor users are updated in step (5). Another test that should be performed is to check if the selected receiver is satisfied in step (6). If so, the selected receiver user is taken out of the receiver user set in step (7). According to step (8), if the receiver user set becomes empty after step (7), the algorithm is able to find a feasible solution as shown in step (9). Note that if the output of step (6) is negative, the algorithm goes to step (10) where the chosen resource is taken out of the reallocation process. Finally, in step (11) we check if there are available resources to be reassigned. If so, the algorithm goes to step (1). Otherwise, the algorithm is not able to find a feasible solution as it is shown in step (12).

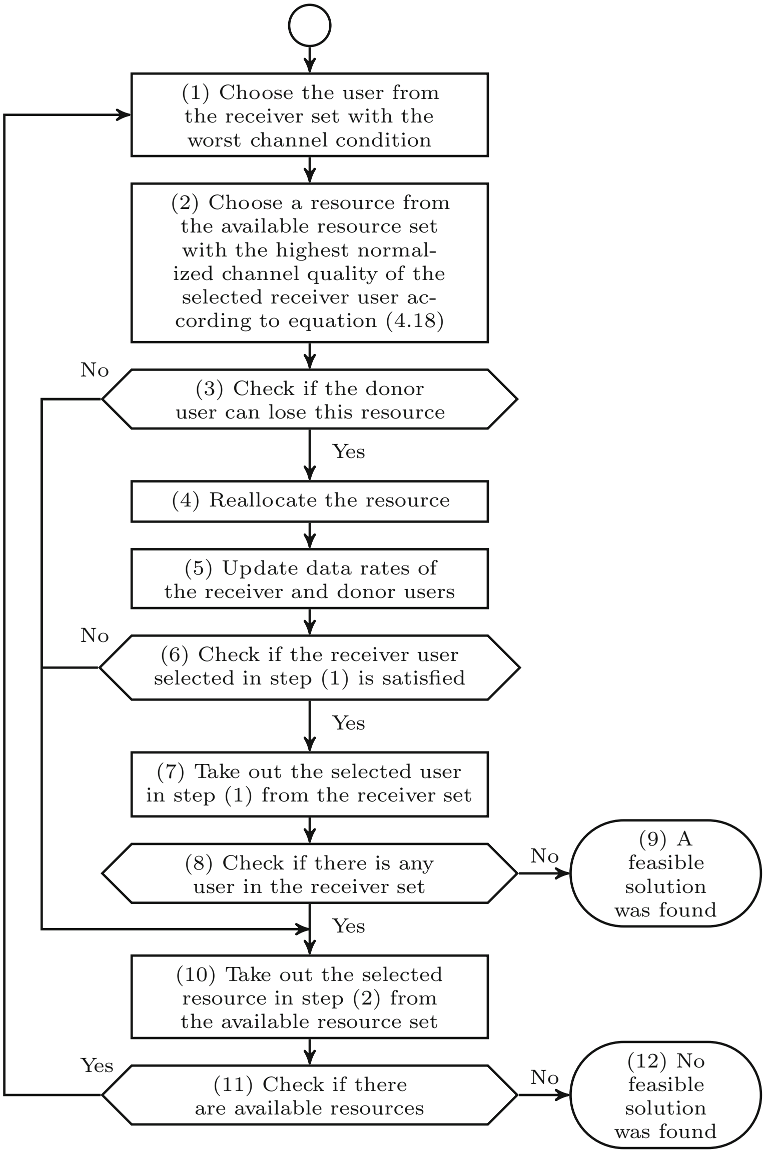
(1) Choose the user from the receiver set with the worst channel condition
(2) Choose a resource from the available resource set with the highest normalized channel quality of the selected receiver user according to equation (4.18)
(3) Check if the donor user can lose this resource
No
Yes
(4) Reallocate the resource
(5) Update data rates of the receiver and donor users
(6) Check if the receiver user selected in step (1) is satisfied
No
Yes
(7) Take out the selected user in step (1) from the receiver set
(8) Check if there is any user in the receiver set
No
(9) A feasible solution was found
Yes
(10) Take out the selected resource in step (2) from the available resource set
(11) Check if there are available resources
Yes
No
(12) No feasible solution was found

Fig. 4.15 Flowchart of the Resource Reallocation part of the CRM technique

As it can be seen in the proposed solution, depending on the system load, channel state, and data rate requirements, the algorithm is not able to find a feasible solution. In these cases, an alternative is to softly decrease the minimum satisfaction constraints and/or the data rate requirements and re-run the proposed solution again.

Table 4.4 Simulation parameters for the evaluation of the CRM technique

Parameter	Value
Number of cells	1
Transmit power per resource	0.35 W
Cell radius	334 m
MT speed	Static
Carrier frequency	2 GHz
Number of subcarriers per resource	12
Path loss[a]	$L = 35.3 + 37.6 \log_{10} d$
Log-normal shadowing standard deviation	8 dB
Noise spectral density	$3.16 \cdot 10^{-20}$ W/Hz
Number of snapshots	3000
Number of services	4
Number of users and required minimum number of satisfied users	See Table 4.5

[a] Distance d in km

4.6.2.4 Simulation Results

In this section we present some simulation results to illustrate the performance of the CRM technique. We consider the downlink of a hexagonal sector belonging to a tri-sectorized cell of a cellular system. In order to get valid results in a statistical sense we perform several independent snapshots. In each snapshot, the terminals are uniformly distributed within each sector, whose BS is placed at its corner. The minimum allocable resource consists in a time-frequency grid composed of a group of 12 adjacent subcarriers in the frequency dimension and 14 consecutive Orthogonal Frequency-Division Multiplexing (OFDM) symbols in the time dimension. We assume that there are 20 resources in the system.

The propagation model includes a distance-dependent path loss model, a log-normal shadowing component, and a Rayleigh-distributed fast fading component. Specifically, we consider that the fast-fading component of the channel gain of a given terminal is independent among resources. We assume that the link adaptation is performed based on the report of 15 discrete Channel Quality Indicators (CQI) used by the LTE system [2]. The SNRs thresholds for MCS switching were obtained by link level simulations from [27]. The main simulation parameters are summarized in Table 4.4.

We assume that there are four different services with three users each. We consider three different cases when the minimum number of satisfied users is concerned (κ_s). The cases are summarized in Table 4.5. As it can be seen in this table we vary the minimum number of users that should be satisfied for services 3 and 4. Case 3 requires that all users from all services should be satisfied while in case 1 two users should be satisfied for services 3 and 4.

In order to assess the relative performance of the proposed solution we simulate also two other RRA solutions. The first solution is the optimal solution of problem

Table 4.5 Assumed cases for simulation of the CRM technique

Cases	J_1	J_2	J_3	J_4	κ_1	κ_2	κ_3	κ_4
1	3	3	3	3	3	3	2	2
2	3	3	3	3	3	3	3	2
3	3	3	3	3	3	3	3	3

(4.16) obtained by ILP solver.[3] We call this solution as SatisOpt. The second is the optimal solution of the relaxed version of problem (4.16), i.e., without constraints from (4.16d) to (4.16f). As explained previously, this solution corresponds to the solution that maximizes the total data rate. We call this solution as MaxRateOpt.

Before presenting the results we first define a metric called success rate. The success rate is defined by the ratio between the number of snapshots in which a given solution was able to solve the problem (4.16) including the minimum satisfaction constraints, and the total number of snapshots. Therefore, the higher the success rate, the better the given algorithm in solving the studied problem.

In Fig. 4.16 we present the success rate for the SatisOpt, MaxRateOpt and the proposed solution in cases 1, 2, and 3. First, we can see that the success rate decreases with the data rate requirement of the users for all algorithms, as expected. Another observation is that the MaxRateOpt solution presents low success rates even for low data rate requirements. The reason for this is that it maximizes the total data rate without any QoS guarantee. Consequently, in general, only few users (with best channel conditions) get most of the system resources and become satisfied.

The relative comparison of cases 1, 2, and 3 shows that all algorithms perform better in case 1 than in cases 2 and 3. In fact, in case 3 it is required that more users should be satisfied than in case 2, that in its turn requires more satisfied users than in case 1.Therefore, the problem to be solved is harder in case 3 than in cases 2 and 1. Looking at the performance of the proposed algorithm, we can observe that its performance is similar to the SatisOpt solution in low and medium data rate requirements. Focusing on the required data rate where the corresponding SatisOpt solution has a success rate of 90 %, the differences in success rate between the proposed solution and SatisOpt are only 1.67, 1.36, and 0.83 % in cases 1, 2, and 3, respectively.

The success rate performance metric shows the capability of the algorithms in finding a feasible solution to our problem. On the other hand, another important information is the objective attained by the different algorithms, i.e., the total achieved data rate. The total data rate consists in the sum of all data rates achieved by all users after resource allocation. It should be noticed that in order to maximize the total spectral efficiency, some users can get allocated data rates much higher than their required data rates. In Table 4.6 and 4.7 we present some percentiles of the total data rate for specific data rate requirements considered in the x-axis of Fig. 4.16 regarding the success rate performance. For a specific case and load, the percentiles

[3] In order to solve ILP problems we used the IBM ILOG CPLEX Optimizer [14].

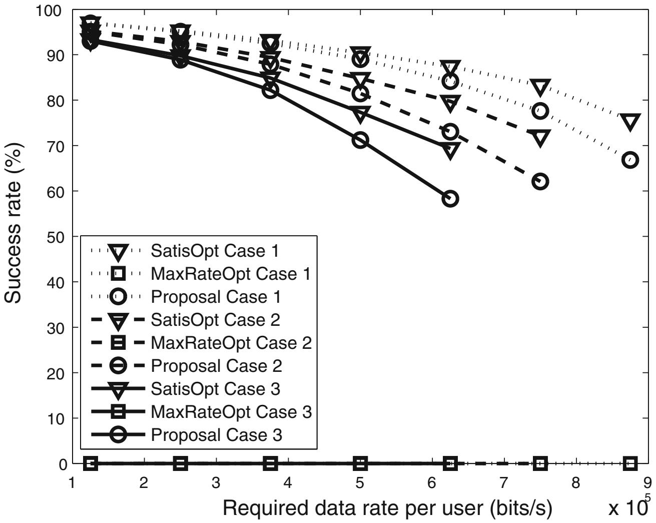

Fig. 4.16 Success rate versus the required data rate per user in cases 1, 2, and 3 for SatisOpt, MaxRateOpt and proposed CRM solution

of all algorithms are built with the samples of the snapshots in which the proposed solution and SatisOpt were able to find a solution. Therefore, it is possible that in many of the samples used to calculate the percentiles for the MaxRateOpt solution, the constraints from (4.16d) to (4.16f) were not fulfilled. The main idea to include results of the MaxRateOpt solution is to show how the problem constraints imposed losses in the total achievable data rate.

In Table 4.6 we present the 25th, 50th, and 75th percentiles of the total data rate for all algorithms in cases 1 and 2 for the required data rate of 250 kbps. In Table 4.7 we present the 25th, 50th, and 75th percentiles of the total data rate for all algorithms in cases 1 and 2 for the required data rate of 750 kbps. Furthermore, in both tables we present the losses in the percentiles of the total data rate comparing MaxRateOpt and SatisOpt as well as SatisOpt and the proposed solution.

We have three comments about both tables. First, the MaxRateOpt algorithm provides the highest total data rates in all percentiles and cases as can be seen in both tables. This comes at the cost of low success rates as shown in Fig. 4.16. Second, the difference in the total data rate between the MaxRateOpt algorithm and SatisOpt increases with the required data rate as can be seen by comparing the fifth column of Tables 4.6 and 4.7. The total data rate of SatisOpt is penalized when the data rate requirement is high since many resources should be assigned to the users in medium and bad channel conditions. Finally, focusing on the performance of the proposed algorithm, we can see that it performs almost optimally at the required data rate of 250 kbps, with performance loss compared to the SatisOpt solution not higher than 1 %. At the required data rate of 750 kbps, the proposed algorithm leads to higher performance losses compared with the ones of Table 4.6. It is important to highlight

Table 4.6 Percentile of the total data rates (Mbps) and performance losses in cases 1 and 2 for SatisOpt, MaxRateOpt, and proposed solution in the required data rate of 250 kbps

	SatisOpt (Mbps)	MaxRateOpt (Mbps)	Proposal (Mbps)	Loss from MaxRateOpt to SatisOpt (%)	Loss from SatisOpt to Proposal (%)
Case 1					
25th percentile	15.68	18.66	15.60	15.99	0.51
50th percentile	16.70	18.66	16.65	10.53	0.30
75th percentile	17.48	18.66	17.46	6.31	0.14
Case 2					
25th percentile	16.37	18.66	16.31	12.30	0.35
50th percentile	17.23	18.66	17.19	7.64	0.27
75th percentile	17.83	18.66	17.78	4.43	0.29

Table 4.7 Percentile of the total data rates (Mbps) and performance losses in cases 1 and 2 for SatisOpt, MaxRateOpt, and proposed solution in the required data rate of 750 kbps

	SatisOpt (Mbps)	MaxRateOpt (Mbps)	Proposal (Mbps)	Loss from MaxRateOpt to SatisOpt (%)	Loss from SatisOpt to Proposal (%)
Case 1					
25th percentile	14.06	18.66	12.79	24.64	9.02
50th percentile	15.47	18.66	14.82	17.10	4.23
75th percentile	16.75	18.66	16.54	10.26	1.24
Case 2					
25th percentile	14.58	18.66	13.48	21.88	7.50
50th percentile	16.06	18.66	15.61	13.95	2.80
75th percentile	17.28	18.66	17.11	7.40	1.01

that this data rate is just considered for emphasizing the sensitivity (degradation) of the proposed algorithm to this parameter, even though this is not a feasible load in terms of user satisfaction.

In summary, from the joint analysis of the results in Fig. 4.16 and Tables 4.6 and 4.7, we can see that our proposed CRM solution performs near optimally considering the problem objective and constraints in low and medium load conditions. According to [24, 25] the worst case computational complexity to obtain the optimal solution by using the BB algorithm is $\mathscr{O}\left(2^{JK}\right)$. The complexity of the proposed heuristic algorithm is $\mathscr{O}\left(K\left(J-\sum_{s\in\mathscr{S}}\kappa_s\right)\left(J+\sum_{s\in\mathscr{S}}\kappa_s\right)\right)$. Therefore, the computational complexity to obtain the optimal solution of problem (4.16) is too high for the short time basis in which resource allocation takes place in current mobile networks. By analyzing the computational complexity and performance of the proposed algorithm we conclude that it leads to a good performance-complexity trade-off when compared to the strategy used to obtain the optimal solution.

4.6.3 Adaptive Throughput-Based Efficiency-Satisfaction Trade-Off

The Adaptive Throughput-based Efficiency-Satisfaction Trade-off (ATES) technique is based on the utility-based RRA framework described in Sect. 4.4 and tries to solve the problem of controlling the trade-off between capacity and satisfaction (QoS).

This section is organized as follows. The RRA problem to be solved is formulated in Sect. 4.6.3.1. The proposed technique is described in Sect. 4.6.3.2, while Sect. 4.6.3.3 shows the performance evaluation of ATES and other classical RRA techniques.

4.6.3.1 Problem Formulation

We claim that it is possible to perform user satisfaction shaping for NRT services with low complexity if we consider a sigmoid utility function in the optimization problem formulated in Sect. 4.4. This utility function should be based on a particular QoS parameter suitable for NRT services.

In this section, we propose a particular case of the RRA framework described in Sect. 4.4 that uses a sigmoid utility function. This framework is called the utility-based sigmoid-rule, and comprises a novel RRA technique called ATES, whose formulation is based on the users' throughput and is suitable for NRT services. It aims to control the trade-off between resource efficiency and user satisfaction using an adaptive utility function and a feedback control loop.

In order to achieve user satisfaction shaping, we propose to use an increasing sigmoid utility function based on the throughput T_j of the user j, as indicated below:

$$U\left(T_j\left[n\right]\right) = \frac{1}{1 + e^{-\sigma\left(T_j[n] - T_j^{\text{req}}\right)}}, \tag{4.19}$$

where σ is a nonnegative parameter that determines the shape of the sigmoid function; and $T_j\left[n\right]$ and T_j^{req} are the current throughput and the throughput requirement of user j, respectively.

The marginal utility given by the utility-based weight plays an important role in the DRA algorithm described in Sect. 4.4. The higher the weight, the higher the priority of the user to get a resource. For the case of the utility function defined by (4.19), the marginal utility is given by

$$w_j = \frac{\partial U\left(T_j\left[n\right]\right)}{\partial T_j\left[n\right]} = \frac{\sigma\, e^{-\sigma\left(T_j[n] - T_j^{\text{req}}\right)}}{\left(1 + e^{-\sigma\left(T_j[n] - T_j^{\text{req}}\right)}\right)^2}. \tag{4.20}$$

Table 4.8 Features of the utility-based sigmoid-rule framework: $U\left(T_j\left[n\right]\right) = \dfrac{1}{1+e^{-\sigma\left(T_j[n]-T_j^{\text{req}}\right)}}$

Techniques	Parameter σ	Weight w_j	Characteristics
RM	$\sigma \to 0$	$\sigma/4$	High resource efficiency and low throughput-based satisfaction
TSM	$\sigma \to \infty$	Impulse at T_j^{req}	Low resource efficiency and high throughput-based satisfaction
ATES	Adaptive	$\dfrac{\sigma\, e^{-\sigma\left(T_j[n]-T_j^{\text{req}}\right)}}{\left(1+e^{-\sigma\left(T_j[n]-T_j^{\text{req}}\right)}\right)^2}$	Dynamic trade-off between resource efficiency and throughput-based satisfaction

Therefore, the corresponding DRA algorithm, which is given by (4.3), must now use the particular expression of w_j presented in (4.20).

Depending on the value of σ, we can achieve a different user satisfaction shaping. If we consider σ as an adaptive parameter, interesting properties of the sigmoid function appear. The higher the value of σ, the closer to a step-shaped function the utility function will be. Otherwise, considering lower values of σ, the utility function becomes more linear. This characteristic can be visualized in Fig. 4.17.

Depending on the value of the controlling parameter σ, the sigmoid-rule framework presented above can be designed to work as different RRA techniques, achieving different performances in terms of resource efficiency and throughput-based satisfaction. The main characteristics of the sigmoid-rule framework and the three particular RRA techniques contemplated by this framework are presented in Table 4.8. When σ approaches to zero, then the RRA technique is the well-known classical rate maximization (RM) technique. On the opposite side, when σ goes to ∞, the throughput-based satisfaction maximization (TSM) technique is achieved. Details about this technique are given in Chap. 2 of this book.

4.6.3.2 Algorithm Description

The ATES technique is an adaptive version of the utility-based sigmoid-rule. It aims to achieve an efficient trade-off between resource efficiency and throughput-based satisfaction planned by the network operator in a scenario with NRT services. This is done by means of the adaptation of the satisfaction controlling parameter σ in the utility function presented in (4.19).

Due to the fact that the shape of the sigmoid utility function is not so sensitive to the variation of σ in linear unit, thus it is assumed that the adaptation of σ is done in dB unit. As it was previously mentioned, the trade-off between satisfaction and efficiency is limited by the TSM and RM policies, respectively. According to Table 4.8, the TSM and RM techniques are associated to σ values tending to ∞ and 0, respectively. However, these are extreme values, and the ATES technique does need

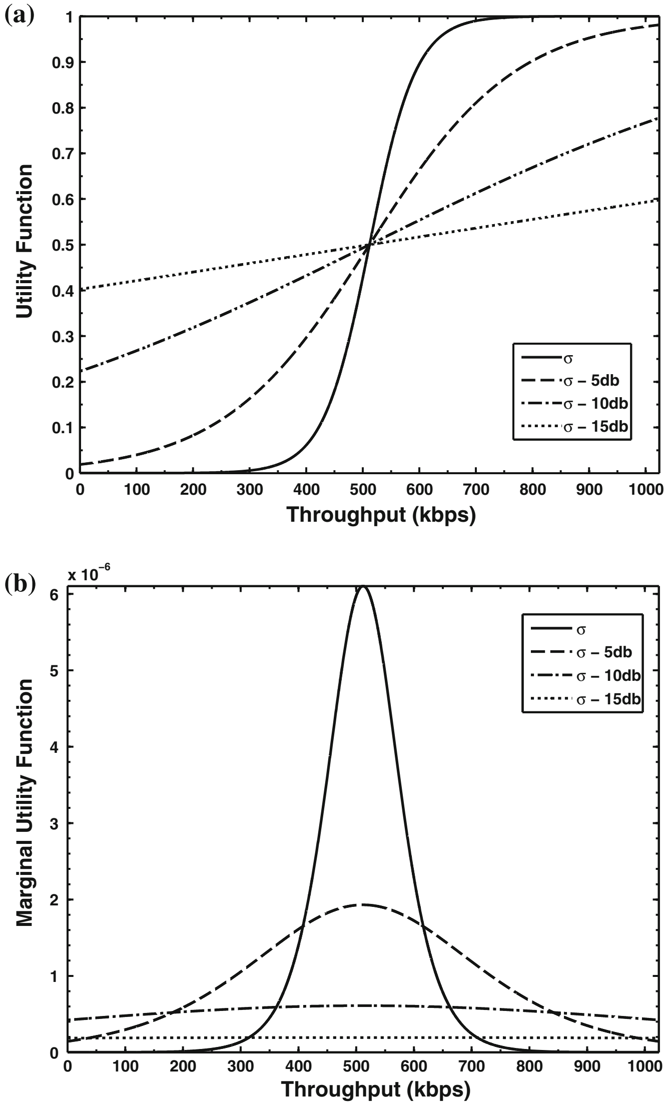

Fig. 4.17 Family of utility functions used in the utility-based sigmoid-rule framework. **a** Utility functions. **b** Marginal utility functions

to cover the whole range. Let us assume that the path between these policies can be characterized by a dynamic range Δ_{dB} in dB unit. The TSM policy is associated with the maximum value of σ. In this work, we assume $\sigma_{\mathrm{TSM}} = 2.441 \times 10^{-5}$, which is suitable for the case of $T_j^{\mathrm{req}} = 512\,\mathrm{kbps}$ (see Table 4.9). We also define $\Delta_{\mathrm{dB}} = 10\log_{10}(\sigma_{\mathrm{TSM}} / \sigma_{\mathrm{RM}})$. Lower values of σ yield utility functions more linear.

We assume that σ_{RM} is associated with a utility function sufficiently linear for our purposes.

The ATES policy needs to calculate the percentage of satisfied users in the system continuously. A user from an NRT service is considered satisfied if its session throughput $T_j\,[n]$ is higher than the requirement T_j^{req}. The percentage of satisfied users is calculated as

$$\Psi^{\mathrm{cell}}\,[n] = \frac{J_{\mathrm{sat}}[n]}{J}, \tag{4.21}$$

where $J_{\mathrm{sat}}[n]$ is the instantaneous number of satisfied users and J is the total number of users in the cell.

The objective of the ATES policy is to assure a strict throughput-based satisfaction distribution among the users, i.e., the instantaneous satisfaction percentage $\Psi^{\mathrm{cell}}\,[n]$ must be kept around a planned value Ψ^{target}. Therefore, the ATES policy adapts the parameter σ in the utility-based sigmoid-rule framework in order to achieve the desired operation point. Aiming at this objective, the new value of the parameter σ is calculated using a feedback control loop of the form:

$$\sigma\,[n] = \sigma\,[n-1] - \eta\left(\Psi^{\mathrm{filt}}\,[n] - \Psi^{\mathrm{target}}\right) \tag{4.22}$$

where $\Psi^{\mathrm{filt}}\,[n]$ is a filtered version of the satisfaction percentage $\Psi^{\mathrm{cell}}\,[n]$ using an exponential smoothing filtering, which is used to suppress short-run fluctuations and smooth time series with slowly varying trends; Ψ^{target} is the target satisfaction, i.e., the desired value for the satisfaction percentage; and the parameter η is a step size that controls the speed of adaptation of σ.

ATES is an iterative technique that is executed every TTI. The technique is able to reach a stable convergence of the target satisfaction percentage. In this way, we are able to manage efficiently the trade-off between system capacity and throughput-based satisfaction. By controlling the network QoS, the network operator can also have a good prediction about the system capacity.

4.6.3.3 Simulation Results

In this section, the performance of the adaptive ATES technique is compared to the RM [18], PF [19], TSM, and Satisfaction-Oriented Resource Allocation for Non-Real Time Services (SORA-NRT)[4] techniques by means of system-level simulations, which took into account the main characteristics of a single-cell LTE-based system. In this simulation scenario, several satisfaction targets were considered for the ATES technique, namely $\Psi^{\mathrm{target}} = [70, 80, 90, 100]\,\%$. The particular simulation parameters used in this analysis are depicted in Table 4.9.

The percentage of satisfied users calculated by (4.21) averaged overall simulation snapshots is depicted in Fig. 4.18 for various system loads. More precisely, the per-

[4] The TSM and SORA-NRT techniques are described and evaluated in details in Chap. 2 of this book.

Table 4.9 Simulation parameters for the evaluation of the ATES technique

Parameter	Value
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 subcarriers	512
Total number of useful subcarriers	300
Subcarrier bandwidth	15 kHz
Number of PRBs	25
Path loss	$L = 128.1 + 37.6 \log_{10} d$
Log-normal shadowing standard deviation	8 dB
Small-scale fading	3GPP typical urban (TU) [1, 17]
AWGN power per sub-carrier	-123.24 dBm
Noise figure	9 dB
Link adaptation	Using link level curves from [27]
SNR threshold of MCS 1 [27]	-6.9 dB
Transmission Time Interval (TTI)	1 ms
NRT traffic model	Full buffer
Throughput filtering constant (f^{thru})	1/1,000
User throughput requirement (T_j^{req})	512 kbps
Parameter σ for TSM	2.441×10^{-5}
Parameter σ for ATES	Adaptive
Maximum σ value	2.441×10^{-5}
Minimum σ value	2.441×10^{-13}
Dynamic range for σ adaptation (Δ_{dB})	80 dB
ATES control time window	1 ms
ATES satisfaction target (Ψ^{target})	Variable
ATES step size (η)	0.1
ATES filtering time constant	10
Number of independent simulation runs	30
Simulation time span	30 s

formance of the adaptive ATES policy is compared to the RM and TSM policies. It can be observed that ATES is successful at achieving its main objective, which is to guarantee a strict satisfaction pattern among the NRT users. This is achieved due to the feedback control loop that dynamically adapts the parameter σ of the utility-based sigmoid-rule framework. In some cases the satisfaction target is not met exactly. This is due to the small number of full-buffer flows considered in the simulations, which does not provide enough granularity in the calculation of the satisfaction percentages.

Notice that the structure of the sigmoid-rule framework bounds the performance of the ATES policy between the performances of the RM and TSM policies. We considered in the simulations a dynamic range of 80 dB for the adaptation of the parameter σ by the ATES policy. When a maximum satisfaction is desired, we con-

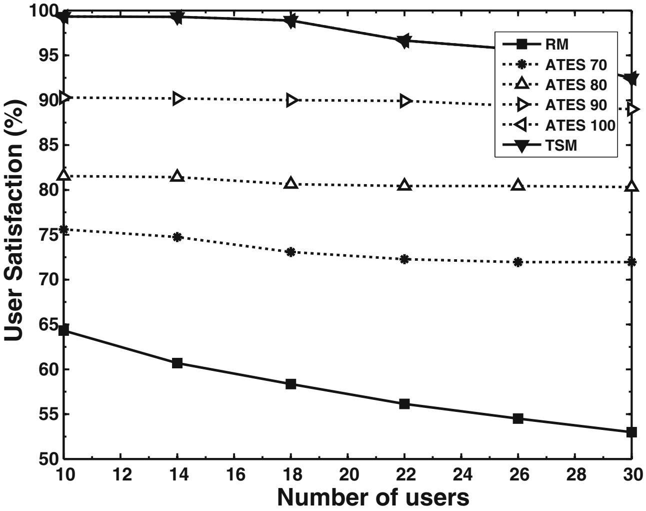

Fig. 4.18 User satisfaction as a function of the number of users for the utility-based sigmoid-rule framework

figure ATES with $\Psi^{\text{target}} = 100\,\%$ and set the upper limit of $\sigma = 2.441 \times 10^{-5}$, which is the same σ as used in the TSM policy. On the other extreme, it is clear that RM works as a lower bound for ATES configured with low values of Ψ^{target}.

It can be concluded that the advantage of the ATES policy compared with RM and TSM strategies is that the adaptive technique can be designed to provide any required satisfaction level between the limits imposed by the sigmoid-rule, while RM and TSM are static and do not have the freedom to adapt themselves and guarantee a specific performance result.

In Fig. 4.19 we analyze how the RRA policies behave in terms of efficiency in the resource usage. We consider the total cell throughput (cell capacity) as the efficiency indicator, which is presented in Fig. 4.19 as a function of the number of users in the system.

As expected, RM was able to maximize the system capacity, while TSM presented the lowest cell throughput, since it is not able to exploit the available resources in the most efficient way possible due to the QoS-dependent component in the resource allocation prioritization. The ATES policy is able to achieve several cell throughput performances depending on the value of the chosen satisfaction target. In this way, we realize that ATES is able to work as a hybrid policy between RM and TSM strategies. Furthermore, one can notice that the capacity gain when lower satisfaction targets are set is not linear. The capacity gain from changing Ψ^{target} from 100 to 90 % is higher than changing it from 90 to 80 %, and so on. We can also see that $\Psi^{\text{target}} = 70\,\%$ already provides a capacity very close to the maximum, with satisfaction levels much higher than RM (see Fig. 4.18).

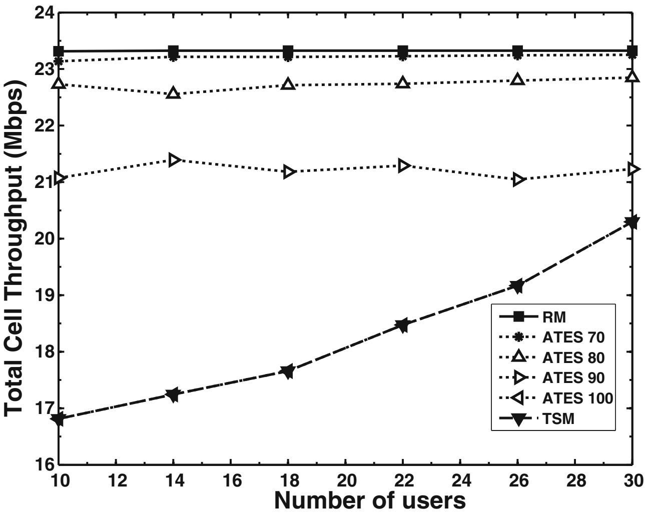

Fig. 4.19 Total cell throughput as a function of the number of users for the utility-based sigmoid-rule framework

Looking at Figs. 4.18 and 4.19, one can clearly see the conflicting objectives of capacity and satisfaction maximization, and how RM and TSM are able to achieve one objective in detriment of the other. A good way to explicitly evaluate the trade-off between resource efficiency and user satisfaction is to combine Figs. 4.18 and 4.19 and plot a 2D plane between total cell throughput (capacity) and the satisfaction percentage. Figure 4.20 presents the plane built from the simulations of the studied RRA policies on a scenario with 26 active NRT flows.

In Fig. 4.20, the static RRA policies (RM, PF, SORA-NRT and TSM) are indicated as single markers, and the adaptive policy ATES is indicated as a solid line. One can clearly see the static behavior of the former policies on the capacity-satisfaction plane. TSM is able to provide maximum satisfaction at the expense of low system capacity, while RM is the most efficient on the resource usage but provides an unfair throughput distribution among users (low satisfaction). The PF and SORA-NRT policies appear as fixed trade-offs between TSM and RM, with intermediate user satisfaction and system capacity.

The ATES policy, which controls the parameter σ adaptively according to (4.22) in order to achieve a desired satisfaction target, is able to cover the whole path between the RM and TSM policies in the capacity-satisfaction plane. Notice in the ATES curves that the satisfaction targets set in the simulations (70, 80, 90 %) are always met. One can observe that the performance of the ATES policy for very low and very high satisfaction ranges converges to the performance of the RM and TSM policies, as expected. In this way, it can be concluded that the ATES policy can adaptively adjust the utility-based sigmoid-rule framework presented in Table 4.8 in order to provide a dynamic trade-off between resource efficiency and user satisfaction. Furthermore,

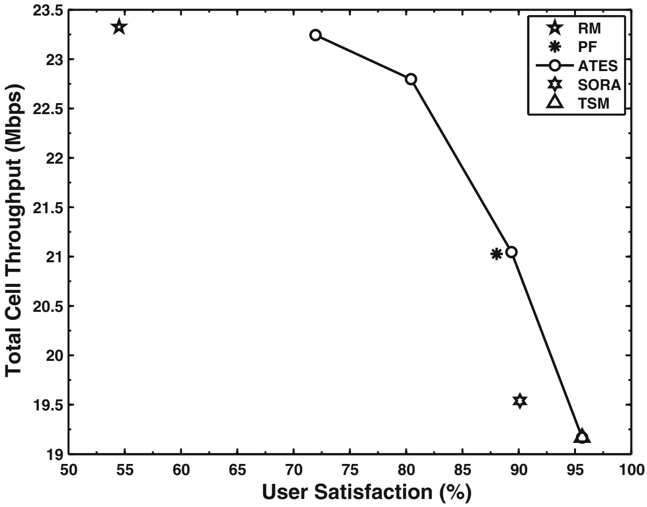

Fig. 4.20 Capacity-Satisfaction plane for the classical policies and the utility-based sigmoid-rule framework considering a system load of 26 users

ATES is able to provide equal or better cell capacity than the static policies for the same satisfaction levels.

4.6.4 Conclusions

We propose to manage the capacity versus QoS trade-off by means of user satisfaction control. CRM and ATES use two different ways to control the satisfaction of the users in the system: instantaneous or average satisfaction control, respectively.

First, we could see that this trade-off can be studied with the optimization problem of maximizing spectral efficiency constrained to minimum satisfaction constraints. We have shown that the optimal solution to this problem can be achieved by ILP solvers that have exponential worst-case computational complexity. According to this, we proposed a low-complexity algorithm following a heuristic framework that first intends to obtain a solution that presents a high spectral efficiency and then, through iterative resource reallocations among users, it fulfills the satisfaction constraints. According to the simulation results, the proposed CRM solution is able to maintain an acceptable performance loss compared to the optimal solution with much lower computational complexity.

Furthermore, an adaptive utility-based RRA framework, which is called sigmoid-rule, was proposed to control this trade-off. The ATES technique, which is derived from this framework, uses an increasing sigmoid function whose shape is determined by a parameter σ that is adapted by a feedback control loop in order to guarantee a given target satisfaction level. The dynamic configuration of the sigmoid function as a linear-shaped or step-shaped function allows several trade-offs between system capacity and user satisfaction. The adaptive ATES technique is able to provide almost the same cell capacity as the static policies (RM, PF, and SORA-NRT) for the same satisfaction levels. Furthermore, it is also able to provide dynamic trade-offs covering the capacity-satisfaction plane. This is a remarkable strategic advantage to the network operators, because they can now control the trade-off between system capacity and user satisfaction and decide in which point on the plane they want to operate.

4.7 Conclusions

In this chapter we studied two important trade-offs in the downlink of wireless mobile networks: capacity versus fairness and capacity versus QoS. Following heuristic and utility-based frameworks for conceiving RRA solutions, we proposed different strategies that are able to achieve adaptive configurations of these trade-offs in a scenario with NRT services.

The use of smart RRA strategies has a great potential to help the network operator to decrease the gap between these opposing design objectives. If these compromises cannot be solved in a “win-win” approach, adaptive RRA strategies are still very useful at finding an appropriate trade-off between these objectives.

Regarding the capacity versus fairness trade-off, we claim that this trade-off can be managed by controlling the fairness in the system. Toward this goal, we proposed two adaptive RRA techniques: FSRM and ATEF. The former uses heuristics to perform an instantaneous fairness control, while the latter uses a dynamic utility-based resource allocation (alpha-rule framework) and a feedback control loop to perform an average fairness control. System-level simulations assuming an LTE-based cellular system demonstrate that both techniques are able to achieve several trade-off operation points according to the network operator’s interests.

We also propose to manage the capacity versus QoS trade-off by guaranteeing certain levels of satisfaction for the users in the system. In this sense, another two adaptive RRA techniques were presented: CRM and ATES. The former is a heuristic technique that solves instantaneously the optimization problem of maximizing the system capacity under minimum satisfaction constraints. The latter performs an average satisfaction control by using the utility-based sigmoid-rule and a feedback control loop that tracks the overall satisfaction of the users and keep it around a desired target value. It was shown by means of system-level simulations of an LTE-based network that the capacity versus QoS trade-off can be successfully controlled by both techniques. This is a strategic advantage to the network operator who is able to design and operate the network according to a planned user satisfaction profile.

Some perspectives for future work are: run simulations with more realistic assumptions like detailed traffic models, mobility, imperfect CSI, etc.; address the same trade-off problems in a mixed traffic scenario with both NRT and RT services;

perform a detailed analysis of the computational complexity and convergence of all techniques; and evaluate how adaptive power allocation algorithms and MIMO technology can help the proposed techniques to achieve even better fairness and satisfaction control.

Appendix: Utility-Based Optimization Formulation for NRT Services

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

$$\max_{\mathscr{K}_j} \sum_{j=1}^{J} U\left(T_j\left[n\right]\right) \tag{4.23a}$$

$$\text{subject to} \qquad \bigcup_{j=1}^{J} \mathscr{K}_j \subseteq \mathscr{K}, \tag{4.23b}$$

$$\mathscr{K}_i \bigcap \mathscr{K}_j = \emptyset, \quad i \neq j, \quad \forall i, j \in \{1, 2, \ldots, J\}, \tag{4.23c}$$

where J is the total number of users in a cell, K is the total number of resources in the system (subcarriers, codes, or the like) to be assigned to the users, $\mathscr{K}$ is the set of all resources in the system, $\mathscr{K}_j$ is the subset of resources assigned to user j, and $U\left(T_j\left[n\right]\right)$ is a monotonically increasing utility function based on the current throughput $T_j\left[n\right]$ of the user j in TTI n. Constraints (4.23b) and (4.23c) 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:

$$T_j\left[n\right] = \left(1 - f^{\text{thru}}\right) \cdot T_j\left[n-1\right] + f^{\text{thru}} \cdot R_j\left[n\right], \tag{4.24}$$

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

Evaluating the objective function in (4.23a) and the throughput expression in (4.24), the derivative of $U\left(T_j\right)$ with respect to the transmission rate R_j is given by:

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

In the case that f^{thru} is sufficiently small, the expression above can be simplified as follows [41]:

$$\frac{\partial U\left(T_j\left[n\right]\right)}{\partial R_j\left[n\right]} \approx f^{\text{thru}} \cdot \left.\frac{\partial U}{\partial T_j}\right|_{T_j=T_j[n-1]}, \tag{4.25}$$

where the previous resource allocation totally determines the current values of the marginal utilities. Using the one-order Taylor formula for the utility function [33, 41] and considering (4.25), we have

$$\begin{aligned}\sum_{j=1}^{J} U\left(T_j\left[n\right]\right) &\approx \sum_{j=1}^{J} U\left(T_j\left[n-1\right]\right)\\ &+ \sum_{j=1}^{J} \left.\frac{\partial U}{\partial T_j}\right|_{T_j=T_j[n-1]} \cdot \left(f^{\text{thru}} \cdot R_j\left[n\right] - f^{\text{thru}} \cdot T_j\left[n-1\right]\right).\end{aligned} \tag{4.26}$$

Notice that maximizing (4.26) leads to the maximization of the original objective function (4.23a). Since f^{thru} is a constant and $T_j\left[n-1\right]$ 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

$$\max_{\mathscr{K}_j} \sum_{j=1}^{J} U^{'}\left(T_j\left[n-1\right]\right) \cdot R_j\left[n\right], \tag{4.27}$$

where $U^{'}\left(T_j\left[n-1\right]\right) = \left.\frac{\partial U}{\partial T_j}\right|_{T_j=T_j[n-1]}$ is the marginal utility (derivative of the utility function) of user j with respect to its throughput in the previous TTI. The objective function (4.27) characterizes a weighted sum rate maximization problem [11], whose weights are adaptively controlled by the marginal utilities.

Notice that we started with an optimization formulation based on throughput given by (4.23a), made some logical assumptions and mathematical simplifications, and ended up with a linear optimization formulation based on instantaneous rates given by (4.27). According to these arguments, we claim that the instantaneous optimization maximizing (4.27) leads to a long-term optimization that maximizes (4.23a).

References

1. 3GPP: Deployment aspects. Technical report TR 25.943 V9.0.0, Third Generation Partnership Project (2009)
2. 3GPP: Evolved universal terrestrial radio access (E-UTRA); physical layer procedures. Technical report TR 36.213 V8.6.0, Third Generation Partnership Project (2009)
3. Agrawal, R., Berry, R., Huang, J., Subramanian, V.: Optimal scheduling for OFDMA systems. In: Proceedings of the 40th Asilomar Conference on Signals, Systems and Computers (ACSSC), pp. 1347–1351 (2006). doi:10.1109/ACSSC.2006.354976
4. Ameigeiras, P., Wigard, J., Mogensen, P.: Performance of packet scheduling methods with different degree of fairness in HSDPA. In: Proceedings of the IEEE 60th Vehicular Technology Conference (VTC-Fall), vol. 2, pp. 860–864 (2004). doi:10.1109/VETECF.2004.1400143

5. Aniba, G., Aissa, S.: Adaptive proportional fairness for packet scheduling in HSDPA. In: Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM), vol. 6, pp. 4033–4037 (2004). doi:10.1109/GLOCOM.2004.1379124
6. Chiu, D.M., Jain, R.: Analysis of the increase and decrease algorithms for congestion avoidance in computer networks. Comput. Netw. ISDN Syst. **17**(1), 1–14 (1989). http://dx.doi.org/10.1016/0169-7552(89)90019-6
7. Doirieux, S., Baynat, B., Begin, T.: On finding the right balance between fairness and efficiency in WiMAX scheduling through analytical modeling. In: Proceedings of the IEEE International Symposium on Modeling, Analysis Simulation of Computer and Telecommunication Systems (MASCOTS), pp. 1–10 (2009). doi:10.1109/MASCOT.2009.5366812
8. 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, Technical University of Berlin (2006)
9. Haider, A., Harris, R.: A novel proportional fair scheduling algorithm for HSDPA in UMTS networks. In: Proceedings of the 2nd International Conference on Wireless Broadband and Ultra Wideband Communications: AusWireless, pp. 1–7 (2007). doi:10.1109/AUSWIRELESS.2007.9
10. Holma, H., Toskala, A. (eds.): WCDMA for UMTS: Radio Access for Third Generation Mobile Communications, 3rd edn. Wiley, New York (2004)
11. Hoo, L.M.C., Halder, B., Tellado, J., Cioffi, J.M.: Multiuser transmit optimisation for multicarrier broadcast channels: asymptotic FDMA capacity region and algorithms. IEEE Trans. Commun. **52**(6), 922–930 (2004)
12. Hosein, P.A.: QoS control for WCDMA high speed packet data. In: Proceedings of the 4th International Workshop on Mobile and Wireless Communications Network, pp. 169–173 (2002). doi:10.1109/MWCN.2002.1045716
13. Hou, H., Zhou, W., Zhou, S., Zhu, J.: Cross-layer resource allocation for heterogeneous traffic in multiuser OFDM based on a new QoS fairness criterion. In: Proceedings of the IEEE 66th Vehicular Technology Conference (VTC-Fall), pp. 1593–1597 (2007). doi:10.1109/VETECF.2007.338
14. IBM: IBM ILOG CPLEX Optimizer. http://www-01.ibm.com/software/integration/optimization/cplex-optimizer/
15. ITU-R: Requirements related to technical performance for IMT-advanced radio interface(s). Technical report M.2134, ITU-R (2008)
16. 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)
17. Jakes, W.C.: Microwave Mobile Communications. Wiley, New york (1994)
18. Jang, J., Lee, K.B.: Transmit power adaptation for multiuser OFDM systems. IEEE J. Sel. Areas Commun. **21**(2), 171–178 (2003)
19. Kelly, F.: Charging and rate control for elastic traffic. Eur. Trans. Commun. **8**, 33–37 (1997)
20. Khan, M.A., Vesilo, R., Collings, I.B., Davis, L.M.: Alpha-rule scheduling for MIMO broadcast wireless channels with linear receivers. In: Proceedings of the Australian Communications Theory Workshop (AusCTW), pp. 110–115 (2009). doi:10.1109/AUSCTW.2009.4805610
21. Kim, I., Lee, H.L., Kim, B., Lee, Y.H.: On the use of linear programming for dynamic subchannel and bit allocation in multiuser OFDM. In: Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM), vol. 6, pp. 3648–3652 (2001)
22. Kivanc, D., Guoqing, L., Hui, L.: Computationally efficient bandwidth allocation and power control for OFDMA. IEEE Trans. Wireless Commun. **2**(6), 1150–1158 (2003)
23. Lee, M., Oh, S.K.: A simple scheduling algorithm capable of controlling throughput-fairness tradeoff performance. In: Proceedings of the IEEE 70th Vehicular Technology Conference (VTC-Fall), pp. 1–4 (2009). doi:10.1109/VETECF.2009.5378861
24. Lima, F.R.M.: Maximizing spectral efficiency under minimum satisfaction constraints on multiservice wireless networks. Ph.D. thesis, Department of Teleinformatics Engineering, Federal University of Ceará, Fortaleza, Brazil (2012)

25. Lima, F.R.M., Maciel, T.F., Freitas, W.C., Cavalcanti, F.R.P.: Resource assignment for rate maximization with QoS guarantees in multiservice wireless systems. IEEE Trans. Veh. Technol. **61**(3), 1318–1332 (2012). doi:10.1109/TVT.2012.2183905
26. Liu, X., Chong, E., Shroff, N.: Opportunistic transmission scheduling with resource-sharing constraints in wireless networks. IEEE J. Sel. Areas Commun. **19**(10), 2053–2064 (2001)
27. Mehlführer, C., Wrulich, M., Ikuno, J.C., Bosanska, D., Rupp, M.: Simulating the long term evolution physical layer. In: EUSIPCO, Glasgow, Scotland (2009)
28. Mo, J., Walrand, J.: Fair end-to-end window-based congestion control. IEEE/ACM Trans. Networking **8**(5), 556–567 (2000). doi:10.1109/90.879343
29. Nemhauser, G., Wosley, L.: Integer and Combinatorial Optimization. Wiley, New York (1999)
30. Palomar, D.P., Fonollosa, J.R.: Practical algorithms for a family of waterfilling solutions. IEEE Trans. Signal Process. **53**(2), 686–695 (2005)
31. Pong, D., Moors, T.: Fairness and capacity trade-off in IEEE 802.11 WLANs. In: Proceedings of the 29th Annual IEEE International Conference on Local Computer Networks, pp. 310–317 (2004). doi:10.1109/LCN.2004.58
32. Rhee, W., Cioffi, J.M.: Increase in capacity of multiuser OFDM system using dynamic subchannel allocation. In: Proceedings of the IEEE 51st Vehicular Technology Conference (VTC), vol. 2, pp. 1085–1089. Spring (2000)
33. Rodrigues, E.B.: Adaptive radio resource management for OFDMA-based macro- and femtocell networks. Ph.D. thesis, Universitat Politècnica de Catalunya, Barcelona, Spain (2011)
34. 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)
35. Rodrigues, E.B., Casadevall, F.: Rate adaptive resource allocation with fairness control for OFDMA networks. In: Proceedings of the 18th European Wireless Conference, pp. 1–8 (2012)
36. Sang, A., Wang, X., Madihian, M., Gitlin, R.D.: A flexible downlink scheduling scheme in cellular packet data systems. IEEE Trans. Wireless Commun. **5**(3), 568–577 (2006). doi:10.1109/TWC.2006.1611087
37. Shakkottai, S., Srikant, R.: Network optimization and control. Found. Trends Networking **2**(3), 271–379 (2007)
38. Shen, Z., Andrews, J.G., Evans, B.L.: Adaptive resource allocation in multiuser OFDM systems with proportional rate constraints. IEEE Trans. Wireless Commun. **4**(6), 2726–2737 (2005)
39. Shi, J., Hu, A.: Maximum utility-based resource allocation algorithm in the IEEE 802.16 OFDMA system. In: Proceedings of the IEEE International Conference on Communications (ICC), pp. 311–316 (2008). doi:10.1109/ICC.2008.65
40. Song, G., Li, Y.G.: Cross-layer optimization for OFDM wireless networks, part I: theoretical framework. IEEE Trans. Wireless Commun. **4**(2), 614–624 (2005)
41. Song, G., Li, Y.G.: Cross-layer optimization for OFDM wireless networks, part II: algorithm development. IEEE Trans. Wireless Commun. **4**(2), 625–634 (2005)
42. Uchida, M., Kurose, J.: An information-theoretic characterization of weighted alpha-proportional fairness. In: Proceedings of the IEEE 28th Conference on Computer Communications (INFOCOM), pp. 1053–1061 (2009). doi:10.1109/INFCOM.2009.5062017
43. Viswanath, P., Tse, D.N.C., Laroia, R.: Opportunistic beamforming using dumb antennas. IEEE Trans. Inf. Theory **48**(6), 1277–1294 (2002). doi:10.1109/TIT.2002.1003822
44. Wong, C.Y., Cheng, R.S., Lataief, K.B., Murch, R.D.: Multiuser OFDM with adaptive subcarrier, bit, and power allocation. IEEE J. Sel. Areas Commun. **17**(10), 1747–1758 (1999)
45. Wong, I.C., Shen, Z., Evans, B.L., Andrews, J.G.: A low complexity algorithm for proportional resource allocation in OFDMA systems. In: Proceedings of the IEEE Workshop on Signal Processing Systems, pp. 1–6 (2004)
46. Yang, L., Kang, M., Alouini, M.S.: On the capacity-fairness tradeoff in multiuser diversity systems. IEEE Trans. Veh. Technol. **56**(4), 1901–1907 (2007). doi:10.1109/TVT.2007.897229
47. Zhang, Y.J., Letaief, K.B.: Multiuser adaptive subcarrier-and-bit allocation with adaptive cell selection for OFDM systems. IEEE Trans. Wireless Commun. **3**(5), 1566–1575 (2004). doi:10.1109/TWC.2004.833501