# Resource allocation in networks via coalitional games

Shams, F.

**Author**: Shams, F.
**Title**: Resource allocation in networks via coalitional games
**Issue Date**: 2016-09-21

## Chapter 2

# A survey on resource allocation techniques in OFDM(A) networks

Interest in orthogonal frequency-division multiplexing (OFDM), has grow steadily, as it appears to be the most efficient air-interface for wireless communications primarily due to its inherent resistance to frequency-selective multipath fading and the flexibility it offers in radio resource allocations. One of the crucial issues in OFDM transmission is the allocation of the power resources to the available subchannels.

This chapter presents a survey on the radio resource allocation techniques in OFDM and orthogonal frequency division multiple access (OFDMA) networks. This problem goes back to 1960s and that is related to properly and efficiently allocate the radio resources, namely subcarriers and power. We start by overviewing the main open issues in OFDM. Then, we describe the problem formulation in OFDMA, and we review the existing solutions to allocate the radio resources. The goal is to discuss the fundamental concepts and relevant features of different radio resource management criteria, including water-filling, max-min fairness, proportional fairness, cross-layer optimization, utility maximization, and game theory, also including a toy example with two terminals to compare the performance of the different schemes. We conclude the survey with a review of the state-of-the-art in resource allocation for next-generation wireless networks, including multicellular systems, cognitive radio, and relay-assisted communications, and we summarize advantages and common problems of the existing solutions available in the literature. The distinguishing feature of this contribution is a tutorial-style introduction to the fundamental problems in this area of research, intended for beginners on this topic.

The use of orthogonal frequency division multiplexing (OFDM), as a modulation, and of orthogonal frequency division multiple access (OFDMA), as a channel access scheme, has grown steadily, as they appear to be the most efficient solutions for wireless communications, primarily due to their inherent resistance to frequency-selective multipath fading and the flexibility they offer to radio resource allocation [86]. One of the crucial issues in OFDM(A) transmission is the allocation of the power to the available subchannels. Even though the OFDM(A) concept is simple in its basic principle, building a practical OFDM(A) system is far from being a trivial task without a well-devised resource allocation algorithm. In the case of a multiple-access network, a typical resource allocation problem in OFDMA is based on assigning a subset of available subcarriers to simultaneously transmitting (and thus interfering) wireless terminals and (possibly jointly) adjusting the power amount over each used subcarrier in order to guarantee the minimum required quality of service (QoS). An efficient algorithm for subcarrier selection can significantly increase the signal-to-interference-plus-noise ratio (SINR), that is necessary to enhance the throughput in a dynamic scenario. Similarly, regulating the transmit power in wireless cellular networks constitutes a key degree of freedom in the management of interference, energy, and connectivity. This motivates us to revisit the relevant criteria for resource management in present and next-generation wireless networks.

This chapter aims at providing a survey on state-of-art research, providing an overview of selected topics in the context of OFDM(A) systems. We start with historical notes in the following section. Then, we provide a description of resource allocation issue in OFDM in Sect. 2.2. In Sect. 2.3, we describe different fashions of allocation of radio resources in OFDMA. We continue with exploration of three classic power allocation solutions of water-filling, max-min fairness, and weighted proportional fairness in Sects. 2.3.1,2.3.2, and 2.3.3, respectively. Sects. 2.3.4 and 2.3.5 discuss about two important resource allocation issues in multi service traffic networks: utility maximization, and cross layer. Different solutions based on game theory are reviewed in Sect. 2.4. Finally, we summarize key features of the existing solutions in Sect. 2.8.

## 2.1 Orthogonal frequency division modulation and multiple access

The basic principle of OFDM is to transmit data by dividing them into several interleaved bit streams, and using these to modulate several carriers. This concept helps reducing the detrimental effects of multipath fading in communication systems. In brief, OFDM is a parallel transmission scheme, where a high-rate serial data stream is split up into a set of low-rate substreams with generally equal bandwidth, each of which is modulated on a separate *subcarrier* (called also *subchannel* or *tone*). Thereby, the bandwidth of the subcarriers becomes small compared with the coherence bandwidth of the channel, so that the individual subcarriers experience flat fading, thus enabling a simple equalization. This implies that the symbol period of the substreams is made long compared to the delay spread of the time-dispersive radio channel. While each subcarrier is separately modulated by a data symbol, the overall modulation operation across all the subchannels (multicarrier modulation) results in a frequency multiplexed signal, so as to accommodate very high throughputs in severe frequency-selective scenarios.

This solution was proposed for the first time by Doelz *et al.* for the U.S. military HF communication applications in 1957 in the pioneering Collins Kineplex system [87]. This led to a few OFDM schemes in the 1960s, which were proposed by Saltzberg [88] and Chang [89]. In the late 1960s, the multicarrier concept was adopted in some military applications, such as KATHRYN [90] and ANDEFT [91]. These systems involved a large hardware complexity, since the parallel data transmission was essentially through a bank of oscillators, each tuned to a specific subcarrier. The first patent on OFDM was granted in 1970 [92]. The major contribution to the OFDM scheme came after the results of Weinstein and Ebert [93], who demonstrated that using discrete Fourier transforms (DFT) to perform the baseband modulation and demodulation considerably increases the efficiency of modulation and demodulation processing. The adoption of OFDM has been finally facilitated by the efficient implementation of fast Fourier transform (FFT) and inverse FFT (IFFT) algorithms in digital signal processing (DSP) chips.

OFDM is extremely effective in a time dispersive environment where signals can have many paths to reach their destinations, resulting in variable time delays. With classical modulations, these time delays cause one symbol to interfere with the next one(s) (giving rise to inter symbol interference, ISI) at high bit-rates. With the OFDM, all of the sinus cardinal (sinc)-shaped subchannel spectra exhibit zero crossings at all the other subcarriers' frequencies, and the subchannel spectra result to be orthogonal to each other. The orthogonality among different tones ensures that the subcarrier signals do not interfere with each other, when communicating over perfectly distortionless channels.

Although the ISI is mitigated by the guard interval between consecutive OFDM symbols and the raised-cosine filtering OFDM imposes, it is not completely eliminated. To attain perfect orthogonality between subcarriers in a time dispersive channel, Peled and Ruiz [94] introduced the notion of *cyclic prefix* (CP): the guard interval is filled with a cyclic extension of each time domain OFDM symbol, in order to overcome the inter-OFDM symbol interference due to the channel memory. The CP performs the circular convolution by the channel under the assumption that the channel impulse response is shorter than the length of the CP, thus preserving the orthogonality of subcarriers. Although adding the CP causes power and spectrum efficiency loss, this deficiency is highly compensated by the ease of receiver implementation that makes OFDM both practical and attractive to the radio link designers.

OFDM has in fact been adopted by many European and American telecommunication standards in the last few decades. In the context of wired environments, OFDM is applied for high speed digital voice services, e.g., asymmetric digital subscriber lines (ADSL) [95] and its faster version, very-high-bit-rate digital subcarrier line (VDSL) [96]. In wireless communications, the OFDM technique is the fundamental building block of the IEEE 802.16 standards and it has been considered as a solution to mitigate multipath propagation in broadband multimedia broadcasting, e.g., digital video broadcasting for terrestrial television (DVB-T) [97], digital audio broadcasting (DAB) [98], and 3G mobile communication (3GPP-LTE) [99]. To summarize, the wide interest in OFDM technique is due the following advantages:

- high spectral efficiency;
- interference suppression capability through the use of the CP;
- protection against narrowband interference and inter carrier interference (ICI);
- efficient implementation using FFT;
- flexible spectrum adaptation; and

- separated subcarrier modulation, which implies that different constellations can be applied on individual subcarriers, thus allowing for several resource allocation strategies.

Even though the concept of multicarrier transmission is simple in its basic principle, the design of practical OFDM systems is far from being a trivial task. Synchronization, channel estimation, and radio resource management are only a few examples of the numerous challenges related to multicarrier technology. As a result of continuous efforts of many researchers, most of these challenging issues have been studied and several solutions are currently available in the open literature. Besides its significant advantages, OFDM suffers from the following disadvantages:

- high peak-to-average power ratio (PAPR), which requires highly linear amplifiers and consequently high power consumption [100];
- sensitivity to Doppler effects and carrier frequency offsets [101];
- sensitivity to phase noise, and time and frequency synchronization problems [102]; and
- loss in data rate due to the guard interval insertion.

OFDM is also good from the standpoint of multiple access opportunities. Compared to single carrier systems, OFDM is a versatile modulation, that can be adopted to provide channel access scheme for multiple access systems, in that it intrinsically facilitates both time-division multiple access (TDMA) and frequency-division (or subcarrier-division) multiple access. In a multiuser scenario, the available bandwidth must be shared among several users. Each user may experience different conditions in terms of path loss and shadowing. Furthermore, each user may have different requirements in terms of QoS. An acceptable design of the network should therefore take into account the different user conditions while providing fairness, without a drastic reduction in the overall spectral efficiency.

To meet these needs, in 1998 a combination of OFDM and frequency division multiple access (FDMA), called OFDMA, was proposed by Sari and Karam for cable television (CATV) networks [103]. OFDMA is a promising multiple access scheme that has attracted interest for wireless metropolitan area networks (MANs), as it inherits the immunity to ISI and frequency selective fading of OFDM. Furthermore, in OFDMA systems different modulation schemes can be employed for different users. For instance, each user, according to its distance from the base station (BS), can invoke different orders of modulation schemes (either high- or low-order modulation) to increase its data rate.

An extension of the multiple-access technique ALOHA [86] over OFDMA was first proposed by Shen *et al.* in [104], and further discussed in [105, 106]. In [107], Qin *et al.* propose a distributed access protocol denoted as *channel-aware* ALOHA. The authors extend this idea in [108] to OFDMA systems where water-filling is performed on subcarriers. It can be easily applied to OFDMA selecting subcarriers for different users. Besides, the authors show that the algorithm can reach the Shannon capacity: the users with best channel conditions are usually transmitting if the channel is invariant over the necessary time to manage collisions.

To conclude, its implementation flexibility, the low complexity equalizer required in the transceiver, as well as the attainable high performance, make OFDMA a highly attractive candidate for high data rate communications over time-varying frequency selective multiuser radio channels. Compared to classical FDMA, OFDMA presents a higher spectral efficiency by avoiding the need for large guard bands between different users' signals. The main advantages of OFDMA are the increased flexibility in resource management and the ability for a dynamic channel assignment. OFDMA can exploit the channel state information (CSI) to provide users with the best subcarriers (in terms of channel condition between transmitter and receiver over different subcarriers) that are available, thereby leading to remarkable gains in terms of achievable data throughput. In terms of architecture complexity, OFDMA systems can now be implemented using powerful integrated circuits optimized to perform FFT operations. Because of its increasingly widespread acceptance as the modulation scheme of wireless networks of the future, it attracts a lot of research attention, in areas like resource allocation, time-domain equalization, PAPR reduction, phase noise mitigation and pulse shaping.

Thanks to its favorable features, OFDMA is widely recognized as the technique that is able to meet the requirements for fourth generation broadband wireless networks, as witnessed by the IEEE 802.16m [86] and LTE-Advanced standards [109]. In the next sections, we will focus on resource allocation techniques which includes subcarriers selection and power allocation, by leveraging on multiuser diversity and channel fading.

## 2.2 Resource allocation in OFDM

The research interest on resource allocation in multicarrier systems was encouraged by the successfully development of ADSL services in the 1990s [95]. This technology employs, for high-speed wireline data transmissions, a digital multitone (DMT) modulation, a particular form of frequency multitone (FMT) technology, which is a frequency-division-based transmission technique. Due to crosstalk from adjacent copper twisted pairs, the ADSL channel is characterized by strongly frequency-selective noise. This scenario is similar to that experienced in the OFDM transmission scheme, which is flexible enough to allocate individual power and modulation on different subcarriers.

In the context of OFDM, different criteria to allocate the available resources can be performed depending upon whether the network is trying to maximize the overall data rate under a total power constraint, or to minimize the overall transmit power given a fixed data rate or bit error rate (BER). The optimal OFDM adaptation algorithm, called the water-filling (WF) criterion [110] and originally derived for DMT systems, tends to allocate most information bits onto the highest signal-to-noise ratios (SNRs) carriers. Note that the number of bits determines the constellation size as follows: 1 bit corresponds to binary phase-shift keying (BPSK) modulation, 2 bits to quadrature phase-shift keying (QPSK) modulation, 4 bits to 16-quadrature amplitude modulation (16-QAM), and so on. In some situations, some subcarriers may even be left unassigned if their SNRs are too low to provide reliable data transmission.

In the literature, the problem of efficient bit allocation on the available subchannels and using the best efficient modulation (to each subcarrier) is equivalently referred to as *bit loading*, *adaptive modulation*, and *link adaptation*. In an OFDM communication, the (unique) transmitter spending power $p_n \leq \overline{p}_n$ over the $n$th subcarrier, with $\overline{p}_n$ being the maximum power constraint on subcarrier $n$, can use a number of bits $b_n$ that is calculated using the Shannon channel capacity formula as [111, Eq. 1]:

$$b_n = \left\lfloor \log_2 \left( 1 + \frac{|H_n|^2 \, p_n}{(\xi + \Gamma) \cdot \sigma_w^2} \right) \right\rfloor \tag{2.1}$$

where $\lfloor \cdot \rfloor$ is the floor operator, $|H_n|^2$ is the amplitude of the (complex) frequency response of subcarrier $n$, $\sigma_w^2$ is the noise power on each subcarrier, $\xi$ is the additional amount of noise that the system can tolerate while achieving the minimum desired

BER requirement [111], even when the noise level is increased by a factor $\xi$, and $\Gamma$ is the SNR gap (also known as the normalized SNR), used to evaluate the relative performance of a modulation scheme versus the theoretical capacity of the channel [112]. Therefore, by increasing the value of $\xi$, we can improve the system robustness against noise, and hence have the new operating point of the constellations at a distance of $10\log_{10}(\xi+\Gamma)$ dB from the Shannon limit.

There are many theoretical works that aim at regulating the transmit powers $\{p_n\}$ to perform the adaptive bit loading (2.1). In the following we cite some of the pioneering and well-known bit loading algorithms in the context of OFDM systems:

- Hughes-Hartogs in 1987 [113] designed a greedy algorithm to approximate the WF (e.g., see [110]) for twisted-pair channels over an additive white Gaussian noise (AWGN) channel with ISI. The goal of this discrete bit loading algorithm is the minimization of the transmit power under a BER and data rate constraints for each tone. This is accomplished by successively assigning bits to carriers, each time choosing the carrier that requires the least incremental power, until the given target rate is reached. Bingham in [114] proposes to apply sinc functions for each individual spectra instead of using quadrature amplitude shift keying (QASK) in [113]. Applying the technique proposed in [114] allows us to separate signals at the receiver using computationally efficient FFT techniques, although the high complexity burden of the proposed algorithm makes it unsuitable for a practical implementation is high-speed wireless networks.

- The principle of adaptive modulation and power over OFDM was recognized in 1989 by Kalet [115], who simulates a twisted-pair OFDM system, in which each subcarrier uses QAM to maximize the bit rate. The power distribution between the subcarriers and the number of bits per symbol per subcarrier is optimized for a given BER, showing that the proposed power allocation achieves similar results to the WF solution. Furthermore, multicarrier QAM performance is about 9 dB worse than the channel capacity, irrespectively of the channel response. Quantitative results for a twisted-pair cable show that multitone QAM transmission outperforms single-tone QAM by more than 40%. The Kalet's algorithm is often referred to as WF in the frequency domain, which is a simpler version of the technique proposed by Cimini in [116] for mobile communication channels.

- Chow *et al.* in [111] proposed an iterative bit loading algorithm which offers significant advantages over the Hughes-Hartogs algorithm [113] and the WF method [115, 116]. The simulation results over high-speed ADSL service using a required BER of $10^{-7}$ show a maximum degradation of only 1.3 dB in terms of SNR compared to [115]. Even though the proposed algorithm is faster than that Hughes-Hartogs one, it is not optimal in terms of number of iterations and computational load.

- Czylwik [117] in 1996 simulates an OFDM transmission system with time-variant channel functions, measured with a wideband channel sounder with fixed carrier frequency antennas. The simulation results show that the proposed subcarrier-adaptive modulation demands a total power consumption at least 5 dB lower (and reaching 15 dB lower, depending on the propagation scenario) than that required by the non-adaptation (fixed modulation) OFDM, by placing a requirement in terms of BER equal to $10^{-3}$. Different modulation formats can be selected so as to minimize the BER under a constant data rate constraint.

- Fischer and Huber in 1996 [118] proposed a bit loading algorithm to reduce the computational complexity of Hughes-Hartogs and Chow algorithms. This algorithm distributes bits and transmit power to maximize the SNR over each carrier. Van-der Perre *et al.* in [119] apply [118] to simulate the performance of OFDM-based high speed wireless LANs (with data rate on the order of 100 Mb/s). Simulation results show that the proposed adaptive loading strategy improves the system performance considerably, with an SNR gain of 6 dB with respect to the fixed QPSK or 16-QAM modulations, under a BER constraint of $10^{-2}$.

As a conclusion, all link adaptation studies reported here have demonstrated that a performance improvement in OFDM systems can be attained by properly adjusting power and data rate over each subcarrier, so as to exploit the channel frequency selectivity. To further increase the capacity of the system, state-of-the-art solutions always adopt *coding techniques*. In the practice, a frequency-selective radio channel may severely attenuate the data symbols transmitted on several subcarriers, leading to bit errors. By spreading the coded bits over the bandwidth of the transmitted system, an efficient coding scheme can correct for the erroneous bits and thereby exploit the wideband channel frequency diversity.

To this aim, all communication systems include forward error correction (FEC) coding techniques [86] to attain the system SNR requirements at low required BER values. In FEC schemes, only the error correction is performed, whereas in automatic repeat request (ARQ) [86] schemes, the retransmission of erroneous blocks is requested whenever the decoded data is labeled as unreliable. OFDM systems that utilize *adapting modulation and coding per subcarrier* are often referred as coded OFDM (COFDM) systems [120]. COFDM increases the data rate and outperforms solutions using only either modulation or channel coding. COFDM does not adapt the data rate of each subcarrier due to the differing SNRs, rather it uses the same high-order modulation on an all subcarriers and uses coding to correct the errors.

Recently, there has been numerous interest on the design of good error-correcting codes achieving near-Shannon performance, particularly low-density parity-check (LDPC) codes [121], that are well suited for OFDM systems [122], as they can reduce the impact of deep channel fades in both the time and the frequency domains. In high data rate wireless OFDM systems, current challenges include the design of LDPC codes with reasonable block length (and thus with feasible encoding and decoding complexity) and overhead delay [123].

In addition to the signal strength, the wireless medium may also affect the original signal through *dispersion*, which includes time dispersion (frequency selective) and frequency dispersion (time selective) fading. While OFDM is immune to the time dispersion effect at the expense of CP, it is not guaranteed whether the signals across different subchannels will not interfere to each other. Hassibbi and Hochwald in [124] pioneered to propose a linear space-time coding, called *linear dispersion coding* and also *linear constellation coding*, for high data rate communications with large number of subcarriers. The codes are designed to optimize mutual information between transmitter and receiver. To reduce the decoding complexity, [125] divides subcarriers into a number of disjoint groups based on criterions in [126] and [127] for reducing multiuser interference and PAPR, respectively. Then, to apply the same linear dispersion coding to subcarriers within each group and to transmit every information symbol over subcarriers within only one group. The idea of grouping subcarriers in smaller groups is used by references [128] and [129] which aims at minimization BER in single user and multi users, respectively. The open problem of linear pre-coding related works is the gap between achieved data rate and the outer region capacity.

## 2.3 Resource allocation in OFDMA

A typical case of a multiple access channel is the uplink of a cellular system. In general, in the OFDMA uplink scenario, each user receives a channel assignment and a power allocation from the BS that consists of a (usually exclusive) subset of subcarriers and power levels on each of them. In an OFDMA network, the BS must optimally allocate power and bits over different subcarriers based on instantaneous channel conditions of different active wireless terminals. The only requirement is that the fading rate is not too fast (compared to the OFDMA symbol time), as instantaneous resource allocation is impractical in the presence of rapidly-varying transmission channels of mobile terminals. Other impairments include interference management and limited resources, such as bandwidth and transmit power. This makes the link adaptation task much more challenging than in single-user systems. Recent exhaustive surveys on these topics include [130] (with emphasis on scheduling schemes), [131] (with focus on ICI mitigation), [132] (with emphasis on relay communications), and [133] (with focus on game-theoretic approaches). In the remainder of this contribution, we aim at introducing the very basics for OFDMA resource allocation by means of detailed problem formulation and numerical examples, which, to the best of our knowledge, is not available in the literature.

It is clear that, compared to a point-to-point single-user OFDM-based connection, a multiuser OFDMA link adaptation is much more complicated and hardly scalable [134]. In particular, one of the main problems with OFDMA is the large amount of feedback required from the users. Since different users can be scheduled over different subcarriers, they must feed the measurement information back about *every* subcarrier to the BS. Consider a network with $K$ active OFDM mobile terminals and $N$ total available subcarriers. The scheduler requires full channel state information (CSI) consisting of $K \cdot N$ complex numbers (the values of the channel frequency response at each subcarrier for every user). This feedback information represents a very large overhead if there are many users and subcarriers in the system. To reduce it, Cimini *et al.* [135] proposed to group adjacent subcarriers into *clusters* and to feed back the information about the best cluster(s) in terms of channel quality. In [136, 137], it is shown that sending back only heavily quantized CSI dramatically reduces the feedback needs without significantly sacrificing the overall performance. In this context, Svedman *et al.* [138] showed that a suitable cluster size, which highly impacts on the performance in terms of achieved downlink throughput, must be selected according to the average channel delay spread of the users.

In this context, let us start to review the main concepts and the main categories behind bit and power loading in OFDMA-based transmissions. Generally, a resource allocation algorithm can either be *centralized* or *distributed*. In centralized schemes, such as [16,139], the algorithm is run by a central unit (in an infrastructure networks, typically the BS) that is aware of the demands and of the channel conditions of all mobile terminals, as described in the previous paragraph. In a distributed model (such as [140]), each mobile terminal tries to accomplish its own (minimum) QoS autonomously, sometimes resorting to cross-layer approaches (e.g., [141]), to reduce the total power consumption and to support different services and traffic classes, mostly for the downlink of an OFDMA system. In general, centralized techniques show better performance at the expense of a higher signaling between terminals and the central unit, and lower scalability than distributed techniques.

Another typical classification of resource allocation techniques for OFDMA networks is based on the objective of the optimization problem. The solutions available in the literature mainly fall into two different categories: *margin-adaptive* and *rate-adaptive* methods. The goal of margin adaptive schemes [142] is to minimize the total transmit power expenditure given a set of fixed user data rates and BER requirements. Algorithms based on the rate-adaptive criterion [143] aim on the contrary at achieving the maximum total (continuous) sum-rate over all users subject to different QoS constraints, e.g., power expenditure. Note that, unlike some broadband systems, e.g., based on code division multiple access (CDMA), ultra-wideband (UWB), and multicarrier CDMA (MC-CDMA), in which the whole bandwidth is shared by all active wireless terminals, OFDMA-based networks do not consider resource allocation strategies based on the *mean-BER minimization*, in which the robustness of the system is enhanced by allocating bits and powers to subcarriers to minimize the error rate of an entire symbol. This scheme is not of major interest for OFDMA systems, since, as will be seen in the next subsection, in (almost) all OFDMA resource allocation techniques each subcarrier is not permitted to be assigned to more than one user. This means that a well devised algorithm to maximize the total data rate also results in minimizing each user's BER.

The first resource allocation strategy presented here is the minimization of the OFDMA system power expenditure for a given target data rate, solving the following margin-adaptive optimization problem:

$$\min_{\mathbf{p},\,\mathcal{N}} \sum_{k=1}^{K} \sum_{n\in\mathcal{N}_k} p_{kn} \tag{2.2a}$$

$$\text{s.t.} \quad \sum_{n\in\mathcal{N}_k} R_{kn} \geq \underline{R}_k \qquad \forall\, k \in \mathcal{K} \tag{2.2b}$$

$$\text{and} \quad \sum_{n\in\mathcal{N}_k} p_{kn} \leq \overline{p}_k \qquad \forall\, k \in \mathcal{K} \tag{2.2c}$$

$$\text{and} \quad \mathcal{N}_k \cap \mathcal{N}_m = \emptyset \qquad \forall\, k,\, m \in \mathcal{K}, k \neq m, \tag{2.2d}$$

where $k \in \mathcal{K} = [1, \ldots, K]$ denotes the index of the wireless terminal which transmits with powers $\mathbf{p}_k = \left[p_{k1}, \ldots, p_{kn}, \ldots, p_{kN}\right]$ over the $N$ subcarriers, which are represented by the set $\mathcal{N} = [1, \ldots, n, \ldots, N]$, and $\mathbf{p} = [\mathbf{p}_1, \ldots, \mathbf{p}_k, \ldots, \mathbf{p}_K]$. Let $\mathcal{N}_k \subset \mathcal{N}$ be the set of subcarriers assigned to user $k$, and $R_{kn}$ be the channel capacity that can be achieved by user $k$ over the $n$th subcarrier. The sets of assigned subcarriers are disjoint, as explicitly stated by (2.2d): this means that each subcarrier is not allowed to be shared by more than one terminal. Each user $k$ wishes to attain its target rate $\underline{R}_k$, as specified in (2.2b), under the constraint $\overline{p}_k$ on its total transmit power, as formulated in (2.2c). It is clear that, for each terminal $k$ and every $n \notin \mathcal{N}_k$, we have $p_{kn} = 0$, and accordingly, $R_{kn} = 0$. The overall data rate of each user is obtained by the Shannon capacity formula as:

$$R_k = \sum_{n\in\mathcal{N}_k} R_{kn} = \sum_{n\in\mathcal{N}_k} \log_2\left(1 + \frac{|H_{kn}|^2\, p_{kn}}{\sigma_w^2}\right) \tag{2.3}$$

in which $|H_{kn}|^2$ denotes the amplitude of the Gaussian-complex path gain experienced by user $k$ on subcarrier $n$, and, similarly to Sect. 2.2, $\sigma_w^2$ is the power of the AWGN zero mean Gaussian noise on each subcarrier.

Two levels of decomposition are necessary to turn this NP-hard problem into the set of subproblems, subcarrier allocation and power control [144]. In fact, the exclusive assignment of subcarriers to users is a way to reduce the complexity computation of the optimization problem (2.2a), as the rate $R_{kn}$ can be computed using (2.3). On the other hand, as users are not allowed to share a common subcarrier, the allocation process boils down to a combinatorial optimization problem, for which no optimal greedy solution exists. Kivanc *et al.* [145] developed a computationally inexpensive method for OFDMA resource assignment which achieves a comparable performance with respect to the solution of the NP-hard problem (2.2) in terms of transmission power and bandwidth efficiency at a reduced computational complexity. However, this approach does not provide a fair apportionment among users, so that some of them may be dominant in terms of resource occupancy even when the minimum rate requirement is not satisfied for the others.

In addition to this limitation, the margin-adaptive formulation that focuses on the minimization of transmit powers is often of lower interest compared to the maximization of the data rates. For this reason, the most common optimization problem for OFDMA systems is the rate-adaptive one, that aims at maximizing the bit rate as follows:

$$\max_{\mathbf{p},\,\mathcal{N}} \sum_{k=1}^{K} \sum_{n\in\mathcal{N}_k} R_{kn} \tag{2.4a}$$

$$\text{s.t.} \quad \sum_{n\in\mathcal{N}_k} R_{kn} \geq \underline{R}_k \qquad \forall\, k \in \mathcal{K} \tag{2.4b}$$

$$\text{and} \quad \sum_{n\in\mathcal{N}_k} p_{kn} \leq \overline{p}_k \qquad \forall\, k \in \mathcal{K} \tag{2.4c}$$

$$\text{and} \quad \mathcal{N}_k \cap \mathcal{N}_m = \emptyset \qquad \forall\, k,\, m \in \mathcal{K}, k \neq m. \tag{2.4d}$$

The objective of this problem is to distribute bits and power among different subcarriers in such a way that the overall data rate of the system is maximized. Most algorithms focus on the downlink scenario, with constraints on the total power transmitted by the radio BS. In the uplink scenario, restrictions apply on an individual basis to each user terminal, and the simplest solution to maximize the channel capacity of mobile devices under a power constraint is the WF criterion [146], described in the following subsection.

### 2.3.1 The water-filling solution

Cheng and Verdú in [147] pioneered the application of the WF solution in an uplink OFDMA network scenario, and derived the capacity region and the optimal power allocation of individual users. In the rate-adaptive optimization, the channel capacity is obtained by maximizing the right-hand side of (2.3) with respect to (2.4c), i.e.,

$$\max_{\mathbf{p}\,,\mathcal{N}} \left\{ \sum_{k=1}^{K} \sum_{n\in\mathcal{N}_k} \log_2\left(1 + \frac{|H_{kn}|^2\, p_{kn}}{\sigma_w^2}\right) \right\}. \tag{2.5}$$

Since the objective function in (2.5) is convex in the variables $\{\mathbf{p}_k\}$, the optimum power allocation under the convex constraints of overall transmit power can be found using Lagrangian methods [148]. The optimal strategy to satisfy (2.4a) is such that each subcarrier $n \in \mathcal{N}$ is assigned to the user with the largest channel gain in a centralized way as follows:

$$k \longleftarrow \arg\max_{\ell \in \mathcal{K}} |H_{\ell n}|^2 \,. \tag{2.6}$$

The resulting optimal power allocation for user $k$ is given by:

$$p_{kn} = \left[ \frac{1}{\lambda_k} - \frac{\sigma_w^2}{|H_{kn}|^2} \right]^+ \tag{2.7}$$

where $[x]^+ = \max\{x, 0\}$, and $\lambda_k$ is the Lagrangian parameter ("water-level"), chosen such that the sum of the allocated powers satisfies the total power constraint $\overline{p}_k$:

$$\lambda_k = |\mathcal{N}_k| \cdot \left( \overline{p}_k + \sum_{n \in \mathcal{N}_k} \frac{\sigma_w^2}{|H_{kn}|^2} \right)^{-1} \,. \tag{2.8}$$

To conclude, the WF is a greedy (centralized) power allocation scheme that increases the channel capacity by assigning every subcarrier to the user with the best path gain, and by distributing the power according to (2.7). Note that the WF solution is highly unfair, since only users with the best channel gains receive an acceptable channel capacity, while users with bad channel conditions (e.g., far users) achieve very low data rates. More information-theoretic discussions on related topics can be found in [149]. To derive fair resource allocation schemes, we resort to other techniques, described in the following subsections.

### 2.3.2 The max-min fairness criterion

In an OFDMA network, one possible approach to overcome the unfairness of WF is described in [150]. This alternative formulation aims at maximizing the minimum data rate across users, thus enforcing the notion of *max-min rate-maximization fairness* that avoids the starvation of some users.

**Definition 17** *A feasible*[1] *rate vector* $\mathbf{R} = [R_1, \ldots, R_k, \ldots, R_K]$ *is defined to be* max-min fair *if any rate $R_k$ cannot be increased without decreasing some other rate $R_m$, $m \neq k$, which is smaller than or equal to $R_k$.* ■

[1]A rate allocation $\mathbf{R}$ is *feasible* if the network resources are enough to provide every user $k$ in

Roughly speaking, in the max-min power control the objective is to optimize the performance of the worst link amongst all users for a fixed QoS-based power control approach. The idea behind the max-min fair approach is to treat all users as fairly as possible, by making all rates as large as possible [151]. The work of Rhee and Cioffi in [150] is an extension of [152], which is a dual problem of minimizing the total transmit power for given data rate requirements. The problem is formulated as the following convex optimization problem [150]:

$$\max_{\mathbf{p},\,\mathcal{N}} \min_{k\in\mathcal{K}} \sum_{n\in\mathcal{N}_k} R_{kn} \tag{2.9a}$$

$$\text{s.t.} \quad \sum_{n\in\mathcal{N}_k} p_{kn} \leq \overline{p}_k \qquad \forall\, k \in \mathcal{K} \tag{2.9b}$$

$$\text{and} \quad \mathcal{N}_k \cap \mathcal{N}_m = \emptyset \qquad \forall\, k,\, m \in \mathcal{K}, k \neq m. \tag{2.9c}$$

The Lagrangian relaxation [148] algorithm proposed in [150, 152] approaches the solution to (2.9a) by slowly increasing the power level for each user. By elaborating on a simple iterative algorithm to compute a suboptimal max-min fair rate vector proposed by Bertsekas and Gallager in [153, p. 527], we can easily extend it for an OFDMA network as follows:

1) *Zero initialization:* Supposing $K \ll N$, the algorithm starts with an all-zero data rate vector, i.e., $R_k = 0$ and $\mathcal{N}_k = \emptyset \quad \forall\, k \in \mathcal{K}$.

2) *Round-robin step:* Assign every user $k \in \mathcal{K}$ the subcarrier $n$ whose channel gain $|H_{kn}|^2$ is the highest among the remaining ones, using a uniform power $\overline{p}_k/N$ as:

$$n \longleftarrow \arg\max_{m\in\mathcal{N}} |H_{km}|^2; \tag{2.10a}$$

$$\mathcal{N}_k = \mathcal{N}_k \cup \{n\}; \tag{2.10b}$$

$$\mathcal{N} \leftarrow \mathcal{N} \setminus \{n\}; \tag{2.10c}$$

$$R_k = R_{kn}^{1/N}, \tag{2.10d}$$

---

the network with rate $R_k$. To the best of the authors' knowledge, the algorithms available in the literature do not propose criteria to assess the a-priori feasibility of a certain vector $\mathbf{R}$. The remainder of this paper is thus based on the assumption that the network resources can guarantee achievable rates $R_k$, e.g., based upon the Shannon capacity [111] and some performance gaps, such as those mentioned in Sect. 2.2 for the single-user scenario.

where

$$R_{kn}^{1/N} = \log_2\left(1 + \frac{|H_{kn}|^2\,\overline{p}_k}{N\sigma_w^2}\right). \tag{2.11}$$

At this point, every user $k \in \mathcal{K}$ is assigned exactly one subcarrier.

3) *Best user rate update:* Find the user $k$ with the smallest attained data rate, i.e., $k \leftarrow \arg\min_{\ell\in\mathcal{K}} R_\ell$, and then assign to it the subcarrier $n \in \mathcal{N}$ with the best channel condition $|H_{kn}|^2$, and update its data rate as:

$$k \leftarrow \arg\min_{\ell\in\mathcal{K}} R_\ell; \tag{2.12a}$$

$$n \longleftarrow \arg\max_{m\in\mathcal{N}} |H_{km}|^2; \tag{2.12b}$$

$$\mathcal{N}_k = \mathcal{N}_k \cup \{n\}; \tag{2.12c}$$

$$\mathcal{N} \leftarrow \mathcal{N} \setminus \{n\}; \tag{2.12d}$$

$$R_k = R_k + R_{kn}^{1/N}. \tag{2.12e}$$

4) *Exit condition:* If there exists some unassigned subcarrier, then go back to step 3, else exit the algorithm.

As can be seen by inspecting the steps of the algorithm, the rationale behind max-min fairness solution, in contrast to the WF result, is to assign more power to users exhibiting poor channel conditions (step 3) so that they can achieve a data rate comparable to that of other users with better channel quality. It is worthwhile to note that the max-min fair rate allocation is unique when the number of resources and flows, i.e., subcarriers and wireless terminals, are both finite [151]. Unfortunately, due to the nonlinear nature of the integer problem (2.9), the algorithm proposed in [150, 152] is computationally very expensive.

In [150, Eq. 2], the formulation (2.9a) is extended to:

$$\max_{\mathbf{p},\,\mathcal{N}} \min_{k\in\mathcal{K}} \sum_{n\in\mathcal{N}_k} t_{kn}\,R_{kn} \tag{2.13a}$$

$$\text{s.t.} \quad \sum_{n\in\mathcal{N}_k} p_{kn} \le \overline{p}_k \qquad \forall\, k \in \mathcal{K} \tag{2.13b}$$

$$\text{and} \quad \sum_{k=1}^{K} t_{kn} \le 1 \qquad \forall\, n \in \mathcal{N}, \tag{2.13c}$$

wherein the positive coefficient $t_{kn} \in [0, 1]$ introduces the percentage of time each subcarrier $n$ is used by a given user $k$. With $t_{kn}$, each subchannel can be shared by different users in a TDMA fashion. Clearly, the assumption behind this approach is that the users' channel responses do not change significantly over a timing interval. However, practical solutions, such as those reported in [150], assume $K \ll N$ and that no subchannel is shared among users, i.e., $t_{kn}$ is a binary value and $\sum_{k \in \mathcal{K}} t_{kn} = 1 \ \forall \ n \in \mathcal{N}$, or, equivalently, (2.9c). In addition, determining the best values for $t_{kn} \in (0, 1)$ and indicating a time-sharing allocation policy is not always feasible for $K > N$, as reported in [154].

In addition, to achieve a max-min fairness data rate vector, Kelly [155] suggested to formulate the problem as:

$$\max_{\mathbf{p},\, \mathcal{N}} \sum_{k=1}^{K} - \left( - \log_2 \left( \frac{R_k}{\beta} \right) \right)^{\rho} \tag{2.14}$$

wherein $\rho > 1$ is a constant parameter, and $\beta$ is a positive constant, measured in bits/s, satisfying $R_k < \beta \ll \infty \ \forall \ k \in \mathcal{K}$. Thus, the collection of utility functions (2.14) provides a priority to smaller data rates, which increases as $\rho$ increases, and becomes absolute as $\rho \to \infty$. Furthermore, instead of choosing the best user in step 3, (2.12a), an alternative criterion is defined in [155]:

$$k \longleftarrow \arg\max_{\ell \in \mathcal{K}} \left\{ \frac{1}{R_\ell} \cdot \left( \log_2 \left( \frac{R_\ell}{\beta} \right) \right)^{\rho - 1} \right\} \tag{2.15}$$

to find the best user $k$ for subcarrier $n$. Note that, for $\rho \to \infty$, the condition (2.15) becomes:

$$k \longleftarrow \arg\min_{\ell \in \mathcal{K}} R_\ell \tag{2.16}$$

which coincides with the original strategy of max-min fairness to allocate a subcarrier to the user with the minimum achieved data rate.

Although the max-min criterion gives priority to the weakest users, thus balancing the near-far effect, this solution cannot be used in the practice, because, in general, the number of allocated bits may not correspond to any practical modulation scheme [156]. Furthermore, the results show that under the max-min fair solution, some users may consume significantly more bandwidth than others [157], at the cost of a reduction in the overall throughput of the network.

### 2.3.3 The weighted proportional fairness criterion

Achieving traffic fairness and efficiency either in the energy or in the spectral domains are two conflicting goals. Hence, the optimization of the radio resource utilization tends to penalize terminals with low SINRs, irrespectively of their traffic level performance. The max-min fairness scheme described in Sect. 2.3.2 is however inappropriate when different users have different priorities. Generally, the problem is how to balance between fairness and utilization of the resources. This led Kelly *et al.* to formulate in [158] the notion of *weighted proportional fairness*. Under a proportional maximization rate constraint, the rate of each user should adhere to a set of predetermined proportionality constants which make a concrete way of assigning priorities to the users as follows:

$$R_1 : \cdots : R_k : \cdots : R_K = \varphi_1 : \cdots : \varphi_k : \cdots : \varphi_K \tag{2.17}$$

where $\{\varphi_k\}$'s are the proportion constants. In the practice, $\varphi_k$ can be interpreted as the amount user $k$ is willing to pay per unit time. At the end, user $k$ receives in return a data rate $R_k$ which is proportional to $\varphi_k$.

**Definition 18** *A vector data rate* $\mathbf{R} = [R_1, \ldots, R_K]$ *is* proportional fair *if it is feasible and, for any other feasible rate vector* $\mathbf{R}^{'} = \left[R_1^{'}, \ldots, R_K^{'}\right]$, *the aggregate of proportional changes is non-positive, i.e.:*

$$\sum_{k=1}^{K} \varphi_k \frac{R_k^{'} - R_k}{R_k} \leq 0. \tag{2.18}$$

■

This method is also useful for service level differentiation, which allows for flexible allocation mechanisms to different classes of users with separable constraints. The proportional-fair objective of (2.18) is continuously differentiable, monotonically increasing, and strictly concave, therewith admitting a convex optimization formulation [148]. In [158], Kelly *et al.* suggested an algorithm that converges to the proportionally fair rate vector, using the maximization of the sum of the (logarithmic) long-run average data rates provided to the users, based on the Kuhn-Tucker conditions for the problem (2.4a). Otherwise stated, a proportional-fairness rate allocation can be achieved by formulating the problem as [158]

$$\max_{\mathbf{p},\,\mathcal{N}} \sum_{k=1}^{K} \varphi_k \log_2 (R_k) \tag{2.19}$$

over all feasible rate allocations. Thus, since the logarithm function is strictly concave, proportional-fair rates are unique [159, Sect. 6.7]. Note that the logarithmic utility function indicates that users with low average rates benefit more in terms of utility from being scheduled than users with high average rates. The iterative algorithm to compute proportionally max-min fair rate vectors is similar to that for max-min fairness, except for the choice of the best user for each unassigned subcarrier $n$ in (2.12a) and (2.12b) (step 3). In this case, it follows the following criterion instead:

$$k \longleftarrow \arg\max_{\ell\in\mathcal{K}} \varphi_\ell \frac{R_{\ell n}^{1/N}}{R_\ell} \tag{2.20}$$

where $R_{\ell n}^{1/N}$ is computed according to (2.11). The rationale behind this approach is the following. Using (2.20), users compete for resources not directly based on their channel conditions, as happens in Sect. 2.3.2, but according to the combination of priorities $\varphi_\ell$ and rates *normalized* by their respective average throughputs, $R_{\ell n}^{1/N}/R_\ell$. In other words, each subcarrier is assigned to a user when its channel, weighted by its priority, is near its own peak in the frequency domain, thus trading off multiuser diversity and fairness.

The update of the data rate $R_k$ can be done in different ways. A low-complexity update equation that also bears low memory requirements is defined in [159, Sect. 6.7], by keeping track of the average throughput $R_k$ of each user in an exponentially-weighted time window of length $T_c$ as follows:

$$\begin{cases} R_k = \left(1 - \frac{1}{T_c}\right) R_k + \frac{1}{T_c} R_{kn}^{1/N} & , \\ R_k = \left(1 - \frac{1}{T_c}\right) R_k & k \neq m, \end{cases} \tag{2.21}$$

where $k$ is the index of the preferred user for the next updating round and $m$ is the selected user for the current round, both selected following (2.16). The update (2.21) is an exponentially weighted filter that, instead of using (2.12e), includes all historical rates in the average rate. Note that using a very large time-scale $T_c$, (2.21) is equivalent to maximization problem (2.14) [159, Sect. 6.7].

In the literature of OFDMA resource allocation, some other instantaneous sum-rate maximization methods with proportional rate constraints have been studied (e.g., [160–162]). In terms of problem formulation, the main emphasis of these works is on the maximization of the data rates with instantaneous proportional rate constraints, exclusive subcarrier assignment, and constrained total transmit power. The solution is achieved by resorting to integer programming methods, with time complexity (i.e., number of time steps in the iterative algorithm) on the order of $\mathcal{O}\left(NK \log_2 N\right)$ or higher.

The notion of weighted proportional fairness has been extended by Mo and Walrand in [163], observing some particular transmission control protocol (TCP)-based network traffics, in which the total throughput of weighted proportional fairness is not optimal in terms of spectral efficiency. To overcome this drawback, the problem is then formulated using the following definition.

**Definition 19** *Let $\alpha$ be a non negative constant, and $\boldsymbol{\varphi} = [\varphi_1, \ldots, \varphi_K]$ be a positive weight vector. A vector data rate $\mathbf{R} = [R_1, \ldots, R_K]$ is $(\boldsymbol{\varphi}, \alpha)$* proportional fair *if it is feasible and, for any other feasible rate vector $\mathbf{R}^{'} = \left[R_1^{'}, \ldots, R_K^{'}\right]$,*

$$\sum_{k=1}^{K} \varphi_k \frac{R_k^{'} - R_k}{{R_k}^{\alpha}} \leq 0. \tag{2.22}$$

■

Obviously, if $\alpha = 1$, (2.22) reduces to the weighted proportional fairness introduced in (2.18). If $\alpha \to \infty$, $\mathbf{R}$ approaches the max-min rate vector [163, Lemma 3]. In other words, this generalization includes arbitrarily close approximation of max-min fairness. Unfortunately, the challenge of choosing the best value of $\alpha$ makes this framework (almost) impractical. Further examinations clarify that $(\boldsymbol{\varphi}, \alpha > 1)$ proportional fairness maximizes [163, Lemma 2]:

$$\sum_{k=1}^{K} \varphi_k \left(1-\alpha\right)^{-1} {R_k}^{1-\alpha} \tag{2.23}$$

over all feasible data rate vectors.

Mathematically, (2.22) is a twice continuously differentiable and strictly concave function. Algorithms for computing $(\boldsymbol{\varphi}, \alpha)$ proportionally fair rates have been developed in [163], where each transmitter adapts its window size based on the total delay between the transmission of a packet and the reception of its acknowledgment. The main drawback of the proportionally fair rate allocation is that utility (maximization) functions are commonly assumed as concave. Lee *et al.* [164] showed that, if the abovementioned algorithms developed for concave utility functions are applied to non concave utility functions, the system can be unstable and cause excessive congestion in the network. Since the rate adaptive functions of some real-time applications are not concave (e.g., a multimedia communication) [165], they cannot be dealt with in this kind of systems.

### 2.3.4 Utility maximization

Max-min fairness (Sect. 2.3.2) and weighted proportional fairness (Sect. 2.3.3) consider the same QoS requirements among network users with a strictly concave rate adaptive function. As mentioned above, in some systems, e.g., multimedia applications, the rate maximization functions are not concave. Furthermore, in such contexts we are not able to formulate real-time constraints, e.g., in terms of delay. In their seminal work [166], Cao and Zegura overcome these disadvantages by introducing the concept of *utility maximization* in terms of application-layer performance, whose aim is to provide individual QoS requirements for each user with a (not necessarily concave) function for rate maximization. More in general, a utility function is a function that can be used to mathematically describe the QoS characteristics of an application, thus allowing the system designers to put the emphasis on specific QoS parameters of the network. Unlike rate-adaptive formulations, in which the objective, as described in the previous subsections, is the sum-rate maximization with constraints in terms of power expenditure, the utility maximization approach can guarantee the application-specific demand which can be characterized by bandwidth, delay, delay jitter, or time spent to complete data deliveries, just to mention a few examples.

In other words, this framework allows for more general resource allocation problems, that can be formulated in different ways according to the goal of the system. For instance, a power control scheme for optimal uplink SNR assignment can be expressed in a centralized way as follows:

$$\max_{\mathbf{p},\,\mathcal{N}} \sum_{k=1}^{K} \sum_{n\in\mathcal{N}_k} u_k(\gamma_{kn}) \tag{2.24a}$$

$$\text{s.t.} \quad \sum_{n \in \mathcal{N}_k} p_{kn} \leq \overline{p}_k \qquad \forall\, k \in \mathcal{K} \tag{2.24b}$$

$$\text{and} \quad \mathcal{N}_k \cap \mathcal{N}_m = \emptyset \qquad \forall\, k,\, m \in \mathcal{K}, k \neq m. \tag{2.24c}$$

where

$$\gamma_{kn} = \frac{|H_{kn}|^2 \, p_{kn}}{\sigma_w^2} \tag{2.25}$$

denotes the SNR of the user $k$ on the $n$th carrier (used by user $k$ only in an exclusive fashion) as measured at the BS, and $u_k(\cdot)$ is user $k$'s individual maximization function that is a function of each user's relevant parameters. The maximization function can also be represented as a greedy function for each user as follows:

$$\max_{\mathbf{p}_k,\, \mathcal{N}_k} \sum_{n \in \mathcal{N}_k} u_k(\gamma_{kn}) \qquad \forall\, k \in \mathcal{K} \tag{2.26}$$

that makes the power control a distributed problem, in which each user $k$ seeks the optimal vectors $\mathbf{p}_k$ and $\mathcal{N}_k$ that maximize its own sum-utility (2.26). However, note that selecting the set $\mathcal{N}_k$ by each user while meeting the exclusive assignment of the subcarriers, i.e., (2.24c), implies a certain amount of feedback information among the users, which, although less demanding in terms of feedback rate than the problem (2.24), makes this scheme not completely distributed.

In the literature, many utility-based resource allocation formulations appeared in the last few years. As already mentioned, [166] proposed the use of a utility function to maximize the performance of the application layer. The drawback in [166] is a high delay in the communication network among users. Cho and Chang *et al.* in [167] extend this formulation to address the limitation in terms of delay, by proposing a control-theoretic utility max-min flow control algorithm, and showing that the algorithm converges to a utility max-min fair rate vector by using Dewey and Jury's stability criterion [168]. Among the others, Huang *et al.* in [169, 170] introduced scheduling and radio resource allocation algorithms in OFDMA-based wireless networks in the downlink and the uplink direction, respectively, using a dual formulation, and showing a complexity $\mathcal{O}\left(KN + N\log_2 N\right)$. In particular, [169] looks at joint scheduling and resource allocation for the downlink by considering several practical constraints largely ignored in the previous literature (e.g., self-noise). Reference [170] aims at maximizing the achieved data rates taking into account the queue length of each user, using an algorithm that can also be applied to downlink transmissions. Zhou *et al.* [171] solve a scheduling and resource allocation problem in an OFDMA system using an approach based on utility functions, that eventually results in a discrete optimization problem with a non-differentiable non-convex objective with minimum data rate constraints. The idea is to transform the discrete problem into a suitable weighted max-min fairness problem which is easier to be implemented. In [172], Kim and Lee present a general utility-based framework for joint uplink/downlink optimization, where the user's satisfaction is modeled by two different utility functions, one for the uplink, and another one for the downlink. The resource allocation is formulated as a maximization problem with an objective based on the sessions' utility functions and allocation probabilities as scheduling constraints that are solved via dual optimization techniques. To investigate radio resource allocation in OFDMA with heterogenous traffic classes, reference [173] defines a utility function as a sinusoidal function that depends on minimum and maximum data rates.

References [174–176] address the problem of energy efficiency maximization subject to power constraints according to the circuit power consumed. Xiong *et al.* in [175] devise a joint uplink/downlink water-filled energy efficient resource allocation under users priority constraints. The WF based iterative algorithm proposed by [175] converges faster than that of [174], while the spectral efficiency of the algorithm proposed by [174] is higher than that of [175].

To summarize, even though the utility maximization approach has made advances in dealing with heterogeneous resource allocation issues, it also exhibits serious limitations. As already mentioned, there exists a tradeoff between average throughput and fairness in the system. Sometimes there also exists a conflict between the QoS balance and the utility maximization. If users select utility functions based on their actual QoS requirements, then the optimal achieved data rate may result in a totally unfair resource allocation within the network. Applying advanced optimization methods of geometric programming [148], majorization theory [177], and fractional programming [178] may achieve an admissible tradeoff between fairness and overall throughput [176, 179, 180]. However, depending on the problem formulation, it is impossible to achieve the desired network performance if the resource allocation scheme operates on the link layer only. To further generalize the problem formulation, and thus to increase its potential, it is worth resorting to the cross-layer approach described in the next subsection.

### 2.3.5 Cross-layer optimization

So far, we talked about the design of an OFDMA system based on classical link-level approach. The wireless link level primarily addresses two challenges that arise from the physical medium: channel fading and multiple access interference (MAI). Advances in link design for wireless channels have led to different modulation and channel coding schemes that provide increased robustness to MAI and multipath and, thereby, enhance the radio band capacity. While OFDMA provides a powerful physical layer engine for broadband communications, applying it without thorough application level considerations may lead to poor results. In high-speed data networks, in which the traffic is in fact highly diverse (i.e., with distinct QoS parameters), and channel conditions that may vary dramatically over a short time scale, the traditional (decoupled) layer design cannot meet such requirements. For instance, if the medium access control (MAC) layer does not interact with the upper layers, it cannot obtain information regarding the type of service and the associated QoS parameters. As a consequence, the MAC has no ability to adjust itself to the variable characteristics of the traffic.

An OFDMA radio allocation module can be designed to be both channel-aware and application-aware through *cross-layer* interactions [181] that break the traditional layered paradigm of communication by relying on the concept of joint optimization across multiple layers. The cross-layer approach allows different layers to be grouped and/or assumes the existence of protocols that work with more than one layer, thus optimizing the protocol stack. With cross-layer techniques, decision making can be more accurate, bringing forth several benefits to the performance of the network. Bohge *et al.* in [182] provides basic definition and knowledge of cross-layer optimization in the context of OFDM and in the downlink direction of OFDMA systems.

In the context of cross-layer design, many joint scheduling-routing-flow control algorithms have been proposed, including multiuser techniques such as: maximization of the rate delivered on the radio channel [183]; a fair allocation of resources among users belonging to the same traffic class [184]; shaping the dynamics of traffic sources by limiting the delay of data packets in the queues [185]; and maximization of the QoS at the application layer [186]. Sometimes the difference between cross-layer and utility maximization schemes blur away, since cross-layer schemes may require to improve their performance by applying a non-concave utility function, that may consist of parameters from different layers. The common idea behind cross-layer schemes is to properly maintain packet queues to dynamically adapt packet transmission as well as rate allocation. Some pioneering works in the field of resource allocation in OFDMA using cross-layer design have appeared in [187–189].

Jiang *et al.* in [190, Ch. 6] present a more general framework of cross-layer OFDMA resource allocation, with [169, 170] as special cases (Sect. 2.3.4). References [141, 191, 192] propose some feasible solutions to maximize the throughput for the downlink of an OFDMA system under QoS constraints, also reducing the computational complexity. This is achieved using the method of Lagrangian relaxation [148], that is effective to provide users with very low SINRs with good performance. However, in the case the channel conditions and the QoS requirements vary significantly between successive frames, a new set of Lagrangian multipliers must be found in each frame, that may reveal to be impractical.

## 2.4 OFDMA resource allocation based on game theory

In the utility maximization (Sect. 2.3.4) and cross-layer (Sect. 2.3.5) schemes, different utility functions apply for different users. Sometimes the interests of wireless terminals are not aligned, so that they compete for the scarce wireless resources, namely bandwidth and power. Each user's interest could also be in conflict with others'. In this situation, the wireless terminals can decide to behave in either an altruistic or a selfish manner. In both cases, the related problems can be formulated applying *game theory* [10], which considers the users as *players* in a *game*. In particular, in an OFDMA network, there are multiple interacting users which occupy a fraction of the whole bandwidth, using a fraction of their available transmit power on each subcarrier based not only on their decisions, but also on the interests of any other mobile terminal in the network. This kind of interactions is just the main field of application of game theory, which thus represents an effective analytical tool not only to extend the optimization methods described in the subsections above (see [133] and references therein), but also to address the problem of *Pareto optimality* [10]. In resource allocation problems, one of the major challenges is in fact to achieve a Pareto-optimal rate vector, i.e., a rate allocation such that each user is provided with a certain performance, and any allocation other than that will degrade the performance of at least one user in the network. Interestingly, Pareto-optimal solutions can be investigated by means of game-theoretic formulations.

In the context of game theory, depending on the interaction rules, there exist various types of games. For instance, if the users are allowed to exchange their proper interests and information before the game starts in order to form coalitions and coordinate their actions, the game is said to be *cooperative* (and thus studied by *coalitional game theory*). If coordination among users is not present, the game is said to be *non-cooperative*, and modeled according to *non-cooperative game theory*. In both frameworks, the players act according to their *strategies*. The strategy of a player can be a single move or a set of moves during the game. For games in wireless communications, each transmitter represents a player whose strategy space covers the choices of modulation level, coding rate, transmit power, transmission frequency, just to mention a few examples. Another factor that identify different types of games is the number of times the users interact. If users play the game over multiple rounds, the game is said to be a *repeated game*. Contexts where the users only interact once are referred to as *static games* [10, 22].

### 2.4.1 Non-cooperative solutions

Non-cooperative game theory has been vastly applied to wireless communication problems, and much progress has been made on distributed power control in Gaussian interference channels. In [193], Wu *et al.* investigates a joint power and (exclusive) subcarrier assignment scheme in single-cell uplink OFDMA systems based on non-cooperative game theory, using the sum-capacity as the utility function to be maximized. This game bears a unique Nash equilibrium (NE), which is a stable outcome of the game (i.e., a stable resource apportionment across users) in which no player has incentive to *unilaterally* (i.e., non-cooperatively) deviate from [10]. In [140], Yu *et al.* apply a different convex utility function to the same scenario, aiming at maximizing the power efficiency of the network. In the utility function, a (transmit) power pricing factor [194] is introduced to overcome the near-far effect, reaching a (nearly) Pareto-optimal NE point. The fairness of both approaches [140, 193] is experimentally showed among a small number of users.

Kwon *et al.* in [195] aim at maximizing the weighted sum-rate of the users in the uplink of a multicell OFDMA scenario. This objective, together with power and rate constraints, defines the non-cooperative game. The simulation results show that the performance of the proposed algorithm strictly depends on the power pricing coefficient, which represents the cost imposed on each BS for the co-channel interference generated by it as well as its power consumption.

Han *et al.* in [16] analyze the previous scenario to maximize the data rates under a constraint in terms of maximum transmit power, showing that the pure non-cooperative game may have some undesirable NE points with low system and individual performance. The authors suggest to introduce a centralized "virtual referee" whose role is to prevent users with high co-channel interference from sharing one subcarrier, or to reduce the demanded transmission rates that prove to be unfeasible. Even though the results significantly outperforms the WF solution in terms of reduced transmit power and increased data rate, the proposed algorithm suffers from high computational complexity. Tan *et al.* in [196] experimentally show that their non-cooperative game-based algorithm achieves a good performance in terms of total data rate, computational complexity, and fairness among users.

The problem of energy-efficient resource allocation for a multicell OFDMA system is studied in [197, 198]. In [197], the authors devise a non-cooperative potential game [199] aimed at maximizing the users' energy efficiency, which proves to bring performance improvements in terms of goodput (error-free delivery) for each unit of energy. In [198], the same purpose is accomplished by a centralized subcarrier allocation procedure and a distributed non-cooperative power control game. The simulation results in a realistic multicell network scenario show that the proposed algorithm achieves an acceptable performance and computational complexity burden.

Non-cooperative game theory is also flexible enough to investigate resource allocation problems for contexts different from the data detection phase, popularly considered in the literature. For instance, in [200] Bacci *et al.* formulate a non-cooperative game to regulate the transmit powers in an OFDMA uplink during the initial, contention-based network association.

### 2.4.2 Cooperative solutions

Recently, several other methods which use various heuristics based on cooperative (coalitional) game theory [10, 22] have been proposed to address the problem of fair resource allocation for OFDMA systems, using either centralized or distributed algorithms. The Nash bargaining solution (NBS) [10] is the most refined technique applied to wireless resource allocation problems in an OFDMA network. The NBS proves the existence and uniqueness of an NE point of the following convex utility function:

$$\max_{\mathbf{p},\,\mathcal{N}} \prod_{k=1}^{K} (R_k - \underline{R}_k) \tag{2.27a}$$

$$\text{s.t.} \quad R_k = \sum_{n \in \mathcal{N}_k} R_{kn} \geq \underline{R}_k \qquad \forall\, k \in \mathcal{K} \tag{2.27b}$$

$$\text{and} \quad \sum_{n \in \mathcal{N}_k} p_{kn} \leq \overline{p}_k \qquad \forall\, k \in \mathcal{K} \tag{2.27c}$$

$$\text{and} \quad \mathcal{N}_k \cap \mathcal{N}_m = \emptyset \qquad \forall\, k,\, m \in \mathcal{K}, k \neq m, \tag{2.27d}$$

In other words, the goal is to maximize the product of the excesses of the transmitters' rates over their own minimum demands $\underline{R}_k$. The NBS guarantees each user to achieve its own demand, thus providing an individual rationality to the resource allocation. The important result of applying NBS is that the final rate allocation vector is Pareto optimal. Taking into consideration the strictly concave increasing property of the logarithm function, we can transform (2.27a) into:

$$\max_{\mathbf{p},\,\mathcal{N}} \sum_{k=1}^{K} \log_2 (R_k - \underline{R}_k) \tag{2.28}$$

Clearly, when $\underline{R}_k = 0$, the NBS fairness scheme reduces to the weighted proportional one, with $\varphi_k = 1$ [201].

Han *et al.* in [139] introduce a distributed algorithm for an OFDMA uplink based on the NBS and the Hungarian method [202] to maximize the overall system rate under individual power and rate constraints. The underlying idea is that once the minimum demands are provided for all users, the rest of the resources are allocated proportionally to different users according to their own conditions. The proposed algorithm shows a complexity $\mathcal{O}\left(K^2 N \log_2 N + K^4\right)$, without considering the (expensive) computational load to solve the (convex) equations of the NBS. In [203], Lee *et al.* solve two subproblems of exclusive subcarrier assignment and power control in an OFDMA network aiming at maximizing the NBS fairness. The simulation results show an overall end-to-end rate between the nodes comparable to that achieved in [139].

One main drawback of applying NBS in resource allocation problems is that this scheme guarantees minimum requirements of the users, but it does not impose any upper bound constraint. In fact, the achieved data rate may be much higher than the initial demands and this is unsatisfactory from the wireless network provider viewpoint. One of the most prominent alternatives to the NBS is the Raiffa-Kalai-Smorodinsky bargaining solution (RBS), defined by Raiffa [13] and characterized by Kalai and Smorodinsky [204]. The RBS requires that a user's payoff data rate should be proportional not only to its minimal rate, but also to its maximal one. Whereas the NBS takes into account the individual gains, RBS emphasizes the importance of one's gain and others' losses. For an OFDMA resource allocation problem, the RBS bargaining outcome is the solution to:

$$\max_{\mathbf{p},\,\mathcal{N}} \prod_{k=1}^{K} \left( R_k - \underline{R}_k + \frac{1}{K-1} \sum_{m\in\mathcal{K},m\neq k} \left(\overline{R}_m - R_m\right) \right) \tag{2.29a}$$

$$\text{s.t.} \quad \underline{R}_k \leq R_k \leq \overline{R}_k \qquad \forall\, k \in \mathcal{K} \tag{2.29b}$$

$$\text{and} \quad \sum_{n\in\mathcal{N}_k} p_{kn} \leq \overline{p}_k \qquad \forall\, k \in \mathcal{K} \tag{2.29c}$$

$$\text{and} \quad \mathcal{N}_k \cap \mathcal{N}_m = \emptyset \qquad \forall\, k,\, m \in \mathcal{K}, k \neq m, \tag{2.29d}$$

wherein $\overline{R}_k$ denotes the upper bound of the transmission rate of the each user. When applying RBS, if the channel quality of a terminal improves, it will get a better capacity without any reduction to that of the other users (individual monotonicity). The existence and uniqueness of RBS can be shown, but a Pareto optimal NE point is not always attained for more than two players, as Roth stated in [205]. By using again the properties of the logarithm function, the utility maximization (2.29a) can be equivalently investigated using the following objective function:

$$\max_{\mathbf{p},\,\mathcal{N}} \sum_{k=1}^{K} \log_2 \left( \frac{R_k - \underline{R}_k}{\overline{R}_k - \underline{R}_k} \right) \tag{2.30}$$

Using this formulation, the RBS is a point at which each individual's gain is proportional to its maximum gain. When $\underline{R}_k = 0\ \forall\, k \in \mathcal{K}$ and $R_1 : \cdots : R_K = \overline{R}_1 : \cdots : \overline{R}_K$, the RBS achieves the same results of the max-min fairness criterion. In RBS formulation, the achieved data rate vector satisfies:

$$\frac{R_1 - \underline{R}_1}{\overline{R}_1 - \underline{R}_1} = \cdots = \frac{R_k - \underline{R}_k}{\overline{R}_k - \underline{R}_k} = \cdots \tag{2.31}$$

In [18], Chee *et al.* propose a centralized algorithm for the OFDMA downlink scenario based on RBS. The results show a good performance only when the gap between the maximum and the minimum rate is (very) large. Even though the subcarriers are assigned in an exclusive manner, the computational complexity of this algorithm is $\mathcal{O}\left(KN + K^2\right)$. Reference [206] investigates the problem of time-space resource allocation in a MIMO-OFDMA network in the downlink direction with aim at maximization data rate of each terminal, without specifying the complexity of the iterative algorithm to solve the NBS convex equation. In Chapter 3, we will attempt to improve the fairness of the solution and to reduce the complexity in the uplink direction of OFDMA-based networks, by deriving a coalition-based algorithm to provide each terminal with *exactly* the desired rate, so as to satisfy both wireless terminals and the network service provider.

Auction methods are another cooperative game scheme which has recently drawn attention in the resource allocation research literature. In [19], Noh proposes a distributed and iterative auction-based algorithm in the OFDMA uplink scenario with incomplete information. The time complexity of the algorithm is experimentally equal to $\mathcal{O}\left(KN \log_2 K\right)$. However, the simulation parameters are not realistic (three users and three subcarriers), and it is thus hard to estimate the computational complexity when using real-world network parameters. Alavi *et al.* in [207] propose an auction-based algorithm to achieve near a proportionally fairness data rate vector, although the computational complexity is not specified. Reference [208] propose a joint downlink/uplink subcarrier allocation (with fixed power) based on stable matching game to maximize data rate of each terminal in downlink and uplink directions, simultaneously.

## 2.5 A toy example with two terminals

In this section, we apply the different problem formulations introduced above to a simplified scenario, namely a network populated by just two terminals ($K = 2$). For the reader's convenience, we report the optimization formulas for this specific case:

$$\text{Max rate:} \quad \arg\max_{\mathcal{U}} \left(R_1 + R_2\right) \tag{2.32a}$$

$$\text{Max-min rate:} \quad \arg\max_{\mathcal{U}} \min_{k=1,2} R_k \tag{2.32b}$$

$$\text{Proportional fairness:} \quad \arg\max_{\mathcal{U}} \left(R_1 + R_2\right) \quad \text{s.t. } R_1 : R_2 = \varphi_1 : \varphi_2 \tag{2.32c}$$

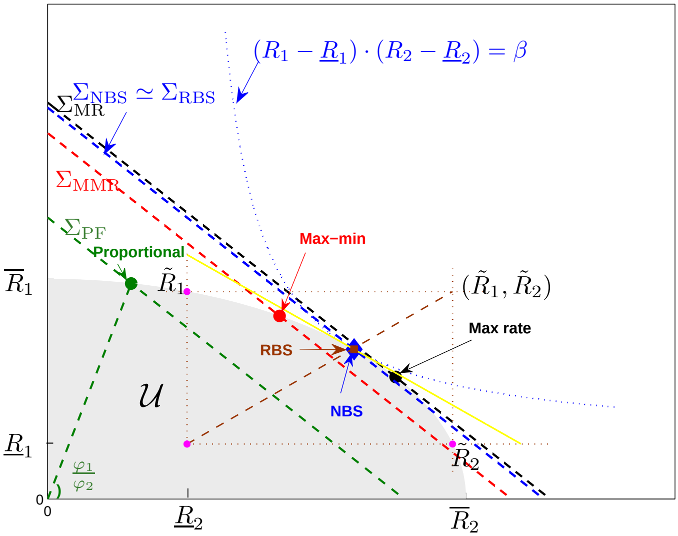


**Fig. 2.1:** *Two-user rate optimization.*

$$\text{NBS fairness:} \quad \arg\max_{\mathcal{U}} \,(R_1 - \underline{R}_1) \cdot (R_2 - \underline{R}_2) \tag{2.32d}$$

RBS fairness:

$$\arg\max_{\mathcal{U}} \left(R_1 - \underline{R}_1 + \left(\overline{R}_2 - R_2\right)\right) \cdot \left(R_2 - \underline{R}_2 + \left(\overline{R}_1 - R_1\right)\right) \tag{2.32e}$$

The (network) overall rate (the sum-rate) can be maximized using the max rate optimization formula (2.32a). However, although globally efficient, this solution is unfair: users with bad channel quality (e.g. cell-edge users) may be completely excluded from communication. On the contrary, max-min rate optimization protects terminals with a low data rate, and to this aim the lowest data rate is maximized according to (2.32b). In this case, a user with bad channel quality limits the system performance, and, if the channel quality of one user improves, all users will achieve a higher data rate. To maximize the sum-rate while balancing the ratios among the rates, we can use the formulation (2.32c).

To better visualize the problem, Fig. 2.1 reports the case in which user 2 experiences a better propagation channel than user 1. This is confirmed by the gray area $\mathcal{U}$

representing the feasible ranges for $R_1$ and $R_2$, that can be computed by assuming that both terminal receivers treat co-channel interference as noise [209]. As can be seen, $\mathcal{U}$ is a convex area, with $\overline{R}_2 > \overline{R}_1$, due to user 2's better channel conditions. By numerically solving (2.32a), we can find the max rate solution, represented by the black circle in Fig. 2.1. Geometrically, it can be obtained by identifying the point at which the Pareto boundary, given by the contour of $\mathcal{U}$, osculates a straight line with slope $-1$, depicted by the black dashed line labeled[2] with $\Sigma_{\text{MR}}$. Note that such line, given by all pairs $(R_1, R_2)$ such that $R_1 + R_2 = \Sigma_{\text{MR}}$, is the only constant-sum-rate line that is tangent to the Pareto boundary: all other lines such that $R_1 + R_2 = \xi < \Sigma_{\text{MR}}$ intersect the Pareto boundary in two points, whereas all lines such that $R_1 + R_2 = \xi > \Sigma_{\text{MR}}$ do not intersect it. Using numerical methods and setting $\varphi_1/\varphi_2 = 3.2$ in this example, we can also find the solutions to (2.32b) and (2.32c), represented by the red and the green circles, respectively. It is worth noting that the such points intersect the red and green dashed lines, corresponding to $R_1 + R_2 = \Sigma_{\text{MMR}}$ and $R_1 + R_2 = \Sigma_{\text{PF}}$, respectively, confirming that the sum rate achieved by such formulations is of course lower than that given by (2.32a), as $\Sigma_{\text{MMR}} < \Sigma_{\text{MR}}$ and $\Sigma_{\text{PF}} < \Sigma_{\text{MR}}$. This result is valid in general, and can be met with equality only under special settings of the network.

In addition to the proportional fair method, we can use two cooperative game-based solutions, namely NBS and RBS, to introduce fairness into our resource allocation problem. The NBS, formulated by the convex formula (2.32d), can be seen as a general case of the weighted proportional fairness, in which all users are guaranteed to receive some resources $\underline{R}_k$. The RBS solution, represented by (2.32e), is a generalization of the max-min solution, in which the achieved rates are bounded between a minimum, $\underline{R}_k$, and a maximum, $\overline{R}_k$, demanded rates. Similarly to the max-min allocation, if the channel quality of a user improves, in the RBS solution he/she will get a higher data rate wihtout any reduction for the other users' rates. In the cooperative-game formulations, we consider a minimum demanded $(\underline{R}_1, \underline{R}_2)$ and a maximum constraint $\left(\overline{R}_1, \overline{R}_2\right)$ as disagreement points. NBS and RBS solutions must then satisfy $\underline{R}_k \leq R_k$ and $\underline{R}_k \leq R_k \leq \overline{R}_k$, respectively. If the demanded rates, ignored in max rate, max-min rate, and proportional fairness solutions, are not met, in NBS and

[2]For the sake of graphical presentation, all line labels $\Sigma_{(\cdot)}$ correspond to the lines whose points are such that $R_1 + R_2 = \Sigma_{(\cdot)}$.

RBS solutions a user would leave the negotiation (hence the name "disagreement point"). As a consequence, the feasible regions for max rate, max-min rate, and proportional fairness solutions is $\mathcal{U}$, whereas it is $\mathcal{U} \bigcap \{[\underline{R}_1, \underline{R}_2] \leq \mathbf{R}\}$ for NBS and $\mathcal{U} \bigcap \{[\underline{R}_1, \underline{R}_2] \leq \mathbf{R} \leq [\overline{R}_1, \overline{R}_2]\}$ for RBS, respectively, with $\mathbf{R} = [R_1, R_2]$.

The NBS and RBS solutions can be found numerically in this example, and are depicted by the blue diamond and the brown square in Fig. 2.1, respectively. As can be seen, the difference between the NBS and RBS pairs is negligible. Although this does not always hold in general, here is due to having only $K = 2$ sources, and placing the same maximum achievable rate $\overline{R}$ for both solutions. Similarly to the max-rate case, we can obtain such points graphically [11, Ch. 35]. The NBS point can be identified as the point of tangency between the Pareto boundary of $\mathcal{U}$ and the hyperbola $(R_1 - \underline{R}_1) \cdot (R_2 - \underline{R}_2) = \beta$, where $\beta > 0$ is chosen properly to ensure only one intersection between the two curves. Note that, if we draw the tangent line to $\mathcal{U}$ at the NBS point, the length of the segment between the NBS solution and the vertical line drawn through $\underline{R}_2$ is equal to the length of the segment between the NBS solution and the horizontal line drawn through $\underline{R}_1$ (see Fig. 2.1, yellow segments).

To obtain the RBS point graphically, we need to identify the "utopian point" $(\tilde{R}_1, \tilde{R}_2)$, where $\tilde{R}_k$ is the maximum achievable rate by user $k$ when the other user demands its minimum one $\underline{R}_m$ (see Fig. 2.1). This point is named utopian, as both terminals cannot achieve such rates simultaneously, as confirmed by the feasible region $\mathcal{U}$. The RBS solution is thus the intersection between the Pareto boundary of $\mathcal{U}$ and the segment connecting the utopian point $(\tilde{R}_1, \tilde{R}_2)$ and the disagreement point $(\underline{R}_1, \underline{R}_2)$. To measure the global efficiency of the cooperative solutions, we can draw the constant-sum-rate lines $\Sigma_{\text{NBS}}$ and $\Sigma_{\text{RBS}}$, depicted by the blue and brown dashed lines, respectively. As expected, the sum-rate achieved by both solutions is nearly the same, and lower than that provided by the max-rate solution, although in this particular example the gap is significantly reduced with respect to the max-min and proportional-fair solutions. Note this is true in general, as NBS and RBS outperform max-min and proportional-fair solutions in terms of achieved sum-rate [11, Ch. 35].

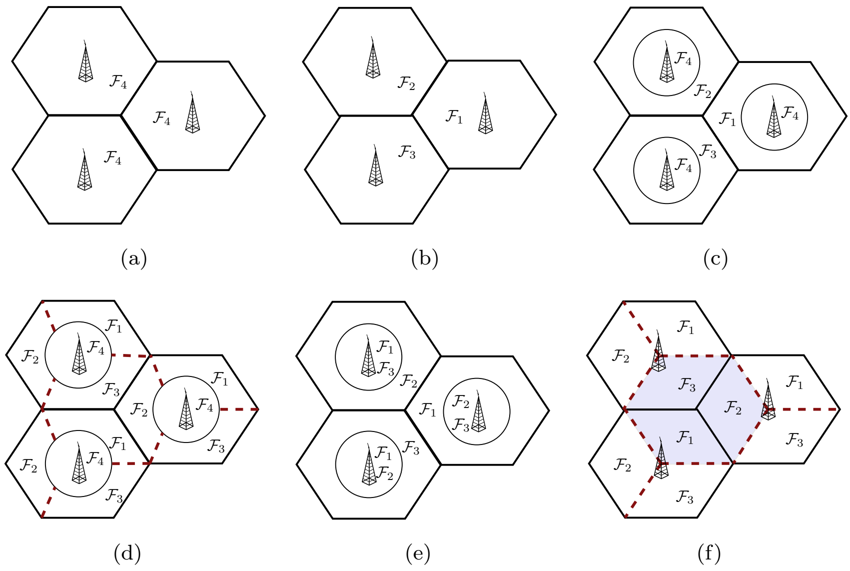


**Fig. 2.2:** *Different frequency reuse schemes.*

# 2.6 Toward 4th generation of wireless networks

## 2.6.1 Multicellular networks

Until now, we focused on a single-cell OFDMA network scenario wherein the BS and the mobile terminals use one set of frequency bands. In a multi-cell network scenario, the same frequency band can be reused by cells which are physically separated far enough to endure mutual interference. A BS that uses the same frequency band begets *co-channel interference*. The allocation of frequency bands to BSs is called *frequency reuse* which has a significant impact on system performance. Existing frequency reuse schemes can be divided into three classes: static frequency channel allocation, dynamic channel allocation, and combined channel allocation. In static frequency reuse, the partitions of frequency band do not adapt to traffic dynamics and interference conditions experienced by users. A subcarrier under a deep fade for one user at a given time may not be in a deep fade for other users. Therefore, every subcarrier may have a good channel response for some users in a multiuser environment. Unevenly loaded traffic results in unbalanced performance over the cells, which leads to degraded overall system performance. Instead of predicting and averaging, dynamic subcarrier allocation takes advantage of multiuser channel and traffic diversity to adjust the channel allocation over time. Although dynamic subcarrier requires higher computation complexity and signaling overhead during operation, its ability to utilize real-time system information leads to higher spectrum efficiency. Basically, dynamic radio resource allocation in a multi-cell network can be performed in centralized or distributed manner. In centralized schemes, the terminals and BSs are responsible for gathering traffic/channel information and feed back the information to the controller and then participate in protocol implementation. Combine channel allocation can be reputed as the combination of static and dynamic channel allocation, where some of the channels are fixed for each BS and others are dynamically assigned to cells.

Static frequency reuse approaches are based on fractional frequency reuse (FFR) where frequency bands are divided into a number of segments. Fractional frequency reuse is divided into two schemes: *strict fractional frequency reuse (FFR)*, and *soft frequency reuse (SFR)*. Typically, in an FFR approach, each segment is reserved for a certain reuse factor and is associated with a particular transmission power profile. When (strict) "frequency reuse 1" (called also universal frequency reuse) is supported, all BSs operate on the same frequency channel (Fig. 2.2(a)). In this case, to maximize frequency efficiency, decreasing the inter-cell interference is a major concern. Since a link will experience co-channel interference from signals from neighboring cells, the SINR for a link between the $b$th BS and the $k$th terminal via the $n$th subcarrier is defined as:

$$\gamma_{kn}^{b} = \frac{h_{kn}^{b} p_{kn}}{\sum_{b \neq c \in \mathcal{B}} h_{kn}^{c} p_{kn} + \sigma_w^2} \tag{2.33}$$

wherein $\mathcal{B}$ denotes the set of cells (BSs). In domain of frequency reuse 1, the result of contribution [210] written by Gjendemsjø *et al.* show that in a two-cells setup scenario, a binary power allocation policy which assigns either full power to both cells or shuts down one cell significantly increases the whole data rate in downlink direction under power constraints per-BS. Venturino *et al.* in [211] use a distributed approach based on optimization problem with aim at maximization downlink rate region subject to power constraints per BS. Yu *et al.* in [212] applied the idea of "interference pricing" which measures the impact of each terminal's interference on its neighbors cells and then BSs dynamically allocate radio resources based on the exchange of these measures.

The scheme called (strict) “frequency reuse 3” divides the frequency band into 3 sub-band and allocates one sub-band to a given cell, so that adjacent cells use different frequency band as illustrated in Fig. 2.2(b). A “hybrid frequency allocation” is proposed which is, in general, a mix of reuse 1 and 3 approaches. A hybrid approach can be applied to avoid interference at cell edges. For example, suppose we have three cells covering a certain area, and there are four frequency segments. Then, frequency segment $\mathcal{F}_4$ can be reserved for cell-interior users (with less interference from other sectors), and frequency segments $\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3$, for cell-edge users (more interference from other sectors) in cells 1, 2, 3, respectively (Fig. 2.2(c)). As a result, we have 1/3 reuse for far users and 1/1 reuse for close users. Reference [213] proposes an algorithm to apply hybrid frequency allocation in an OFDMA-based network that maximizes the total cell throughput. Reference [214] introduces an FFR optimization technique wherein the edge-region of each is divided into three sectors as illustrated in Fig. 2.2(d). The optimal configuration of the proposed algorithm is based on maximizing the average sector data rate subject to a minimum cell-edge data rate. In a SFR deployment, the cell-interior users are allowed to share frequency bands with edge-users in other cells (Fig. 2.2(e)). Typically, cell-interior users in SFR transmit at lower power than cell-edge users [215]. While SFR outperforms FFR in terms of spectral efficiency, it results in more interference to users [216]. Some studies in the literature (e.g., [217, 218]) suggest that power control does not always yield significant performance gain in OFDM systems compared to the complexity it adds to the operations of the system. Accordingly, these studies adopt a simple binary power control model in which either a subcarrier can be assigned to a terminal with maximum power or not. In the scheme proposed by Kwon *et al.* in [217], resource allocation is managed independently at each “pseudo-cell” composed of the major-interfering sectors belonging to the neighboring cells (Fig. 2.2(f)) in order to reduce the signaling and computation overhead. For each subcarrier, the transmission power is fixed, while varying the transmission rate by using adaptive modulation. The subcarriers are dynamically divided into different groups according to their load condition and for each group it is applied either frequency reuse 1 or frequency reuse $N_s$, where $N_s$ denotes the number of sectors within one pseudo-cell.

For OFDMA networks, centralized FFR approaches available in references [219, 220] show results with equal/unequal power levels and adaptive power levels as well. These works consider fixed partitioning of radio resources for cell center and edge users.

The parameters such as distance and SINR to partition cell center and edge regions are used as fixed thresholds. References [221–226] investigate the problem of game-theoretic radio resource allocation in OFDMA-based multi-cell networks. Reference [131] surveys different techniques for ICI mitigation in OFDMA-based multicellular networks.

There are few open issues in applying existing channel assignment approaches to OFDMA-based multi-cell networks. First, traditional frequency assignment assumes a predefined SINR threshold, which is rather suitable for homogenous services, but cannot be adopted for multimedia services. Secondly, traditional frequency assignment deals with flat fading channels and consequently overhead and computational complexity due to the measurement and signaling is associated to one frequency band. However, an OFDMA network needs to exchange information on all of the subcarriers which dramatically increases the complexity of measurements. As a result, fully centralized schemes are often too heavy for implementation as all the interference information on all channels has to be gathered and calculated at a central controller. On the other hand, fully distributed schemes have difficulties dealing with uneven and instantaneous loaded traffic.

### 2.6.2 Pico- and Femto-cell networks

One of the key expectations for the future wireless system is to provide ubiquitous high data rate coverage. But with the traditional cellular architecture, increasing the capacity together with the coverage requires the deployment of a large number of BSs, which is very costly. Multi-tier networks consist of a conventional cellular network overlaid with low-power and small range micro- and pico-BS (e.g. femto-cells, distributed antennas, or wired/wireless relays) which offer a cost effective way to enhance cellular system performance. Small BSs act as access points which extend network coverage and enhance capacity without incurring in the cost of backhaul connections. Femto-cells can be installed by end-users to improve the indoor coverage and capacity in residential and office environments. On the other hand, a pico-cell can be installed by a wireless service provider in public spaces or large buildings. In multicellular networks, macro-, micro-, and pico-cells suffer from a problem of co-channel interference between them in the neighboring cells. In radio resource allocation the problem of co-channel interference must be taken into consideration to evaluate the improvement of system throughput and economical feasibility by employing the different types of BSs. The coverage of a pico-/femto-cell becomes smaller when it is closer to a high power macro BS [227]. Under the pre-allocated frequency band within a macro-cell through the FFR optimally, a radio resource management allocates sub-bands the small BSs efficiently to consider macro-cell having a priority over micro-/femto-cells and total/edge throughputs.

References [228–230] propose spatial reuse schemes to mitigate the co-channel interference for OFDMA femto-cell networks and increasing femto-cells throughput as well. Ko *et al.* in [231] present a self-organizing femto-cell networks where users optimize their performance in a distributed manner. References [232, 233] investigate spectrum allocation techniques for femto-cells, based on Markov modeling and Q-learning [79] , respectively. Contributions [234–236] propose game theoretic algorithms to adjust the transmit powers of femto-BSs to mitigate interference, improve total capacity, and approach fairness among users. An efficient power allocation to femto-cells to cover specific terminals is presented by [237]. References [227, 238] propose algorithms to adjust pico BSs power as a tradeoff between cell coverage and cell throughput. Reference [239] devises a cross layer design for joint resource allocation and admission control in a two-tier OFDMA based network. The admission control is optimally designed based on Markov decision and the power of femto BSs is efficiently allocated using non-cooperative game theory.

### 2.6.3 Relay assisted networks

Cooperative communication with intermediate relay stations is an emerging technology to improve the performance of a wireless communication system. A relay is used to improve the transmission quality between a source node and a destination, and it can operate in either the amplify-and-forward (AF), the decode-and-forward (DF), or compress-and-forward (CF) mode. In the AF mode, the relay simply retransmits its received signal, including the interference and its local additive noise. A DF relay decodes its received signal before retransmission. Since the interference generated in the source-to-relay transmission is eliminated, the DF mode improves the effective SINR considerably, but at the cost of an increased hardware complexity. When the relay is not able to perfectly decode the received signal, the CF strategy is used to estimate the transmitted signal by the source node.

The most important factor which impacts on performance of relay-aided communications is radio resource management, i.e., how different subcarriers of OFDMA sources/relays should take part into a relayed transmission. References [240, 241] study the capacity of relaying in OFDMA applying AF and DF strategies. They also propose algorithm to subcarrier assignment with fixed-power. The algorithm proposed by Hammerström *et al.* in [242] provides a power allocation on the different subcarriers at the AF relay and the transmitter node, in a dual-hop OFDM relay communication scenario. The goal of the proposed power allocation is the maximization of the channel capacity with respect to the separate constraints on the transmitted power on the relay and source. In [243] and [244] L. Vandendorpe *et al.* propose some power allocation techniques in an OFDMA dual-hop relay communication applying the DF strategy. In particular, [243] places a constraint on the sum of the power consumption by the relay and sender, whereas [244] uses two individual constraints. References [245–248] investigate the resource allocation problem in an OFDMA-based point-to-point communication assisted by multiple relays to maximize the data rate under a power constraint.

Relaying is one of the enabling techniques for the next generation wireless networks. The first commercial relay-assisted OFDMA network has been standardized by IEEE 802.16j [249]. In single-cell scenarios, references [250, 251] evaluate the performance of a relay-assisted OFDMA network for a specific setup with three relay stations in a cell with and without intracell frequency reuse. The numerical results show that the relay-assisted system significantly outperforms the conventional cellular system with respect to system capacity and coverage. References [252, 253] present radio resource approaches in the downlink and uplink direction, respectively, assisted by multiple relays. The optimization is formulated as a spectrally efficient maximization, and the results show good performance in terms of spectral efficiency and fairness. Game theoretic frameworks for the best relay selection policy and efficient spectral usage in relayed OFDMA networks are presented in [254–257]. These schemes significantly increase the system performance and achieve a fair achieved data rate vector.

In the multicellular environment, the deployment of relay stations in the co-channel cells can be jointly optimized to maximize the overall spectral efficiency. Joung *et al.* in [258] introduce a power efficient radio resource allocation algorithm in relay multicellular networks in the downlink direction. Reference [259] present a dynamic FFR, like Fig. 2.2(e), in a multicellular relay-assisted network in the downlink direction to mitigate ICI. For the maximization system data rate in multicellular relay assisted OFDMA-based networks, different resource management approaches are presented by [260–266] under the constraint on the total power consumption. For further details, in the context of relay assisted OFDMA networks, the reader is referred to the survey provided by Salem *et al.* [132].

### 2.6.4 Cognitive radio

IEEE 802.22 [267], a continuously developing standard, employs OFDMA for the physical layer with cognitive radio technology in the MAC layer, thus opening new topics of radio resource management for OFDMA-based cognitive radio. In a cognitive radio system, a secondary user (SU) identifies available or unused licensed parts of the spectrum and exploit them with the goal of maximizing the throughput while minimizing the interference to primary users (PUs). The inherent FFT operation and the capability of assigning disjoint subsets of subcarriers to different secondary users make OFDMA adaptive for spectrum sensing in frequency domain and spectrum shaping, respectively [268]. The SUs can change the assigned subcarriers, the transmit powers, and the modulation and coding over each subcarrier according their channel gains. This flexibility helps OFDMA terminals to adapt to the environment with goal of minimizing BER and maximizing throughput. The limiting factor of the system performance is ICI between PUs and SUs. The interference can be limited not only by a proper radio resource allocation, but also usingthe channel state information. In [269], Weiss *et al.* provide a quantitative evaluation of the mutual interference between SUs and PUs that is caused by non-orthogonality of their respective transmitted signals. The “adjacent channel interference” can be mitigated by providing a flexible guard bands between PUs and SUs. They show that, although cyclic prefix and postfix help to reduce interference, they result in a reduced system throughput. To this aim, the authors propose dynamic deactivation of subcarriers lying adjacently to PU allocated sub-bands.

Generally, the resource allocation problem in OFDMA-based cognitive radio systems is formulated as the maximization of the total transmission rates of SUs by adjusting the power of selected subcarrier while the interference introduced to the PU must be kept below a tolerable threshold, and the total power of subcarrier does not exceed the total power constraint. An SU may not able to detect existence a PU when there is a large distance between them, but they can interfere when the SU is transmitting. In order to avoid unacceptable interference to the PUs that may not be detected by a SU, the SU should limit its transmit power even when no PU is detected [270]. As a result, the traditional WF approach is not suitable to maximize the capacity of an SU in an OFDMA-based cognitive radio system. For this purpose, [271–273] reformulate the WF problem and propose iterative algorithms based on convexity programming. Mao *et al.* in [274] introduce an iterative energy efficient WF-based power allocation which converges after few time steps. To maximize the sum rate in OFDMA-based cognitive radio multicellular networks, [275–278] propose algorithms based on convex geometric programming to optimize allocation of radio resources. Ma *et al.* in [276] propose a (reconfigured) WF solution to achieve weighted sum rate of SUs over multiple cells. The centralized and iterative algorithm proposed by [277] outperforms the one introduced in [275] in terms of computational complexity and spectral efficiency. The computational complexity of algorithm MLWF proposed by [278] is lower than that of ELCI proposed by [277]. Wang *et al.* in [279] propose an algorithm to achieve a proportional fairness data rate vector with a good sum data rate.

Applying game theory in the context of cognitive radio systems where OFDMA terminals can be useful to sense the environment and adaptively adjust their transmission power over the best selected subcarriers. To this aim, both non-cooperative (e.g., in [280, 281]) and cooperative (e.g., [282]) approaches are used to address this problem.

## 2.7 Summary

In this chapter, we have presented a survey on the state-of-the-art techniques to apportion the resources in both single-user systems, based on the OFDM modulation, and multiuser networks, based on the OFDMA channel access scheme. While resource allocation in OFDM systems is a mature field of investigation, with current trends of research attempting to further increase the efficiency towards near-Shannon performance, similar techniques for OFDMA still represent a hot topic of research. Many optimization have been proposed in the literature, focusing on both energy- and spectral-efficient approaches, using the margin- and the rate-adaptive formulations, respectively. The latter optimization problem has received most attention, appearing to be the most appealing one from both the centralized and the distributed point of view, and many practical solutions have been proposed in the last two decades. Important issues such as fairness and algorithmic complexity have also been included into the loop. To improve the performance of the proposed schemes, cross-layer formulations and several optimization tools, including game theory, have been recently adopted. Nevertheless, many drawbacks that limit the application of state-of-the-art algorithms to practical contexts still hold, mainly due to the high computational complexity and the weak scalability of the proposed techniques, that often make real-time solutions intractable problems, as they also require a considerable amount of feedback information across the network nodes. In this respect, we have tried to summarize the most relevant state-of-the-art techniques, also outlining current open problems in this active area of research, and including an overview of current challenges for next-generation networks, such as macro/micro multicellular planning, relaying and cognitive communications.

## 2.8 Discussion

It is a matter of controversy whether the OFDMA resource allocation techniques in the literature are actually usable in the practice. All the mentioned schemes, which represent, to the author's knowledge, the most relevant algorithms for OFDMA resource allocation with cooperative game theory, exhibit a good trade-off between overall system rate and fairness. The fairness schemes in the solutions based on cooperative game theory are extended approaches of that in classic solutions of: max-min fairness, and weighted proportional fairness schemes. Unfortunately, they also present a number of common problems:

1) In almost all algorithms the utility function is restricted to either be convex or strictly concave;

2) Most algorithms are based on non-linear programming, which is computationally intensive and hardly scalable when considering thousands of subcarriers and tens of users. Thus, they are not suitable for a cost-effective real-time implementation by network designers;

3) Although the resource apportionment turns out to be fair from the users point of view, the achieved QoS may be much larger than demanded. This implies a waste

of network resources from a network service provider perspective, which is often overlooked by previous works;

4) To reduce the computational complexity, each subcarrier is allocated to mobile terminals in an exclusive manner, although this may limit the number of concurrent connections in the uplink channel;

5) To reduce the computational complexity, the power constraint is usually defined as the overall energy consumption of each user over all subcarriers rather than individual limitation on each subcarrier, and this may result in impractical spectral power distribution.

We reviewed the concepts of coalitional game theory in Chapter 1. In Chapter 3 we will introduce an algorithm based on cooperative games to overcome most of the above mentioned disadvantages of the existing schemes. We aim at designing a low-complexity algorithm that achieves each users QoS requirement in terms of target transmit rates, with the best utilization of the network resources, so as to satisfy both the users and the network service provider.