EURASIP Journal on Wireless Communications and Networking
a SpringerOpen Journal

RESEARCH Open Access

Unified approach to cross-layer scheduling and resource allocation in OFDMA wireless networks

Guillem Femenias*, Borja Dañobeitia and Felip Riera-Palou

Abstract

Orthogonal frequency division multiple access (OFDMA) has been selected as the core physical layer access scheme for state-of-the-art and next-generation wireless communications standards. In these systems, scheduling and resource allocation algorithms, jointly assigning transmission data rates, bandwidth and power, become crucial to optimize the resource utilization while providing support to multimedia applications with heterogeneous quality of service (QoS) requirements. In this article, a unified framework for channeland queue-aware QoS-guaranteed cross-layer scheduling and resource allocation algorithms for heterogeneous multiservice OFDMA wireless networks is presented. The framework encompasses different types of traffic, uniform and continuous power allocation, discrete and continuous rate allocation, and protocols with different amounts of channel- and queue-awareness. System parameters and QoS requirements are projected into utility functions and the optimization problem is then formulated as a constrained utility maximization problem. Optimal solutions for this problem are obtained for the uniform power allocation schemes, and novel quasioptimal algorithms are proposed for the adaptive power allocation strategies. Remarkably, these techniques exhibit complexities that are linear in the number of resource units and users. Simulation results demonstrate the validity and merits of the proposed cross-layer unified approach.

1 Introduction

Due to its high spectral efficiency, inherent robustness against frequency-selective fading and flexibility in resource allocation, orthogonal frequency division multiple access (OFDMA), combined with multiple-input multiple-output (MIMO) strategies, has been chosen as the multiple access technique for state-of-the-art and next-generation wireless communications standards such as IEEE 802.16e/m-based WiMAX systems [1] and Third Generation Partnership Project (3GPP) technologies based on the long-term evolution (LTE) and LTE-advanced (LTE-A)[a] [2]. These systems have been designed with different quality of service (QoS) frameworks and strategies to allow the delivery of the wide range of emerging Internet multimedia applications with diverse QoS requirements [3]. In this context, scheduling and resource allocation algorithms jointly assigning transmission data rates (AMC–adaptive modulation and coding), subcarriers, time slots and power become crucial for maximizing the resource utilization while providing satisfactory service delivery to end users.

The instantaneous characteristics of the transmission channel used by wireless MIMO-OFDMA networks are inherently varying in time and frequency due to multipath propagation, changing positions of mobile stations (MS) relative to the base station (BS), and nonsta-tionary environment. Consequently, the result is that different users sharing a BS experience different channel conditions at the same time and frequency. This phenomenon, referred to as multiuser diversity, constitutes the basis of opportunistic or channel-aware scheduling algorithms. The goal of these strategies is to jointly allocate resources (i.e., power, subcarriers and/or time slots) in order to either minimize the weighted sum of powers under a prescribed minimum rate budget [4-6] or maximize the weighted sum-rate under a prescribed power budget [7-9]. Nevertheless, greedy-opportunistic schedulers serving only the users with favorable channel quality conditions raise the issue of fairness, as those users experiencing bad channel quality conditions may suffer from starvation. Therefore, besides channel state information (CSI), fairness is also an important issue that has

* Correspondence: guillem.femenias@uib.es
Mobile Communications Group, University of the Balearic Islands (UIB), Ctra. de Valldemossa Km. 7.5, 07122 Palma, Spain

been taken into account when designing scheduling algorithms for OFDMA-based multiservice networks [10-16]. Fairness, however, may lead to low spectral efficiency, and this may become an issue when facing real-time services with stringent QoS requirements in terms of delay and error tolerance. Thus, beyond channel quality conditions and fairness, another important issue that should be considered to maximize users' satisfaction is the one raised by the wide range of QoS requirements of heterogeneous applications supported by emerging OFDMA-based wireless networks.

In order to tackle all previously mentioned issues, the data link layer (DLC) bursty packet arrivals and queueing behavior should be jointly taken into consideration, in a cross-layer fashion, with the physical layer (PHY) channel conditions when designing scheduling and resource allocation algorithms. In this context, publications such as [17-23], reporting optimal and suboptimal cross-layer algorithms for very specific wireless multiuser OFDMA network configurations, lack a complete overview of the full problem, making it difficult to extract general conclusions. Song and Li [17,18] present a framework for cross-layer optimization of downlink multiuser single-cell OFDMA systems, where the interactions between the physical and DLC layers are modeled using a utility function that trades fairness for throughput efficiency. This work assumes, however, that the system has an infinite number of subcarriers and proposes suboptimal allocation algorithms for practical realization. A cross-layer scheduling scheme for OFDMA wireless systems with heterogeneous delay requirements taking into account both queueing theory and information theory in modeling the system dynamics is presented in [19]. The objective of maximizing system throughput with constraints on the delay and the maximum transmitted power is formulated as a mixed convex and combinatorial optimization problem. Mohanram and Bhashyan [20] propose a sub-optimal joint subcarrier and power allocation for channel- and queue-aware schedulers aiming at the maximization of the global average long term throughput. This scheduler, however, seems to be only applicable for traffic types without any constraint on delays. In [21] the authors propose a QoS-aware proportional fairness (QPF) scheduling policy based on a cross-layer design where the scheduler is aware of both the channel and the queue state information. The proposed approach, however, apart from using suboptimal modified greedy multicarrier proportional fairness algorithms, only considers Shannon's capacity-based data rate allocation schemes and uniform power allocation (UPA) in the frequency domain. In [22], Song et al. propose a joint channel-and queue-aware scheduler, which is called the max-delay-utility (MDU) scheduling, designed to efficiently support delay-sensitive applications. However, this scheduler is only effective for traffic types without explicit constraints on the minimum achievable average data rate and/or the maximum allowable absolute delay. Furthermore, only suboptimal sorting-search algorithms for the subcarrier (subband) allocation problem and greedy algorithms for the power allocation problem are proposed. Finally, Zhou et al. [23] propose a packet-dependent adaptive cross-layer design for downlink multiuser OFDMA systems, designed to maximize the weighted sum capacity of users with multiple heterogeneous traffic queues and based on the suboptimal algorithms proposed in [19].

Scheduling and resource allocation based on cross-layer principles can be regarded as a multi-objective optimization problem taking into account not only the system throughput but also the transmitted power, the QoS constraints on traffic delay and minimum and maximum data rates, the priority levels of different traffic classes and amount of backlogged data in the queues. In general, there is not a single optimal solution to a multi-objective optimization problem, however, using tools from information theory, queueing theory, convex optimization, and stochastic approximation [24], a unified framework for channel- and queue-aware QoS guaranteed scheduling and resource allocation for heterogeneous multiservice OFDMA wireless networks is proposed in this article. To this end, this study introduces a framework able to account for different types of traffic (e.g., best effort, non-real-time and real-time), different allocation strategies (e.g., continuous and discrete rate allocation (DRA), uniform and adaptive power allocation (APA)), protocols with different amounts of channel- and queue-awareness, and different utility functions measuring user's satisfaction in terms of, for instance, throughput, queue length and/or service time (waiting time in the queues). Channel state, physical-layer characteristics, queueing delay and/or QoS requirements are projected into utility functions and the multi-objective optimization problem is then formulated as a constrained utility maximization problem, where the objective function is the maximization of the user services' utility functions. The constraints are related to the specifications of the network and offered services under consideration, namely, power limitations, per-service rate limits, and exclusive chunk (frequency/time resource unit) assignment. The unified algorithmic framework presented in this article generalizes results presented in, for instance, [19,21,22,25,26]. The proposed approach is based on dual decomposition optimization [27] and stochastic approximation techniques [24] exhibiting complexities that are linear in the number of resource units and users, and that achieve negligible duality gaps in numerical simulations based on current

standards-like scenarios. Algorithms presented in this article optimize non-static utility functions based on the temporal evolution of throughput and/or waiting time of packets in the queues. Stochastic approximation techniques are used that allow these strategies to be implemented in real time.

This article is organized as follows. Section 2 presents a brief description of the system model under consideration alongside with the key assumptions made in the formulation of the optimization problem. A thorough description of the single-cell scenario, transmitter and receiver architectures, as well as of the channel model employed is also provided. As part of the cross-layer unified framework, the variables involved in the optimization problem are described in Section 3. Next, Section 4 presents a unified framework for constrained channel- and queue-aware QoS guaranteed scheduling and resource allocation for heterogeneous multi-service OFDMA wireless networks. Both continuous (Shannon-capacity-based) and discrete (AMC-based) strategies are considered, and solutions based on dual-optimization techniques are provided. In Section 6, numerical results illustrating the different performance/complexity tradeoffs of the proposed unified optimization framework are presented. Special emphasis is paid to efficiency, fairness and the fulfillment of QoS requirements. Finally, Section 7 summarizes the contributions of this article, and outlines the most interesting avenues for further research.

This introduction ends with a notational remark. Vectors and matrices are denoted by lower- and uppercase bold letters, respectively. The K-dimensional identity matrix is represented by $\boldsymbol{I}_K$. The symbols $\mathbb{R}_+$and $\mathbb{C}$ serve to denote the set of non-negative real numbers and the set of complex numbers, respectively. Superscripts $(\cdot)^T$ and $(\cdot)^\dagger$ are used to denote the transpose and the conjugate transpose (hermitian) of a matrix. Finally, and for the sake of clarity, a list of the most important symbols (in order of appearance) is also provided in Table 1.

2 System model and assumptions

Let us consider the downlink of a time-slotted MIMO-OFDMA wireless packet access network as the one depicted in Figure 1. In this setup, a BS with a total transmit power P_T and equipped with N_T transmit antennas provides service to N_m active MS, each equipped, without loss of generality, with an equal number of receive antennas, denoted by N_R.

Transmission between the BS and active MSs is organized in time slots of a fixed duration T_s, assumed to be less than the channel coherence time. Thus, the channel fading can be considered constant over the whole slot and it only varies from slot to slot, i.e., a slot-based block fading channel is assumed. Each of these slots consists of a fixed number N_o of OFDM symbols of duration $T_o + T_{\mathrm{CP}} = T_s / N_o$, where T_{CP} is the cyclic prefix duration. Slotted transmissions take place over a bandwidth B, which is divided into N_b orthogonal subbands, each consisting of N_{sc} adjacent subcarriers and with a bandwidth $B_b = B/N_b$ small enough to assume that all subcarriers in a subband experience frequency flat fading. One subband in the frequency axis over one slot in the time axis forms a basic resource allocation unit. Active MS and frequency subbands in a given slot are indexed by the sets $\mathcal{N}_m = \{1, \ldots, N_m\}$ and $\mathcal{N}_b = \{1, \ldots, N_b\}$, respectively.

Without loss of generality, and in order to simplify the mathematical notation of the problem, only one service data flow (also known as connection or session) per active MS will be assumed. Depending on the traffic type, three classes of service and the associated QoS requirements and priorities must be accounted for in wireless communications [24]:

- *Best effort* (BE) low priority services with a prescribed maximum allowable error rate but without specific requirements on rate or delay guarantees. Examples of best-effort services include applications such as e-mail or HTTP web browsing.

- *Non-real-time* (nRT) services entail applications such as file transfers (FTP). They do not impose any constraint on delays but, in addition to a maximum allowable error rate, they require sustained throughput guarantees.

- *Real-time* (RT) high priority services are used for applications such as video conferencing and streaming entailing QoS guarantees on maximum allowable error rate, minimum throughput, and maximum delay.

Traffic flows arriving from higher layers are buffered into the corresponding N_m first-in first-out (FIFO) queues at the DLC layer. At the beginning of each scheduling time interval, based on the available joint channel- and queue-state information (CSI/QSI), the cross-layer scheduling and resource allocation algorithms select some packets in the queues for transmission, which are then forwarded to the OFDM transmitter, at a rate $R_m(t)$ for all $m \in \mathcal{N}_m$, where they are adaptively modulated and channel encoded (AMC), and are allocated power and subbands, just before MIMO processing.

2.1 PHY layer modeling

2.1.1 Transmitter

Multiple-input multiple-output technology provides a great variety of techniques to exploit the multiple propagation paths between the N_T transmit antennas and the N_R receive antennas. Notably when CSI is available at the transmitter and receiver sides, and multiplexing in the spatial domain is not used, the joint use of maximum ratio transmission (MRT) [28] at the transmitter

Table 1 List of selected symbols (in order of appearance)

Symbols	Description
P_T	BS transmit power
N_T, N_R	Number of transmit and receive antennas
N_m	Number of active MSs
T_s	Time slot duration
T_o	OFDM symbol duration (without the cyclic prefix)
T_{CP}	OFDM cyclic prefix duration
N_o	Number of OFDM symbol per slot
B	System bandwidth
B_b	Subband bandwidth
N_b	Number of orthogonal subbands
Nsc	Number of subcarriers per subband
$\mathcal{N}_m$	Set of active MSs
$\mathcal{N}_b$	Set of frequency subbands
$p_{m,\ b}(t)$	Power allocated to MS m on subband b during the time slot t
$\delta_{m,\ b}(t)$	Equivalent MRT/MRC channel gain for MS m on subband b during the time slot t
σ_v^2	Noise variance
$Q_m(t), \hat{Q}_m(t), \bar{Q}_m(t)$	Measured, predicted and sample average queue length of MS m at the beginning of time slot t
$A_m(t)$	Number of arriving bits to the queue of MS m during time slot t
λ_m	Average arrival data rate for MS m
$r_m(t)$, $R_m(t)$	Allocated and effective data rates of user m during time slot t
$\hat{W}_m(t), \bar{W}_m(t)$	Predicted and sample average delay (waiting time) for MS m at the beginning of time slot t
$\bar{T}_m(t)$	Throughput sample average for MS m at the beginning of time slot t
$\tau_m^{(A)}(t), \hat{\tau}_m^{(A)}(t)$	Measured and predicted arrival time of the HOL packet in the queue of MS m
$W_{\mathrm{HOL},m}(t),\ \hat{W}_{\mathrm{HOL},m}(t)$	Measured and predicted HOL delay for MS m at the end of time slot t
$\boldsymbol{p}(t)$	Vector of power allocation values during time slot t
$\boldsymbol{\mathcal{P}}$	Set of allowed power allocation vectors
$\rho m,\ b(t)$	PHY layer transmission rate of MS m over subband b during time slot t
$\mathcal{N}_k$	Set of available MCSs when using discrete-rate AMC
N_k	Number of available MCSs when using discrete-rate AMC
$\rho_m^{(k)}$	Data rate characterizing MCS k when MS m uses discrete-rate AMC
$\Gamma_m^{(k)}$	Instantaneous SNR boundaries defining MCS selection intervals for MS m
Λ_m	Coding gap for MCSs used by MS m
$\boldsymbol{\theta}_m(t)$	Set of quantitative QoS measures used to characterize the satisfaction of user m
$\check{\boldsymbol{\Omega}}_m$	Set of QoS requirements for MS m
$\check{\varepsilon}_m, \check{D}_m, \check{\xi}_m$	Maximum tolerable BER, absolute delay and outage delay probability for MS m
$\check{T}_m$	Minimum sustainable throughput for user m
$\check{W}_m$	Delay threshold related to $\check{T}_m$ in MDU scheduling rule
ϕ_m	Set of constants used by the MDU scheduler to differentiate between heterogeneous services
$w_m(t)$	Weighing (prioritization) coefficient for MS m during the time slot t
$U(\cdot)$	Utility function used to express the satisfaction of user m
$\mathcal{L}(\cdot)$	Lagrangian of an optimization problem
μ	Lagrange multiplier
$g(\cdot)$	Dual problem

and maximal ratio combining (MRC) at the receiver is known to provide optimum performance in the sense of maximizing the received signal-to-noise ratio (SNR).

Let us assume that subband b has been allocated to MS m and that the BS uses an MRT scheme to exploit the spatial diversity provided by the MIMO channel. In this case, bits from the queue of MS m are channel encoded and mapped onto a sequence of symbols drawn from the allocated normalized unit energy complex constellation (e.g., BPSK, QPSK, 16QAM, 64QAM). Furthermore, before the usual OFDM modulation steps on each transmit antenna (IFFT, cyclic prefix appending and up-conversion), the symbols are allocated power and are processed in accordance with the MRT

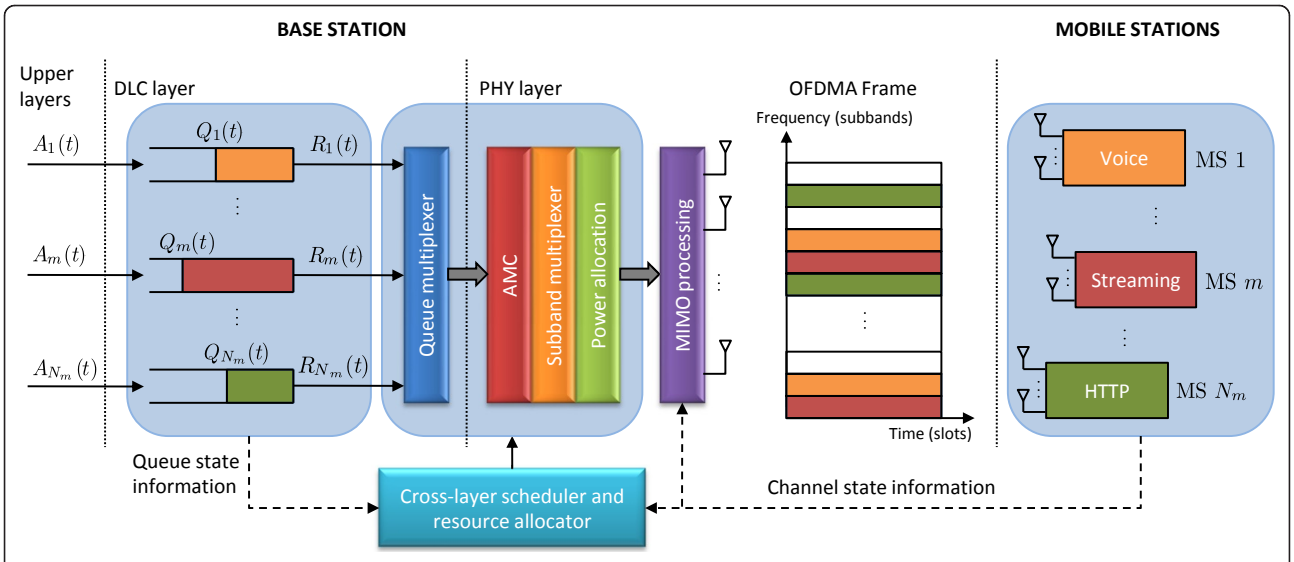


Figure 1 System model for cross-layer downlink scheduling and resource allocation over a OFDMA wireless network.

transmission scheme. Denoting by $d_{m,b}^{(c,o)}(t)$ the symbol to be sent to MS m over subcarrier $c \in \{1, ..., N_{sc}\}$ of subband b and OFDM symbol $o \in \{1, ..., N_o\}$ during time slot t, then the corresponding $N_T \times 1$ transmitted vector can be expressed as

$$\boldsymbol{x}_{m,b}^{(c,o)}(t) = \sqrt{\frac{p_{m,b}(t)}{N_{sc}}}\boldsymbol{v}_{m,b}(t)d_{m,b}^{(c,o)}(t), \tag{1}$$

where $p_{m,\,b}(t)$ is the power allocated to MS m on subband b during the time slot t (in a given subband, power is uniformly allocated to subcarriers), and $\boldsymbol{v}_{m,b}(t) \in \mathbb{C}^{N_T \times 1}$ denotes the unit energy linear transmit filter used by the MRT transmission system.

2.1.2 Channel model

The propagation channel between the BS and MS m is characterized by a power delay profile [29], common to all pairs of transmit and receive antennas, that can be expressed as

$$S_m(\tau) = \sum_{l=0}^{L_p-1} \sigma_{m,l}^2 \delta(\tau - \tau_l), \tag{2}$$

where L_p denotes the number of independent propagation paths, and $\sigma_{m,l}^2$ and τ_l are, respectively, the power and delay of the lth propagation path. Hence, assuming that the channel coherence time is greater than T_s, the channel impulse response between transmit antenna n_T and the receive antenna n_R of MS m, over the whole frame period t, can be written as

$$h_m^{n_R,n_T}(t;\tau) = \sum_{l=0}^{L_p-1} h_{m,l}^{n_R,n_T}(t)\delta(\tau - \tau_l), \tag{3}$$

where $E\{\left|h_{m,l}^{n_R,n_T}(t)\right|^2\} = \sigma_{m,l}^2$. The corresponding frequency response, when evaluated over subband b (with center frequency f_b), can be safely approximated by

$$H_{m,b}^{n_R,n_T}(t) = \sum_{l=0}^{L_p-1} h_{m,l}^{n_R,n_T}(t)e^{-j2\pi f_b \tau_l}. \tag{4}$$

Accordingly, the MIMO channel between the BS and MS m, for subband b and over the whole time slot period t, will be characterized by the complex valued $N_R \times N_T$ matrix

$$\boldsymbol{H}_{m,b}(t) = \begin{bmatrix} H_{m,b}^{1,1}(t) & \cdots & H_{m,b}^{1,N_T}(t) \\ \vdots & & \vdots \\ H_{m,b}^{N_R,1}(t) & \cdots & H_{m,b}^{N_R,N_T}(t) \end{bmatrix}. \tag{5}$$

2.1.3 Receiver

At the receiver side, as usual, ideal synchronization and sampling processes, and an OFDM cyclic prefix duration greater than the maximum delay spread of the channel impulse response are assumed. In this case, the received samples at the output of the N_R FFT processing stages of MS m over subcarrier c of subband b and OFDM symbol o during time slot t are given by the $N_R \times 1$ complex valued vector

$$\boldsymbol{y}_{m,b}^{(c,o)}(t) = \boldsymbol{H}_{m,b}(t)\boldsymbol{x}_{m,b}^{(c,o)}(t) + \boldsymbol{v}_{m,b}^{(c,o)}(t), \tag{6}$$

where $\boldsymbol{v}_{m,b}^{(c,o)}(t) \in \mathbb{C}^{N_R \times 1}$ is a noise vector with elements modeled as independent identically distributed (i.i.d.) zero-mean complex circular-symmetric Gaussian random variables, with covariance matrix $E\left\{\left\|\boldsymbol{v}_{m,b}^{(c,o)}(t)\right\|^2\right\} = \sigma_v^2 \boldsymbol{I}_{N_R}$.

According to the MRT strategy [28], the transmission filter $\boldsymbol{v}_{m,\,b}(t)$ that maximizes the SNR at the receiver side is the right singular vector of matrix $\boldsymbol{H}_{m,\,b}(t)$ associated with its largest singular value, denoted as $\sigma_{\max}(\boldsymbol{H}_{m,\,b}(t))$. In this case, the instantaneous SNR experienced by all subcarriers in subband b at the output of the maximal ratio combiner (MRC) used at the receiver side can be expressed as

$$\gamma_{m,b}(t) = \frac{p_{m,b}(t)\delta_{m,b}(t)}{N_{\mathrm{sc}}\sigma_v^2}, \tag{7}$$

where $\delta_{m,b}(t) = \sigma_{\max}^2(\boldsymbol{H}_{m,b}(t))$, which coincides with the largest eigenvalue of the $N_R \times N_R$ positive semidefinite Hermitian matrix $\boldsymbol{H}_{m,b}(t)\boldsymbol{H}_{m,b}^{\dagger}(t)$.

2.2 DLC layer modeling

In order to characterize the queueing behavior at the DLC layer, a slightly modified version of the model proposed by Kong et al. [[21], Section IV.A] is assumed. At the beginning of time slot t, MS m is assumed to have $Q_m(t)$ bits in the queue. If there are $A_m(t)$ bits arriving during time slot t, the queue length at the end of this time slot, assuming queues of infinite capacity, can then be expressed as

$$Q_m(t+1) = Q_m(t) + A_m(t) - R_m(t)N_oT_o, \tag{8}$$

where

$$R_m(t) = \min\left\{r_m(t), \frac{Q_m(t)}{N_oT_o}\right\}, \tag{9}$$

with $r_m(t)$ denoting the data rate allocated to user m during time slot t. A cross-layer resource allocation strategy that, in order to avoid the waste of resources, selects a transmission rate

$$r_m(t) \le \frac{Q_m(t)}{N_oT_o} \tag{10}$$

is said to fulfill the *frugality constraint* (FC) [22].

As will be shown in Section 4.1, most of the schedulers and resource allocation schemes that have been proposed in the literature can be interpreted as decision making algorithms that, at the beginning of time slot t estimate or predict the future behavior of QoS quantitative performance measures such as the throughput, average delay, queue length and/or head-of-line delay, and decide which users will be granted a transmission opportunity and the amount of resources that they will be allocated.

2.2.1 Predicting the queue length

As $A_m(t)$ is unknown at the beginning of time slot t, and assuming that the DLC layer only knows the average arrival data rate λ_m, then a prediction of the queue length at the end of this time slot can be obtained from (8) as

$$\begin{aligned}\hat{Q}_m(t+1) &= E_{A_m}\{Q_m(t+1)\} \\ &= Q_m(t) + \lambda_m T_s - R_m(t)N_oT_o,\end{aligned} \tag{11}$$

where $E_x\{\cdot\}$ denotes the statistical expectation operator with respect to the random variable x.

2.2.2 Predicting the average waiting time

Using standard stochastic approximation recursions, a recursive estimate of the slot-by-slot queue length sample average can be obtained as [24]

$$\bar{Q}_m(t+1) = (1-\beta_t)\,\bar{Q}_m(t) + \beta_t\, Q_m(t+1), \tag{12}$$

where the step-size $\beta_t \in (0, 1)$ implements a forgetting factor in the averaging and can be selected to be either constant (i.e., $\beta_t = \beta$) or asymptotically vanishing (e.g., $\beta_t = 1/t$). Little's law [30] asserts that with stable queues the average delay at the end of time slot t can be obtained as

$$\overline{W}_m(t+1) = \frac{\overline{Q}_m(t+1)}{\lambda_m}. \tag{13}$$

Using (11) and (12), this in turn leads to a recursive prediction of the slot-by-slot average delay via

$$\begin{aligned}\hat{W}_m(t+1) &= \frac{E_{A_m}\{\overline{Q}_m(t+1)\}}{\lambda_m} \\ &= (1-\beta_t)\overline{W}_m(t) + \beta_t\frac{\hat{Q}_m(t+1)}{\lambda_m}.\end{aligned} \tag{14}$$

2.2.3 Estimating the average throughput

Stochastic approximation tools can also be used to obtain a recursive estimate of the frame-by-frame throughput sample average as

$$\overline{T}_m(t+1) = (1-\beta_t)\overline{T}_m(t) + \frac{\beta_t R_m(t)N_oT_o}{T_s}. \tag{15}$$

2.2.4 Predicting the head-of-line delay

The HOL delay of user m at the beginning of time slot t (or equivalently, the end of time slot $(t - 1)$) can be written as $W_{\mathrm{HOL},m}(t) = tT_s - \tau_m^{(A)}(t)$, where $\tau_m^{(A)}(t)$ denotes the arrival time of the HOL packet in the queue of user

m. Hence, a prediction of the HOL delay at the end of time slot t can be readily obtained as

$$\begin{aligned}\hat{W}_{\mathrm{HOL},m}(t+1) &= (t+1)T_s - \hat{\tau}_m^{(\mathrm{A})}(t+1)\\ &= (t+1)T_s - \left(\tau_m^{(A)}(t) + \frac{R_m(t)N_oT_o}{\lambda_m}\right)\\ &= W_{\mathrm{HOL},m}(t) + T_s - \frac{R_m(t)N_oT_o}{\lambda_m}.\end{aligned}$$

3 Optimization variables

3.1 Power allocation

Let $\boldsymbol{p}_b(t) = [p_{1,b}(t) \ \ldots \ p_{Nm,\ b}(t)]^T$ denote the vector of power allocation values for subband b and time slot t. For a given set of constraints, the scheduling and resource allocation algorithm will be in charge of determining the power allocation vector

$$\boldsymbol{p}(t) = \left[(\boldsymbol{p}_1(t))^T \cdots (\boldsymbol{p}_{N_b}(t))^T\right]^T \tag{16}$$

optimizing a prescribed objective function. In addition to determining the power allocation values, the resource allocation algorithms should also allocate subbands and transmission rates. Nevertheless, as it will be shown next, the power allocation vector $\boldsymbol{p}(t)$ can also be used to represent the allocation of all these resources, thus simplifying the formulation of the optimization problem [9].

3.2 Subband allocation

As usual, it is assumed that subband allocation is exclusive, that is, only one MS is allowed to transmit on a given subband. Hence, the subband allocation constraints can be captured by constraining the power allocation vectors as

$$\boldsymbol{p}_b(t) \in \boldsymbol{\mathcal{P}}_b, \tag{17}$$

where

$$\boldsymbol{\mathcal{P}}_b \triangleq \{\boldsymbol{p}_b \in \mathbb{R}_+^{N_m} : p_{m,b}p_{m',b} = 0, \quad \forall m' \neq m\}, \tag{18}$$

with $\mathbb{R}_+$ denoting the set of all non-negative real numbers. Hence, the power allocation vector satisfies

$$\boldsymbol{p}(t) \in \boldsymbol{\mathcal{P}} = \boldsymbol{\mathcal{P}}_1 \times \cdots \times \boldsymbol{\mathcal{P}}_{N_b} \subset \mathbb{R}_+^{N_mN_b}, \tag{19}$$

where $\times$ denotes the Cartesian product (or product set).

3.3 Rate allocation

In the downlink of multi-rate systems based on AMC, a channel estimate is obtained at the receiver of each MS and it is then fed back to the BS so that the transmission scheme, comprising a modulation format and a channel code, can be adapted in accordance with the channel characteristics.

If MS m is allocated subband b over time slot t, then the BS selects a modulation and coding scheme (MCS) that can be characterized by a transmission rate $\rho_{m,\ b}(t)$ (measured in bits per second). As each subband contains N_{sc} subcarriers, the aggregated data rate allocated to MS m over time t will be given by

$$r_m(t) = N_{\mathrm{sc}} \sum_{b=1}^{N_b} \rho_{m,b}(t). \tag{20}$$

Transmission rate $\rho_{m,\ b}(t)$ can be related to bit error rate (BER) observed by MS m, denoted as ε_m, and instantaneous SNR $\gamma_{m,\ b}(t)$ as [[31], Chapter 9] (see also [24])

$$\varepsilon_m(\gamma_{m,b}(t), \rho_{m,b}(t)) = \kappa_1 \exp\left(-\frac{\kappa_2\gamma_{m,b}(t)}{2^{T_{o\rho m,b}(t)} - 1}\right), \tag{21}$$

where κ_1 and κ_2 are modulation-and code-specific constants that can be accurately approximated by exponential curve fitting. This expression is general enough to obtain the BER performance of any transmission system for which the joint effects of transmission filters, channel coefficients and reception filters can be represented through an instantaneous SNR $\gamma_{m,\ b}(t)$. For the special case of MRT/MRC scheme, $\gamma_{m,\ b}(t)$ is defined by (7).

3.3.1 Discrete-rate AMC

Realistic AMC strategies can only use a discrete set $\mathcal{N}_k = \{0, 1, \ldots, N_k\}$ of MCSs that can differ for different MSs. Each MCS is characterized by a particular transmission rate $\varrho_m^{(k)}$, with $\varrho_m^{(1)} < \cdots < \varrho_m^{N_k}$, and $\varrho_m^{(0)} = 0$ denoting the case where MS m does not transmit.

Given $p_{m,\ b}(t)$, $\delta_{m,\ b}(t)$ and the noise variance σ_ν^2, we can use (7) to find $\gamma_{m,\ b}(t)$ and then, considering the maximum allowable BER $\check{\varepsilon}_m$, employ (21) to select the most adequate MCS scheme as the one with transmission rate

$$\rho_{m,b}(t) = \max\{\varrho_m^{(k)} : \varepsilon(\gamma_{m,b}(t), \varrho_m^{(k)} \leq \check{\varepsilon}_m\}, \tag{22}$$

In fact, the transmission rate $\rho_{m,\ b}(t)$ can be expressed using the staircase function

$$\rho_{m,b}(t) = \begin{cases} \varrho_m^{(0)}, & 0 \leq \gamma_{m,b}(t) < \Gamma_m^{(1)} \\ \varrho_m^{(1)}, & \Gamma_m^{(1)} \leq \gamma_{m,b}(t) < \Gamma_m^{(2)} \\ \quad\vdots & \\ \varrho_m^{(N_k)}, & \Gamma_m^{(N_k)} \leq \gamma_{m,b}(t) < \infty \end{cases} \tag{23}$$

where $\{\Gamma_m^{(k)}\}_{k=1}^{N_k-1}$, with $\Gamma_m^{(k)} \leq \Gamma_m^{(k+1)}$, are the instantaneous SNR boundaries defining the MCS intervals,

which can be obtained from (21) as

$$\Gamma_m^{(k)} = \frac{2^{T_{o\varrho_m}^{(k)}} - 1}{\kappa_2^{(k)}} \ln \frac{\kappa_1^{(k)}}{\check{\varepsilon}_m}. \tag{24}$$

3.3.2 Continuous-rate AMC

A useful abstraction when exploring rate limits is to assume that each user's set of MCSs is infinite. In this case, the maximum allowable transmission rate fulfilling the prescribed BER constraint with equality can be obtained from (21) as

$$\rho_{m,b}(t) = \frac{1}{T_o} \log_2 \left(1 + \frac{\gamma_{m,b}(t)}{\Lambda_m}\right), \tag{25}$$

where $\Lambda_m = \kappa_2^{-1} \ln(\kappa_1/\check{\varepsilon}_m) \geq 1$ represents the coding gap due to the utilization of a practical (rather than ideal) coding scheme. With $\Lambda_m = 1$ this expression results in the Shannon's capacity limit and allows the comparison of practical AMC-based schemes against fundamental capacity-achieving benchmarks.

4 Problem formulation

The main objective of cross-layer scheduling and resource allocation algorithms over a wireless network is the establishment of effective policies able to optimize metrics related to spectral/energy efficiency and fairness, while satisfying prescribed QoS constraints. The issues of efficient and fair allocation of resources have been intensively investigated in the context of economics, where utility functions have been used to quantify the benefit obtained from the usage of a pool of resources. In a similar way, utility theory can be used in wireless communication networks to evaluate the degree up to which a given network configuration can satisfy users' QoS requirements [17,18].

Utility functions are used to map the resources (e.g., bandwidth, power, ...), performance criteria (e.g., throughput, delay,...) and QoS requirements (e.g., maximum tolerable error rate, maximum absolute delay, maximum allowable outage delay probability, ...) into the corresponding user's satisfaction. Different applications can be characterized by different utility functions and/or even different performance quantitative measures and QoS requirements. For instance, utility functions for BE applications are typically characterized in terms of throughput, whereas those for nRT or RT delay-sensitive applications are characterized in terms of queuing delay with QoS requirements on the sustainable throughput, and/or the average or absolute delay. Thus, in general, the satisfaction of MS m at time t can be expressed by a utility function $U_m(\boldsymbol{\theta}_m(t), \check{\boldsymbol{\Omega}}_m)$, where $\boldsymbol{\theta}_m(t) = \{\theta_m^1(t), \ldots, \theta_m^{N_z^{(m)}}(t)\}$ is the set of quantitative QoS measures used to characterize the satisfaction of MS m (e.g., throughput $\overline{T}_m(t)$, average delay $\overline{W}_m(t)$, queue length $Q_m(t)$ or HOL delay $W_{\mathrm{HOL,m}}(t)$) and $\check{\boldsymbol{\Omega}}_m = \left\{\check{\Omega}_m^1, \ldots, \check{\Omega}_m^{N_y^{(m)}}\right\}$ is the set of QoS requirements for user m (e.g., maximum tolerable error rate $\check{\varepsilon}_m$, maximum tolerable absolute delay $\check{D}_m$, maximum outage delay probability $\breve{\xi}_m$). Hence, assuming the availability of perfect CSI/QSI, the utility-based cross-layer scheduling and resource allocation scheme can be formulated as the following optimization problem,

$$\begin{aligned} &\max_{\boldsymbol{p}(t)\in\boldsymbol{\mathcal{P}}} \sum_{m=1}^{N_m} U_m(\boldsymbol{\theta}_m(t), \check{\boldsymbol{\Omega}}_m) \\ &\text{subject to} \sum_{m=1}^{N_m}\sum_{b=1}^{N_b} p_{m,b}(t) \leq P_T. \end{aligned} \tag{26}$$

4.1 Gradient-based scheduling and resource allocation

The first order Taylor's expansion of $U_m(\boldsymbol{\theta}, \check{\boldsymbol{\Omega}}_m)$ in a neighborhood of $\boldsymbol{\theta} = \boldsymbol{\theta}_m(t)$ can be written as

$$\begin{aligned} U_m(\boldsymbol{\theta}, \check{\boldsymbol{\Omega}}_m) \simeq\ & U_m(\boldsymbol{\theta}_m(t), \check{\boldsymbol{\Omega}}_m) \\ &+ (\boldsymbol{\theta} - \boldsymbol{\theta}_m(t))^T \nabla_{\boldsymbol{\theta}} U_m(\boldsymbol{\theta}_m(t), \check{\boldsymbol{\Omega}}_m), \end{aligned} \tag{27}$$

where ∇_θ denotes the vector differential operator or gradient function with respect to $\boldsymbol{\theta}$. Thus, using this approximation, the variation of utility for MS m during time slot t is given by

$$\begin{aligned} &U_m(\boldsymbol{\theta}_m(t+1), \check{\boldsymbol{\Omega}}_m) - U_m(\boldsymbol{\theta}_m(t), \check{\boldsymbol{\Omega}}_m) \\ &\quad\simeq \sum_{z=1}^{N_z^{(m)}} \frac{\partial U_m(\boldsymbol{\theta}_m(t), \check{\boldsymbol{\Omega}}}{\partial \theta_m^z(t)} [\theta_m^z(t+1) - \theta_m^z(t)]. \end{aligned} \tag{28}$$

Using this result, the cross-layer long-term optimization problem in (26) can be rewritten, as shown in [17,18,32], as the instantaneous gradient-based optimization problem

$$\begin{aligned} &\max_{\boldsymbol{p}(t)\in\boldsymbol{\mathcal{P}}} \sum_{m=1}^{N_m}\sum_{z=1}^{N_z^{(m)}} \frac{\partial U_m(\boldsymbol{\theta}_m(t), \check{\boldsymbol{\Omega}}_m)}{\partial \theta_m^z(t)} [\theta_m^z(t+1) - \theta_m^z(t)] \\ &\text{subject to} \sum_{m=1}^{N_m}\sum_{b=1}^{N_b} p_{m,b}(t) \leq P_T. \end{aligned} \tag{29}$$

Although utility functions based on QoS quantitative performance measures other than the throughput, the average delay, the queue length and/or the HOL delay could be devised, most practical utility functions used in state-of-the-art wireless communications are based on either one of these performance measures or a

combination of them. Therefore, let us assume hereafter that $\boldsymbol{\theta}_m(t) = \{\theta_m^1(t), \theta_m^2(t), \theta_m^3(t), \theta_m^4(t)\} = \{\overline{T}_m(t), \overline{W}_m(t), Q_m(t), W_{\mathrm{HOL},m}(t)\}$. In this case, using (11)-(2.2.4), then

$$\overline{T}_m(t+1) - \overline{T}_m(t) = \beta_t \left[R_m(t) \frac{N_o T_o}{T_s} - \overline{T}_m(t) \right], \tag{30}$$

$$\begin{aligned} &\hat{W}_m(t+1) - \overline{W}_m(t) \\ &\quad = -\beta_t \left[\frac{R_m(t) N_o T_o}{\lambda_m} - T_s - \frac{Q_m(t)}{\lambda_m} + \overline{W}_m(t) \right], \end{aligned} \tag{31}$$

$$\hat{Q}_m(t+1) - Q_m(t) = \lambda_m T_s - R_m(t) N_o T_o, \tag{32}$$

and

$$\hat{W}_{\mathrm{HOL},m}(t+1) - W_{\mathrm{HOL},m}(t) = T_s - \frac{R_m(t) N_o T_o}{\lambda_m}. \tag{33}$$

Finally, using these expressions in (29) and eliminating constants not affecting the optimization process yields

$$\begin{aligned} &\max_{p(t) \in \mathcal{P}} \sum_{m=1}^{N_m} w_m(t) R_m(t) \\ &\text{subject to } \sum_{m=1}^{N_m} \sum_{b=1}^{N_b} p_{m,b}(t) \le P_T. \end{aligned} \tag{34}$$

where the weighing (prioritization) coefficient for MS m during the time slot t is given by

$$\begin{aligned} w_m(t) = &\frac{\beta_t}{T_s} \frac{\partial U_m(\boldsymbol{\theta}_m(t), \breve{\boldsymbol{\Omega}}_m)}{\partial \overline{T}_m(t)} - \frac{\beta_t}{\lambda_m} \frac{\partial U_m(\boldsymbol{\theta}_m(t), \breve{\boldsymbol{\Omega}}_m)}{\partial \overline{W}_m(t)} \\ &- \frac{\partial U_m(\boldsymbol{\theta}_m(t), \breve{\boldsymbol{\Omega}}_m)}{\partial Q_m(t)} - \frac{1}{\lambda_m} \frac{\partial U_m(\boldsymbol{\theta}_m(t), \breve{\boldsymbol{\Omega}}_m)}{\partial W_{\mathrm{HOL},m}(t)}. \end{aligned} \tag{35}$$

4.2 Marginal utility functions

4.2.1 Max-sum-rate (MSR) rule

The MSR scheduler [33] is based on a channel-aware scheduling rule that, using

$$w_m(t) = 1, \ \forall m, \tag{36}$$

maximizes the slot-by-slot aggregated transmission rate

$$U(t) = \sum_{m=1}^{N_m} R_m(t). \tag{37}$$

However, as stated by Song et al. [22], although it maximizes the spectral efficiency, it can lead to unfairness and queue instability, especially for nonuniform traffic patterns and MSs operating in uneven channel conditions. Furthermore, since MSs with unfavorable channel conditions can experience long deep fading periods, long delays are expected and consequently, the MSR rule is not able to support delay-sensitive applications.

4.2.2 Proportional fair (PF) rule

The PF scheduler [34] is based also on a channel-aware scheduling rule aiming at maximizing the logarithmic-sum-throughput of the system, that is

$$U(t) = \sum_{m=1}^{N_m} \ln(\overline{T}_m(t)). \tag{38}$$

Thus, the gradient-based PF scheduling algorithm is effected by using

$$w_m(t) = \frac{1}{\overline{T}_m(t)}, \forall m. \tag{39}$$

Nevertheless, although PF rule can trade off spectral efficiency and fairness among users belonging to the same QoS class, it cannot cope with MSs with disparate QoS requirements, especially those supporting delay-sensitive applications. Particularly, long deep fading starvation periods are not solved by this rule.

It is worth pointing out that for incoming low-rate data flows it is quite common that for some users $\overline{T}_m(t) = \lambda_m$ no matter how good their average channel condition is; as a result, for those users, $\overline{T}_m(t)$ is not a good measure of the actual amount of resources allocated to them and so, it is better to use [35],

$$w_m(t) = \frac{1}{\overline{r}_m(t)}, \forall m. \tag{40}$$

4.2.3 Modified largest weighted delay first (M-LWDF) rule

The M-LWDF scheduler was proposed by Andrews et al. [36] for single-carrier CDMA networks with a shared downlink channel and was proved to be *throughput optimal*.[b] It is based on a channel- and queue-aware scheduling rule that considers the waiting time in the queues, the instantaneous potential transmission rates and the maximum tolerable delay requirements. At each time slot t, the M-LWDF scheduler aims at choosing the best combination of queueing delay and potential transmission rate, serving the users that maximize the sum of marginal utility functions given by [36]

$$U(t) = \sum_{m=1}^{N_m} \chi_m(t) W_{\mathrm{HOL},m}(t) \frac{R_m(t)}{\overline{r}_m(t)}.$$

That is,

$$w_m(t) = \chi_m(t) W_{\mathrm{HOL},m}(t) / \overline{r}_m(t), \ \forall m, \tag{41}$$

where $\chi_m(t)$ are arbitrary positive constants that can be used to set different priority levels between traffic flows. The M-LWDF scheduling rule remains

throughput optimal if for all or some queues, the head-of-line delay $W_{\mathrm{HOL},m}(t)$ is replaced by the queue length $Q_m(t)$. Thus, the scheduler can be easily implemented by time stamping arriving data packets of all MSs, and/or keeping track of the corresponding queue lengths.

In order to guarantee that users with absolute delay requirement $\check{\mathrm{D}}_m$ and maximum outage delay probability requirement $\breve{\xi}_m$ will be satisfied, the authors of [36] propose to *properly* set the values of $\chi_m(t)$ as

$$\chi_m(t) = -\frac{\log(\breve{\xi}_m)}{\check{D}_m}, \tag{42}$$

providing in this way QoS differentiation among user's flows.

As stated by Andrews et al. [37], services with QoS constraints on the minimum sustainable throughput $\check{T}_m$ can also be supported by the M-LWDF scheduling rule, provided that the scheduler is used in conjunction with a token bucket control. In this case, each queue m is associated to a *virtual* token bucket, with tokens arriving at a *constant* rate $\check{T}_m$. At each time slot, queues are served according to the M-LWDF rule, with $W_{\mathrm{HOL},m}(t)$ denoting the delay of the head-of-line token in bucket m instead of the head-of-line packet delay for queue m. After serving a given queue m, the number of tokens in the corresponding bucket must be reduced by the actual amount of data served.

4.2.4 Exponential (EXP) rule

The EXP scheduler, proposed by Shakkottai and Stolyar [38], is also based on a channel-and queue-aware throughput optimal scheduling rule that considers the waiting time in the queues, the instantaneous potential transmission rates and the maximum tolerable delay requirements. It was proposed for single-carrier CDMA networks with a shared downlink channel but, similarly to M-LWDF, it can easily be extended to multichannel scenarios. In this case, at each time slot t, the EXP scheduler serves the users maximizing the sum of marginal utility functions given by [35]

$$U(t) = \sum_{m=1}^{N_m} \chi_m(t)\frac{R_m(t)}{\bar{r}_m(t)} \exp\left(\frac{\chi_m(t)W_{\mathrm{HOL},m}(t) - \overline{\chi W}}{1+\sqrt{\overline{\chi W}}}\right).$$

That is,

$$w_m(t) = \frac{\chi_m(t)}{\bar{r}_m(t)} \exp\left(\frac{\chi_m(t)W_{\mathrm{HOL},m}(t) - \overline{\chi W}}{1+\sqrt{\overline{\chi W}}}\right)$$

for all m, with

$$\overline{\chi W} = \frac{1}{N_m}\sum_{m=1}^{N_m} \chi_m(t)W_{\mathrm{HOL},m}(t). \tag{43}$$

As in the M-LWDF scheduler, the head-of-line delay $W_{\mathrm{HOL},m}(t)$ can be replaced, for all or some queues, by the queue length $Q_m(t)$ without affecting the throughput optimality of this strategy. Furthermore, if providing a minimum throughput $\check{T}_m$ to a given flow is a goal, the EXP rule can also be modified by introducing a *virtual* token queue, where tokens arrive at a constant rate $\check{T}_m$, and serving the queue according to the EXP rule, with $W_{\mathrm{HOL},m}(t)$ denoting the delay of the head-of-line token in bucket m.

4.2.5 MDU rule

To efficiently support delay-sensitive applications, Song et al. [22] proposed another joint channel-and queue-aware scheduling approach, known as MDU rule, which maximizes the total utility with respect to average delays or average waiting times in the queues. Generalizing the marginal utility functions proposed by [39], the MDU scheduling rule can be treated in the unified optimization framework defined in (34) by setting

$$w_m(t) = \begin{cases} \dfrac{T_s\overline{W}_m^{\varphi m,1}(t)}{\lambda_m}, & \overline{W}_m(t) \le \tilde{W}_m \\ \dfrac{T_s\left[\overline{W}_m^{\varphi m,2}(t) - \tilde{W}_m^{\varphi m,2} + \tilde{W}_m^{\varphi m,1}\right]}{\lambda_m}, & \overline{W}_m(t) > \tilde{W}_m, \end{cases}$$

where $\phi_m = \{\phi_{m,1}, \phi_{m,2}\}$ is a set of constants used to differentiate between heterogeneous services and $\tilde{W}_m$ is a delay threshold related to the maximum tolerable delay $\check{T}_m$. In [39], based on the corresponding required QoS, these parameters were set to $\phi_m = \{1, 1.5\}$ and $\tilde{W}_m = 25\,\mathrm{ms}$ for packet-switched voice with end-to-end delay required to be less than 100 ms, $\phi_m = \{0.6, 1\}$ and $\tilde{W}_m = 100\,\mathrm{ms}$ for good-quality streaming transmission requiring end-to-end delays between 150-400 ms and, finally, $\phi_m = \{0.5, 0\}$ and $\tilde{W}_m = 100\,\mathrm{ms}$ ms for BE traffic. Actually, using these settings the MDU scheduling for the best-effort traffic becomes the PF scheduling.

4.2.6 Other scheduling rules

Although not treated in this article, the unified cross-layer optimization approach defined in (34) can also be extended to scheduling rules such as those proposed in [16,23,40-42]. Notably, Al-Manthari et al. [41] propose the use of utility functions that are based on both the throughput and the average delay.

5 Unified optimization framework

The optimization problem formulated in (34) is general enough to account for different power and rate allocation strategies, either with or without FC. For clarity of presentation, the following list of acronyms will be used: UPA, APA, continuous rate allocation (CRA), DRA and FC. Furthermore, since optimization is performed on a slot-by-slot basis, from this point onwards the time dependence (i.e., (t)) of all the variables will be dropped.

5.1 UPA without FC

Let us assume a system where the scheduling and rate allocation schemes do not consider the FC. In this case, problem (34) can be rewritten as

$$\begin{aligned} &\max_{\boldsymbol{p}\in\mathcal{P}} \sum_{m=1}^{N_m} w_m r_m \\ &\text{subject to} \sum_{m=1}^{N_m}\sum_{b=1}^{N_b} p_{m,b} \le P_T. \end{aligned} \tag{44}$$

Let us also assume that the BS transmit power P_T is uniformly allocated to all subbands. In this case, if subband b is allocated to user m_b^*, then the subband exclusive allocation constraint (i.e., $\boldsymbol{p} \in \mathcal{P}$) forces that

$$p_{m,b} = \begin{cases} P_T/N_b, & m = m_b^* \\ 0, & m \neq m_b^*, \end{cases} \tag{45}$$

for all b. Thus, using (20) in (44) it is straightforward to show that subband b must be allocated to MS m_b^* satisfying

$$m_b^* = \arg\max_{m\in\mathcal{N}_m} \{w_m \rho_{m,b}\}, \forall b, \tag{46}$$

with $\rho_{m,\,b}$ obtained as in either (23), for the DRA case, or (25), for the CRA case.

5.2 APA without FC

The objective function in (44) is concave, but $\mathcal{P}$ is a highly non-convex discrete constraint space. Fortunately, problem (44) is separable across the subbands and, as stated in [9,27], it can be approached by using Lagrange duality principles. With μ denoting the Lagrange multiplier associated with the power constraint, the Lagrangian of (44) can be expressed as

$$\mathcal{L}(\boldsymbol{p},\mu) = \sum_{m=1}^{N_m} w_m r_m + \mu\left(P_T - \sum_{m=1}^{N_m}\sum_{b=1}^{N_b} p_{m,b}\right), \tag{47}$$

and the dual problem can then be written as [43]

$$\begin{aligned} g(\boldsymbol{p},\mu) &= \min_{\mu\ge 0}\left\{\max_{\boldsymbol{p}\in\mathcal{P}} \mathcal{L}(\boldsymbol{p},\mu)\right\} \\ &= \min_{\mu\ge 0}\left\{\max_{\boldsymbol{p}\in\mathcal{P}}\left[\sum_{m=1}^{N_m}\left(w_m r_m - \mu\sum_{b=1}^{N_b} p_{m,b}\right)\right] + \mu P_T\right\}, \end{aligned} \tag{48}$$

Now, using the subband exclusive allocation constraint and the separability of power variables across subbands, the dual problem can be simplified as [26]

$$\begin{aligned} g(\boldsymbol{p},\mu) &= \min_{\mu\ge 0}\left\{\max_{\boldsymbol{p}\in\mathcal{P}} \mathcal{L}(\boldsymbol{p},\mu)\right\} \\ &= \min_{\mu\ge 0}\left\{\sum_{b=1}^{N_b}\max_{m\in\mathcal{N}_m}\left\{\max_{p_{m,b}\ge 0}\{w_m N_{\mathrm{sc}}\rho_{m,b} - \mu p_{m,b}\}\right\} + \mu P_T\right\}. \end{aligned} \tag{49}$$

The solution to the simplified dual problem is given by optimizing (49) over all $(\boldsymbol{p}, \mu) \succcurlyeq 0$. This optimization can be done iteratively and coordinate-wise, starting with the $\boldsymbol{p}$ variables and continuing with μ.

5.2.1 Optimizing the dual function over p

CRA: In case of using $\rho_{m,\,b}$ as defined in (25), and for a given value of μ, the innermost maximization in (49) provides a multilevel water-filling closed-form expression for the optimal power allocation given by

$$p_{m,b}^* = \left[\frac{N_{\mathrm{sc}} w_m}{\mu T_o \ln 2} - \frac{N_{\mathrm{sc}}\Lambda_m \sigma_v^2}{\delta_{m,b}}\right]^+, \tag{50}$$

where $[x]^+ \triangleq \max\{0, x\}$. Now, using (50) in (49) yields

$$g(\mu) = \min_{\mu\ge 0}\left\{\sum_{b=1}^{N_b}\max_{m\in\mathcal{N}_m}\left\{w_m N_{\mathrm{sc}}\rho_{m,b}^* - \mu p_{m,b}^*\right\} + \mu P_T\right\}, \tag{51}$$

with

$$\rho_{m,b}^* = \frac{1}{T_o}\log_2\left(1 + \frac{p_{m,b}^*\delta_{m,b}}{\sigma_v^2 N_{\mathrm{sc}}\Lambda_m}\right). \tag{52}$$

Hence, for a fixed dual variable μ, the subband b will be allocated to MS m_b^* satisfying

$$m_b^* = \arg\max_{m\in\mathcal{N}_m}\left\{w_m N_{\mathrm{sc}}\rho_{m,b}^* - \mu p_{m,b}^*\right\}, \forall b. \tag{53}$$

DRA: In this case $\rho_{m,\,b}$ is a non-derivable discontinuous function. However, the approach proposed in [[9] Chapter 3] can be applied to arrive at the optimal solution. Using (23) the set of non-negative real numbers (i.e., $\mathbb{R}^+$) can be subdivided, for each MS m and subband b, into N_k segments

$$\mathcal{R}_{m,b,k}^+ = \left[\frac{N_{\mathrm{sc}}\sigma_v^2\Gamma_m^{(k)}}{\delta_{m,b}}, \frac{N_{\mathrm{sc}}\sigma_v^2\Gamma_m^{(k+1)}}{\delta_{m,b}}\right), \quad k \in \mathcal{N}_k. \tag{54}$$

Furthermore, given that μ and $p_{m,\,b}$ belong to $\mathbb{R}^+$, if a power allocation $p_{m,\,b}$ is used such that $\Gamma_m^{(k)} \le \gamma_{m,p} < \Gamma_m^{(k+1)}$ then

$$\begin{aligned} w_m N_{\mathrm{sc}}\rho_{m,b} - \mu p_{m,b} &= w_m N_{\mathrm{sc}}\varrho_m^{(k)} - \mu p_{m,b} \\ &\le w_m N_{\mathrm{sc}}\varrho_m^{(k)} - \mu\frac{N_{\mathrm{sc}}\sigma_v^2\Gamma_m^{(k)}}{\delta_{m,b}}. \end{aligned} \tag{55}$$

As a consequence, there only exist N_k candidate power allocations

$$p_{m,b}^* \in \left\{\frac{N_{\mathrm{sc}}\sigma_v^2\Gamma_m^{(0)}}{\delta_{m,b}}, \ldots, \frac{N_{\mathrm{sc}}\sigma_v^2\Gamma_m^{(N_{k-1})}}{\delta_{m,b}}\right\} \tag{56}$$

from which the one maximizing $w_m N_{\mathrm{sc}}\varrho_m^{(k)} - \mu\frac{N_{\mathrm{sc}}\sigma_v^2\Gamma_m^{(k)}}{\delta_{m,b}}$ must be selected, that is,

$$p^*_{m,b} = N_{\mathrm{sc}}\sigma_\nu^2 \Gamma_m^{(k^*_{m,b})}/\delta_{m,b}, \tag{57}$$

where

$$k^*_{m,b} = \arg\max_{k\in\mathcal{N}_k}\{N_{\mathrm{sc}} w_m \varrho_m^{(k)} - \mu N_{\mathrm{sc}}\sigma_\nu^2 \Gamma_m^{(k)}/\delta_{m,b}\}. \tag{58}$$

Furthermore, as in the CRA case, given μ and $p^*_{m,b}$, the subband b must be allocated to MS m^*_b satisfying (53).

5.2.2 Optimizing the dual function over μ

Once known the optimal vector $\boldsymbol{p}^*$ for a given μ, the dual optimization problem (49) reduces to

$$g(\mu) = \min_{\mu\geq 0}\left\{\sum_{b=1}^{N_b}\left(w_m N_{\mathrm{sc}}\rho^*_{m^*_b,b} - \mu p^*_{m^*_b,b}\right) + \mu P_T\right\}. \tag{59}$$

Using standard properties of dual optimization problems [9,27], it can be shown that this problem is convex with respect to μ, and thus, derivative-free line search methods like, for example, Golden-section or Fibonacci, can be used to determine μ^*. Once μ^* has been found, it can be used to obtain optimal power, subband and rate allocation for each of the data flows in the system.

Algorithm 1 Resource allocation for UPA/APA with FC

1: $i = 1$; {Initialize iteration counter}
2: $\mathcal{N}_b^{(1)} = \mathcal{N}_b$ {Initialize set of non allocated subbands}
3: $Q_m^{(1)}(t) = Q_m(t)\ \forall m$ {Initialize queue lengths}
4: $P_T^{(1)}(t) = P_T$ {Initialize available power (APA)}
5: **while** $\mathcal{N}_b^{(i)} \neq \emptyset$ and $\sum_{m=1}^{N_m} Q_m^{(i)} \neq 0$ do
6: {Allocate resources using (61)-(63) or (65)-(69)}
7: $\mathcal{N}_b^{(i+1)} = \left\{\mathcal{N}_b^{(i)} \backslash b\right\}$ {Update non allocated subbands}
8: $Q_{m^*_b}^{(i+1)} = Q_{m^*_b}^{(i)} - N_{\mathrm{sc}}\rho_{m^*_b,b}N_o T_o$ {Update queue}
9: $P_T^{(i+1)} = P_T^{(i)} - p_{m^*_b}$ {Update available power (APA)}
10: $i = i + 1$; {Update iteration counter}
11: **end while**

5.3 UPA and APA with FC

When considering the so-called *FC*, the unified optimization problem in (34) can be rewritten as

$$\begin{aligned}&\max_{\boldsymbol{p}\in\boldsymbol{\mathcal{P}}}\sum_{m=1}^{N_m} w_m \min\left\{N_{\mathrm{sc}}\sum_{b=1}^{N_b}\rho_{m,b}, \frac{Q_m}{N_o T_o}\right\}\\ &\text{subject to}\sum_{m=1}^{N_m}\sum_{b=1}^{N_b} p_{m,b}(t) \leq P_T.\end{aligned} \tag{60}$$

This problem belongs to the class of nonlinear integer optimization programs, which have no general global optimal solution. In an attempt to provide a fast and efficient subopti-mal solution to the joint scheduling and resource allocation problem, an iterative searching algorithm providing quasi-optimal solutions is proposed in Algorithm 1. Our approach is based on a modified version of the optimal solutions presented in Sections 5.1 and 5.2. If necessary, the proposed algorithm allocates a subband $b \in \mathcal{N}_b$ per iteration i, assuming that queue length and available transmit power (APA cases only) are updated, in each iteration, by taking into account the data rate and power (APA cases only) allocated to subband b.

5.3.1 UP A with FC

When implementing UPA strategies, if subband b is allocated to user m^*_b in iteration i, then the subband exclusive allocation constraint (i.e., $\boldsymbol{p} \in \boldsymbol{\mathcal{P}}$) forces that

$$p_{m,b} = \begin{cases} P_T/N_b, & m = m^*_b \\ 0, & m \neq m^*_b, \end{cases} \tag{61}$$

for all b. Thus, in iteration i, subband b is allocated to MS m^*_b satisfying

$$m^*_b = \arg\max_{\substack{m\in\mathcal{N}_m \\ b\in\mathcal{N}_b^{(i)}}}\left\{w_m \min\left\{N_{\mathrm{sc}}\rho_{m,b}, \frac{Q_m^{(i)}}{N_o T_o}\right\}\right\}, \tag{62}$$

where $Q_m^{(i)}$ is the *updated* queue length of user m at the beginning of this iteration, with $Q_m^{(1)} = Q_m$ and

$$Q_{m^*_b}^{(i+1)} = Q_{m^*_b}^{(i)} - N_{\mathrm{sc}}\rho_{m^*_b,b}N_o T_o. \tag{63}$$

The per-subband data rate $\rho_{m,\ b}$ is obtained as in either (23), for the DRA case, or (25), for the CRA case.

5.3.2 APA with FC

When implementing APA strategies, assuming that vector $\boldsymbol{p}^{*(i)}$ for a given $\mu^{(i)}$ fulfils the FC, then the corresponding dual optimization problem, for iteration i, reduces to

$$g(\mu^{(i)}) = \min_{\mu^{(i)}\geq 0}\left\{\sum_{b\in\mathcal{N}_b^{(i)}}\left(w_m N_{\mathrm{sc}}\rho^{*(i)}_{m^*_b,b} - \mu^{(i)} p^{*(i)}_{m^*_b,b}\right) + \mu^{(i)} P_T^{(i)}\right\}, \tag{64}$$

which, as previously stated, can be solved by using derivative-free line search methods. Once $\mu^{*(i)}$ has been found, it can be used to obtain power, subband and rate allocation.

In the APA/CRA scheme the optimal power allocation in (50) must be redefined in order to fulfil the FC. Thus,

$$p^{*(i)}_{m,b} = \min\left\{\left[\frac{N_{\mathrm{sc}}w_m}{\mu^{*(i)}T_o\ln 2} - \frac{N_{\mathrm{sc}}\Lambda_m\sigma_\nu^2}{\delta_{m,b}}\right]^+, \pi^{(i)}_{m,b}\right\}, \tag{65}$$

where

$$\pi^{(i)}_{m,b} = \frac{N_{\mathrm{sc}}\Lambda_m\sigma_\nu^2}{\delta_{m,b}}\left(2^{Q_m^{(i)}/N_{\mathrm{sc}}N_o} - 1\right), \tag{66}$$

which has been obtained from (25), is the minimum power required to fulfill the FC in iteration i. Subband b will be allocated to MS m_b^* satisfying

$$m_b^* = \arg \max_{\substack{m \in \mathcal{N}_m \\ b \in \mathcal{N}_b^{(i)}}} \left\{ w_m N_{\text{sc}} \rho_{m,b}^{*(i)} - \mu^{(i)} p_{m,b}^{*(i)} \right\}. \quad (67)$$

In the APA/DRA scheme, the power allocation in iteration i is obtained as

$$p_{m,b}^{*(i)} = N_{\text{sc}} \sigma_v^2 \Gamma_m^{(k_{m,b}^{*(i)})} / \delta_{m,b}, \quad (68)$$

where $k_{m,b}^{*(i)}$ is obtained by redefining (58) as

$$k_{m,b}^{*(i)} = \arg \max_{\substack{k \in \mathcal{N}_k \\ b \in \mathcal{N}_b^{(i)}}} \left\{ w_m \min \left\{ N_{\text{sc}} \varrho_m^{(k)}, \frac{Q_m^{(i)}}{N_o T_o} \right\} - \mu^{*(i)} \frac{N_{\text{sc}} \sigma_v^2 \Gamma_m^{(k)}}{\delta_{m,b}} \right\}. \quad (69)$$

Furthermore, as in the APA/CRA case, given $\mu^{*(i)}$ and $p_{m,b}^{*(i)}$, the subband b must be allocated to MS m_b^* satisfying (67).

6 Numerical results

Based on the unified cross-layer framework previously described, this section is devoted to the performance comparison of different scheduling and resource allocation algorithms in the downlink of a MIMO-OFDMA network. The following performance metrics will be discussed:

- *Average system throughput:* average number of transmitted bits per second by the BS. It is obtained as the average sum-rate achieved by the whole set of users connected to the BS.

- *Average delay:* average amount of time the bits spent in the queue at the BS in addition to transmission time. Notice that delay can be interpreted as an indirect throughput measure. If a given MS is not allocated enough resources, the achievable throughput is below the traffic arrival rate, the corresponding queue gets unstable and delay grows toward infinity. On the other hand, if the MS is overprovisioned, the traffic arrival rate is below the maximum achievable throughput, the queue is stable, and the mean delay remains bounded.

- *Fairness:* Jain's fairness index [44] will be used to calculate fairness among users of the same class of service. With Ω_m denoting the performance metric for user m (i.e., throughput or average delay), then Jain's fairness index is calculated as

$$\text{JFI}_{\mathcal{C}} = \frac{\left(\sum_{m \in \mathcal{C}} \Omega_m\right)^2}{|\mathcal{C}| \sum_{m \in \mathcal{C}} \Omega_m^2}, \quad \Omega_m \geq 0 \ \ \forall m, \quad (70)$$

where $\mathcal{C}$ is the set of users belonging to a given class of service and $|\mathcal{C}|$ denotes the cardinality of this set. The Jain's fairness index is constrained to the set of values $\text{JFI}_{\mathcal{C}} = 1$. If all the users in $\mathcal{C}$ get the same Ω_m, then $\text{JFI}_{\mathcal{C}} = 1$ and maximum fairness is achieved. Lower Jain's fairness index values indicate a higher variance in their achieved QoS, revealing unfairness in scheduling and resource allocation.

- *Service coverage:* percentage of users who achieve their QoS requirements in terms of minimum throughput or maximum allowable average or absolute delay.

6.1 Simulation configuration

Let us consider a single-cell downlink scenario where the BS, transmitting with a power of $P_T = 37$ dBm over a carrier frequency $f_0 = 2$ GHz, is assumed to be located at the center of a circular coverage area with a radius $R = 500$ m. This BS serves a set of N_u MSs that are uniformly distributed over the whole coverage area. Unless otherwise specified, a default 2 × 2 MIMO configuration will be assumed. The entire system bandwidth is $B = 5.6$ MHz, and is divided into $N_b = 64$ orthogonal subbands, each with a bandwidth $B_b = 87.5$ kHz and consisting of $N_{\text{sc}} = 8$ adjacent subcarriers. Transmission between the BS and active MSs is organized in time slots of duration $T_s = 2.0571$ ms, and each of these slots consists of $N_o = 20$ OFDM symbols of duration (without considering the cyclic prefix) $T_o = 91.4286$ μs. Thus, the basic resource allocation unit is formed by 8 adjacent subcarriers and 20 OFDM symbols. We would like to point out that, without loss of generality, most of the chosen parameters are very much aligned with those considered in the Mobile WIMAX standard (see, for instance, [[45] Table 2.3]).

When using DRA strategies, the set of achievable transmission rates in bits/symbol has been fixed to {0, 0.5, 1, 1.5, 2, 3, 4, 4.5}, the coding gap has been set to $\Lambda_m = 3$, and the switching thresholds between transmission modes have been obtained as $\Gamma_m^{(k)} = \Lambda_m (2^{\varrho^{(k)}} - 1)$. In contrast, a coding gap of $\Lambda_m = 1$ (Shannon's capacity limit) has been set when using CRA strategies, whose performance serves as a benchmark against which practical DRA strategies can be measured.

The channel model describing the path-losses, shadowing effects and frequency-, time-and space-selective fading experienced by the transmitted signal on its way from the BS to the MSs, has been implemented by using Stanford University Interim (SUI) channel model 4 [46] with a shadow fading standard deviation of 6 dB. The power delay profile of this model is characterized by $L_p = 3$ Rayleigh distributed paths with power gains $\sigma_{m,0}^2 = 0\,\text{dB}, \sigma_{m,1}^2 = -4\,\text{dB}$ and $\sigma_{m,2}^2 = -8\,\text{dB}$, and corresponding delays $\tau_0 = 0$ μs, $\tau_1 = 1.5$ μs and $\tau_2 = 4$ μs. Moreover, a per subcarrier AWGN power of $\sigma_v^2 = -163.6\,\text{dBW}$ has been assumed at the receiver front-end.

To demonstrate the ability of our proposed unified framework to schedule and allocate resources to service flows with different QoS requirements, three traffic classes are considered, i.e., real time (RT), non real time (nRT) and BE. As in [22], traffic arrivals have been modeled as Poisson random variables, with a mean that depends on the average arrival rate per flow (measured in bits/s). Without loss of generality, the maximum tolerable delays ($\check{D}_m$) for each traffic class have been set to 100 ms (RT), 2 s (nRT) and 20 s (BE), and the outage delay probabilities (ξ_m) to 0.01 (RT), 0.01 (nRT) and 0.1 (BE).

All the numerical results presented in this article have been obtained by averaging the outcomes of a dynamic discrete event simulation performed over 60 scenarios, each with a particular random distribution of MSs over the coverage area, and transmitting 15,000 slots per scenario. To guarantee that the presented results correspond to the steady-state of the system, initial transitory periods of 1,000 slots per scenario, which are not accounted for in the performance evaluation process, have been used.

6.2 Comparing scheduling rules

The performance metrics versus traffic load for MSR, PF, EXP and MLWDF scheduling rules are compared in Figure 2 for an adaptive MIMO-OFDMA system serving N_u = 20 RT users with the same average arrival rate. The use of uniform power and CRA strategies over a 2 × 2 MIMO system has been considered. Furthermore, since the FC can be implemented with all scheduling rules, performance results have been obtained for each scheduling rule either with or without FC.

Without FC, the EXP and MLWDF rules provide the best joint results in terms of throughput, delay, Jain's fairness indexes, and service coverage, with MLWDF achieving a slightly higher throughput, lower delay and better service coverage that EXP, at the cost of lower throughput and delay fairness indexes.[c] The PF scheduler, although achieves a quite good result in terms of average throughput per flow, fails in providing QoS requirements. In fact, the PF rule can only guarantee a 99% service coverage for average arrival rates per flow less than 0.3 Mbps compared to the 0.8 and 1 Mbps that can be guaranteed by EXP and MLWDF rules, respectively. The MSR scheduling rule, which only considers CSI as a quality indicator, allocates all the resources to the users with favorable channel quality conditions, and those users experiencing bad channel quality conditions suffer from starvation. Hence, as it wastes resources, MSR rule is not capable of achieving queue stability and presents a very low average throughput and an *infinite*[d] average delay per flow, irrespective of the average traffic arrival rate.

Except for a slight increase in delay Jain's fairness index, which is only perceptible for light or moderate traffic loads, the effect of implementing FC on the performance of EXP and MLDF scheduling rules is very small. This can be explained by the fact that, when calculating the weighting coefficients $w_m(t)$, the EXP and MLDF schedulers use QSI and thus, the performance gains provided by the introduction of the FC are just incremental. On the contrary, the performance improvement induced by the implementation of FC is considerable for the PF rule, and specially important for the MSR scheduler, which do not use QSI when calculating $w_m(t)$. In fact, even though the PF and MSR rules provide poorer Jain's fairness indexes than those delivered by the EXP and MLWDF rules, they can guarantee a 99% service coverage for average arrival rates per flow less than approximately 0.8 Mbps, which is almost the same that can be guaranteed when using the EXP scheduler.

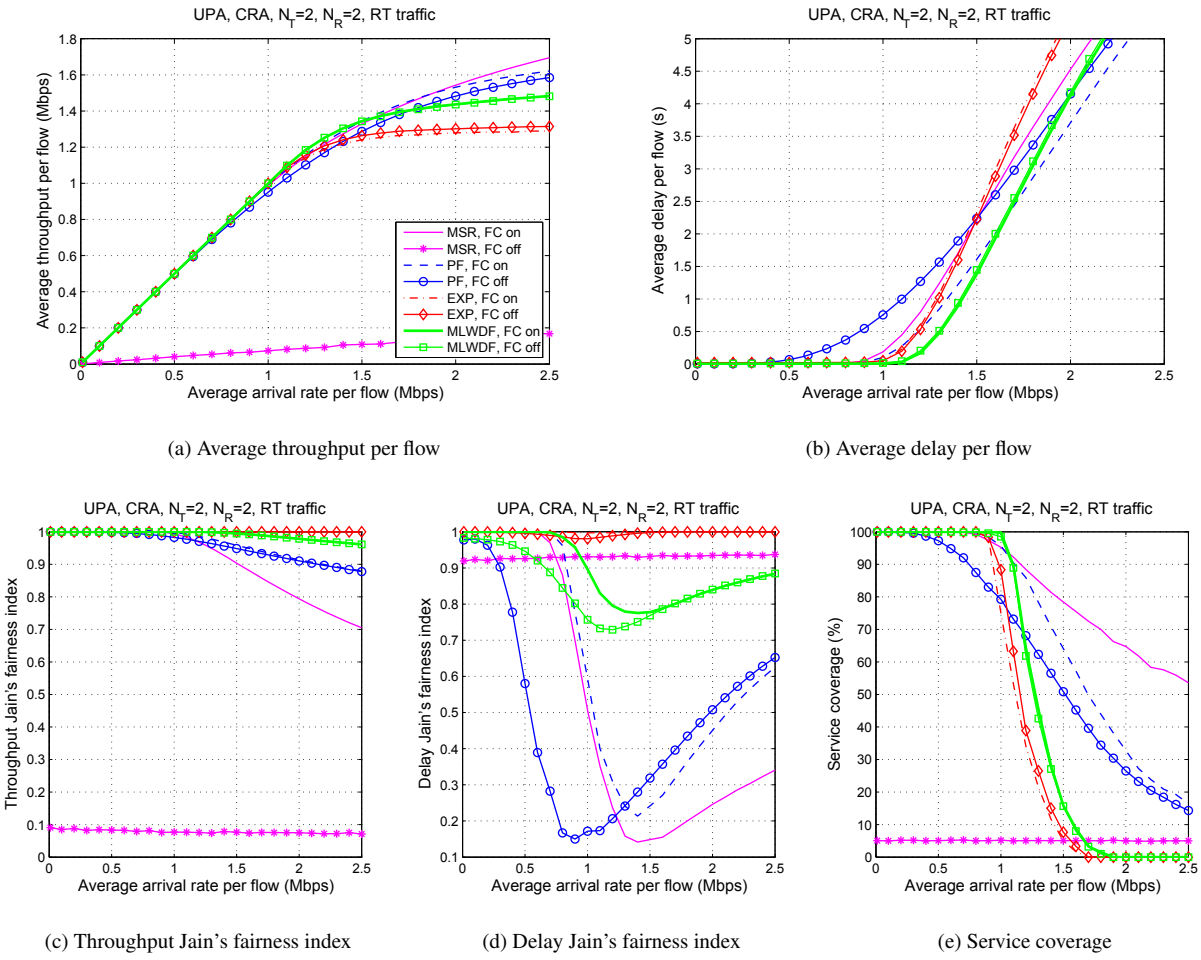
6.3 Comparing allocation strategies

Figure 3 shows the performance metrics versus traffic load for a MIMO-OFDMA system using MLWDF scheduling rule and different combinations of UPA, APA, DRA and CRA strategies, with and without FC. A set of N_u = 20 RT users with the same average arrival rate has been assumed. As it can be observed, APA-based strategies improve the performance of UPA-based ones. Nevertheless, this performance improvement, although noticeable for discrete rate-based systems, becomes almost negligible when using continuous rate-based schemes. This result suggests that using AMC schemes with a large set of modulation formats combined with powerful channel codes with adaptive coding rates can make unnecessary the use of power allocation strategies.

The effect of implementing FC on the system performance metrics is practically identical irrespective of the power and rate allocation strategies implemented at the cross-layer resource allocation unit. The average throughput per flow, delay, throughput JFI and service coverage are basically unaffected, and only an improvement in delay JFI is obtained with light and moderate traffic loads. Although not shown in the graphs, when implementing APA strategies, the use of FC also introduces a decrease in power consumption. This is due to the fact that resources (power and subbands) are only allocated when necessary, that is, when there is enough information in the queues ready to be transmitted.

6.4 Comparing MIMO configurations

The effects of using different $N_T \times N_R$ MRT/MRC MIMO configurations on the average throughput per flow and service coverage are depicted in Figure 4. Results have been obtained for an adaptive MIMO-OFDMA system using the MLWDF scheduler, uniform



(a) Average throughput per flow

(b) Average delay per flow

(c) Throughput Jain's fairness index

(d) Delay Jain's fairness index

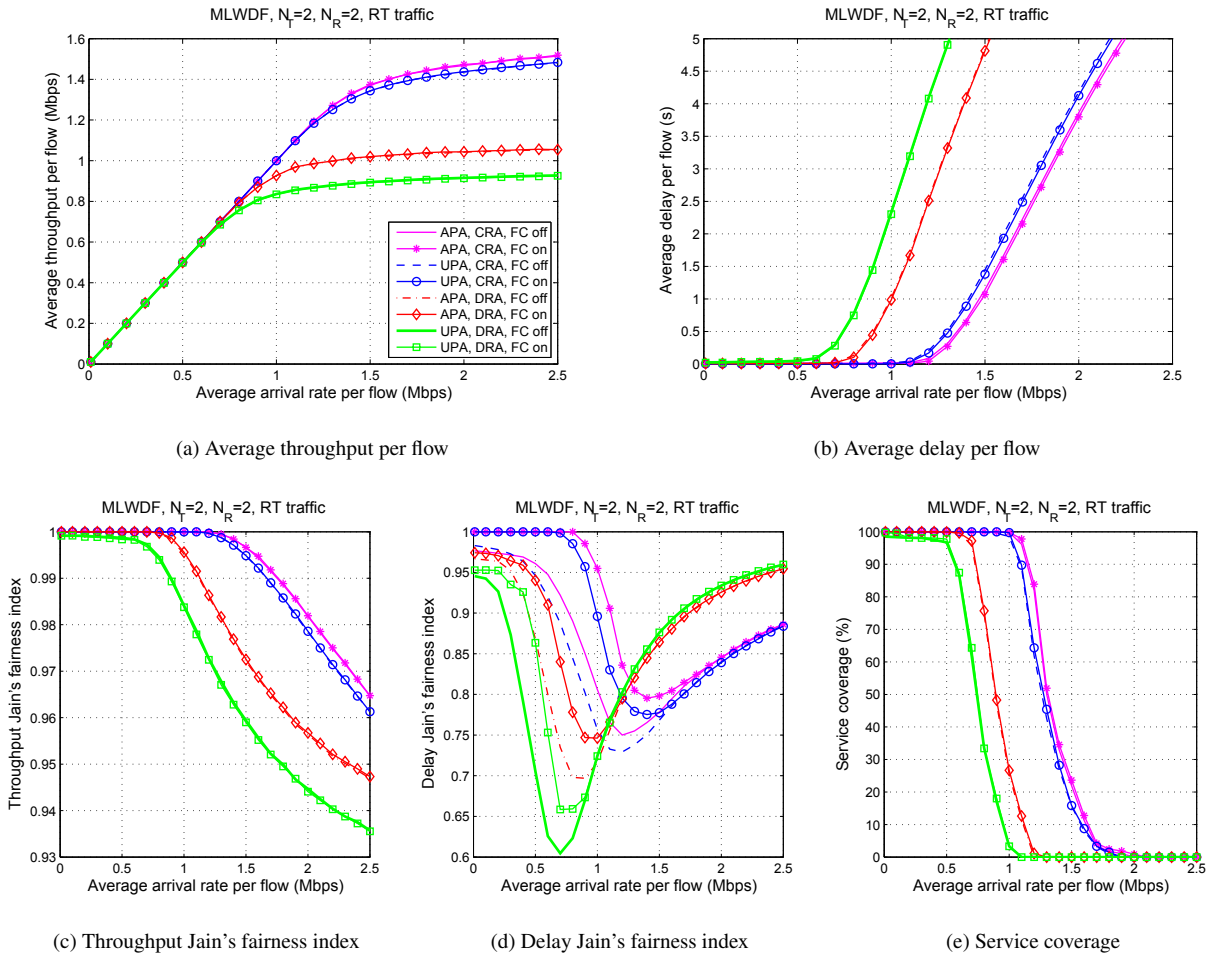
(e) Service coverage

Figure 2 Comparing schedulers.

power and CRA strategies, without FC, to serve N_u = 20 RT users with the same average arrival rate. As it can be observed, increasing the number of transmit and/or receive antennas at the PHY can significantly improve the system capacity. In fact, the increase of N_T and/or N_R translates into a widening of the stability region, which proves the convenience of employing MIMO spatial diversity at the PHY to support statistical QoS for upper layer protocols. For instance, Figure 4b shows that, using this particular configuration, a single transmit/receive antenna system can only guarantee a 99% service coverage for average arrival rates per flow less than 0.45 Mbps, compared to the 0.95 or 1.5 Mbps that can be guaranteed by using 2 × 2 or 4 × 4 MIMO configurations, respectively.

6.5 Performance results for heterogeneous traffic scenarios

Figure 5 shows the performance metrics versus traffic load for a MIMO-OFDMA system serving a set of heterogeneous traffic flows. The behavior of three scheduling rules, namely, MLWDF, EXP and MDU, are compared for a system implementing UPA and CRA, without FC. Simulations have been performed assuming that $N_m^{(RT)} = N_m^{(nRT)} = N_m^{(BE)} = 10$ users are always active in the system. Furthermore, based on the required QoS of the different traffic flows, the parameters of the MDU scheduler have been set to ϕ_m = {1, 1.5} and $\tilde{W}_m$ = 25 ms ms for RT users, ϕ_m = {0.6, 1} and $\tilde{W}_m$ = 500 ms for nRT users and, finally, ϕ_m = {0.5, 0} and $\tilde{W}_m$ = 500 ms for BE users. As it can be observed, cross-layer scheduling and resource allocation strategies are able to fairly allocate resources among traffic classes, according to the assigned priorities $\chi_m(t)$, obtained from the QoS requirements. Obviously, RT users, which exhibit stringent absolute delay requirements, tend to be allocated more resources than nRT and BE users as the arrival data rates increase. For the same reasons, nRT users are allocated more resources than BE users. The result is that, although for light traffic arrivals the three



(a) Average throughput per flow

(b) Average delay per flow

(c) Throughput Jain's fairness index

(d) Delay Jain's fairness index

(e) Service coverage

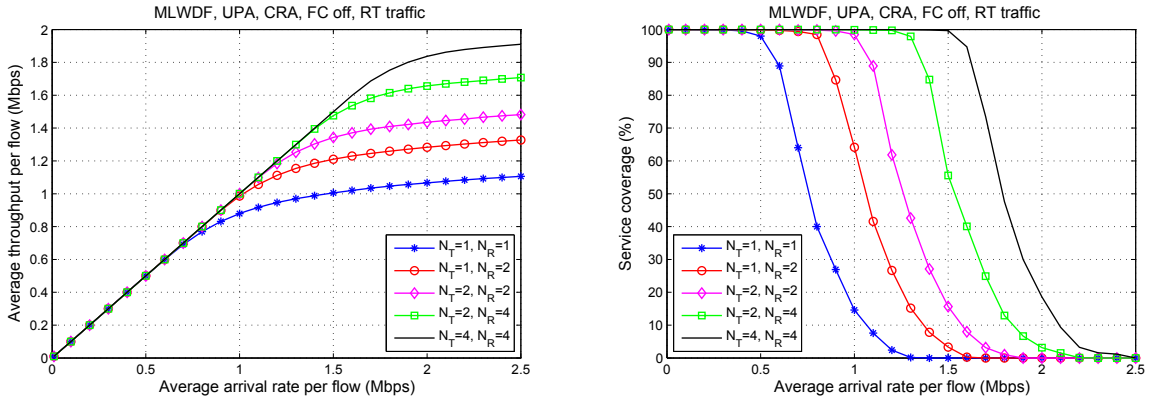
Figure 3 Comparing allocation strategies.

classes of service can achieve good performance figures, for moderate traffic arrivals, RT and nRT users can only maintain acceptable performance at the cost of a decrease in the performance of BE users. Furthermore, for heavy traffic arrivals, the performance of RT users can only be maintained by sacrificing that of nRT and BE users.

In this particular scenario, except for very heavy traffic arrivals, the MLWDF scheduling rule provides the best performance results in terms of average throughput per flow and service coverage at the cost of a worse behavior of the delay Jain's fairness index. The MDU scheduler provides the best results in terms of both throughput and delay Jain's fairness indexes for RT and nRT traffic classes, but such a fair behavior is obtained at the cost of service coverage. The EXP scheduler sacrifices the average throughput and delay per flow of BE users to obtain a good trade-off between service coverage and delay Jain's fairness index.

7 Conclusions

The emergence of state-of-the-art and next-generation wireless communications networks based on adaptive MIMO-OFDMA PHY access schemes, will enable the support of a wide range of multimedia applications with heterogeneous QoS requirements. In order to optimize the resource utilization while maintaining the QoS provided to as many users as possible, these systems require of adaptive scheduling and resource allocation algorithms able to grant a proper trade off between efficiency and fairness. In this context, using tools from information and queueing theories, mathematical convex programming, and stochastic approximation, a unified framework for channel- and queue-aware QoS-guaranteed cross-layer scheduling and resource allocation algorithms has been developed in this article. The proposed unified framework generalizes previous work on this topic by encompassing different types of traffic, different utility functions measuring user's satisfaction, uniform



(a) Average throughput per flow

(b) Service coverage

Figure 4 Comparing MRT/MRC MIMO configurations.

and adaptive power allocation, continuous and DRA, and protocols with different amounts of channel- and queue-awareness. System parameters and QoS requirements have been projected into utility functions, which have then been used to formulate a unified constrained utility maximization problem, whose main aim is to balance the efficiency and fairness of resource allocation. Optimal solutions for this problem have been obtained for the UPA schemes, and novel quasi-optimal algorithms have been proposed for the APA strategies, exhibiting complexities that are linear in the number of resource units and users.

The proposed unified optimization framework allows for a fair performance comparison of different scheduling rules, different allocation strategies, different MRT/MRC-based MIMO configurations, and different traffic scenarios. Simulation results presented in this article have shown that:

- Without FC, the EXP and MLWDF rules provide the best joint performance results, with MLWDF achieving a slightly higher throughput, lower delay and better service coverage that EXP, at the cost of lower throughput and delay fairness indexes. The PF and MSR scheduling rules, which only consider CSI as a quality indicator, fail in providing QoS. However, although implementing FC has a negligible impact on the performance of EXP and MLDF scheduling rules, the performance improvement induced by FC is remarkable for the PF and MSR schedulers.

- APA-based strategies improve the performance of UPA-based ones. Nevertheless, this performance improvement, although noticeable for DRA systems, becomes almost negligible when using CRA schemes. Thus, using AMC schemes with a large set of modulation and coding formats can make unnecessary the use of power allocation strategies.

- Increasing the number of transmit and/or receive antennas at the PHY translates into a widening of the stability region, proving in this way the convenience of employing MIMO spatial diversity to support statistical QoS provision to upper layer protocols.

- Channel- and queue-aware cross-layer scheduling and resource allocation strategies can fairly allocate resources among heterogenous traffic classes, with different scheduling policies (e.g., EXP, MDU and MLWDF) providing different trade-offs between efficiency, delay, fairness and service coverage.

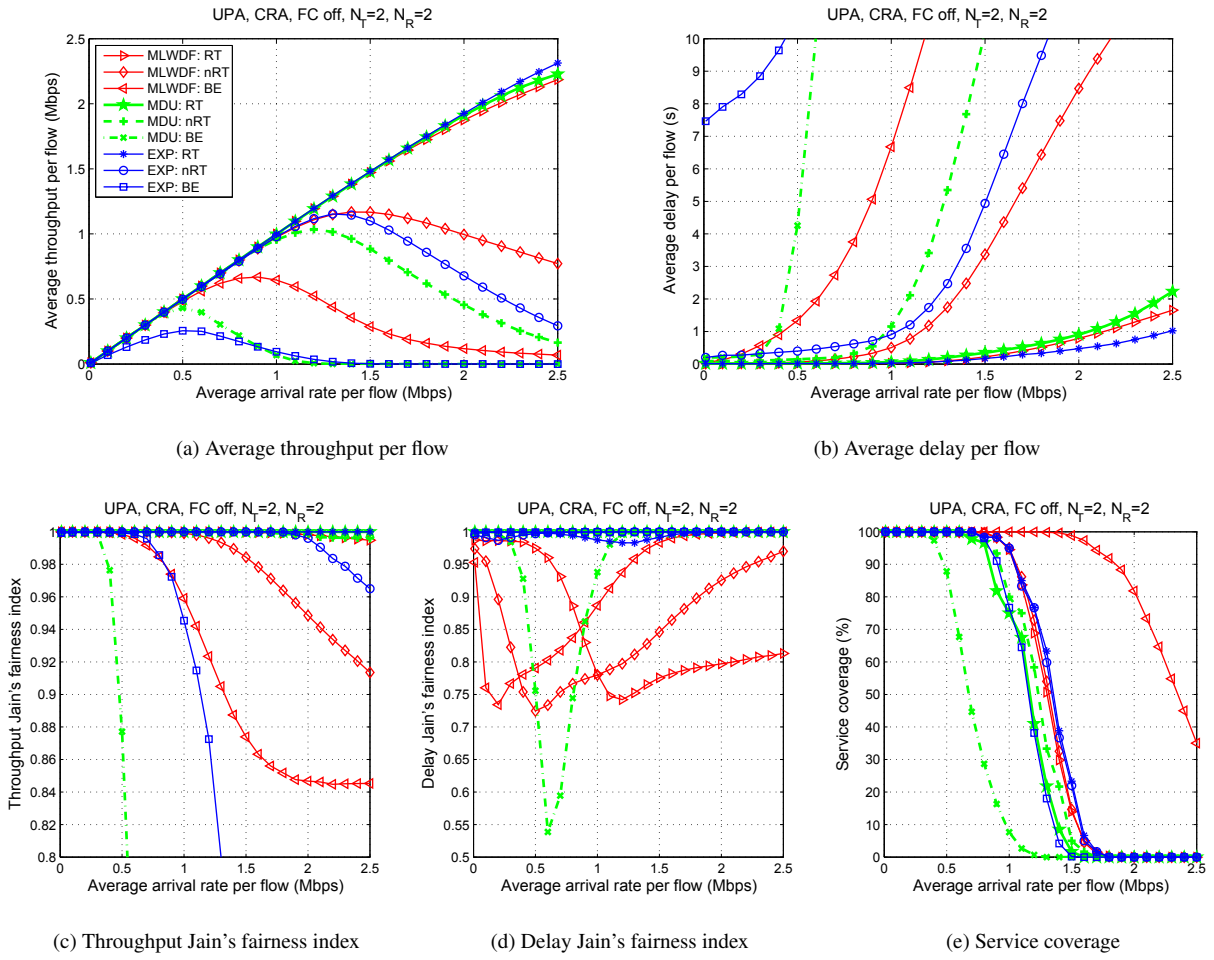
Simulation results have demonstrated the validity and merits of the proposed cross-layer unified approach. However, the optimization problem treated in this article is only applicable to single cell scenarios using MRT/MRC-based MIMO techniques. Therefore, to widen its application scope, current work focusses on extending the cross-layer unified approach to distributed scheduling and resource allocation in generalized MIMO-OFDMA multicellular wireless heterogeneous networks, possibly including more sophisticated MIMO techniques, one- and two-way relays, shared relays, femto-cells and/or clusters of coordinated BSs.

Endnotes

[a]LTE was introduced in 3GPP Releases 8 and 9 as a major step forward for UMTS-based networks, and LTE-Advanced is the fourth generation (4G) LTE standard in 3GPP Release 10.

[b]A scheduling algorithm is said to be throughput optimal if it can keep all the queues stable if this is at all feasible to do.

[c]Typically, the delay Jain's fairness index is high for light traffic arrival rates because, in this case, all the flows can be served after very low average waiting times in the queues. For moderate traffic arrival rates, the



(a) Average throughput per flow

(b) Average delay per flow

(c) Throughput Jain's fairness index

(d) Delay Jain's fairness index

(e) Service coverage

Figure 5 Performance metrics for heterogeneous traffic.

variance of per flow waiting times in the queues increases and, consequently, the delay Jain's fairness index decreases. Heavy traffic arrivals tend to cause queue instability, with almost all the flows experiencing large average delays, thus producing again an increase of the delay fairness index.

[d]The average delay per flow would be infinite if simulations were performed over an infinite period of time.

Acknowledgements

This was supported in part by the MEC and FEDER under projects COSMOS (TEC2008-02422) and AM3DIO (TEC2011-25446), and in part by the Govern de les Illes Balears through a Ph.D. scholarship.

Competing interests

The authors declare that they have no competing interests.

Received: 23 September 2011 Accepted: 17 April 2012
Published: 17 April 2012

References

1. I Fu, Y Chen, P Cheng, Y Yuk, Y Kim, J Kwak, Multicarrier technology for 4G WiMax system. IEEE Commun Mag. **48**(8), 50–58 (2010)
2. A Ghosh, R Ratasuk, B Mondal, N Mangalvedhe, T Thomas, LTE-advanced: next-generation wireless broadband technology. IEEE Wirel Commun. **17**(3), 10–22 (2010)
3. M Alasti, B Neekzad, J Hui, R Vannithamby, Quality of service in WiMAX and LTE networks. IEEE Commun Mag. **48**(5), 104–111 (2010)
4. CY Wong, RS Cheng, KB Letaief, RD Murch, Multiuser OFDM with adaptive subcarrier, bit, and power allocation. IEEE J Sel Areas Commun. **17**(10), 1747–1758 (1999). doi:10.1109/49.793310
5. BS Krongold, K Ramchandran, DL Jones, Computationally efficient optimal power allocation algorithms for multicarrier communication systems. IEEE Trans Commun. **48**(1), 23–27 (2000). doi:10.1109/26.818869
6. D Kivanc, G Li, H Liu, Computationally efficient bandwidth allocation and power control for OFDMA. IEEE Trans Wirel Commun. **2**(6), 1150–1158 (2003). doi:10.1109/TWC.2003.819016
7. J Jang, KB Lee, Transmit power adaptation for multiuser OFDM systems. IEEE J Sel Areas Commun. **21**(2), 171–178 (2003). doi:10.1109/JSAC.2002.807348
8. LMC Hoo, B Halder, J Tellado, JM Cioffi, Multiuser transmit optimization for multicarrier broadcast channels: asymptotic FDMA capacity region and algorithms. IEEE Trans Commun. **52**(6), 922–930 (2004). doi:10.1109/TCOMM.2004.829570
9. IC Wong, B Evans, *Resource allocation in multiuser multicarrier wireless systems*, (Springer New York, 2008)
10. M Ergen, S Coleri, P Varaiya, QoS aware adaptive resource allocation techniques for fair scheduling in OFDMA based broadband wireless access systems. IEEE Trans Broadcast. **49**(4), 362–370 (2003). doi:10.1109/TBC.2003.819051

11. IC Wong, Z Shen, B Evans, J Andrews, A low complexity algorithm for proportional resource allocation in OFDMA systems, in *Proc IEEE Workshop on Signal Processing Systems*, Austin Texas USA, pp. 1–6 (2004)
12. H Kim, Y Han, A proportional fair scheduling for multicarrier transmission systems. IEEE Commun Lett. **9**(3), 210–212 (2005). doi:10.1109/LCOMM.2005.03014
13. Z Shen, JG Andrews, BL Evans, Adaptive resource allocation in multiuser OFDM systems with proportional rate constraints. IEEE Trans Wirel Commun. **4**(6), 2726–2737 (2005)
14. Z Han, Z Ji, K Liu, Fair multiuser channel allocation for OFDMA networks using Nash bargaining solutions and coalitions. IEEE Trans Commun. **53**(8), 1366–1376 (2005). doi:10.1109/TCOMM.2005.852826
15. Y Ma, Rate maximization for downlink OFDMA with proportional fairness. IEEE Trans Veh Technol. **57**(5), 3267–3274 (2008)
16. EB Rodrigues, F Casadevall, Control of the trade-off between resource efficiency and user fairness in wireless networks using utility-based adaptive resource allocation. IEEE Commun Mag. **49**(9), 90–98 (2011)
17. G Song, Y Li, Cross-layer optimization for OFDM wireless networks-part I: theoretical framework. IEEE Trans Wirel Commun. **4**(2), 614–624 (2005)
18. G Song, Y Li, Cross-layer optimization for OFDM wireless networks-part II: theoretical framework. IEEE Trans Wirel Commun. **4**(2), 625–634 (2005)
19. D Hui, V Lau, W Lam, Cross-layer design for OFDMA wireless systems with heterogeneous delay requirements. IEEE Trans Wirel Commun. **6**(8), 2872–2880 (2007)
20. C Mohanram, S Bhashyam, Joint subcarrier and power allocation in channel-aware queue-aware scheduling for multiuser OFDM. IEEE Trans Wirel Commun. **6**(9), 3208–3213 (2007)
21. Z Kong, Y-K Kwok, J Wang, A low-complexity QoS-aware proportional fair multicarrier scheduling algorithm for OFDM systems. IEEE Trans Veh Technol. **58**(5), 2225–2235 (2009)
22. G Song, Y Li, L Cimini, Joint channel- and queue-aware scheduling for multiuser diversity in wireless OFDMA networks. IEEE Trans Commun. **57**(7), 2109–2121 (2009)
23. N Zhou, X Zhu, Y Huang, H Lin, Low complexity cross-layer design with packet dependent scheduling for heterogeneous traffic in multiuser OFDM systems. IEEE Trans Wirel Commun. **9**(6), 1912–1923 (2010)
24. X Wang, G Giannakis, A Marques, A unified approach to QoS-guaranteed scheduling for channel-adaptive wireless networks. Proc IEEE. **95**(12), 2410–2431 (2007)
25. B Dañobeitia, G Femenias, F Riera-Palou, Resource allocation in MIMO-OFDMA wireless systems based on linearly precoded orthogonal space-time block codes, in *Proceedings of EUNICE*, Barcelona, Spain, pp. 127–134 (2009)
26. B Dañobeitia, G Femenias, Dual methods for channel- and QoS-aware resource allocation in MIMO-OFDMA networks, in *IFIP Wireless Days*, Venice, Italy, (2010)
27. W Yu, R Lui, Dual methods for nonconvex spectrum optimization of multicarrier systems. IEEE Trans Commun. **54**(7), 1310–1322 (2006)
28. T Lo, Maximum ratio transmission. IEEE Trans Commun. **47**(10), 1458–1461 (1999). doi:10.1109/26.795811
29. JG Proakis, *Digital Communications*, 4th edn. (McGraw Hill, New York, 2001)
30. L Kleinrock, in *Queueing Systems*, vol. I. (Wiley, New York, 1975). *Theory*
31. AJ Goldsmith, *Wireless Communications*, (Cambridge University Press, Cambridge, 2005)
32. A Stolyar, On the asymptotic optimality of the gradient scheduling algorithm for multiuser throughput allocation. J Operat Res Soc. **53**, 12–25 (2005). doi:10.1287/opre.1040.0156
33. R Knopp, P Humblet, Information capacity and power control in single-cell multiuser communications, in *IEEE International Conference on Communications (ICC)*, vol. 1. Seattle, Washington, USA, pp. 331–335 (1995)
34. F Kelly, A Maulloo, D Tan, Rate control for communication networks: shadow prices, proportional fairness and stability. J Operat Res Soc. **49**(3), 237–252 (1998)
35. S Shakkottai, A Stolyar, *Scheduling algorithms for a mixture of real-time and non-real-time data in HDR*, (Bell Laboratories, Lucent Technologies, Murray Hill, 2000)
36. M Andrews, S Borst, F Dominique, P Jelenkovic, K Kumaran, K Ramakrishnan, P Whiting, Dynamic bandwidth allocation algorithms for high-speed data wireless networks. Bell Labs Tech J. **3**(3), 30–49 (1998)
37. M Andrews, K Kumaran, K Ramanan, A Stolyar, P Whiting, R Vijayakumar, Providing quality of service over a shared wireless link. IEEE Commun Mag. **39**(2), 150–154 (2001). doi:10.1109/35.900644
38. S Shakkottai, A Stolyar, Scheduling for multiple flows sharing a time varying channel: the exponential rule, (Bell Labs Technical Report, 2000)
39. G Song, Cross-layer resource allocation and scheduling in wireless multicarrier networks, (School of Electrical and Computer Engineering, Georgia Institute of Technology, 2005) PhD Thesis
40. K Sundaresan, X Wang, M Madihian, Scheduler design for heterogeneous traffic in cellular networks with multiple channels, in *Proceedings of the Third International Conference on Wireless Internet (WICON)*, Austin, Texas, USA, p. 12. 1-12:10 (2007)
41. B Al-Manthari, H Hassanein, N Ali, N Nasser, Fair class-based downlink scheduling with revenue considerations in next generation broadband wireless access systems. IEEE Trans Mobile Comput. **8**, 721–734 (2009)
42. M Mehrjoo, MK Awad, M Dianati, XS Shen, Design of fair weights for heterogeneous traffic scheduling in multichannel wireless networks. IEEE Trans Commun. **58**(10), 2892–2902 (2010)
43. S Boyd, L Vandenberghe, *Convex Optimization*, (Cambridge University Press, Cambridge, 2004)
44. R Jain, DM Chiu, W Hawe, A quantitative measure of fairness and discrimination for resource allocation in shared systems, (DEC Research Report, 1984) TR-301
45. JG Andrews, A Gosh, R Muhamed, *Fundamentals of WiMAX. Undestanding Broadband Wireless Networking*, (Prentice-Hall, Upper Saddle River, 2007)
46. V Erceg, K Hari, M Smith, D Baum, K Sheikh, C Tappenden, J Costa, C Bushue, A Sarajedini, R Schwartz, D Branlund, T Kaitz, D Trinkwon, Channel models for fixed wireless applications. IEEE 802.16 Broadband Wireless Access Working Group, Tech Rep (2001)

doi:10.1186/1687-1499-2012-145
Cite this article as: Femenias *et al.*: **Unified approach to cross-layer scheduling and resource allocation in OFDMA wireless networks.** *EURASIP Journal on Wireless Communications and Networking* 2012 **2012**:145.