A cross-layer scheduling strategy for the downlink of a MIMO-OFDMA system with heterogeneous traffic

Velio Tralli *, Ana Perez-Neira #,$, Pol Henarejos #

Universitat Politecnica de Catalunya (UPC)- Barcelona - Spain
$ *Centre Tecnologic de Telecomunicacions de Catalunya (CTTC)- Barcelona - Spain*
* *ENDIF, University of Ferrara - CNIT -Italy*

***Abstract*—In this paper we propose and investigate a cross-layer multiuser scheduling strategy for the support of heterogeneous traffic in the downlink of a MIMO-OFDMA system. It jointly considers different objectives: maximize the sum-rate on the radio channel, ensure a fair allocation of resources among users belonging to the same traffic class, consider the dynamics of traffic sources by looking at the delay of data packets in the queues, contribute to maximize quality of service figures at the application level. To exploit temporal diversity and to reduce complexity, the ergodic weighted sum-rate is maximized and dual optimization with stochastic approximation is applied to derive on-line algorithms. The numerical results show the capability of the scheduler to allocate physical layer resources according to rate constraints imposed for each different traffic class and with fairness inside each class, even in presence of different channels conditions and different network loads.**

I. Introduction

In the development of fourth generation wireless networks, e.g LTE or WiMax, OFDM and MIMO technologies will be heavily exploited to support the transmission of multiple users. In fact, Multi-User MIMO (MU-MIMO) spatial multiplexing schemes are a promising way to increase system throughput and there is a growing interest on the topic as [1,2,3] shows. Recently, attention has been paid to the combination of spatial diversity multiple access systems and frequency domain packet scheduling [4,5,6,7,8]. Specifically, in [6] the authors present a low complexity sum-power constraint iterative waterfilling that is capacity achieving and probably convergent. In [8] the authors address the problem of feedback reduction. On the other hand, future wireless networks need to provide connectivity to heterogeneous users offering different data traffic types, e.g. voice, video, web browsing, etc. This poses several constraints and additional challenges which can be faced within the framework of the cross-layer design [9,10].

In this paper we propose and investigate a cross-layer multiuser scheduling strategy for MIMO-OFDMA systems which jointly considers different objectives: maximize the sum-rate on the radio channel, ensure a fair allocation of resources among users belonging to the same traffic class, consider the dynamics of traffic sources by looking at the delay of data packets in the queues, contribute to maximize quality of service figures at the application level. In the proposed solution we look for a low computation complexity and a reduced feedback. In contrast to [7], in order to further reduce complexity for on line implementation we follow a dual decomposition strategy and a stochastic approximation. In order to reduce feedback load the paper resorts to opportunistic strategies that solve the spatial scheduling. In summary, this paper proposes a joint spatial and frequency scheduler that allows on-line implementation and only requires partial feedback and a low-complexity implementation.

The paper is organized as follows. After having introduced the radio interface model in Sect. II, the radio resource allocation problem is formulated in Sect. III. The dual optimization framework that allows on-line implementation is illustrated in Sect. IV and the allocation algorithm follows in Sect. V. The architecture of the system applying the proposed solution is described in Sect. VI and, finally, results are presented in Sect. VII.

II. Radio interface model

We consider an OFDMA scenario with M subcarriers and K users. Each user k is single antenna and receives simultaneously up to N_T signals, which can come from different spatial locations, antennas or beams. Only one of the N_T signals is intended for user k. Each signal is characterized by a spatial signature, i.e. a beamforming vector $\mathbf{b}_{m,q}$, $m = 1, \ldots, M$ $q = 1, \ldots, N_T$. Subcarriers and spatial signatures are shared by all the users at each time slot. A binary allocation variable $\{a_{k,m,q}\}$ indicates whether a space-frequency resource is used by user k, i.e. $a_{k,m,q} = 1$ if user k is scheduled on frequency m and beam q, $a_{k,m,q} = 0$ otherwise, with the constraint $\sum_{k=1}^{K} a_{k,m,q} \leq 1$. The composite signal received by user k on the subcarrier m is therefore given by

$$y_{k,m} = \mathbf{h}_{k,m}^T \sum_{s=1}^{K} \sum_{q=1}^{N_T} a_{s,m,q} \sqrt{p_{m,q}} \mathbf{b}_{m,q} x_{s,m,q} + w_{k,m} \qquad (1)$$

where $\mathbf{h}_{k,m}$ is the N_T-dimensional vector of channel gains, $x_{s,m,q}$ is the transmitted complex symbol of user s, $p_{m,q} \geq 0$ is the power of the q-th transmitted signal and $w_{k,m}$ is the additive white Gaussian noise with variance σ_n^2. From the viewpoint of information theory, the problem could correspond to a broadcast channel.

In spite of the big gains in spectral efficiency that can be obtained by incorporating multiantenna transmission to a multicarrier system, an evident drawback of this scenario is the increased design complexity. In order to keep feedback and computation complexity low in the optimization of the PHY layer parameters we consider in this paper opportunistic beamforming (OB) technique [11]. However, the scheduling strategy considered here can be also extended to a more general framework of spatial precoding. According to OB, the transmitter generates orthogonal spatial signatures randomly for each subcarrier. Based on partial CSI feedback, the scheduler and resource allocator only handle the set of binary allocation variables $\mathbf{a} = \{a_{k,m,q}\}$ and the set of powers $\mathbf{p} = \{p_{m,q}\}$. The CSI is the set of equivalent channel power gains $c_{k,m,q}$ seen by each user k at frequency m with respect to the q-th beam, which are given by $c_{k,m,q} = |\mathbf{h}_{k,m}^T \mathbf{b}_{m,q}|^2 / \sigma_n^2$.

III. Radio resource allocation problem

The aim of resource allocation is to dynamically assign radio interface resources $\mathbf{a}$ and $\mathbf{p}$ to the different users in order to achieve the best tradeoff among different objectives:

- to maximize the sum-rate on the radio channel
- to ensure a fair allocation of resources among users belonging to the same traffic class and/or guarantee a minimum amount of resources to some users or classes

- to consider the dynamics of traffic sources and control the delay of data packets in the queue according to requirements of specific traffic classes.
- to contribute at the maximization of quality of service figures at the application level

Motivated by the need of flexibility in considering the outlined objectives, we formulate the resource allocation problem at a given time instant as a problem of ergodic weighted sum rate maximization with constraints on the average value of the rate provided to users as follows.

$$\max_{\mathbf{a},\mathbf{p}} \sum_{k=1}^{K} \mathbb{E}\left\{w_k R_k(\mathbf{a},\mathbf{p})\right\} \tag{2}$$

$$\begin{aligned} s.t. \qquad & \mathbb{E}\left\{\sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{q=1}^{N_T} a_{k,m,q}p_{m,q}\right\} \le P \\ & \mathbb{E}\left\{R_k(\mathbf{a},\mathbf{p})\right\} = R_{0k}, \quad k \le K_0 \\ & \mathbb{E}\left\{R_k(\mathbf{a},\mathbf{p})\right\} \ge \phi_k \sum_{s=K_0+1}^{K} \mathbb{E}\left\{R_s(\mathbf{a},\mathbf{p})\right\}, \quad k > K_0 \end{aligned} \tag{3}$$

where

$$R_k(\mathbf{a},\mathbf{p}) = \sum_{m=1}^{M}\sum_{q=1}^{N_T} \log_2(1+\gamma_{k,m,q}(\mathbf{a},\mathbf{p})) \tag{4}$$

is the rate provided to user k and

$$\gamma_{k,m,q}(\mathbf{a},\mathbf{p}) = \frac{a_{k,m,q}p_{m,q}c_{k,m,q}}{1+\sum_{s=1,s\neq q}^{N_T}\sum_{r\neq k} a_{r,m,s}p_{m,s}c_{k,m,s}} \tag{5}$$

is the SINR of user k at frequency m and beam q. This approach extends the techniques outlined in [13] for OFDM with single antenna.

The coefficients w_k are the weights that allow prioritizing the users according to the service class and status of the queue buffers; they are not fixed constants, but are randomly changing parameters. The parameters R_{0k}, $k = 1,\ldots,K_0$ and ϕ_k, $k = K_0+1,\ldots,K$ are fixed parameters that define a constraint on the rate provided to users with

$$\sum_{k=K_0+1}^{K} \phi_k = 1 \tag{6}$$

Without loosing in generality, the first K_0 users have a fixed average rate constraint R_{0k}, whereas the remaining users have a proportional rate constraint depending on coefficients ϕ_k, which may be different for different traffic classes. We assume that $\gamma_{k,m,q}(\mathbf{a},\mathbf{p})$ are known by the N_T transmitters by means of partial channel feedback. For instance, this would be the case of a broadcast channel where the Base Station has perfect SINR feedback.

It is important to underline that rate and sum power constraints are referred in this problem to average values. The reason is twofold: i) these constraints relax the instantaneous constraints leading to a reduction in the complexity of the resulting optimization algorithm and ii) it incorporates the time dimension in the resulting resource allocation by using the ergodicity assumption. For systems with hard instantaneous power constraint, the solution of the problem needs to be suitably adapted by using power rescaling, as shown later.

IV. Dual optimization and adaptive implementation

The proposed algorithm is based on a dual optimization framework, based on a Lagrangian relaxation of power and rate constraints. This relaxation retains the subcarrier assignment exclusivity constraints, but dualizes the power/rate constraints incorporating them into the objective function, thereby allowing us to solve the dual problem instead. This dual optimization is much less complex.

To derive the dual problem we first write the Lagrangian:

$$\begin{aligned} L(\mathbf{a},\mathbf{p},&\lambda,\boldsymbol{\mu},\boldsymbol{\nu}) = \sum_{k=1}^{K}\mathbb{E}\left\{w_k R_k(\mathbf{a},\mathbf{p})\right\} + \lambda P \\ &-\lambda\mathbb{E}\left\{\sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{q=1}^{N_T} a_{k,m,q}p_{m,q}\right\} + \sum_{k=1}^{K_0}\nu_k\mathbb{E}\left\{R_k(\mathbf{a},\mathbf{p})\right\} - \sum_{k=1}^{K_0}\nu_k R_{0k} \\ &+ \sum_{k=K_0+1}^{K}\mu_k\mathbb{E}\left\{R_k(\mathbf{a},\mathbf{p})\right\} - \sum_{k=K_0+1}^{K}\mu_k\phi_k \sum_{k=K_0+1}^{K}\mathbb{E}\left\{R_k(\mathbf{a},\mathbf{p})\right\} \end{aligned} \tag{7}$$

where the dual variables $\lambda, \boldsymbol{\mu}, \boldsymbol{\nu}$ relax the cost function. It can be rewritten as

$$\begin{aligned} L(\mathbf{a},\mathbf{p},&\lambda,\boldsymbol{\mu},\boldsymbol{\nu}) = \lambda P \\ &+\sum_{k=1}^{K_0}(\mathbb{E}\left\{w_k R_k(\mathbf{a},\mathbf{p})\right\} + \nu_k(\mathbb{E}\left\{R_k(\mathbf{a},\mathbf{p})\right\} - R_{0k}) - \lambda\mathbb{E}\left\{P_k(\mathbf{a},\mathbf{p})\right\}) \\ &+\sum_{k>K_0}^{K}(\mathbb{E}\left\{w_k R_k(\mathbf{a},\mathbf{p})\right\} + (\mu_k - \boldsymbol{\mu}^{\mathbf{T}}\phi)\mathbb{E}\left\{R_k(\mathbf{a},\mathbf{p})\right\} - \lambda\mathbb{E}\left\{P_k(\mathbf{a},\mathbf{p})\right\}) \end{aligned} \tag{8}$$

where $P_k(\mathbf{a},\mathbf{p}) = \sum_{m=1}^{M}\sum_{q=1}^{N_T} a_{k,m,q}p_{m,q}$ is the power of user k.

The dual objective of problem (3) is defined as

$$g(\lambda,\boldsymbol{\mu},\boldsymbol{\nu}) = \max_{\mathbf{a},\mathbf{p}} L(\mathbf{a},\mathbf{p},\lambda,\boldsymbol{\mu},\boldsymbol{\nu}) = L(\mathbf{a}^*,\mathbf{p}^*,\lambda,\boldsymbol{\mu},\boldsymbol{\nu}) \tag{9}$$

where $\mathbf{a}^*, \mathbf{p}^*$ are the optimal solutions which maximize the Lagrangian for each feasible value of $\lambda, \boldsymbol{\mu}, \boldsymbol{\nu}$. Hence, the dual problem can then be written as

$$\begin{aligned} &\min_{\lambda,\boldsymbol{\mu},\boldsymbol{\nu}} g(\lambda,\boldsymbol{\mu},\boldsymbol{\nu}) \\ &s.t. \lambda \ge 0, \ \boldsymbol{\mu} \ge \mathbf{0} \end{aligned} \tag{10}$$

which is a convex optimization problem (even though the primal problem is not a concave maximization problem) with $K+1$ variables. Since the dual objective may be not differentiable, an iterative subgradient method can be used to update the $K+1$ solutions of the dual problem.Starting from initial solutions λ^0 and $\boldsymbol{\mu}^0$, $\boldsymbol{\nu}^0$, the update equations are:

$$\begin{aligned} \lambda^{i+1} &= [\lambda^i - \delta_\lambda g_\lambda^i]^+ \\ \boldsymbol{\mu}^{i+1} &= [\boldsymbol{\mu}^i - \delta_\mu \mathbf{g}_\mu^i]^+ \\ \boldsymbol{\nu}^{i+1} &= [\boldsymbol{\nu}^i - \delta_\nu \mathbf{g}_\nu^i] \end{aligned} \tag{11}$$

where g_λ^i is the subgradient of function $g()$ with respect to λ, i.e.

$$g_\lambda^i = P - \sum_{k=1}^{K}\mathbb{E}\left\{P_k(\mathbf{a}^{*\mathbf{i}},\mathbf{p}^{*\mathbf{i}})\right\}, \tag{12}$$

$\mathbf{g}_\mu^i$ and $\mathbf{g}_\nu^i$ are the subgradients of function $g()$ with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$, i.e.

$$\begin{aligned} \mathbf{g}_{\nu,k}^i &= (\mathbb{E}\left\{R_k(\mathbf{a}^{*\mathbf{i}},\mathbf{p}^{*\mathbf{i}})\right\} - R_{0k}), \quad k \le K_0 \\ \mathbf{g}_{\mu,k}^i &= \mathbb{E}\left\{R_k(\mathbf{a}^{*\mathbf{i}},\mathbf{p}^{*\mathbf{i}})\right\} - \phi_k \sum_{k=K_0+1}^{K}\mathbb{E}\left\{R_k(\mathbf{a}^{*\mathbf{i}},\mathbf{p}^{*\mathbf{i}})\right\}, \quad k > K_0 \end{aligned} \tag{13}$$

and δ_λ, δ_μ, δ_ν are positive step-size parameters; $\mathbf{a}^{*\mathbf{i}}, \mathbf{p}^{*\mathbf{i}}$ are the optimal solutions of (9) at the i-th iteration. An useful advantage of the iterative method is that it keeps low the computation complexity. In the practical applications, the adaptive implementation is suggested, where the iterations are performed along time (i becomes the time

index), and the evaluation of the average power and rate can be done through a stochastic approximation: if $f^i(\mathbf{a}^*, \mathbf{p}^*)$ is the generic rate or power at time i, its average $\mathbb{E}\left\{f^i(\mathbf{a}^*, \mathbf{p}^*)\right\}$ can be approximated with the current value of $f^i(\mathbf{a}^*, \mathbf{p}^*)$ or with a weighted time average $\sum_{j\geq 0} \alpha^j f^{i-j}(\mathbf{a}^*, \mathbf{p}^*)$, where α is a tuning parameter. The dynamic behavior of the adaptive method, i.e the convergence speed and the residual error, depends on the choice of step-size parameters. A discussion is provided in [13].

A. Convergence issues

Concerning the global convergence of the proposed algorithm, the system utility in eq.(2) is nonconcave. Therefore, it appears that the proposed algorithm will not converge or will converge to only a locally optimal power/rate allocation, because it is based on solving the dual problem and yet the duality gap can be strictly positive. Note that spatial power allocation for downlink sum-rate optimization is a non-convex problem and its solution is still open in the literature [14]. This fact has motivated our suboptimal approach leading to eq.(19) in the next Section, which precludes convergence to the global solution. However, simulations that we have done so far present: a stable behavior in stationary scenarios, a good trade-off complexity versus performance and good tracking capabilities. Moreover, the stochastic approximation procedure when applying the subgradient in order to solve the dual optimization problem, is quite well studied in the literature [13], [15]. We omit its analysis for the sake of clarity.

V. Allocation algorithm

We are now deriving the allocation algorithm, i.e. the algorithm which provides at each time slot i, the optimal values of $\mathbf{a}^*, \mathbf{p}^*$ given $\lambda, \boldsymbol{\mu}, \boldsymbol{\nu}$. By keeping as main task the low complexity, we seek for a suboptimal solution. By using ergodic approximation[1] the Lagrangian can be maximized with respect to $\mathbf{a}, \mathbf{p}$, given the dual variables $\lambda, \boldsymbol{\mu}, \boldsymbol{\nu}$. The result can be expressed as follows. Let $\mathbf{u}_m = [u_{m,1}, \ldots, u_{m,N_T}]$ with $u_{m,q} \in \{0, 1, \ldots, K\}$ a vector of N_T user indexes (index $u_{m,q} = 0$ has the meaning of no user in position q), and $\mathbf{v} = [v_1, \ldots, v_{N_T}]$ with $v_q \geq 0$ a vector of N_T power values. The solution is written for each frequency m as

$$\mathbf{u}_m^* = \arg\max_{\mathbf{u}_m} M^*(\mathbf{u}_m) \tag{14}$$

with

$$M^*(\mathbf{u}_m) = \max_{\mathbf{v}\geq \mathbf{0}} M(\mathbf{u}_m, \mathbf{v}) \tag{15}$$

and

$$M(\mathbf{u}_m, \mathbf{v}) = \sum_{q=1, u_{m,q}\neq 0}^{N_T}$$

$$\begin{cases} (\nu_{u_{m,q}} + w_{u_{m,q}}) F_{m,q}(\mathbf{u}_m, \mathbf{v}) - \nu_{u_{m,q}} R_{0k}/M - \lambda v_q & u_{m,q} \leq K_0 \\ (\mu_{u_{m,q}} + w_{u_{m,q}} - \boldsymbol{\mu}^{\mathbf{T}}\phi) F_{m,q}(\mathbf{v}) - \lambda v_q & u_{m,q} > K_0 \end{cases} \tag{16}$$

where $F_{m,q}(\mathbf{u}_m, \mathbf{v}) = \log_2(1 + \gamma_{u_{m,q},m,q}(\mathbf{v}))$,

$$\gamma_{u_{m,q},m,q}(\mathbf{v}) = \frac{v_q c_{u_{m,q},m,q}}{1 + \sum_{s\neq q, u_{m,s}\neq 0} v_s c_{u_{m,q},m,s}} \tag{17}$$

and $\tilde{\mathbf{v}} = \tilde{\mathbf{v}}(\mathbf{u}_m)$ is the solution (15). When the optimal solution $\mathbf{u}_m^*, \tilde{\mathbf{v}}(\mathbf{u}_m^*)$ is found, then

$$a^*_{u*_{m,q},m,q} = 1 \text{ if } u^*_{m,q} \neq 0, \quad a^*_{k,m,q} = 0 \text{ if } k \neq u^*_{m,q} \tag{18}$$

$$p^*_{u*_{m,q},m,q} = \tilde{v}_q(\mathbf{u}_m^*) \tag{19}$$

[1]This means that the statistical average is replaced by the time average

We distinguish between the scheduling phase (variable $\mathbf{a}$) and the power allocation (variable $\mathbf{p}$). The problem in (15) is the power allocation problem for a given set of users that are spatially multiplexed on frequency m. This problem is usually analytically untractable and it is difficult to find a direct solution. As a suboptimal solution we will provide the water-filling solution evaluated by assuming constant uniform power allocation for the interfering beams:

$$\tilde{v}_q = \left[\frac{\nu_{u_{m,q}} + w_{u_{m,q}}}{\lambda \log(2)} - \frac{V}{\gamma_{u_{m,q},m,q}(\mathbf{V})}\right]^+ \quad u_{m,q} \leq K_0$$

$$\tilde{v}_q = \left[\frac{\mu_{u_{m,q}} + w_{u_{m,q}} - \boldsymbol{\mu}^{\mathbf{T}}\phi}{\lambda \log(2)} - \frac{V}{\gamma_{u_{m,q},m,q}(\mathbf{V})}\right]^+ \quad u_{m,q} > K_0 \tag{20}$$

where $\mathbf{V} = [V, V, \ldots, V]$, and the power V is a parameters which estimates the power of interfering beams.

The equation (14) represents a discrete optimization problem which is referred as space-frequency allocation. This problem could be further simplified, by making it independent of power allocation, through the replacement of $\tilde{\mathbf{v}}$ with $\mathbf{V}$ and $\gamma_{u_{m,q},m,q}(\tilde{\mathbf{v}})$ with $\gamma_{u_{m,q},m,q}(\mathbf{V})$. This simplified allocation can enjoy the possibility of a simplified SNIR feedback, which, on the other hand, does not allow the evaluation of the exact achievable rate to be used for adaptive modulation and coding.

Space-frequency allocation requires in general an exhaustive search in the space of all possible vectors $\mathbf{u}_m$. We may describe this space by using the number $Q \in \{1, \ldots, N_T\}$ of allocated beams (those with $u(n, m, q) \neq 0$), the disposition index $j \in \{1, \ldots, N_T!/((N_T - Q)!Q!)\}$ (there are up to $N_T!/((N_T - Q)!Q!$ dispositions of Q allocated beams out of N_T), the combination index $h \in \{1, \ldots, K!/(K - Q)!\}$ (there are up to $K!/(K - Q)!$ combinations of Q users over each disposition of Q allocated beams. This huge search space can be reduced by using by using suboptimal algorithms. One of these algorithms is the one considered in [12] which can be applied when $\tilde{\mathbf{v}}$ is replaced with $\mathbf{V}$. A slightly different formulation which emphasizes the role of a scheduler as an entity that selects users according to a given metric is summarized here. It iterates the following operations until all the available MN_T space-frequency resources have been allocated:

- for each user k find the best resource (m_k, q_k) which maximizes $\gamma_{k,m,q}(\mathbf{V})$
- for each user k find the best other-user combination able to share frequency m_k with user k. This requires the evaluation, for each candidate combination of users sharing frequency m_k, of the metric $M^*(\mathbf{u}_{m_k})$. A suboptimal greedy search is implemented-
- (*) select for allocation the user combination (among the K evaluated before, one for each user) with the best metric and mark the just allocated frequency as unavailable.

The step marked with (*), as discussed later, can be executed by an entity denoted as scheduler, whereas the other steps are executed iteratively by an entity denoted as resource allocator[2]

VI. System architecture

This section discusses the main architectural elements of a system that applies the proposed scheduling strategy to the transmission of multi-user heterogeneous traffic flows in the downlink of a wireless system. In this discussion we refer to the scheme in Fig. 1 and the cross-layer aspects are emphasized.

[2]These steps need only, at each iteration, a simple update which takes into account the new status of the available resources.

A. Applications

We consider 3 different types of applications:

- Application generating VoIP traffic. The source has active periods where a constant bit-rate data flow is generated and silence periods. The traffic has stringent delay requirements and requires a minimum rate during active periods. Data transmitted has a timestamp indicating its deadline D which is used by the scheduler. Voice activity information is useful for scheduling to avoid assignment of unuseful resources. For this traffic, the minimum rate R_{0k} is equal to R_0 in the active periods and switched to 0 during silence period
- Application generating streaming video traffic. The source generates a variable rate data flow and the video packet have different roles and priorities inside the video stream. The traffic has delay requirements: according to playout delay at destination a deadline is fixed and packets arriving after deadline are discarded. Deadline D is used by the scheduler. The quality of the transmission is related to the amount of data arriving at the decoder within the deadline and to their importance in video reconstruction. Side information on the importance of video packet is sent to the queue manager. For this type of traffic the available rate is shared with fairness among users of the same class by tuning the parameters ϕ_k.
- Application generating FTP data traffic. In this case the traffic has not delay requirements. For this type of traffic the available rate is shared with fairness among users of the same class by tuning the parameters ϕ_k. The available rate in this case is lower than the rate assigned to streaming traffic.

The applications provides the lower layers with information on how to set parameters R_0 and ϕ_k, on the the maximum tolerable delay, on the deadline of packets and on the type of video packets.

B. Queue buffers

Data packets coming from the application are placed in a queue. There is a queue for each traffic flow. The queue manager provides the scheduler with the information on the time to deadline TD_k of the Head of Line (HOL) packet of each queue. The queue manager also implements a buffer management policy for each buffer. We consider here the following policies:

- Dropping of expired packets (DXP): packets with expired deadline are dropped to avoid waste of radio resource. This can be applied for VoIP traffic.
- Dropping based on packet priority and dependency (DPD): low priority packets with time to deadline below a given threshold are dropped from the buffer when the size of the queue is too large. This can be applied to streaming video traffic to prevent packet loss for late delivery in case of peaks in the source rate. This policy requires the knowledge of side information on the priority of the packets in the video stream.

C. Scheduler

This is the entity that decides which user or set of users is scheduled for transmission on part of the next available radio resources. This decision is a part of the iterative process for space-frequency allocation described in the previous Section and is based on the comparison of the metrics $M(\mathbf{u}_{m_k})$. These metrics depend on the weights w_k and on the resources that the resource allocator intends to assign. The weights are dynamically set up by the scheduler to make its decision dependent on the status of the queues. The following two strategies are proposed for real time traffic:

$$w_k = 1 + be^{-aTD_k/\tau_{max}} \tag{21}$$

and

$$w_k = 1 + \beta_k e^{-aTD_k/\tau_{max}} \tag{22}$$

where $\beta_k = b(\nu_k + 1)$, if $k \leq K_0$, and $\beta_k = b\mu_k$, if $k > K_0$. Parameters a and b are used to suitably shape the functions of the time to deadline TD_k of the HOL packet in the queue and τ_{max} is the maximum tolerable delay of the application. For both strategies the weight w_k increases exponentially as the TD_k decreases; in the second strategy the increase is proportional to dual variable μ_k or ν_k to improve fairness. The scheduler, by using the parameters R_{0k} and ϕ_k and the information on rate allocated to user provided by the resource allocator, keeps track of rate constraints and updates the values of dual variables μ_k,ν_k. It also sends to resource allocator the updated values of w_k and μ_k,ν_k. This entity does not interact with the physical layer.

D. Resource allocator

This is the entity which decides which resources (space, frequency and power) can be assigned to a given set of users. This decision is a part of the iterative process for space-frequency allocation described in the previous Section and is based on the CSI provided by the physical layer, and the information provided by the scheduler. It evaluates for the scheduler the metrics $M(\mathbf{u}_{m_k})$ and the rate achievable with the assigned resources. When a set of users is scheduled for transmission with a set of assigned resources, the resource allocator sends to physical layer information on the power to use and on the modulation and coding format suitable to realize the assigned rate. The resource allocator is air-interface aware, but does not interact with the queues. The resource allocator also keeps track of power constraint and updates the value of λ.

E. Physical layer

Performs the transmission of scheduled data, according to the resources assigned by the resource allocator. It also handles CSI feedback, which is collected and periodically sent to resource allocator.

VII. Numerical Results

Results are obtained for a scenario which incorporates some characteristic aspects of practical applications in next generation wireless systems. We are considering here a single cell of the downlink of an OFDMA wireless system with M=128 subcarriers working on a bandwidth of 1.25 Mhz. Base station is equipped with multiple antennas. The system is TDD and it is assumed that 2/5 of frame interval is used for downlink transmission. The CSI coming from users is updated every 10ms and for rate assignment a signal-to-noise-ratio gap of 3dB is adopted. The users have a position which is uniformly distributed in circular area of radius 500m. The simulation time is 100s.

Channel model includes path loss, correlated shadowing (not present in the second option for user distributions) and time and frequency correlated fast fading. Path loss is modeled as a function of distance as $L(dB) = k_0 + k_1 \log(d)$ (k_1=40, k_2=15.2 for results). Shadowing is superimposed to path-loss, with classical lognormal model (std. deviation: 6 dB) and exponential correlation in space (correlation distance equal to 20m). Fast fading on each link of the MIMO broadcast channel is complex Gaussian, independent across antennas and is modeled according to a 3GPP Pedestrian model. This model has a finite number of complex multipath components with fixed delay (delay spread of 2.3 microseconds) and power (average

normalized to 1). Time correlation is obtained according to a Jakes' model with given Doppler bandwidth (6 Hz in the results). At the base station orthogonal beamforming is adopted, where beam vectors change randomly at each frame. In the simulated system the total average power constraint is fixed to 1W. To obtain the results we also assume that channel variations in time due to Doppler effects have a negligible impact on the feedback quality. The effects of outdated feedback would require a proper investigation which is out of the scope of the paper.

The first two Tables I and II have the aim of investigating the behavior of the scheduling algorithm with respect to different design options for the weights w_k, the step size δ_ν and the update of variables ν_k for voice users, the use of buffer managements strategies. The performance figures considered are the average rate assigned to users (efficiency at the physical layer), the achieved PSNR for video transmission and the drop-rate for VoIP transmission (quality at the application). The fixed parameters are: R_{0k} =64 kbit/s, $\phi_k = 1/12$, $\alpha = 0.75$, $\delta_\lambda = 0.01$, $\delta_\mu = 0.00005$. The first remark is the substantial fairness achieved among the different video users in spite of the very different channel conditions. We also found the following elements. First, the best way to set up weights w_k is as in eq. (21) which exploits information on packet deadline coming from the queues to prioritize in the short term packets approaching deadlines. This is important for variable rate traffic flows. Second, the periodical update of variables ν_k during the silence periods of voice transmission is very important to avoid data dropping[3] in the first part of voice talk periods where subgradient algorithms needs some time to converge to the new values of ν_k. The update of variables ν_k according to eqs.(10)-(12) requires a virtual assignment of rate which is not used in silence periods. Third, the use of suitable packet dropping strategies at buffer level is important to improve achieved quality at the application level: dropping packets that will be useless if transmitted avoids waste of resource; a content-aware dropping of selected packets in a video stream improves the quality at the decoder.

Figures 2 and 3 explore the dynamic behavior of the scheduling algorithm for the system configuration corresponding to the last line of Table II. Fig. 2 shows the behavior of the power which the resource allocator would assign according to the constraint on the average total power as in the problem (2). Since the variations of instantaneous power around the average are reasonably limited, it si possible to apply instantaneous power scaling after power assignments to satisfy the constraint P even for the instantaneous power. when needed, and this does not destroy the efficiency of the scheduling algorithm. Fig. 3 shows, for one video user, the behavior of variable μ_k and of weight w_k according to the state of the channel and to the status of the queue (which is illustrated through the graph of the normalized time-to deadline of the HOL packet) It is interesting to note that w_k reacts to the state of the queue in favor of situations where packets approach the deadline, whereas μ_k follows the state of the entire system with its slow channel variations[4], trying to accommodate average rate constraints.

Figure 4 has the aim of showing the capability of the scheduler to allocate physical layer resources according to rate constraints imposed in eq. (2). In the system there are 4 different classes of users with 4 different rate requirements (1 class with constant rate constraint, 3 classes with proportional rate constraints). In spite of different channel conditions the constraints are respected with fairness inside each class.

[3]the residual dropping rate of one voice user in this set of simulated results is due to a very deep shadowing lasting for nearly 10s.

[4]fast fading is handled by the space-frequency scheduler

TABLE I

SCENARIO WITH 12 VIDEO USERS. RATE ASSIGNED BY THE SCHEDULER AND PSNR FOR DIFFERENT CHOICES OF WEIGHTS w_k. THE RESULTS IN THE LOWER PART OF THE TABLE REFER TO THE USE OF DPD IN THE BUFFER.

param.	Tot. rate [kbit/s]	Rate [kbit/s]	av. PSNR [dB]	PSNR range [dB]
w_k=1,	2035	160-175	31.1	2.5
(20), $b = 2$, $a = 2$	2240	170-210	33.3	6
(21), $b = 4$, $a = 2$	2069	167-177	32.3	2
w_k=1,	2030	160-175	33.3	1.8
(20), $b = 2$, $a = 2$	2175	168-198	34.8	3.3
(21), $b = 4$, $a = 2$	2067	165-178	34.1	1.8

Finally, the last Table investigate how the assigned rate and quality of service indicators change by increasing the number of user in the system. In this case, the users are assumed to be placed at the same distance $d = 250$m from the base station and do not experience shadowing effects. It is shown that when the user load increases the system is able to preserve the quality of voice users, whereas the quality perceived by the FTP user and the video user decreases, without loosing fairness.

VIII. CONCLUSIONS

In this paper we presented a cross-layer multiuser scheduling strategy for a MIMO-OFDMA system which tries to maximize sum-rate whilst supporting with fairness different traffic classes with different constraints. It exploits temporal diversity by means of ergodic maximization, it allows low complexity on-line implementation and uses a reduced feedback thanks to opportunistic beamforming. We have shown in the numerical results that the algorithm is able to allocate channel resources according to the constraints of each different traffic class and with fairness inside each class, even in presence of different channels conditions and with different network loads.

The proposed scheduling strategy may be easily adapted to a LTE scenario [16] where spatial multiplexing is supported by means of a set of predefined precoding matrices whose colums are ortogonal beamforming vectors. In this case, it is also worth noting that subcarriers are assigned in groups of 12 leading to a complexity reduction in terms of feedback and algorithms.

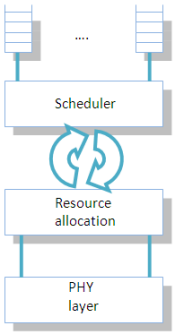

Fig. 1. Simplified architecture of the system that implements the proposed scheduling strategy.

REFERENCES

[1] J.Brehmer, W. Utschick, "Nonconcave Utility Maximisation in the MIMO Broadcast Channel," *Eurasip JASP, vol. 2009.*

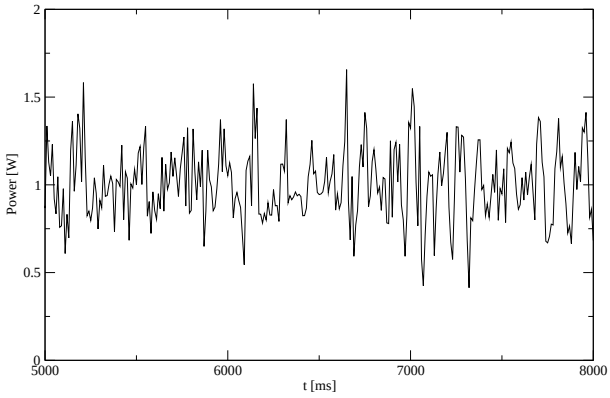

Fig. 2. Dynamic behavior of the total power assigned by the scheduler without power scaling. Scenario with 12 video users and 3 voice users, parameters as in the last line of Table II.

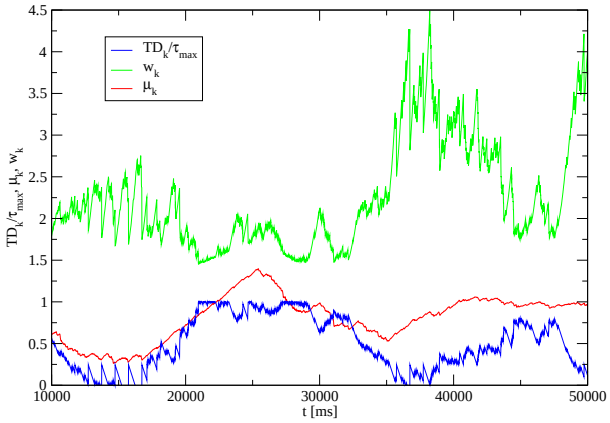

Fig. 3. Dynamic behavior of normalized time-to-deadline of the HOL packet in the queue, the dual variable μ_k and weight w_k for one video user. Scenario

[2] Fuchs, M.; Del Galdo, G.; Haardt, M.; Low-Complexity SpaceTimeFrequency Scheduling for MIMO Systems with SDMA *Vehicular Technology, IEEE Transactions on* Vol. 56, no. 5, Sept. 2007, pp.: 2775-2784

[3] S. Viswanath, N. Jindal, and A. Goldsmith, "On the Capacity of Multiple Input Multiple Output Broadcast Channels," *IEEE Trans. Inf. Theory, vol. 51, pp. 1570-1580, april 2005.*

[4] E. Jorswieck, A. Sezgin, B. Ottersten, A. Paulraj, "Feedback reduction in uplink MIMO OFDM Systems by Chunk Optimization," *Eurasip JASP, vol. 2008.*

[5] N. Wei, A. Pokhariyal, T. B. Sorensen, T. E. Kolding, P. E. Mogensen, "Performance of Spatial Division Multiplexing MIMO with Frequency Domain Packet Scheduling: From Theory to Practice" *IEEE Journal on Sel. Areas in Comm., vol.26, no. 6, august 2008.*Page(s):890 - 900

[6] M. Codreanu, M. Juntti, M. Latva-Aho, "Low-complexity iterative algorithm for finding the MIMO-OFDM broadcast channel sum capacity", *IEEE Trans. on Comm., vol. 55, no. 1, january 2007.* Page(s):48 - 53

[7] Issam Toufik, Marios Kountouris, "Power allocation and feedback reduction for MIMO-OFDMA opportunistic Beamforming," *VTC Spring 2006, pp. 2568-2572.*

[8] G. Liu; J. Zhang; F. Jiang; W. Wang, "Joint Spatial and Frequency Proportional Fairness Scheduling for MIMO OFDMA Downlink," *IEEE WiCom 2007, 21-25 Sept. 2007, pp. 491-494.*

[9] Q. Liu, X. Wang, and G. B. Giannakis, "A cross-layer scheduling algorithm with QoS support in wireless networks," *IEEE Trans. Veh. Technol., vol. 55, pp. 839-847, May 2006.*

[10] V. Corvino, V. Tralli, and R. Verdone, "Cross layer radio resource allocation for multi carrier air interfaces, in multi-cell multi-user environments," *IEEE Trans. Veh. Technol.*, Vol. 58, no. 4, May 2009, pp.: 1864-1875

[11] M. Sharif and B. Hassibi, "On the Capacity of MIMO Broadcast Channel with Partial Side Information," IEEE Trans. on Inform. Theory, vol. 51,no.

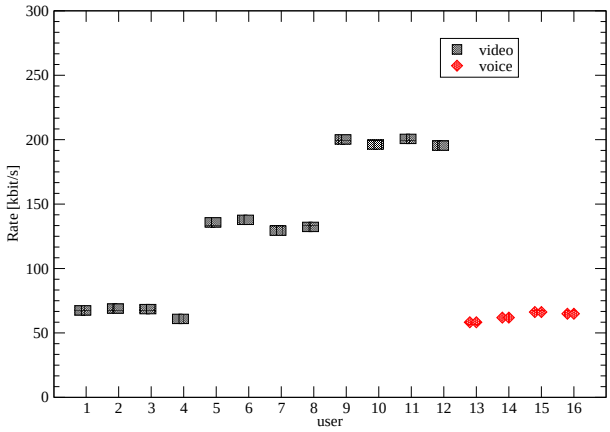

Fig. 4. The different rates assigned by the scheduler in a scenario with 3 different classes of 4 video users ($\phi_k \in \{0.5\phi, \phi, 1.5\phi\}$, with $\phi = 1/12$) and 4 voice users, parameters as in the last line of Table II.

TABLE II

SCENARIO WITH 12 VIDEO USERS AND 3 VOICE USERS. RATE ASSIGNED BY THE SCHEDULER AND QUALITY (PSNR FOR VIDEO AND DROP-RATE FOR VOICE) FOR DIFFERENT CHOICES OF WEIGHTS w_k, PARAMETERS b AND δ_ν, USE OF THE UPDATE (UP) OF DUAL VARIABLES ν_k IN SILENCE PERIODS OF VOICE TRANSMISSION.

param.	Rate video [kbit/s]	Rate voice [kbit/s]	PSNR [dB]	drop-rate [%]
w_k=1, $\delta_\nu = 0.0002$	155-171	40, 63, 75	31.4-34.4	64, 40, 19
w_k=1, $\delta_\nu = 0.001$	145-168	58, 60, 70	31.4-34.4	48, 27, 18
(20), $b = 2$, $\delta_\nu = 0.0002$	160-181	44, 61, 77	31.9-35.3	53, 20, 16
(21), $b = 4$, $\delta_\nu = 0.0002$	153-170	48, 61, 75	32.8-34.8	42, 14, 2.6
(21), $b = 4$, $\delta_\nu = 0.001$	149-165	62, 62, 68	32.6-34.7	25, 5.1, 4.1
(21), $b = 4$, $\delta_\nu = 0.0002$, UP: 4 frames	152-165	73, 87, 102	32.7-34.9	34, 1.6, 0.1
(21), $b = 4$, $\delta_\nu = 0.001$ UP: 4 frames	146-159	79, 88, 93	32.4-34.4	19, 0, 0
(21), $b = 4, 8$, $\delta_\nu = 0.001$ UP: 10 frames	143-158	80, 81, 95	32.5-34.2	17, 0.4, 0

TABLE III

QUALITY FIGURES AS A FUNCTION OF THE NUMBER OF VIDEO USERS IN A SCENARIO WITH 3 OTHERS VOICE USERS AND 1 FTP USER DOWNLOADING 2MB OF DATA.

users	Tot.rate [kbit/s]	PSNR [dB]	drop-rate [%]	max. delay [s]
10	184	36.3	0	5.33
12	198	35.9	0	9.17
14	209	35.5	0	14.85
16	217	34.9	0	20.64
18	220	34.1	0	32.06
20	223	33.5	0.5	41.22
20 (d=300m)	191	32.5	0	50.57
20 (d=350m)	160	30.8	0	66.07

2, pp. 506-522, Feb 2005.
[12] Ana I. Perez-Neira, Pol Henarejos, Velio Tralli, Miguel A. Lagunas, "A low complexity space-frequency multiuser resource allocation algorithm", in Proceedings of WSA 2009, Berlin, Germany, February 16-18, 2009
[13] I. Wong, and B. Evans, "Resource Allocation in Multiuser Multicarrier Wireless Systems," *2008 Ed. Springer*
[14] Mung Chiang, Chee Wei Tan, Daniel P. Palomar, Daniel O'Neill, and David Julian, "Power Control By Geometric Programming," *IEEE Trans on Wireless Comm.*, VOL. 6, NO. 7, JULY 2007
[15] Y. Ermoliev, R. Wets, "Numerical Techniques for Stochastic Optimization," Springer-Verlag, 1988.
[16] 3GPP TS 36.211 Technical Specification: "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Channels and Modulation".