Throughput Maximizing Multiuser Scheduling with Adjustable Fairness

Stefan Schwarz, Christian Mehlfluehrer and Markus Rupp
Institute of Telecommunications, Vienna University of Technology
Gusshausstrasse 25/389, A-1040 Vienna, Austria
Email: {sschwarz, chmehl, mrupp}@nt.tuwien.ac.at

Abstract—We address the problem of downlink multiuser scheduling in practical wireless networks under a desired fairness constraint. Wireless networks such as LTE, WiMAX and WiFi provide partial channel knowledge at the base station/access point by means of quantized user equipment feedback. Specifically in 3GPP’s LTE, the Channel Quality Indicator (CQI) feedback provides time-frequency selective information on achievable rates. This knowledge enables the scheduler to achieve multiuser diversity gains by assigning resources to users with favorable channel conditions. However, focusing only on the possible diversity gains leads to unfair treatment of the individual users. To overcome this situation we propose a method for multiuser scheduling that operates on the boundary of the achievable multiuser rate region while guaranteeing a desired long term average fairness. Our method is based on a sum utility maximization of the $\alpha$-fair utility functions. To obtain a given fairness, quantified with Jain’s fairness index, it is necessary to find an appropriate $\alpha$, which we obtain from the observed CQI probability mass function.

Index Terms—Multiuser scheduling, adaptive resource allocation, utility maximization, fairness, LTE, OFDMA

I. INTRODUCTION

The wireless channel provides frequency diversity caused by multi path propagation and temporal diversity due to the movement of obstacles or users. In addition, a multiuser channel provides multiuser diversity because of the statistical independence of the individual users’ fading processes. Exploiting this diversity by scheduling users on favorable resources is one of the major opportunities to increase the system capacity.

Network providers face the problem of delivering services to users with strongly varying channel quality. System capacity needs to be traded off against user satisfaction, requiring a fairness metric to be considered in the resource allocation process. This problem is mostly tackled by applying proportional fair scheduling [1]. This method is a special case of the more general framework of maximizing the $\alpha$-fair utility functions [2] for $\alpha = 1$. Applying these utility functions for multiuser scheduling, allows obtaining any fairness ranging from the most fair solution, that is, the max. min. scheduler for $\alpha \rightarrow \infty$, to the most unfair but throughput-maximizing solution, the max. throughput scheduler for $\alpha = 0$ [1].

In the literature many approaches to quantify fairness are proposed, one of the most prominent being Jain’s fairness index [3]. However, it is not trivial to apply the $\alpha$-fair utility functions to obtain a predefined fairness, due to the missing relation between $\alpha$ and the fairness measures. The same $\alpha$ can lead to very distinct fairness values in different situations.

We therefore propose a scheduling method that is based on the $\alpha$-fair utility functions but adapts the parameter $\alpha$ to the channel statistics to obtain a desired fairness. In Section II we provide a short motivation explaining the proposed method in the rate region and highlight the advantages. Next, in Section III we present a general formulation of our proposed method and specialize it to Long Term Evolution (LTE). In Section IV we investigate the performance of our method in an LTE system by means of link level simulations employing the Vienna LTE Link Level Simulator [4], [5].

II. MOTIVATION

Consider a hypothetical two-user rate region as shown in Figure 1. User 1 has a high Signal to Noise Ratio (SNR) that allows him to achieve a rate of up to 2 bits per channel use while User 2’s maximum rate equals 1 bit per channel use. Due to the fading nature of the channel it is possible to achieve a sum rate higher than the single user rate by exploiting multiuser diversity. The maximum sum rate is marked with a cross in Figure 1. The other extreme, leading to equal rates of the users, is the solution to the max. min. scheduling problem (marked with a plus symbol) [6]. Let us next impose a fairness constraint applying Jain’s fairness index [3] given by:

$$J(T) = \frac{\left(\sum_{k=1}^{K} T(k)\right)^2}{K \sum_{k=1}^{K} T(k)^2}$$  \hspace{1cm} (1)
Here, $\overline{T}$ is a vector of expected user throughputs (expectation over channel realizations), $\overline{T}(k)$ is the $k$th entry of $\overline{T}$ and $K$ is the number of users. Jain’s fairness index ranges from $\frac{1}{K}$ (only one user is served) to 1 (all users are served at the same rate). The constraint $J(\overline{T}) \geq J_0$ ($J_0$ is the desired fairness) can be reformulated as a Second Order Cone Constraint (SOCC) [7]:

$$\left(\sum_{k=1}^{K} T(k)\right)^2 \geq J_0 \iff 1^T \overline{T}(k) \geq \sqrt{J_0 K} \|\overline{T}(k)\|_2$$

(2)

Two examples are shown in Figure 1 as the two shaded cones. The constraint $J \geq 0.7$ corresponds to the whole cone (ranging from the line with an angle of approximately $80^\circ$ to the line with $15^\circ$). The constraint $J \geq 0.85$ is shown as the inner cone. Increasing $J_0$ shrinks the aperture of the cone until it converges to a line for $J_0 = 1$ going through the max. min. solution.

The scheduling problem is to allocate resources (transmission time, bandwidth) to users such that the rate tuple that maximizes the sum throughput while lying in the intersection of the rate region and the cone given by the fairness constraint is achieved. This suggests to formulate the scheduling problem as a Second Order Cone Program (SOCP) [7]:

maximize: $$\sum_{k=1}^{K} T(k)$$
subject to: $$\sqrt{J_0 K} \|\overline{T}(k)\|_2 \leq 1^T \overline{T}(k)$$

(3)

The difficulty in this formulation is that the resource allocation must be performed for the whole time-frequency interval of interest at once in order to obtain an optimal solution. Typically this is not possible because the channel qualities are not known in advance, just instantaneous values are fed back. This allows solving the SOCP for every time instance, leading to a loss of temporal diversity (see Section IV).

In order to benefit from temporal diversity, another approach to the problem is required. By solving a weighted sum rate maximization problem, points on the boundary of the multiuser achievable rate region can be achieved [8], [9]. Varying the individual users weights, it is possible to trace out the complete boundary. Directly applying such a weighted sum rate maximization in our context, with a desired fairness constraint, requires to find the weights of all users that fulfill the fairness constraint and simultaneously maximize the sum rate. To avoid the difficulties of finding so many weights, we alternatively apply a utility maximization. A user’s utility rate determines his satisfaction with respect to the resource allocation. The task then is to find appropriate utility functions that allow to trade off network fairness versus network throughput. This problem is tackled in Section III.

III. ADJUSTABLE FAIRNESS SCHEDULING

In this section we formulate the utility maximization problem in a general context and specialize it to the requirements of 3GPP LTE by applying the framework presented in [6]. We moreover focus on finding the appropriate utility functions in order to maximize the network throughput and attain a desired fairness specified in terms of Jain’s fairness index.

A. General Formulation

Denote by $C$ the achievable rate region and by $J$ the region corresponding to the SOCC. The sum utility maximization problem is then given by:

maximize: $$\sum_{k=1}^{K} U_\alpha(\overline{T}(k))$$
subject to: $$\overline{T} \in C \cap J$$

(4)

The utility of a user is rated with the $\alpha$-fair utility functions [2]:

$$U_\alpha(x) = \begin{cases} \frac{x^{1-\alpha}}{1-\alpha}, & \alpha \geq 0, \quad \alpha \neq 1 \\ \log(x), & \alpha = 1 \end{cases}$$

(5)

It depends on the user’s average throughput and the parameter $\alpha$. The solution of problem (4) achieves special points on the boundary of the rate region [2]. By varying $\alpha$ from 0 to $\infty$, the line segment in between the max. throughput solution and the max. min. solution is traced out (see Figure 1). Points on this line segment are guaranteed to be Pareto optimal [2], meaning that other points achieving the same fairness do not achieve a larger sum rate. E.g. in Figure 1 the two points marked a) and b) both fulfill the constraint $J_0 = 0.85$, but only the point marked b) attains the maximum sum rate. The points marked c) and d) both satisfy $J_0 = 0.7$, but neither of the two is optimal because a higher sum rate is possible at an even higher fairness, given by the max. throughput solution. In both cases the optimal point lies on the considered line segment. With an appropriate value of $\alpha$, the $\alpha$-fair utility functions yield the desired trade-off between fairness and throughput.

The utility maximization problem (4) depends on the expected user throughputs $\overline{T}(k)$. In [9] it is shown how such a problem is reformulated to an online algorithm that converges to the true solution using standard stochastic approximation techniques. For that purpose the expected throughput $\overline{T}$ is replaced with an average throughput $\overline{T}_n$ at time instance $n$ using the stochastic approximation recursion

$$\overline{T}_n = \left(1 - \frac{1}{\beta}\right) \overline{T}_{n-1} + \frac{1}{\beta} \overline{T}_n$$

(6)

with $\overline{T}_n$ being the instantaneous throughput and the step size $\beta > 1$. The recursion is equal to averaging utilizing an exponentially decaying window function. The parameter $\beta$ determines the decay rate of the window. Substituting (6) into (4) and approximating (4) with a Taylor expansion of first order (see [9]) leads to the online utility maximization problem at time instant $n$

maximize: $$\sum_{k=1}^{K} U_\alpha^*(\overline{T}_{n-1}(k))\overline{T}_n(k)$$
subject to: $$J(\overline{T}) \geq J_0$$

(7)
which has to be maximized with respect to the resource allocation. Jain’s fairness index still depends on the expected throughput. The first derivative $U'_{\alpha}$ of the utility is given by:

$$U'_{\alpha}(x) = \frac{1}{x^{\alpha}} \quad \alpha \geq 0$$  \hspace{1cm} (8)

For $\alpha = 1$ and ignoring the fairness constraint in (7) this problem reduces to the well known formulation of proportional fair scheduling. The difficulty arising from the stochastic approximation is the question of how to deal with the fairness constraint. Simply performing a sum rate maximization as in (3) and substituting (6) for the expected throughput in the fairness constraint enforces the constraint at each time instance and therefore leads to a loss of multiuser diversity (see Section IV). In the following we will not consider the fairness constraint explicitly in the maximization problem anymore but choose the parameter $\alpha$ such that the constraint is met.

**B. Specialization to LTE**

3GPP UMTS/LTE [10] is an Orthogonal Frequency Division Multiple Access (OFDMA) system that imposes several constraints to be considered in multiuser scheduling. LTE divides the time-frequency grid spanned by OFDM into Resource Blocks (RBs) which consist of several contiguous OFDM samples. Different RBs can be assigned to different users. Users provide an RB selective Channel Quality Indicator (CQI) feedback [11], which informs the base station about the achievable rates on every RB. Scheduling decisions can be carried out every subframe (1 ms duration) for the next subframe. A user utilizes the same Adaptive Modulation and Coding (AMC) scheme on all resources he is scheduled onto [10]. The power allocation for all resources is equal and already accounted for in the CQI feedback. In this work, we utilize an approximate linear framework that takes all LTE specific constraints into account [6].

The number of available RBs per subframe is denoted $R$. With $c_{\text{AMC}}^{(n)} \in \mathbb{R}^{R \times 1}$ we refer to the vector of rates $i$ achievable by user $k$ on the $R$ RBs at subframe $n$. The vector $c_{\text{AMC}}^{(n)}$ is computed from the CQI feedback vector CQI$^{(n)}$ (see [6]). Denote by $b_{\text{AMC}}^{(n)} \in \{0, 1\}^{R \times 1}$ the resource allocation vector of user $k$ at subframe $n$. $b_{\text{AMC}}^{(n)}(i) = 1$ means that resource $i$ is allocated to user $k$ at subframe $n$. Excluding multi user MIMO, the resource allocation vectors of the users are mutually orthogonal. With this notation the stochastically approximated utility maximization problem in (7) can be specialized to

$$\{b_{\text{AMC}}^{(1)n}, \ldots, b_{\text{AMC}}^{(K)n}\} = \arg\max_{\{b_{\text{AMC}}^{(1)n}, \ldots, b_{\text{AMC}}^{(K)n}\}} K \sum_{k=1}^{K} c_{\text{AMC}}^{(k)T} b_{\text{AMC}}^{(k)} \quad \text{subject to:} \quad b_{\text{AMC}}^{(i)T} b_{\text{AMC}}^{(j)} = 0 \quad \forall i, \forall j \neq i$$  \hspace{1cm} (9)

with $T_{\text{AMC}}^{n}(k)$ being given by

$$T_{\text{AMC}}^{n}(k) = \left(1 - \frac{1}{\beta}\right) T_{\text{AMC}}^{n-1}(k) + \frac{1}{\beta} c_{\text{AMC}}^{(k)T} b_{\text{AMC}}^{(k)}.$$ \hspace{1cm} (10)

Introducing an equivalent matrix formulation as in [6] allows to reformulate the problem as a Linear Program (LP). Alternatively the reduced complexity proportional fair scheduling strategy for OFDMA systems proposed in [12] can be generalized to arbitrary $\alpha$.

**C. Evaluation of the Parameter $\alpha$**

In order to achieve the goal of a desired fairness, we need to adapt the parameter $\alpha$ to the channel statistics. The channel statistics are a-priori unknown but can be learned online by the scheduler simply by observing the User Equipment (UE) feedback. Specifically we need to learn the probability mass function (pmf) of the achievable rates per resource (RB in LTE). In a practical wireless system there is just a finite number of achievable rates per resource, $N_{\text{AMC}}$, corresponding to the different supported AMC schemes (each scheme has a specific spectral efficiency), which we subsume in the vector $c \in \mathbb{R}^{N_{\text{AMC}} \times 1}$. The achievable rate pmf allows predicting the expected throughput of each user for a given $\alpha$. Then we only need to adapt $\alpha$ such that the predicted expected throughputs meet the fairness constraint.

In order to learn the pmf of the achievable rates per resource of user $k$, $p_{\text{AMC}}^{(k)}$, we utilize the same recursive stochastic approximation as in (6)

$$\hat{p}_{\text{AMC}}^{(k)}(i) = \left(1 - \frac{1}{\beta}\right) \hat{p}_{\text{AMC}}^{(k)}(i) + \frac{1}{\beta} p_{\text{AMC}}^{(k)}.$$ \hspace{1cm} (11)

For LTE the length of the stochastically approximated pmf vector $p_{\text{AMC}}^{(k)} \in [0, 1]^{N_{\text{AMC}} \times 1}$ is equal to the number of different CQI values (16 including CQI zero). The vector $p_{\text{AMC}}^{(k)}$ is an instantaneous approximation of the pmf computed from the feedback values at time instant $n$. For LTE this vector can be computed from CQI$^{(k)}$ by

$$p_{\text{AMC}}^{(k)}(i) = \frac{1}{R} \sum_{r=1}^{R} [\text{CQI}^{(k)}(r) = i] \quad i \in \{0, \ldots, N_{\text{AMC}}\}$$  \hspace{1cm} (12)

with $[i = j]$ being the indicator function for $i = j$.

We next provide an iterative algorithm that computes the parameter $\alpha$ from the approximated pmfs. Utilizing the stochastic approximation (7) the adjustable fairness scheduler allocates a resource to the user for which the ratio of achievable rate and average rate raised to the $\alpha$-th power is largest (this maximizes the sum in (7)). This allows to predict the expected throughput of user $k$ at iteration instant $m + 1$ in the following way:

$$\mathbb{E}(T'_{m+1}(k)) = \sum_{i=1}^{N_{\text{AMC}}} P[k \text{ is served}|i] \cdot c(i) \cdot \hat{p}_{\text{AMC}}^{(k)}(i).$$  \hspace{1cm} (13)

Note the slightly changed notation compared to [6] because the time index is explicitly included.

2Note the difference between $n$ and $m$: the online algorithm described in Sections III-A and III-B uses the index $n$ and is executed at every scheduling instant. The algorithm described now for computation of the $\alpha$ value does not need to be run online. It is only necessary to run it if the pmfs change over time, requiring the computation of a new $\alpha$ value.
Algorithm 1 Algorithm to compute the parameter $\alpha$

\[
T_{\alpha}(l) = T_{\alpha}(l) \forall l \in \{1, \ldots, K\}
\]

$\alpha_1 = \alpha$

$J_1 = J_0(T_{\alpha})$

$m = 1$

stop = false

while stop = false do

for $k = 1$ to $K$ do

Compute $E(T_{m+1}(k))$ utilizing (13) and (14)

Update $T_{m+1}(k)$ utilizing (6)

end for

Compute Jain’s fairness index $J_{m+1}(T_{m+1})$ utilizing (1)

Update $\alpha$ utilizing (15)

if $J_{m+1}$ has converged to $J_0$ then

stop = true

else

$m = m + 1$

end if

end while

$\alpha = \alpha_m$

The probability that user $k$ is served given he experiences an achievable rate $i$, $P[k$ is served$|i]$, can be further developed:

\[
P[k$ is served$|i] = \prod_{l=1}^{K} \sum_{j=1}^{N_{\text{max}}} \left[ \frac{c(j)}{T_{m+1}(l)} < \frac{c(i)}{T_{m+1}(k)} \right] \tilde{p}_i(l)
\]

(14)

From (13) we compute the update $T'_{m+1}$ utilizing the recursive stochastic approximation (6). Then we compute Jain’s fairness index $J$ at iteration instant $m + 1$, $J_{m+1}(T'_{m+1})$, and update $\alpha$ according to

\[
\alpha_{m+1} = \alpha_m + \mu(J_0 - J_{m+1})
\]

(15)

with step size $\mu$. This update rule is obtained by applying the gradient descent algorithm to the cost function $C(\alpha) = (J_0 - J_{m+1}(\alpha))^2$. The step size $\mu$ incorporates the term $\frac{\partial J_{m+1}(\alpha)}{\partial \alpha}$, which cannot be computed in closed form but could be approximated numerically. For simplicity we employ a fixed step size of $\mu = 1$ in our simulations. The complete algorithm, including the initialization and the termination condition, is summarized in Algorithm 1. We apply the experienced throughput and fairness values as initialization.

It is also possible to compute an online update for $\alpha$ using (15) at every scheduling instant $n$ from the experienced fairness $J_n(T_n)$. But we observed that such an algorithm converges much too slowly especially for high fairness values.

IV. SIMULATION RESULTS

In this section we present simulation results obtained with a standard compliant LTE link level simulator [4] implemented in MATLAB. The simulator is publicly available [5]. All simulation results can be reproduced by calling prepared scripts (Schwarz_2011_ICC2011) in the folder “paper scripts”, part of the “Vienna LTE Link Level Simulator” version 1.5.

In the first simulation we consider a two user SISO system with the two users experiencing different average SNRs. The main simulation parameters are summarized in Table I. We do not adapt $\alpha$ to obtain a certain fairness, but fix $\alpha$ to a value in the range $[0, 1000]$. The performance of the $\alpha$-fair utility based scheduler is compared to several other standard schedulers [6].

The Best CQI scheduler is equal to the max. throughput solution discussed earlier. The resource fair scheduler guarantees the same amount of resources for all users while trying to maximize the sum rate. The round robin scheduler allocates to all users the same amount of resources according to a fixed pattern. The SOCP scheduler implements the SOCP described in Equation (3) with the expected throughput $T(k)$ replaced by the average throughput $T_n(k)$. The fairness constraint $J_0$ for this scheduler varies between 0.5 and 0.975. Figure 2 shows that the Best CQI scheduler yields the largest throughput, but the worst fairness, just slightly above the minimum possible fairness of 0.5 (for two users). The SOCP based scheduler, employing a fairness constraint of 0.5, achieves the same performance. The corresponding cone in the rate region equals the non negative orthant. The $\alpha$-fair scheduler reaches that performance for $\alpha = 0$. The highest fairness, $J = 1$, is attained by the max. min. scheduler. The $\alpha$ fair scheduler obtains the same fairness if $\alpha$ is chosen sufficiently large. But, due to the large SNR difference of the two users, even at $\alpha = 1000$ the fairness is only $J = 0.99$. Above this value the throughput decreases strongly due to the large SNR difference. The SOCP based scheduler performs worse than the $\alpha$-fair scheduler because it does not exploit any temporal diversity. The worst scheduler is the round robin type, which neither realizes any temporal nor any frequency diversity.

Next we demonstrate the performance of the pmf based

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>System bandwidth</td>
<td>1.4MHz</td>
</tr>
<tr>
<td>Number of RBs $R$</td>
<td>12</td>
</tr>
<tr>
<td>Number of users $K$</td>
<td>2</td>
</tr>
<tr>
<td>Average SNR of user 1</td>
<td>15 dB</td>
</tr>
<tr>
<td>Average SNR of user 2</td>
<td>0 dB</td>
</tr>
<tr>
<td>$\alpha$ range</td>
<td>[0,1000]</td>
</tr>
<tr>
<td>Channel Model</td>
<td>3GPP TU</td>
</tr>
<tr>
<td>Antenna configuration</td>
<td>1 transmit, 1 receive (1 x 1)</td>
</tr>
<tr>
<td>Receiver</td>
<td>Zero Forcing ZF</td>
</tr>
</tbody>
</table>

Table I: Simulation parameters

Fig. 2. Jain’s fairness index versus sum throughput obtained with several schedulers in a two user scenario.

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>Throughput [Mbit/s]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2.5</td>
</tr>
<tr>
<td>0.5</td>
<td>2.0</td>
</tr>
<tr>
<td>1</td>
<td>1.5</td>
</tr>
<tr>
<td>2</td>
<td>1.0</td>
</tr>
<tr>
<td>3</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Throughput [Mbit/s]

Jain’s fairness index
TABLE II  SIMULATION PARAMETERS

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>System bandwidth</td>
<td>1.4 MHz</td>
</tr>
<tr>
<td>Number of RBs R</td>
<td>12</td>
</tr>
<tr>
<td>Number of users K</td>
<td>5</td>
</tr>
<tr>
<td>Average SNR of users 1 to 5</td>
<td>[15,12,10,5,0] dB</td>
</tr>
<tr>
<td>Step size β</td>
<td>100</td>
</tr>
<tr>
<td>Channel Model</td>
<td>3GPP TU</td>
</tr>
<tr>
<td>Antenna configuration</td>
<td>1 transmit, 1 receive (1 x 1)</td>
</tr>
<tr>
<td>Receiver</td>
<td>Zero Forcing ZF</td>
</tr>
<tr>
<td>Schedulers</td>
<td>α-fair utility-based</td>
</tr>
</tbody>
</table>

Fig. 3. Temporal behavior of user throughputs and fairness for a correlated channel.

parameter α evaluation. We consider a five user scenario, with simulation parameters summarized in Table II. We consider two extreme cases for the correlation of consecutive channel realizations of the block fading channel model. Firstly, the channels of all users stay constant over time. In that case, the achievable rate pmf obtained from the first UE feedback values is already valid and α needs to be computed only once. We impose the fairness constraint $J_0 = 0.95$. Figure 3 shows the throughput of the individual users and the achieved fairness over subframes. The throughput is computed with a moving average filter of length 200 (explaining the ramp up at the beginning). The obtained long term average fairness (averaged over all subframes) equals $J = 0.954$. The slight deviation from the desired fairness is caused by the non zero Block Error Ratio (BLER) that is not taken into account in the α estimation. Next we consider a block fading channel model with independent channel realizations. The estimate of α improves over time as the observed pmf converges to the true value. The offline α computation algorithm is called every 50 subframes. Already after 100 subframes (the second call of the α computation algorithm) the α value stays almost constant. The obtained long term average fairness equals $J = 0.955$.

V. CONCLUSION

In this paper we propose an adjustable fairness scheduler that allows to specify a desired fairness constraint. We propose a general formulation and specialize it to the needs of LTE. The scheduler solves a sum utility maximization problem. The utility of a user is measured by means of the α-fair utility functions. In order to set a desired fairness we propose an algorithm that predicts the required α from the observed pmf of achievable user rates per resource. We demonstrate the performance of this prediction scheme by means of LTE link level simulations. Already after few channel realizations the predicted α converges, closely achieving the desired fairness.

ACKNOWLEDGEMENT

This work has been funded by A1 Telekom Austria AG and the Institute of Telecommunications, Vienna Univ. of Techn. We thank Dr. S. Lambbotharan (Loughborough Univ.) for stimulating discussions.

REFERENCES