Proportional Fairness for MIMO Multi-User Schedulers with Traffic Arrival Process

Masoomeh Torabzadeh and Wessam Ajib
Department of Computer Sciences
University of Quebec at Montreal (UQAM)
torabzadeh_khorasani.masoomeh@courrier.uqam.ca, ajib.wessam@uqam.ca

Abstract—Packet scheduling at the data link layer may impact significantly the overall performance of a wireless system using multiple antennas. In this paper, we propose a novel packet scheduling scheme based on proportional fairness that considers the traffic arrival process with different packet lengths for the downlink of multiple-input multiple-output (MIMO) multi-user systems. We also provide an analysis of fairness for the new scheme in terms of time and service allocation. The scheduler, referred to as clock-time proportional fairness (C-T PF), performs at the packet level and can provide low average packet transmission delay as well as time and service fairness to users. It is work conserving and can also take into consideration different users guarantees (heterogeneous users). We investigate an ideal service fair scheduler called C-T max-min for MIMO systems as well. We compare the performance of C-T PF with other MIMO schedulers. For the time and service fairness comparison of MIMO schedulers, we also propose time and service indexes. Simulations that consider the traffic characteristics and the mobility of users show the low average packet transmission delay and demonstrate the time and service fairness capabilities of C-T PF.

I. INTRODUCTION

It is widely accepted that the utilization of multiple antennas at the physical layer of wireless systems would improve significantly its performances [1]. Anyhow, an efficient scheduling algorithm at the data link layer is needed in multi-user system to efficiently exploit the benefits of multiple-input multiple-output (MIMO) technology in a wireless environment.

In the literature, different schedulers are already proposed. The max-min fairness scheduler [2] picks the users with the smallest mean throughputs at each time-slot. In [3], the authors introduced a maximum carrier-to-interference ratio (max-C/I) scheduler that maximizes the MIMO system capacity without providing fairness. A fair scheduling (named proportional fairness, PF) is implemented for Qualcomm’s HDR system, where the number of transmission channels is one [4]. A PF scheduling promises a trade-off between the maximization of average throughput and user fairness. At each time instant, the user experiencing the highest instantaneous rate with respect to its average rate is scheduled. One feature of proportional fair scheduling is that if queues are not infinitely backlogged, then the algorithm has to define how to deal with empty queues [5]. For the case of finite queues, different versions can be designed depending on which users are eligible for service and how the average throughput of each user is updated. None of the previous schedulers was designed for multiple antenna systems in function of different packet lengths and hence they are not able to provide both fairness and low average packet transmission delay.

In this paper, we propose a novel proportional fairness scheduler, called clock-time proportional fairness (C-T PF), for the downlink of a MIMO multi-user system. The scheduler considers the transmission times of served packets by using a technique, called clock-time (C-T) to be fair time, and takes into consideration the users’ guarantees to provide service fairness. We demonstrate by simulations that the proposed scheduler provides better performances in comparison with other schedulers when traffic characteristics and mobility of users are considered. We also provide fairness analysis for the scheduling in MIMO multi-user systems. We develop an ideal service fair scheduler called C-T max-min, which is a practical implementation of max-min [2] for the downlink of MIMO systems considering the traffic arrival process.

The reminder of this paper is organized as follows. In the second section, we discuss the system model. In section III, we present our scheduling algorithms and we provide the numerical results in section IV. The fairness issue is discussed in the fifth section and we conclude in section VI.

II. SYSTEM MODEL

We consider the downlink of a MIMO multi-user system, where a single base station (BS) having \( N_R \) antennas is communicating with \( K \) mobile users (MSs) and each MS has \( N_T \) antennas. At the transmitter, data packets are loaded onto the transmit antennas using spatial multiplexing [6]. Each packet is entirely transmitted from one transmit antenna to the destination user. Packets destined to different users can be transmitted simultaneously from different BS antennas. For simplicity, the transmit power is assumed equally divided among the transmit antennas. We assume that the channels between the different users and the BS (and consequently the users’ rates since we are assuming adaptive modulation) are invariant during each time-slot of length \( T_s \) and change over the time-slots. The channel realizations are timely correlated (dependent over the time-slots) because of the mobility of users.

The vector signal received by user \( k \) at time \( t \) can be given by

\[
y_k(t) = H_k s(t) + n_k(t),
\]

where \( s(t) \) is the \( N_T \times 1 \) vector of signals sent by the BS, \( n_k(t) \) is a \( N_R \times 1 \) vector of additive white complex Gaussian noise experienced by user \( k \) \((k = 1, \ldots, K)\), and \( H_k \) is the \( N_R \times N_T \) channel matrix between the BS and user \( k \).
The covariance matrix of the noise is \( \sigma_n^2 I_N \), where \( I_N \) is the \( N \times N \) identity matrix. Each user channel encounters path loss, log-normal shadow fading, and multi-path fading. We also assume that the fading at the receiver antennas are spatially uncorrelated. The channel matrix \( H_k \) between the BS and user \( k (k = 1, \ldots, K) \) for each time-slot can be given as

\[
H_k = \sqrt{\text{SNR}_0(l_k / D)^\beta} \times 10^{5/10} \times G_k ,
\]

where \( \text{SNR}_0 \) is the median signal to noise ratio (SNR), the path loss exponent, \( \beta \) is the path loss exponent, \( S_\text{k} \) is a real Gaussian random variable with zero mean and unit variance and represent Rayleigh-distributed multi-path fading, \( G_k \) is the distance between the base station and user \( k \) at the start of the time-slot and \( D \) is the cell radius \((l_k \leq D)\).

In order to mitigate the multi stream interference (MSI) at the receiver, the maximum likelihood method [7] is known to be optimal but it is of high computational complexity mainly when high number of antennas or high order modulation is used. The linear sub-optimal receivers, such as zero-forcing (ZF) and the widely-used minimum mean square error (MMSE), provide considerably less computational complexity. They are often preferred for their simplicity [6]. In our system model, we assume the utilization of MMSE receivers at the BSs since it is relatively simple and balances the MSI mitigation with the noise enhancement [7]. Note that the scheduler proposed in this paper can be used with any of the spatial multiplexing receiver structures. We assume that the post-detection signal-to-interference and noise ratio (SINR) for each BS antenna is error-free fed back by MSs to the BS at the start of each time-slot.

If the scheduler assigns the transmit antenna \( n \ (1 \leq n \leq N_t) \) to user \( k \ (1 \leq k \leq K) \) at time \( t \), (for simplicity, we say the channel \( n \) is assigned to user \( k \)); then the post-detection vector of signals can be obtained by

\[
z_k(t) = W_k y_j(t) = W_k H_k s(t) + W_k n_k(t) ,
\]

where \( W_k \) is the MMSE weight matrix that is invariant for each time-slot. It is given by

\[
W_k = (H_k^H H_k + (\sigma_n^2 N_t / P_t) I_{N_t})^{-1} H_k^H ,
\]

where \( (\cdot)^H \) denotes the conjugate transpose and \( I_{N_t} \) is the \( N_t \times N_t \) identity matrix.

The post-detection SINR introduced for MIMO single user systems in [8] can be extended for MIMO multi-user systems [9]. In our system, the post-detection SINR for served user \( k \) on channel \( n \) can be given by

\[
\gamma_{k,n} = \frac{||W_k H_k l_{n}||^2}{\sigma_n^2 N_t \sum_{m=1}^{N_t} ||W_k H_k l_{m}||^2 + \sum_{m=1}^{N_t} ||W_k H_k l_{m}||^2} ,
\]

where \( P_t \) is the total transmitted signal power.

Furthermore, we assume an adaptive modulation in conjunction with the MIMO technique: i.e. each transmit signal uses a separately adaptive modulation, which is matched to the instantaneous channel condition. When we use uncoded \( M \)-ary \((M = 2^j, j = 1, 2 \ldots 8) \) modulation schemes for the packet transmissions, the rate of channel \( n \) of the transmitter assigned to user \( k \) can be calculated as [10]

\[
r_{k,n} = B \times \min(8, \lfloor \log_2 (1 + \gamma_{k,n} / \Omega) \rfloor) ,
\]

where \( B \) is the bandwidth in terms of Hz and \( 10 \log_{10} \Omega = 8 \) dB. We notice that for uncoded M-QAM the envelope of the throughput curves is parallel to the Shannon capacity curve, \( \log_2 (1 + \gamma_{k,n}) \), with a fixed offset of about 8 dB. Therefore, similar to the Shannon capacity, the envelope can be expressed as \( \log_2 (1 + \gamma_{k,n} / \Omega) \) [10]. We assume that the spectral efficiency of 8 for each stream is the maximum that can be achieved using a practical modulation scheme, since it corresponds to uncoded 256-QAM.
III. PACKET SCHEDULERS

Some scheduling algorithms in MIMO systems consider that a whole time-slot on each transmit channel is allocated to one user [11]. We use the packet-based MAC framework shown in Fig. 1 to be able to assign packets of different users to the same BS antenna at each time-slot. We assume that the MSs send their channel information to the BS and the packet scheduling is done based on this information at the start of each time-slot. At the start of each time-slot, the scheduling algorithm schedules the head of line (HOL) packets of users to be served during the current time-slot. At each packet scheduling epoch, the scheduler selects a pair of user and BS antenna that maximizes the utility function which is defined later. Since the time-slot duration is fixed whereas the transmission times for packets depend on the packet length and the user transmission rate, there can be unused fractions of time at the end of each time-slot, as shown in the example of Fig. 1. Anyhow, this shortcoming can be overcome by packet fragmentation but for simplicity, we assume no fragmentation. Note that our scheduler design considers the minimization of unused time fractions by selecting as many packets as possible.

After that, the BS broadcasts the scheduling results including the selected users as well as their allocated times and their assigned modulation mode (i.e., their rates). Finally, the scheduled packets are sent on downlink channels and selected users can receive their data at the allocated times considering the assigned modulation mode.

A. Clock-time proportional fairness

The new scheduler, called clock-time proportional fairness (C-T PF), modifies the proportional fairness [4] in order to consider different user packet lengths and the transmission times of served packets. C-T PF uses the proportional fairness utility function for user and antenna selection, while the average users’ throughputs used in the utility function are updated taking into consideration the length and transmission rate of the served packet proportional to those of the last $T$ served packets. In this case, the scheduler can improve the fairness as well as the average throughput.

The scheduler selects a pair of user and antenna, and then the head of line (HOL) packet of the selected user can be assigned to the selected transmit antenna. A user is called backlogged if its queue is not empty and non-backlogged otherwise. Let $\mathcal{B}$ denote the set of backlogged users. The C-T PF scheme determines at the start of each time-slot whether there are backlogged users or not. If $\mathcal{B}$ is empty, the scheme goes to the next time-slot. Otherwise, a user $k^*$ is assigned to the transmit antenna $n$ as

$$ (k^*, n^*) = \arg \max_{k \in \mathcal{B}, n = 1 \ldots N} \frac{r_{k,n}}{\bar{r}_k}, $$

(7)

where $r_{k,n}$ is the transmittable data rate of user $k$ assigned to transmit antenna $n$ at the current time-slot and $\bar{r}_k$ is the average throughput of user $k$. The scheduler checks if the finish time for the packet provided by antenna $n^*$ is over the current time-slot, and if so, it tries to select another user and antenna pair by re-computing (7). Otherwise, the total transmission times of the last $T$ packets, called $C_T$, is updated by

$$ C_T = C_T - \frac{L_{k^*}^{(T)}}{r_{k^*, n^*}} + \frac{L_{k^*}^{(T)}}{r_{k^*, n^*}}. $$

(8)

where $L_{k^*}$ is the length of the HOL packet of user $k^*$, and $L_{k^*}^{(T)}$ and $r_{k^*, n^*}$ are the length and transmission rate of the packet served $T$ scheduling epochs ago, respectively.

![Fig. 2. Clock-time technique](image)

The C-T PF scheme uses a clock-like buffer such that the clock hand turns clockwise one unit to update $C_T$. We call this technique clock-time (C-T) (Fig. 2). Units in the clock equal the transmission times of the last $T$ packets served in the system, and the clock cycle (the sum of units) shows the amount of $C_T$. The amount of the first unit next to the clock’s hand, which is the transmission time of the packet served at $T$ scheduling epochs ago, is updated with the transmission time of the HOL packet of user $k^*$ by turning the hand. The $\bar{R}_k$’s are updated by

$$ \bar{R}_k = \frac{\gamma L_{k'} \sum_{n \in B} W_{n}}{C_{T'} W_{k'}} + (1 - \gamma) \beta_k \bar{R}_k, $$

(9)

where the weight factor $w_{n'}$ is a predefined parameter of user $k'$ proportional to its service share, $\gamma$ equals 1 if $k = k^*$ and zero if not, and $\beta_k$ equals 1 if user $k$ is backlogged and zero if not (note that $\beta_k$ equals 1). Furthermore, the scheduler assigns the HOL packet of user $k^*$ to antenna $n^*$ of the base station. The scheduler tries to serve as many packets as possible by repeating the procedure. Finally, it serves the assigned packets at the current time-slot and goes to the next time-slot.

B. Clock-time max-min

We also propose the C-T max-min scheduler that can be an ideal service fair scheduler and a practical implementation of max-min [2] for the downlink of a MIMO system, by taking into consideration the traffic arrival process. At each scheduling epoch, a user $k^*$ is selected according to

$$ k^* = \arg \min_{k \in \mathcal{B}} \bar{R}_k. $$

(10)

The HOL packet of user $k^*$ is transmitted from the antenna that can provide the highest rate for that user. If the finish time of the packet is over the current time-slot for the selected antenna, the next higher rate antenna can be selected. Note that if none
of the base station antennas can provide a finish time for the packet within the current time-slot, another user will be selected by (10). After that, the $\overline{R}_k$'s are updated by (9).

IV. SIMULATION EXPERIMENTS

A. System parameters and user behaviours

Our simulation is based on a MIMO system with $N_T = 4$ and $N_R = 4$. In (2) and (5), we assume $SNR_B = 10$ dB, the path loss exponent $\beta = 3.7$ dB, the log standard deviation of shadow fading $\sigma_T = 8$ dB, the cell radius $D = 1$ Km, and $P_f / \sigma_n^2 = 10$ dB. In addition, we have the time-slot duration $T_f = 2.4$ ms, $T = 2000$ packets in (8), and $T_{pf} = 2000$ in (13). In (2), $l_k$ is the distance between the base station and user $k$ and it has been randomly initialized between zero and 1 Km. Users approach the base station or leave it at various constant velocities uniformly distributed between zero and 60 Km/h; if the distance of a mobile user becomes 1 Km from the base station, it changes its direction to the opposite direction with the same velocity. We assume $B = 5$ MHz in (6). Using (6), we can calculate users’ rates that are between zero and 40 Mbps.

For modeling the traffic of each MS, we break down the overall data volume of each user into the four classes of conversational, streaming, interactive, and background [12].

For the conventional class, we adopt the ON/OFF voice traffic model with exponentially distributed ON/OFF durations. The mean duration of the ON and OFF periods are set to 2.5 sec and 0.5 sec, respectively. During each ON period, packets are generated at a fixed rate of 64 kbps with a fixed packet size of 80 Bytes. We generate video streaming traffic for the streaming class. Each frame of video data arrives at a regular interval of 100 ms and it is segmented into 10 packets. The size and interval of these packets are distributed as a Pareto with a shape parameter and minimums of 40 Bytes and 2.5 ms, respectively. The interactive class is represented by a WWW traffic model with active and inactive periods. During an active period, we have active ON and OFF periods. The number of ON periods during an active period equals the number of files to be downloaded, which are distributed as a Pareto with a shape parameter of 2.43 and a minimum of 1 file. The ON period is a function of the file size distributed as a Pareto with the shape parameter of 1.2 and a minimum of 1000 Bytes. Each inactive OFF period has a Weibull distribution with $a$ and $b$ parameters that are 0.328 and 1.46, respectively. Finally, we use a FTP traffic model for the background class. The ON period is a function of the file size distributed as a Pareto with a shape parameter of 1.1 and a minimum of 1000 Bytes. The OFF period is distributed as a Weibull with $a$ and $b$ parameters that are 0.328 and 1.46, respectively. In the ON period of both the WWW and FTP, the packet interval is 2.5 ms and the distribution of the packet size is discretely between 40 and 1500 Bytes. The number of mobile users is set to $K = 20$ and the traffic of each of them consists of different numbers of voice sessions, video sessions, FTP sessions, and WWW sessions, except for users 5 and 8, which only contain voice sessions. In Fig. 3 we plot the total arrival traffic rate of the system generated for 12 minutes with average rate of 108.7 Mbps. We can notice the burstiness of the traffic which is useful to evaluate the performance of our scheduler.

B. System performance

For comparison purposes, we also evaluate the performance of traditional schedulers, i.e., max-C/I [3] and PF [4], at the packet level. We implement max-C/I as

$$\left(k^*, n^* \right) = \arg \max_{k \in B, n = 1, ..., N_T} w_k \times r_{k,a}.$$  \hspace{1cm} (11)

Whereas, the PF scheduler is implemented using the two following equations:

$$\left(k^*, n^* \right) = \arg \max_{k \in B, n = 1, ..., N_T} \frac{w_k \times r_{k,a}}{\overline{R}_k},$$  \hspace{1cm} (12)

$$\overline{R}_k = \begin{cases} \frac{1}{T_{pf}}((T_{pf} - 1)\overline{R}_k + r_{k,a}), & k \in B \land k = k^* \\ \frac{T_{pf}}{(T_{pf} - 1)\overline{R}_k + r_{k,a}}, & k \in B \land k \neq k^* \\ \overline{R}_k, & k \notin B \end{cases}.$$  \hspace{1cm} (13)

Fig. 4 shows system throughputs (total user throughputs) for the average throughput of each user is the average amount of bytes per second served for the user. Note that each user packet is served with the transmission rate computed by (6). Fig. 4 shows system throughputs (total user throughputs) for each class.
various schedulers under various traffic loads averaged in 12 minutes simulation. We notice that our proposed C-T PF scheduler outperforms the PF and the C-T max-min schedulers. The throughput provided by the C-T PF scheduler is very close to the one provided by the max-C/I scheduler. It is to be highlighted that, as we will show later in this paper, the max-C/I provides poor fairness performances. As expected, the C-T max-min scheduler provides poor throughput performances. PF scheduler, since they only contain voice sessions with small packet sizes. We remind that the PF scheduler does not take into consideration different packet sizes of users. Anyhow, C-T PF and max-C/I provide low long-term average delays for users. We notice that having a low long-term average delay is not sufficient for QoS provisioning and we need to have a low short-term average delay as well. This can be achieved using a fair scheduler. In the next section, we investigate the fairness issues.

V. FAIRNESS ANALYSIS

In this section, we analyze the fairness of MIMO multi-user schedulers. Assuming \( B(\tau_1, \tau_2) \) is the set of continuously backlogged users \( k \) during time interval \([\tau_1, \tau_2]\), a scheduler is defined fair if the amount of resource \( G_k(\tau_1, \tau_2) \) received by each user \( k \) is proportional to its fair share of resource \( G(\tau_1, \tau_2) \) in \([\tau_1, \tau_2]\), i.e.,

\[
G_k(\tau_1, \tau_2) = \frac{w_k G(\tau_1, \tau_2)}{\sum_{m \in B(\tau_1, \tau_2)} w_m}, \forall k \in B(\tau_1, \tau_2). \tag{14}
\]

where \( w_k \) is the weight of user \( k \). In another word, (14) can be written as

\[
\frac{G_k(\tau_1, \tau_2)}{w_k} = \frac{G_v(\tau_1, \tau_2)}{w_{v}}, \forall k, k' \in B(\tau_1, \tau_2). \tag{15}
\]

And, after normalizing to one we get

\[
\frac{G_k(\tau_1, \tau_2)}{w_k G(\tau_1, \tau_2)} / \sum_{m \in B(\tau_1, \tau_2)} w_m = 1, \forall k \in B(\tau_1, \tau_2). \tag{16}
\]

The index of fairness during \([\tau_1, \tau_2]\) can be expressed as

\[
I(\tau_1, \tau_2) = C_v \left( \frac{G_k(\tau_1, \tau_2)}{w_k G(\tau_1, \tau_2)} \sum_{m \in B(\tau_1, \tau_2)} w_m \right), X = \{ k : k \in B(\tau_1, \tau_2) \}. \tag{17}
\]

where function \( C_v(.) \) expresses the coefficient of variation. For a fair scheduler, the index of fairness goes to zero. The equation can be extended for a time interval \([\tau_1, \tau_2]\) when users may not be continuously backlogged. For each served packet \( p \) in \([\tau_1, \tau_2]\), if user \( k \) is backlogged we have to count a portion of resource \( G^p \) as the fair share of user \( k \). We also calculate the amount of resource \( G_k(t_1, t_2) \) received by user \( k \) in \([t_1, t_2]\).

The index of fairness during \([t_1, t_2]\) is written by

\[
I(t_1, t_2) = C_v \left( \sum_{p=1}^{N(t_1, t_2)} \frac{G_k(t_1, t_2)}{B^p_w G^p} \right), X = \{ k : 1 \leq k \leq K \}. \tag{18}
\]

where \( N(t_1, t_2) \) is the total number of served packets during \([t_1, t_2]\) and \( K \) is the total number of users in the system. At the \( p \)th scheduling epoch in \([t_1, t_2]\), \( B^p_w \) is the set of backlogged users and flag \( B^p_w \) is 1 if user \( k \) is backlogged and zero if not. The term “resource” can be as “transmission time” or “serving bytes” that a scheduler has to share among users in a fair way. Therefore, our index of time fairness during a time interval \([t_1, t_2]\) can be calculated as.


\[ I_s(t_1, t_2) = C_s \left( \frac{1}{N_t(t_1, t_2)} \sum_{p=1}^{N_t(t_1, t_2)} \sum_{m=1}^{N_t(t_1, t_2)} b_{m}(t_1) \sum_{n=1}^{N_t(t_1, t_2)} w_{n}(t_1) \right), X = \{ k : 1 \leq k \leq K \}, \]

where \( S_0(t_1, t_2) \) is the total time in which user \( k \) was served by transmit antennas during \([t_1, t_2]\). At the \( p^{th} \) scheduling epoch in \([t_1, t_2]\), \( L^{(p)} \) and \( r^{(p)} \) are the length and transmission rate of the served packet. Our index of service fairness during \([t_1, t_2]\) is defined as

\[ I_s(t_1, t_2) = C_s \left( \frac{1}{N_t(t_1, t_2)} \sum_{p=1}^{N_t(t_1, t_2)} \sum_{m=1}^{N_t(t_1, t_2)} b_{m}(t_1) \sum_{n=1}^{N_t(t_1, t_2)} w_{n}(t_1) \right), X = \{ k : 1 \leq k \leq K \}, \]

where \( b_{m}(t_1, t_2) \) is the number of bytes of user \( k \) served in \([t_1, t_2]\).

A low index of fairness indicates a better fairness performance.

![Fig. 7. Time fairness index for 108.7 Mbps traffic load](image7)

![Fig. 8. Service fairness index for 108.7 Mbps traffic load](image8)

Fig. 7 plots time fairness indexes of various schedulers in 0.5-second periods for the 108.7 Mbps traffic load. It shows that C-T PF has the lowest time fairness index and is the fairest scheduler in terms of time. On the other hand, the max C/I scheduler with the highest time fairness index can provide the lowest time fairness. The time fairness indexes for C-T max-min and PF schedulers are higher than that of the C-T PF scheduler as well. In Fig. 8, we show service fairness indexes of various schedulers in 0.5-second periods for the 108.7 Mbps traffic load. We notice that C-T max-min is an ideal service fair scheduler; since it serves the user with the lowest average throughput at each scheduling epoch. C-T PF index of service fairness is lower than those of PF and max C/I. The max C/I scheduler has the highest service fairness index and hence the worst service fairness. It provides no fairness since it tries to serve packets belonging to higher rate users. This leads to starvation for low rate users and obviously high short-term average delay for those users. Note that fairness is a very important issue for short-term QoS provisioning.

VI. CONCLUSION

We addressed the problem of practical implementation of proportional fairness scheduling for MIMO multi-user systems, by taking into consideration the traffic arrival process with different packet lengths. We proposed the clock-time proportional fairness (C-T PF) scheduler as a modification to the proportional fairness to take into consideration different packet sizes. We also investigated the clock-time max-min (C-T max-min) scheduler, which is an ideal service fair scheduler. In addition, we provided analysis for comparing time/service fairness of MIMO packet schedulers. Simulation results considering traffic characteristics and the mobility of users demonstrate the gains obtained by using the C-T PF scheduling scheme in terms of delay, throughput, and fairness.

REFERENCES


