Is It Possible to achieve the optimum throughput and Fairness Simultaneously in a MIMO Broadcast Channel?

Alireza Bayesteh
Dept. of Electrical Engineering
University of Waterloo
Waterloo, ON, N2L 3G1
Email: alireza@shannon2.uwaterloo.ca

Mehdi Ansari Sadrabadi
Dept. of Electrical Engineering
University of Waterloo
Waterloo, ON, N2L 3G1
Email: mehdi@shannon2.uwaterloo.ca

Amir K. Khandani
Dept. of Electrical Engineering
University of Waterloo
Waterloo, ON, N2L 3G1
Email: khandani@shannon2.uwaterloo.ca

***Abstract*—In this paper, a MIMO Broadcast Channel (MIMO-BC) with large (K) number of users is considered. It is assumed that all users have a hard delay constraint D. We propose a scheduling algorithm for maximizing the throughput of the system, while satisfying the delay constraint for all users. It is proved that by using the proposed algorithm, it is possible to achieve the maximum throughput and maximum fairness in the network, simultaneously, in the asymptotic case of $K \to \infty$. We introduce a new performance metric in the network, called "Minimum Average Throughput", and prove that the proposed algorithm is capable of maximizing the *minimum average throughput* in a MIMO-BC, in the asymptotic case of $K \to \infty$. Finally, it is established that the proposed algorithm reaches the boundaries of the *capacity region* and *stability region* of the network, simultaneously, in the asymptotic case of $K \to \infty$.**

I. Introduction

With the development of personal communication services, one of the major concerns in supporting data applications is providing quality of service (QoS) for all subscribers. In most real-time applications, high data rates and small transmission delays are desired. Most data-scheduling schemes proposed for current systems have concentrated on the system throughput by exploiting multiuser diversity [1]–[5]. In cellular networks, by applying multiuser diversity, the time-varying nature of the fading channel is exploited to increase the spectral efficiency of the system. It is shown that transmitting to the user with the highest signal to noise ratio (SNR) provides the system with maximum sum-rate throughput [6]. The opportunistic transmission is proposed in Qualcomm's High Data Rate (HDR) system [2].

Although applying multiuser diversity through the scheme in [6] achieves the maximum system throughput, QoS demands, including fairness and delay constraints, provoke designing more appropriate scheduling schemes. The schemes that consider delay constraints have been studied extensively in [1], [7]–[21]. In [7], the authors propose an algorithm which maintains a balance between the throughput maximization, delay, and outage probability in a multiple access fading channel. The tradeoff between the average delay and the average transmit power in fading environments is analyzed in [8]. In [9], [10], authors propose scheduling metrics that combine multiuser diversity gain with the delay constraints. In [11], the scheduling scheme is designed based on maximizing the effective capacity [22] which is characterized by data rate, delay bound, and delay-bound violation probability triplet. The throughput-delay tradeoff of the multicast channel is analyzed for different schemes in a single cell system [12]. This trade-off has been obtained for more general network topologies in [13]. In the static random network with n nodes, the results of [13] show that the optimal tradeoff between throughput T_n and delay D_n is given by $D_n = \Theta(nT_n)$. They also show that the same result is achieved in random mobile networks, when $T_n = O(1/\sqrt{n \log n})$. The first studies on achieving a high throughput and low delay in ad-hoc wireless networks are framed in [4], [14], and [15]. This line of work is further expanded in [13], [16], [17] by using different mobility models such as the random walk and the Brownian mobility models. Neely and Modiano [17] consider the delay-throughput tradeoff only for mobile ad-hoc networks. They investigate the delay characteristics by using the redundant packets transmission through multiple paths. In [18], the authors have proposed and compared different scheduling achemes based on the users' channel qualities and their remaining job times, in the downlink of a MIMO wireless cellular packet data system in fast and slow channel variation scenarios. In [19], the authors have analytically characterized the scheduling gain achieved by opportunistic schedulers with both single-user and multi-user multiplexing, and showed that the average delay grows double-exponentially with the overall throughput, with any opportunistic (single-user time-sharing or multi-user multiplexing) scheduling. In [20], the authors consider a wireless downlink communication system, where the channels are characterized by frequency-selective fading, modeled as a set of M parallel block-fading channels, and a frequency-flat distance-dependent path loss. They compare delay-limited systems (which impose hard fairness) with variable-rate systems (which impose proportional fairness), in terms of the achieved system spectral efficiency C (bit/s/Hz) versus E_b/N_0, and find simple iterative resource allocation algorithms that converge to the optimal delay-limited throughput for orthogonal (FDMA/TDMA) and optimal (superposition/interference cancellation) signaling. In the limit of large K and finite M, the authors find closed-form expressions for C as a function of E_b/N_0 and show that in this limit, the optimal allocation policy consists of letting each user transmit on its best subchannel only.

In [21], the delay is defined as the minimum number of channel uses that guarantees all n users successfully receive m packets. Reference [21] studies the statistical properties of the underlaying delay function. However, the delay constraint is assumed to be *soft*, meaning that this scheme aims to minimize the total *average* network delay and there is not any delay constraints for the individual users.

In this paper, we consider a *hard* delay constraint D for each user, which is enforced by the application or physical limitations (e.g. buffer size). We define a dropping event as the event that there exists a user who does not meet the desired delay constraint. We propose a scheduling scheme for maximizing the throughput of the system, while satisfying the delay constraint for all users. The proposed scheduling algorithm is a variant of the Random Beam-Forming scheme proposed in [23], which works based on setting a threshold on the Signal to Interference plus Noise Ratio (SINR) of the users on each transmitted beam. Among the users with channel gains above the threshold, the user with the minimum *Packet Expiry Countdown*s (PEC), which is defined as the remaining time to the expiration of that users' packet, is served. By doing asymptotic analysis, it is proved that by selecting the threshold level properly, the proposed scheduling algorithm achieves the maximum throughput, maximum fairness, and minimum delay in the network, simultaneously, in the asymptotic case of $K \rightarrow \infty$. The analysis is based on characterizing the probability mass function of PEC in terms of K, D , and the threshold value, and evaluating the network dropping probability accordingly. Moreover, we introduce a new notion of performance in the network, called "Average Throughput", which is defined as the product of the packet arrival rate and the amount of information per channel use in each packet, and prove that the proposed algorithm maximizes the *Minimum Average Throughput* in a MIMO-BC.

The rest of the paper is organized as follows. In section II, the system model is introduced and the proposed algorithm is described. Section III is devoted to the asymptotic analysis of the proposed algorithm. Finally, section IV concludes the paper.

Throughout this paper, the Hermitian operation is denoted by $(.)^H$, notation "$\log$" is used for the natural logarithm, and the rates are expressed in *nats*. For any functions $f(N)$ and $g(N)$, $f(N) = O(g(N))$ is equivalent to $\lim_{N\to\infty}\left|\frac{f(N)}{g(N)}\right| < \infty$, $f(N) = o(g(N))$ is equivalent to $\lim_{N\to\infty}\left|\frac{f(N)}{g(N)}\right| = 0$, $f(N) = \Theta(g(N))$ is equivalent to $\lim_{N\to\infty}\frac{f(N)}{g(N)} = c$, where $0 < c < \infty$, $f(N) \sim g(N)$ is equivalent to $\lim_{N\to\infty}\frac{f(N)}{g(N)} = 1$, and $f(N) \gtrsim g(N)$ is equivalent to $\lim_{N\to\infty}\frac{f(N)}{g(N)} \geq 1$.

II. System Model and Proposed Algorithm

A. System Model, Assumptions, and Definitions

In this paper, a downlink environment in which a Base Station (BS), equipped with M antennas, communicates with a large number (K) single-antenna users, is considered. We assume a homogeneous network, where the channel between each user and the BS is modelled as a zero-mean complex Gaussian random variable (Rayleigh fading). The received signal at the kth terminal can be written as

$$y_k = \mathbf{h}_k\mathbf{x} + n_k, \tag{1}$$

where $\mathbf{x} \in \mathbb{C}^{M\times 1}$ is the transmitted signal with the power constraint $\mathbb{E}\{\mathbf{x}^H\mathbf{x}\} \leq P$ [1], $\mathbf{h}_k \in \mathbb{C}^{1\times M} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$ is the channel vector, $n_k \sim \mathcal{CN}(0, 1)$ is AWGN, and y_k is the received signal by the kth user.

We assume that block coding for error free transmission is performed over frames, where the information content of a frame is called packet. In addition, we assume that the frame length is constant (unit of time), while the information content of a frame can potentially vary depending on the capacity of the corresponding channel realization. As we will see later, the proposed method results in almost equal information content (packet length in bits) for all the frames. It is also assumed that M users are served during each frame. The channel coefficients are assumed to be constant for the duration of a frame, and change independently at the start of the next frame (block fading model). The frame itself is assumed to be long enough to allow communication at rates close to the capacity. This model is also used in [21] and [23].

It is assumed that the users have stringent delay constraint D. In other words, the delay between two consecutive received packets should not be greater than the duration of D frames. Otherwise, the transmitted packet will be dropped. The *network dropping event*, denoted by $\mathscr{B}$, is defined as the event that dropping occurs for any user in the network. We define a parameter ν for each user, which denotes the *Packet Expiry Countdown (PEC)* of that user's packet, i.e., the remaining time to the expiration of the packet. ν is expressed in terms of an integer multiple of the frame length. At the end of each frame, the PEC of each user is decremented by one, except for the user which is served during that frame. For this user, the PEC is set to D at the start of the next frame. Therefore, for all users $\nu \leq D$ (Fig. 1). Since the channel model is independent block fading, and the network topology and the proposed scheduling algorithm are symmetric with respect to the users, it can be easily shown that there exists a steady state for the system (no matter what the initial state is), in which the statistical behavior of the users' PECs is independent of the time index. All the results derived in this paper are based on the assumption that the system is in the steady state.

In this paper, we are interested in maximizing the *throughput* and *fairness* in the network. First, we give the definitions of *throughput* and *fairness*:

[1] Note that the power constraint here is *per frame*, i.e, is independent of the channel realizations.

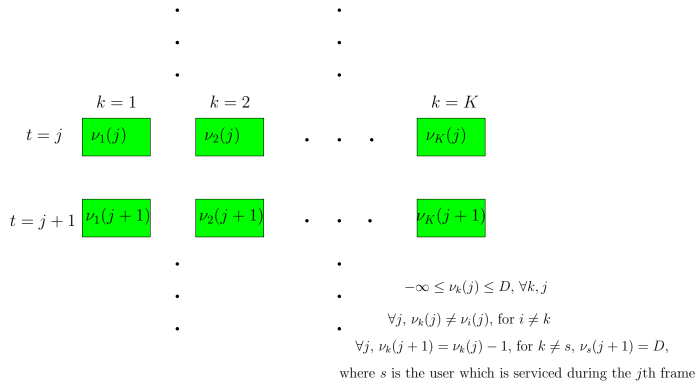
$k=1$
$k=2$
$k=K$
$t=j$
$\nu_1(j)$
$\nu_2(j)$
$\nu_K(j)$
$t=j+1$
$\nu_1(j+1)$
$\nu_2(j+1)$
$\nu_K(j+1)$
$-\infty \le \nu_k(j) \le D, \forall k, j$
$\forall j, \nu_k(j) \neq \nu_i(j)$, for $i \neq k$
$\forall j, \nu_k(j+1) = \nu_k(j) - 1$, for $k \neq s$, $\nu_s(j+1) = D$,
where s is the user which is serviced during the jth frame

Fig. 1. A Schematic figure for the *expiry countdown*.

Definition 1 *The* ***throughput*** *is defined as the average sum-rate of the system, when the average is computed over all the channel realizations.*

Definition 2 *Consider a scheduling $\mathfrak{S}$. Then, the* ***Fairness Factor*** *(FF) for this scheduling is defines as*

$$FF(\mathfrak{S}) \triangleq \frac{MD_{\min}(\mathfrak{S})}{K}, \tag{2}$$

where $D_{\min}(\mathfrak{S})$ denotes the minimum value of D such that Pr$\{\mathscr{B}\} \to 0$, using scheduling $\mathfrak{S}$.

Definition 3 *A scheduling $\mathfrak{S}$ is said to achieve the maximum fairness, if $FF(\mathfrak{S}) = 1$ [2].*

B. Proposed Scheduling Algorithm

The proposed scheduling algorithm is described as follows:

1) Set the threshold Υ.
2) The BS selects M orthogonal unit vectors, denoted by $\mathbf{\Phi}_1, \cdots, \mathbf{\Phi}_M$, randomly, and sends it to all users.
3) Among each of the following sets:

$$\mathcal{S}_m = \{k| \quad \text{SINR}_k^{(m)} > \Upsilon\}, \quad m = 1, \cdots, M, \tag{3}$$

the BS serves the user with the minimum PEC. In the above equation, $\text{SINR}_k^{(m)} \triangleq \frac{\frac{P}{M}|\mathbf{h}_k \mathbf{\Phi}_m^H|^2}{1+\sum_{j\neq m}\frac{P}{M}|\mathbf{h}_k \mathbf{\Phi}_j^H|^2}$ is the received Signal to Interference plus Noise Ratio (SINR) on the mth transmitted beam, by the kth user.

As can be observed, this algorithm is a variant of Random-Beam-Forming scheme proposed in [23], where the PEC is considered in the scheduling.

III. Asymptotic Analysis

In this section, we analyze the network dropping probability, denoted as $\Pr\{\mathscr{B}\}$, in terms of the number of users K, and the delay constraint D, for the proposed scheduling. We consider the asymptotic case of $K \to \infty$ and derive the condition for D such that $\Pr\{\mathscr{B}\} \to 0$. To this end, the probability mass function (pmf) of ν, denoted as $f_\nu(.)$, is characterized in terms of D, K, and Υ. It is interesting to investigate the possibility of achieving the maximum throughput and fairness of the system, simultaneously, which is performed in the following theorem:

Theorem 1 *Using the proposed algorithm, for the values of Υ satisfying*

$$\frac{P}{M}\left[\log K - (M+1)\log\log K\right] < \Upsilon < \frac{P}{M}\left[\log K - (M+0.5)\log\log K\right], \tag{4}$$

one can simultaneously achieve:
I- Maximum Throughput:

$$\lim_{K\to\infty} \mathcal{C}_{\text{sum}} - \mathcal{R} = 0, \tag{5}$$

in which $\mathcal{C}_{\text{sum}}$ denotes the maximum achievable sum-rate in the MIMO-BC and $\mathcal{R}$ denotes the achievable sum-rate of the proposed algorithm, and
II- Maximum Fairness:

$$\lim_{K\to\infty} \frac{MD}{K} = 1, \tag{6}$$

while Pr$\{\mathscr{B}\} \to 0$ (or equivalently, $\lim_{K\to\infty} FF = 1$).

Proof - The steps of the proof are as follows: in Lemma 1, we study the behavior of $f_\nu(l)$ and derive a difference equation satisfied by $f_\nu(l)$. In Lemma 2, we derive an explicit solution for this difference equation. Based on this solution, in Lemma 3, we present a sufficient condition such that the conditions $\lim_{K\to\infty}\frac{MD}{K} = 1$ and $\Pr\{\mathscr{B}\} \to 0$ are satisfied simultaneously. Finally, the theorem is proved by deriving a lower-bound on the achievable sum-rate based on the threshold level given in (4), which is performed in Lemma 4. For the proof of the lemmas, the reader is referred to [24].

Lemma 1 *Defining $D_0 = D - \sqrt{K}n_0(n_0-1)$, where $n_0 = 3(\log K)^2$, for $D_0 \le l \le D$, we have $f_\nu(l) = \frac{M}{K}\left[1 + o(1/K)\right]$, and for $l < D_0$, $f_\nu(l)$ satisfies the following difference equation:*

$$f_\nu(l) - f_\nu(l-1) \quad = \quad \eta f_\nu(l) e^{-(K-1)pF_\nu(l)}\left[1 + O(1/\sqrt{K})\right], \tag{7}$$

where $p = \frac{e^{-\frac{M}{P}\Upsilon}}{(1+\Upsilon)^{M-1}}$, $\eta \triangleq Mp$, and $F_\nu(.)$ denotes the CDF of ν.

Sketch of the Proof - The key step in the proof of Lemma 1 is the following equation, which is proved in [24]:

$$f_\nu(l-1) = f_\nu(l)\left(1 - \Pr\{\mathscr{X}_k|\nu_k = l\}\right), \tag{8}$$

where $\mathscr{X}_k$ denotes the event that user k is served. In the region $D_0 \le l \le D$, it can be shown that $\Pr\{\mathscr{X}_k|\nu_k = l\} = o(\frac{1}{K})$.

[2] This definition is motivated by the fact that for Round-Robin scheduling (which is known to be the most fair scheduling), assuming that M users are served in each frame, $D_{\min} = \lceil\frac{K}{M}\rceil$.

In the region $l < D_0$, $\mathfrak{Q}(l) \triangleq \Pr\{\mathscr{X}_k | \nu_k = l\}$ can be written as

$$\mathfrak{Q}(l) \quad = \quad M \sum_{n=1}^{K} \binom{K-1}{n-1} \left(\frac{q}{M}\right)^n \left(1 - \frac{q}{M}\right)^{K-n} \mathfrak{P}(n,l), \tag{9}$$

where $g_l(n,l) \leq \mathfrak{P}(n,l) \leq g_u(n,l)$, and

$$g_u(n,l) = \left\{ \begin{array}{ll} \prod_{i=1}^{n-1} \left(G_\nu(l-1) + \frac{iM}{K} \right), & n \leq n_0 \\ 1 & n > n_0 \end{array} \right. , \tag{10}$$

$$g_l(n,l) = \left\{ \begin{array}{ll} \prod_{i=1}^{n-1} \left(G_\nu(l) - \frac{iM}{K} \right), & n \leq n_0 \\ 0 & n > n_0 \end{array} \right. , \tag{11}$$

where $G_\nu(l) \triangleq 1 - F_\nu(l)$, the complementary CDF of ν. Substituting the above upper-bound and lower-bounds in (9), after some manipulations Lemma 1 is proved. (for the detailed proof the reader is referred to [24]).

Lemma 2 *The solution to the difference equation (7), in the asymptotic case of $K \to \infty$, is*

$$f_\nu(l) = \frac{\frac{\varphi}{(K-1)p} e^{(K-1)p} e^{\varphi(l-D_0)}}{1 + e^{(K-1)p} e^{\varphi(l-D_0)}} \quad l < D_0, \tag{12}$$

for some $\varphi = \eta \left[1 + O\left(\frac{1}{\sqrt{K}}\right)\right]$.

Lemma 3 *Setting $D_0 = \frac{p}{\varphi}(K-1) + \frac{\log K}{\varphi}$, for some φ such that $\varphi = \eta \left[1 + O\left(\frac{1}{\sqrt{K}}\right)\right]$, yields Pr$\{\mathscr{B}\} \to 0$, while satisfying $\lim_{K \to \infty} \frac{MD}{K} = 1$.*

Lemma 4 *The achievable sum-rate of the proposed algorithm can be lower-bounded as*

$$\mathcal{R} \geq M \log(1 + \Upsilon) \left[1 - \left|O\left(e^{-(\log K)^{1.5}}\right)\right|\right]. \tag{13}$$

Noting the facts that $\mathcal{C}_{\text{sum}} = M \log(1 + \frac{P}{M} \log K + O(\log\log K))$ [23], and $\Upsilon > \frac{P}{M}[\log K - (M+1)\log\log K]$, we have

$$\lim_{K \to \infty} \mathcal{C}_{\text{sum}} - \mathcal{R} \quad = \quad 0. \tag{14}$$

Combining the above equation with Lemma 3 completes the proof of Theorem 1.

■

Theorem 2 *Consider a MIMO-BC, in which the information data delivered to the users are put in packets, which are stored in the transmitter buffer and each packet is mapped to a coded frame, consisting of n channel uses, and transmitted over the channel. Assume that the Packet Arrival Rate (PAR) for user k to be fixed and equal to r_k (measured as the number of arrived packets per unit time, i.e., one frame duration), the amount of information in each packet of that user to be $n\mathcal{R}_k$, and the transmitter has the buffer size of one packet for each user. Let us define the "average throughput" of user k (normalized per channel use) as* [3]

$$\mathfrak{T}_k \triangleq r_k \mathcal{R}_k. \tag{15}$$

Then, for any scheduling scheme, any rate vector $\mathbf{R} = (\mathcal{R}_1, \cdots, \mathcal{R}_K)$ inside the capacity region (decoding error approaches zero), and for any PAR vector $\mathbf{r} = (r_1, \cdots, r_K)$ inside the stability region [25] (Pr$\{\mathscr{B}\} \to 0$), one has

$$\mathfrak{T}_{\min} \triangleq \min_k \mathfrak{T}_k \quad \lesssim \quad \frac{M \log\log K}{K}, \tag{16}$$

which is achievable by the proposed algorithm.

Proof - *Necessary Condition* - Consider a long interval of time T. Defining $\mathcal{A}_k(t)$ as the indicator variable taking one when the user k is served during the frame t, and taking zero otherwise, we have

$$\sum_{k=1}^{K} \mathcal{A}_k(t) \mathcal{R}_k \leq \mathcal{C}_{\text{sum}}, \quad \forall t, 1 \leq t \leq T. \tag{17}$$

The above equation comes from the fact that the rates $(\mathcal{R}_1, \cdots, \mathcal{R}_K)$ must be inside the capacity region of MIMO-BC. Taking the summation with respect to t, we can write

$$\sum_{t=1}^{T} \sum_{k=1}^{K} \mathcal{A}_k(t) \mathcal{R}_k \leq \mathcal{C}_{\text{sum}} T. \tag{18}$$

Since Pr$\{\mathscr{B}\} \to 0$, the arrival rate of the packets must be less than or equal to their service rate, over a long period of time, almost surely. In other words, $\sum_{t=1}^{T} \mathcal{A}_k(t) \gtrsim T r_k$, $\forall k, 1 \leq k \leq K$, with probability one. Substituting in the above equation yields

$$\begin{aligned} \sum_{k=1}^{K} \mathfrak{T}_k = \sum_{k=1}^{K} r_k \mathcal{R}_k \quad & \lesssim \quad \mathcal{C}_{\text{sum}} \\ & \overset{(a)}{\sim} \quad M \log(\frac{P}{M} \log K), \end{aligned} \tag{19}$$

where (a) comes from [23]. Combining (15) and (19), yields

$$\begin{aligned} \mathfrak{T}_{\min} \quad & \leq \quad \frac{\sum_{k=1}^{K} \mathfrak{T}_k}{K} \\ & \lesssim \quad \frac{M \log\log K}{K} + \frac{M \log(\frac{P}{M})}{K} \\ & \sim \quad \frac{M \log\log K}{K}. \end{aligned} \tag{20}$$

Sufficient Condition - Consider the proposed algorithm, with the condition of Theorem 1, i.e., $\frac{P}{M}[\log K - (M+1)\log\log K] < \Upsilon < \frac{P}{M}[\log K - (M+0.5)\log\log K]$. It is realized from Lemma 3 that selecting $r_k = \frac{1}{D}$ for all users, where D is obtained as follows:

$$D = \frac{p}{\varphi}(K-1) + \frac{\log K}{\varphi} + 9\sqrt{K}[\log K]^4 \sim \frac{K}{M},$$

[3] This definition is motivated by the fact that there is a time delay of $\frac{1}{r_k}$ between two consecutive packets of user k, and as a result, the average amount of information per channel use delivered to user k is equal to $r_k \mathcal{R}_k$.

guarantees $\Pr\{\mathscr{B}\} \to 0$. Furthermore, the channel can support the rate

$$\mathcal{R}_k = \log\left[1 + \frac{P}{M}\left(\log K - (M+1)\log\log K\right)\right],$$

for all users, with probability one. Hence,

$$\begin{aligned} \mathfrak{T}_{\min} &\geq \frac{\log\left[1 + \frac{P}{M}(\log K - (M+1)\log\log K)\right]}{D} \\ &\sim \frac{M \log\log K}{K}. \end{aligned} \tag{21}$$

■

In the above theorem, the *minimum average throughput*, denoted by $\mathfrak{T}_{\min}$, is defined as the measure of performance. The average throughput itself can be interpreted as the average amount of information (per channel use) delivered to a user over a long period of time. This measure is suitable for the real-time applications, where the packets have certain amount of information and certain arrival rates. Note that in Theorem 2, we have assumed that the users have the buffer size of one, which is a very restrictive assumption in wireless networks. For the realistic scenarios, this constraint is more relaxed. However, since we have shown the optimality of our proposed scheduling for this assumption, it easily follows that this optimality holds for more relaxed assumptions, as well.

Remark - An interesting observation of Theorem 2 is that the proposed algorithm reaches the boundaries of the *capacity region* and *stability region* of the network (on the line $r_1 = r_2 = \cdots = r_K$), simultaneously, in the asymptotic case of $K \to \infty$.

IV. Conclusion

In this paper, a MIMO Broadcast Channel (MIMO-BC) with large (K) number of users has been considered. It is assumed that all users have a hard delay constraint D. We have proposed a scheduling algorithm for maximizing the throughput of the system, while satisfying the delay constraint for all users. It is proved that by using the proposed algorithm, it is possible to achieve the maximum throughput and maximum fairness in the network, simultaneously, in the asymptotic case of $K \to \infty$. We have introduced a new performance metric in the network, called "Minimum Average Throughput", and proved that the proposed algorithm is capable of maximizing the *minimum average throughput* in a MIMO-BC, in the asymptotic case of $K \to \infty$. Finally, it is established that the proposed algorithm reaches the boundaries of the *capacity region* and *stability region* of the network, simultaneously, in the asymptotic case of $K \to \infty$.

References

[1] P. Viswanath, D.N.C. Tse, R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Trans. Inform. Theory*, vol. 48, pp. 1277–1294, June 2002.

[2] P. Bender, P. Black, M. Grob, R. Padovani, N. Sindhushayana, and A. Viterbi, "CDMA/HDR: A bandwidth efficient high-speed wireless data service for nomadic users," *IEEE Communications Magazine*, pp. 70–77, July 2000.

[3] A. Jalali, R. Padovani, and R. Pankaj , "Data throughput of CDMA/HDR: A high efficiency, high data rate personal wireless system," in *Proc. IEEE Vehicular Tech. Conference*, vol. 3, pp. 1854–1858, May 2000.

[4] X. Liu, E. K. Chong, and N. B. Shroff, "Opportunistic transmission scheduling with resource-sharing constraints in wireless networks," *IEEE JSAC*, vol. 19, pp. 2053–2064, Oct. 2001.

[5] S. Borst, "User-level performance of channel-aware scheduling algorithms in wireless data networks," *IEEE/ACM Trans. Networking*, vol. 13, pp. 636–647, June 2005.

[6] R. Knopp and P. A. Humblet, "Information capacity and power control in single-cell multiuser communications," *IEEE ICC'95*, vol. 1, pp. 331–335, June 1995.

[7] I. Bettesh and S. Shamai, "A low delay algorithm for the multiple access channel with Rayleigh fading," *in Proc. IEEE Personal, Indoor and Mobile Radio Commun.*, vol. 3, pp. 1367–1372, Sept. 1998.

[8] R. A. Berry and R. G. Gallager, "Communication over fading channels with delay constraints," *IEEE Trans. Inform. Theory*, vol. 48, pp. 1135–1149, May 2002.

[9] M. Andrews, K. Kumaran, K. Ramanan, A.L. Stoylar, R. Vijayakumar and P. Whiting, "Providing quality of service over a shared wireless link," *IEEE Commun. Mag.*, vol. 39, pp. 150–154, Feb. 2001.

[10] R. Srinivasan and J. S. Baras, "Understanding the trade-off between multiuser diversity gain and delay - an analytical approach," *IEEE Vehicular Technology Conference*, vol. 5, pp. 2543–2547, May 2004.

[11] D. Wu and R. Negi , "Utilizing multiuser diversity for efficient support of quality of service over a fading channel," *IEEE Trans. Vehicular Techn.*, vol. 54, pp. 1198–1206, May 2005.

[12] P.K. Gopala and H. El Gamal, "On the throughput-delay tradeoff in cellular multicast," *International Conference on Wireless Networks, Communications and Mobile Computing*, vol. 2, pp. 1401–1406, June 2005.

[13] A. El Gamal, J. Mammen, B. Prabhakar, and D. Shah, "Optimal throughput-delay scaling in wireless networks - part I: The fluid model," *IEEE Trans. on Information Theory*, vol. 52, no. 6, pp. 2568–2592, June 2006.

[14] N. Bansal and Z. Liu, "Capacity, delay and mobility in wireless ad-hoc networks," *in Proc. IEEE INFOCOM*, April 2003, pp. 1553–1563.

[15] S. Toumpis and A. J. Goldsmith, "Large wireless networks under fading, mobility, and delay constraints," *in Proc. IEEE INFOCOM*, March 2004, pp. 609–619.

[16] G. Sharma L. Xiaojun, R. R. Mazumdar, and N. B. Shroff, "Degenerate delay-capacity tradeoffs in ad-hoc networks with brownian mobility," *IEEE Trans. on Information Theory*, vol. 52, no. 6, pp. 2777–2784, June 2006.

[17] M. J. Neely and E. Modiano, "Capacity and delay tradeoffs for ad hoc mobile networks," *IEEE Trans. on Information Theory*, vol. 51, no. 6, pp. 1917–1937, June 2005.

[18] M. Airy, S. Shakkottai and R. Heath, "Limiting queuing models for scheduling in multi-user MIMO systems," *IASTED Conference on Communications, Internet and Information Technology*, Scottsdale, AZ, November 17–19, 2003.

[19] Manish Airy, Sanjay Shakkottai and Robert W. Heath Jr, "Scheduling for the MIMO Broadcast Channel: Delay-Capacity Tradeoff," Preprint.

[20] Giuseppe Caire, Ralf R. Mller and Raymond Knopp, "Hard Fairness Versus Proportional Fairness in Wireless Communications: The Single-Cell Case," *IEEE Trans. on Information Theory*, vol. 53, no. 4, pp. 1366–1385, April 2007.

[21] Masoud sharif and Babak Hassibi , "A delay analysis for opportunistic transmission in fading broadcast channels," in *Proc. IEEE, INFOCOM*, vol. 4, pp. 2720–2730, March 2005.

[22] D. Wu and R. Negi, "Effective capacity: a wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, vol. 2, pp. 630–643, July 2003.

[23] M. Sharif and B. Hassibi, "On the capacity of MIMO broadcast channel with partial side infonnation," *IEEE Trans. on Inform. Theory*, vol. 51, pp. 506–522, Feb. 2005.

[24] A. Bayesteh, Mehdi A. Sadrabadi, and Amir K. Khandani, "Throughput and Fairness maximization in Wireless Downlink Systems," *Submitted to IEEE Trans. on Inform. Theory*, Sept. 2007, also Available online at http://cst.uwaterloo.ca/pubjour.html.

[25] D. Bertsekas and R. Gallagher, *Data Networks*. Prentice Hall, Englewood Cliffs, NJ, 2nd edition, 1991.