Delay Considerations for Opportunistic Scheduling in Broadcast Fading Channels

Masoud Sharif and Babak Hassibi

***Abstract*— We consider a single-antenna broadcast block fading channel with n users where the transmission is packet-based. We define the (packet) delay as the minimum number of channel uses that guarantees *all* n users successfully receive m packets. This is a more stringent notion of delay than average delay and is the worst case (access) delay among the users. A delay optimal scheduling scheme, such as round-robin, achieves the delay of mn. For the opportunistic scheduling (which is throughput optimal) where the transmitter sends the packet to the user with the best channel conditions at each channel use, we derive the mean and variance of the delay for any m and n. For large n and in a homogeneous network, it is proved that the expected delay in receiving one packet by all the receivers scales as $n \log n$, as opposed to n for the round-robin scheduling. We also show that when m grows faster than $(\log n)^r$, for some $r > 1$, then the delay scales as mn. This roughly determines the time-scale required for the system to behave fairly in a homogeneous network. We then propose a scheme to significantly reduce the delay at the expense of a small throughput hit. We further look into the advantage of multiple transmit antennas on the delay. For a system with M antennas in the transmitter where at each channel use packets are sent to M different users, we obtain the expected delay in receiving one packet by all the users.**

***Index Terms*— Broadcast channel, fading, opportunistic scheduling, packet delay, longest queue.**

I. Introduction

Resource allocation in wireless systems aims for two conflicting goals, firstly providing quality of service such as delay and fairness to users, and secondly maximizing the throughput of the system. A fundamental property of wireless channels is their time variation due to multi-path effects and the mobility of the users. This implies that at each channel use some users have favorable channel conditions and other users incur deep fades. In fact, assuming a block fading model for the channel and having full CSI in the transmitter, it can be shown that sending to the user with the best channel conditions maximizes the sum rate (or throughput) of the single antenna broadcast channel.

In order to exploit this multiuser diversity, the base station (or the transmitter) has to know the channel state information (CSI) of all the users. In fact, this opportunistic way of transmission has been proposed in Qualcomm's High Data Rate (HDR) system (1xEV-DO). Other variations of this scheduling that do not require full CSI in the transmitter are studied in [1], [2].

However, there is a price to pay for maximizing the throughput which is fairness among users and delay in sending packets. Assuming users have different signal to noise ratios, the throughput optimal scheduling will provide much less service to the user with the lowest signal to noise ratio (SNR) compared to that of the user with the highest SNR. Even in a homogeneous network where users have equal SNRs and so the system is long-term fair, there is no delay guarantee for transmitting a packet to a specific user as the transmission is probabilistic, i.e., at each channel use each user will be chosen with some probability. The other extreme would be to use a round robin type scheduling that fairly gives service to all users and can guarantee a fixed delay for transmitting a packet to each user. In applications with delay constraints, one may wonder how bad the worst case delay (or the delay for the most unfortunate user) for the throughput optimal strategy is.

In this paper, we consider a broadcast channel with n users in which users' messages are independent. The transmission is packet based and the channel is assumed to be block Rayleigh fading and changes independently from one block to the other. We also assume packets are dropped if outage occurs, i.e., the instantaneous capacity goes below the amount of information in the packet. Given the probability of outage P_e, we assume packets carry a fix amount of information C_0 which only depends on the scheduling. For example, opportunistic scheduling is the one that maximizes the throughput given P_e. This will be further discussed in Section 2.

We define the delay as the minimum number of transmissions that guarantees all the users will receive m packets successfully. This notion of delay is clearly stronger than the average delay in the sense that it guarantees the reception of m packets by *all* users. This definition of the delay is specially useful for applications with deadline [16]. Disregarding the throughput and if the users are back-logged, the minimum delay of mn can be achieved by round-robin scheduling[1]. However, the throughput optimal strategy has to contend with delay hits. The overriding question in this paper is to

Manuscript received January 30, 2006; revised July 14, 2006 and July 31, 2006; accepted August 16, 2006. The associate editor coordinating the review of this paper and approving it for publication was J. Zhang. This work is presented in part at IEEE Inter. Symp. on Info. Theory 2004 and IEEE INFOCOM 2005. This work is supported in part by the National Science Foundation under grant no. CCR-0133818, by the office of Naval Research under grant no. N00014-02-1-0578, and by Caltech's Lee Center for Advanced Networking.

M. Sharif is with the Department of Electrical and Computer Engineering, Boston University, Boston, MA 02215 (email: sharif@bu.edu).

B. Hassibi is with the Department of Electrical Engineering, California Institute of Technology, Pasadena, CA 91125 (email: hassibi@systems.caltech.edu).

Digital Object Identifier 10.1109/TWC.2007.06067.

[1] If the users are not backlogged, there is a chance that the chosen user has an empty queue. This probability must be taken into account (see Section 3.1)

characterize the delay for the throughput optimal strategy, e.g. to determine its mean and other moments. Finally, we propose an algorithm to reduce the delay at the expense of a little hit in the throughput of the system. The results in this paper imply that opportunistic transmission increases the delay by a factor of $\log n$ compared to that of delay optimal strategies.

Previously, the question of the delay-throughput trade off has been addressed by several authors in different contexts [15]. In single link systems, the problem of how to optimally allocate the power among channel uses such that the capacity is maximized while guaranteeing the delay for sending bits remains bounded has been considered in [3], [4]. Also, the trade off between average power and delay has been addressed by Berry and Gallager for single link systems [5]. In multiuser channels, traditionally delay and throughput were considered separately and therefore, access schemes such as ALOHA [6] were proposed to avoid collisions without exploiting multiuser diversity. As noted later in [7], [8], there has been a large body of work to combine the physical layer and multiple access layer (see [9], [10], [11], [17], [18] and references there in). For multiple access channels, a decentralized variation of ALOHA algorithm is proposed that exploits multiuser diversity [1]. In [19], the authors consider the problem of characterizing the capacity region under a stability condition for queues. Stability here is in the sense that the probability of the queue overflow can be made arbitrary small by making the buffer size sufficiently large [19].

Scheduling in broadcast channels has been also considered by several authors [20], [21], [22], [23], [12], [14], [13], [15], [16]. In [21], stabilizing parallel queues in the transmitter is considered, where the connectivity of queues are random to capture deep fades in the wireless channel. In [23], the authors incorporate the channel state information in their scheduling while providing delay constraints for packets. Analyzing the average delay (over the users) can be also done using the results for the general independent input/output (GI/GI/1) queues and it can be shown that the average delay is of the order of the number users [24], [25]. However, in order to provide delay guarantee for all users, we have to study the delay for *the most unfortunate user* in the system. Clearly the worst case delay is a function of the number of users and their SNRs (or the probability of being chosen as the best user at each channel use). While these works give many insights and algorithms, they leave open the question of how large the worst case delay is as a function of the number of users and their SNRs for using throughput optimal strategies. This is the main goal of this paper.

This paper is organized as follows. Section II introduces our channel model and our notation. Section III deals with characterizing the delay for single antenna broadcast fading channels. Section IV generalizes the results of Section III to multi-antenna broadcast channels. Finally Section V proposes an algorithm to reduce the delay at the expense of a little reduction in the throughput and Section VI concludes the paper.

II. System Model and Assumptions

In this paper we consider a single antenna broadcast channel with n receivers. We assume a block fading channel with a coherence interval of T, and where the channel changes independently after T seconds. The transmission is assumed to be packet based and the length of each packet is T [2].

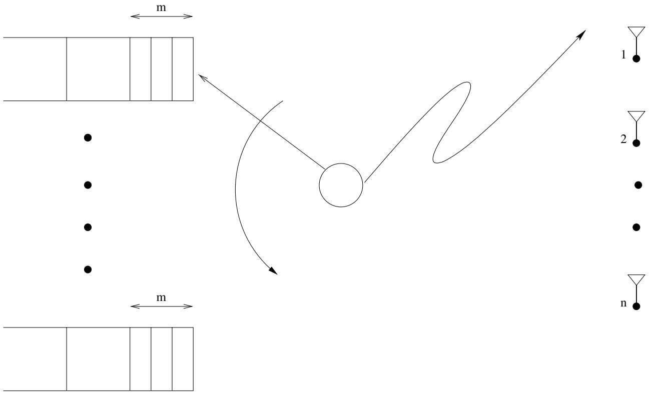

Fig. 1. n parallel queues in the transmitter corresponding to n users; we are interested in the behavior of the longest queue.

For each block of length T, the received signal at the i'th user at time t can be written as,

$$y_i(t) = \sqrt{\rho_i} h_i(t) S(t) + n_i(t), \qquad i = 1, \ldots, n, \tag{1}$$

where $h_i(t)$ is the effect of channel and $n_i(t)$ is additive white noise and that both are i.i.d. circularly symmetric complex Gaussian distributed with zero mean and variance of one. Here ρ_i is the SNR of the i'th user and $S(t)$ is the transmitted symbol at time t. We further assume independent memoryless channel which implies that the channel changes independently to another value after the coherence interval of T.

In the transmitter we assume there are n queues corresponding to each receiver and receiver's messages are independent[3]. For most of our analysis, we will assume that there is always a packet available to be transmitted to any user (i.e., backlogged users) [4] . Fig. 1 illustrates the arrangement of queues in the transmitter. In fact, the main challenges for the scheduler are to first balance the service among all the users and to second exploit the multiuser diversity in the channel in order to maximize the throughput of the system. Any scheduling strategy implies a probability for choosing each user at each channel use that may depend on the signal to noise ratio (SNR) of all users, the length of the queue of users, and the statistics of the channel (see [14], [15], [16]). For the throughput optimal strategy, this probability only depends on the SNR of the user and the channel statistics. For i.i.d channels, it is clear that these probabilities are only functions of users' SNRs.

Assuming that all packets have C_0 information bits for a homogeneous network (i.e., $\rho_i = \rho$), we consider a packet to be dropped if outage occurs, i.e., if the instantaneous capacity C goes below C_0 at the time of the transmission [26]. The

[2] If the length of the packet is smaller than T, the results in this paper can be easily generalized.

[3] Broadcast channels, in full generality, include transmission of common messages between receivers. Here we consider the special scenario in which the transmitter is sending independent messages to the receivers.

[4] In most practical situations, packets have finite arrival rates and so this assumption may not be valid. In section 3.1, we show how our result can be extended to the non-backlogged case.

instantaneous capacity however depends on the scheduling. For the round-robin scheduling, $C = \log(1 + \rho|h_i|^2)$ which does not depend on n. For the throughput optimal strategy [5], C however is the maximum of $\log(1 + \rho|h_i|^2)$ over $1 \leq i \leq n$, i.e., $C = \max_{1 \leq i \leq n} \log(1 + \rho|h_i|^2)$. We assume if a packet is dropped, the transmitter will be notified and the packet will be considered for re-transmission whenever the corresponding user has the best channel conditions.

If we assume that the error probability is simply the outage probability (a reasonable assumption for long packets [4]), we have $P_e = \Pr(C < C_0)$. The throughput is therefore $R = C_0(1 - P_e) = C_0 \Pr(C \geq C_0)$. Given P_e, any scheduling would lead to a different C_0. Note that for any value of C_0, the throughput optimal strategy is to send to the best user as this would minimize P_e. Conversely, for any fixed value of P_e, sending to the strongest user maximizes the throughput as this would allow for the largest possible C_0. It is also worth mentioning that the maximum of n i.i.d. exponential random variables (the $|h_i|^2$) behaves almost surely as $\log n$. Therefore for large n, we do not need to use power control to compensate for the channel variation as the maximization automatically prevents having deep fades for large number of users with high probability. Thus, for the throughput optimal scheduling, it is reasonable to assume that all the packets have the same amount of information, i.e., C_0 roughly about $\log(1+\rho \log n)$, independent of the time and channel condition.

In this paper, we define the (packet) delay in the broadcast channel as the number of channel uses (denoted by $D_{m,n}$) required to guarantee that all the users will receive m packets successfully. It is clear from the definition of $D_{m,n}$ that this notion of delay refers to the worst case delay among users (or the delay for the most unfortunate user). Of course, $D_{m,n}$ is a random variable and depends on the number of users n, the number of packets m and also the scheduling algorithm. A delay-optimal strategy is round-robin scheduling which clearly achieves the optimal (packet) delay of mn (when there is no error). However, round-robin is not throughput optimal which requires transmitting to the user with the best channel conditions at each channel use. Throughput optimal strategies, on the other hand, will have to contend with delay hits. The following section deals with the delay for the throughput optimal scheduling.

It should be also mentioned that our definition of the delay with backlogged users suffers from the weakness that it does not account for the queueing delay. However, our delay is a lower bound for the overall worst case delay in a system with random arrivals. The lower bound should be also tight when the system is highly loaded.

III. Delay Analysis for Single-Antenna Broadcast Channels

Opportunistic transmission is a probabilistic scheduling which implies that each user will be given service with some given probability. Assuming that the outage probability P_e is given, the opportunistic scheduling maximizes the throughput or equivalently C_0 (the amount information bits per packet). Analyzing the average delay (over all the users) can be done as the queue of each user can be considered as an i.i.d. input/output queue [24]. In particular, it can be shown that the average delay is of the order n [25]. However analyzing the worst case delay (or the delay for the most unfortunate user in the system) requires considering n parallel queues of n users all-together [27]. In this section, assuming that at each channel use the transmitter sends to the i'th user with the probability p_i, which only depends on the SNR of all users, and drops the packet with probability P_e, we obtain the moment generating function of the random variable $D_{m,n}$.

We first consider the simple case in which the network is homogeneous and $P_e = 0$. Then we generalize the result to the case where we have a non-zero P_e and/or a heterogeneous network where users are chosen with different probabilities. We obtain the mean and variance of the delay $D_{m,n}$ for any m and n. We further look into the asymptotic behavior of $D_{m,n}$ for different regions of m and n at the end of this section.

A. A Study of the Delay for Users with Poisson Arrival

Before delving into an analysis of $D_{m,n}$ for the backlogged case, let us remark on the more realistic case where we have a poisson arrive for the packets with fixed rate $\lambda \leq \frac{1}{n}$. In this case, there is a non-zero probability that the user with the best channel condition has an empty queue. Two courses of action can be taken: one is to not transmit anything, the other is to transmit to the user with the best channel condition whose channel is non-empty. The latter is a more reasonable action, but seems very difficult to analyze.

In this section, we study the effect of having random arrivals for each queue and find the delay incurred by the scheduling in which no transmission is done if the chosen user has no packet. In order to analyze the delay, we would need to find the probability of having no packet at each queue in the steady state.

Each queue has a poisson arrival process with intensity λ and the service has a binomial distribution, i.e., with probability $\frac{1}{n}$ the queue will be served at each time slot. Therefore the characteristic function for the length of time that the queue has not been served can be written as,

$$S(z) = \sum_{i=1}^{\infty} z^i (1 - 1/n)^{i-1} 1/n = \frac{z/n}{1 - z(1 - 1/n)}. \quad (2)$$

Therefore using known results for the M/G/1 queue, the moment generating function for the random variable N denoting the number of packets in the queue can be written as,

$$G_N(z) = \frac{(1 - \lambda(n-1))(1 - z)}{1 - \frac{z}{S((1-z)\lambda)}} \quad (3)$$

where $S(z)$ is as defined in (2). Thus, the probability of having an empty queue is $G_N(0) = 1 - \lambda(n - 1)$. It is worth noting that in order to have all the n queues in the system to be stable, $\lambda \leq \frac{1}{n}$.

Now assuming that the base station will not transmit any packet if the selected queue is empty, we can easily find the expected delay using the same trick as we used to analyze the probability of dropping a packet. In particular, we may assume that there is a probability of $1 - \lambda(n - 1)$ that the packet is

[5] In this paper, we use the terms opportunistic scheduling and throughput optimal strategy interchangeably.

being dropped. Therefore, the expected delay will be $\frac{1}{\lambda(n-1)}$ times more than the delay for the case of backlogged users.

However if we choose to transmit to the strongest user with a non-empty queue, then the analysis becomes quite formidable. The above result, however, is a simple upper bound for the delay in this case. It is also clear that the delay for the backlogged system can be served as a lower bound for the delay in a more realistic setting where users have random arrivals. In fact upper and lower bounds are tight as the system becomes highly loaded.

B. Homogeneous Network with No Dropping Probability

When users are homogeneous and assuming throughput optimal scheduling, the transmitter chooses the i'th user with probability $\frac{1}{n}$ from the pool of n users since it is equally likely for each user to have the best channel condition. The random variable $D_{m,n}$ is basically the minimum number of channel uses to guarantee all n users have been chosen at least m times.

This problem can be restated as the coupon collector problem [28] which is studied by several authors in the mathematics literature (see also chapter 6 of [29]). To be more precise, users can be seen as people carrying coupons and the transmitter is the collector that chooses randomly and uniformly from the n people and collects his/her coupon. The question is how many times should the collector choose to guarantee that everybody has given at least m coupons. In fact we can state the mean value of $D_{m,n}$ based on a result found in [30].

Theorem 1. *(Newman and Shepp [30]) Consider a homogeneous broadcast system with n users. We assume that at each channel use, the transmitter sends to the user with the best channel condition. Then, we have,*

$$E(D_{m,n}) = n\int_0^{\infty}\left(1-\left(1-S_m(t)e^{-t}\right)^n\right)dt, \qquad (4)$$

for any m and n where $S_m(t)=\sum_{k=0}^{m-1}\frac{t^k}{k!}$.

Proof: Since the network is homogeneous, the probability of choosing the i'th users is $\frac{1}{n}$. Therefore, the problem is the same as the problem considered by Newman and Shepp [30]. See [30] for the proof. ■

Inspired by the proof of Theorem 1, we can derive the moment generating function of $D_{m,n}$ defined as

$$F(z)=\sum_{i=0}^{\infty}z^i\Pr\{D_{m,n}>i\}=\sum_{i=0}^{\infty}z^ib_i. \qquad (5)$$

Using the generating function $F(z)$ in (5), we can obtain all the moments of $D_{m,n}$ with a little effort and by taking higher derivatives of $F(z)$ at $z=1$ [31]. For example, using the definition of $F(z)$ in (5), we can write,

$$\begin{aligned} E(D_{m,n}) &= F(1) \\ \sigma^2(D_{m,n}) &= 2F'(1)+F(1)-(F(1))^2. \end{aligned} \qquad (6)$$

Next Theorem obtains $F(z)$ and generalizes the result of Theorem 1.

Theorem 2. *Considering the setting of Theorem 1, we can write the moment generating function of $D_{m,n}$ defined in (5) as,*

$$F(z)=\frac{n}{z}\int_0^{\infty}e^{-\frac{n}{z}t}\left(e^{nt}-(e^t-S_m(t))^n\right)dt. \qquad (7)$$

Proof: We evaluate $F(z)$ by the same trick as [30] in which the mean of $D_{m,n}$ is derived. In fact, $F(z)$ can be evaluated by noting that b_i is the probability of failure in obtaining m packets at all the n users up to and including the i'th trial. Therefore, b_i is simply the polynomial $(\frac{1}{n}x_1+\ldots+\frac{1}{n}x_n)^i$ evaluated at $x_1=\ldots=x_n=1$ after excluding all terms which have all x_i's with exponent larger than $m-1$. Therefore, we may write

$$F(z)=\sum_{i=0}^{\infty}z^i\frac{\left\{(x_1+\ldots+x_n)^i\right\}}{n^i} \qquad (8)$$

where $\{\cdot\}$ denotes the operator that removes all the terms which have all x_i's with exponent less than $m-1$. Considering the following identities [30],

$$\begin{aligned} \frac{z^i i!}{n^i} &= \frac{n}{z}\int_0^{\infty}e^{-\frac{n}{z}t}t^i dt, & (9)\\ \left\{e^{x_1+\ldots+x_n}\right\} &= \sum_{i=0}^{\infty}\frac{\left\{(x_1+\ldots+x_n)^i\right\}}{i!} \\ &= e^{x_1+\ldots+x_n}-\prod_{i=1}^{n}\left(e^{x_i}-S_m(x_i)\right), & (10) \end{aligned}$$

where the first equality in Eq. (10) is the definition of the exponential function and the second equality follows by noting that the second term in the right hand side just subtracts out the terms with all x_i's larger than m. We may then replace the integral form for $\frac{1}{n^i}$ using (9) in (8) to get,

$$\begin{aligned} F(z) &= \sum_{i=0}^{\infty}\int_0^{\infty}\frac{n}{z}e^{-\frac{n}{z}t}t^i dt\times\frac{\left\{(x_1+\ldots+x_n)^i\right\}}{i!} \\ &= \frac{n}{z}\int_0^{\infty}e^{-\frac{n}{z}t}\sum_{i=0}^{\infty}\frac{\left\{(x_1+\ldots+x_n)^i\right\}}{i!}dt \\ &= \frac{n}{z}\int_0^{\infty}e^{-\frac{n}{z}t}\left(e^{tx_1+\ldots+tx_n}-\prod_{i=0}^{n}\left(e^{tx_i}-S_m(tx_i)\right)\right)dt \\ &= \frac{n}{z}\int_0^{\infty}e^{-\frac{n}{z}t}\left(e^{nt}-(e^t-S_m(t))^n\right)dt. \qquad (11) \end{aligned}$$

where we replaced $x_i=1$ for $i=1,\ldots,n$ and we used (10) to get the second equality and we replaced $x_i=1$ for $i=1,\ldots,n$ to obtain the last equation. ■

It is now quite straightforward to derive the variance of $D_{m,n}$ using $F(z)$ and (6) as shown in (12) on the next page.

C. Heterogeneous Network with Dropping Probability

For the special case of a homogeneous network, we derived the moment generating function of $D_{m,n}$ in Theorem 2. In what follows, we generalize the results to a more general setting in which users may have different SNRs and also a packet may be dropped if outage occurs. We assume the transmitter will be notified in case a packet is dropped and it will be considered for re-transmission whenever the corresponding user has the best SNR. Here, we assume a

$$\sigma^2(D_{m,n}) = 2n^2\int_0^\infty t\left(1-\left(1-S_m(t)e^{-t}\right)^n\right)dt - E(D_{m,n}) - (E(D_{m,n}))^2 \tag{12}$$

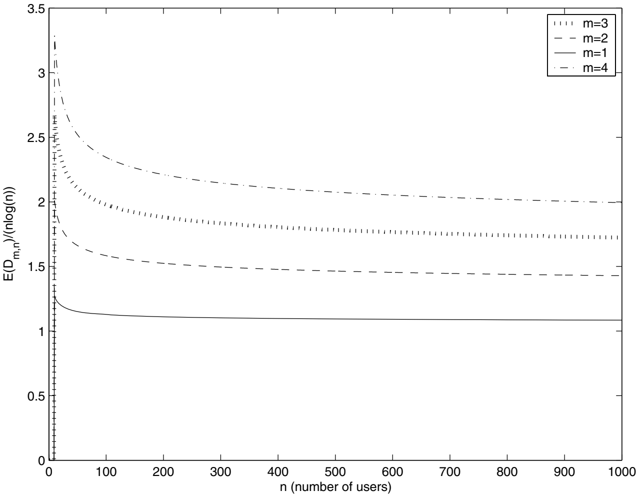

Fig. 2. Expected delay $\frac{E(D_{m,n})}{n\log n}$ for different values of m and n.

memoryless i.i.d. channel and that the transmitter chooses the i'th user with probability p_i that depends on the SNR of all users and their channel conditions for the throughput optimal strategy. Assuming that all the packets have the same length, the packet for the i'th user is dropped with probability of P_{e_i}.

The following Theorem states the mean and variance of $D_{m,n}$ for this general setting and for any m and n. The Theorem is a generalization of the result of Newman and Shepp [30] stated in Theorem 1.

Theorem 3. *Suppose we have n users such that the probability of choosing the i'th user is $p_i = \frac{\alpha_i}{n}$ and the probability of dropping a packet is P_{e_i}. Then the moment generating function for $D_{m,n}$ defined in (5) is,*

$$F(z) = \frac{n}{z}\int_0^\infty e^{-\frac{n}{z}t}\left(e^{nt} - e^{\sum_{i=1}^n P_{e_i}t}\prod_{i=1}^n\left(e^{t\beta_i} - S_m(t\beta_i)\right)\right)dt, \tag{13}$$

where $\beta_i = (1-P_{e_i})\alpha_i$. In particular, assuming $S_m(t)$ is as defined in Theorem 1, we have

$$E(D_{m,n}) = n\int_0^\infty\left(1-\prod_{i=1}^n\left(1-S_m(\beta_i t)e^{-\beta_i t}\right)\right)dt, \tag{14}$$

and (15) on the next page.

Proof: The proof is a generalization of Theorem 2 and we omit it for the sake of brevity. ∎

For example, as a simple consequence of (14), we can obtain the expected delay for the case where n users are equally likely and that the probability of dropping a packet is P_e, as

$$E(D_{m,n}) = \frac{1}{1-P_e}(n+1)\int_0^\infty\left(1-\left(1-S_m(x)e^{-x}\right)^n\right)dx, \tag{16}$$

by a simple change of variable in the integral stated in (14).

Fig. 2 shows the expected delay for $m = 1, 2, 3, 4$ and for different number of users for a homogeneous network. It is clear that when n is large and $m = 1$, the growth in the expected delay is like $n\log n$. Also Fig. 2 implies that the expected delay does not grow linearly with m (for small values of m). In fact it converges to $n\log n$ although the convergence seems to be quite slow. The next subsection deals with the asymptotic analysis of the delay for different regions of m and n.

Remark 1: It is worth mentioning that we can consider the delay in sending m_i packets to the i'th user for $i = 1, \ldots, n$. In particular, considering the setting of Theorem 3, and we are interested in sending m_j packets to the j'th user for $j = 1, \ldots, i$ where $i \leq n$. Defining $\mathbf{m} = (m_1, \ldots, m_i)$ and $D_{\mathbf{m}}$ as the minimum number of channel uses guarantees the receive of m_j packets at the j'the user for $j = 1, \ldots, i$, we can write the moment generating function for $D_{\mathbf{m}}$ as shown in (17) on the next page.

D. Asymptotic Analysis of the Moments of $D_{m,n}$

In the previous subsection, we obtained the moments of $D_{m,n}$ for a general setting and for any m and n in closed form. However, it is hard to speculate how the mean and variance of the delay behave as functions of m and n. In order to get a better insight into the behavior of the delay, we derive some asymptotic results for the moments of $D_{m,n}$ and for different regions of m and n.

Theorem 4. *Assuming a homogeneous network and that a packet will be dropped with probability P_e,*

1) *For m fixed and $n \to \infty$, we have (18) and (19) on the next page.* [6]
2) *For $m = \log n$ and $n \to \infty$, we have*

$$E(D_{m,n}) = \alpha\frac{1}{1-P_e}n\log n + O(n\log\log n). \tag{20}$$

where $\alpha = 3.146$ is the solution to the equation $\alpha - \log\alpha = 2$.

3) *For $m = (\log n)^r$ where $r > 1$ is fixed and $n \to \infty$, then*

$$\begin{aligned} E(D_{m,n}) &= \frac{1}{1-P_e}n(\log n)^r + o(n(\log n)^r) \\ &= \frac{1}{1-P_e}mn + o(mn). \end{aligned} \tag{21}$$

4) *For n fixed and $m \to \infty$,*

$$E(D_{m,n}) = \frac{1}{1-P_e}nm + o(m). \tag{22}$$

Proof: Here we present the sketch of the proof for the first part and omit the proof for the other cases for the sake of brevity. The interested reader can refer to [31] for the complete proofs.

[6] This case has been also proved in [30], however we present other proof which leads to results for another regions of m and n as well.

$$\sigma^2(D_{m,n}) = 2n^2 \int_0^\infty t \left(1 - \prod_{i=1}^{n} \left(1 - S_m(\beta_i t) e^{-\beta_i t}\right)\right) dt - E(D_{m,n}) - \left(E(D_{m,n})\right)^2 \tag{15}$$

$$F(z) = \sum_{i=0}^{\infty} z^i \mathrm{Pr}(D_{\mathbf{m}} > i) = \frac{n}{z} \int_0^\infty e^{-\frac{n}{z}t} \left(e^{nt} - e^{nP_e t + \sum_{k=i+1}^{n} \beta_k t} \prod_{p=1}^{i} (e^{t\beta_p} - S_{m_p}(t\beta_p)) \right) dt \tag{17}$$

$$\begin{aligned} E(D_{m,n}) &= \frac{1}{1-P_e} n \log n + n(m-1) \log\log n + o(n \log\log n) && (18) \\ \sigma^2(D_{m,n}) &= O(n^2) && (19) \end{aligned}$$

$$E\{\max_{1 \le i \le n} x_i\} = \int_0^\infty x f_{max}(x) dx = \int_0^\infty (1 - F_{max}(x)) dx = \int_0^\infty (1 - F^n(x)) dx \tag{23}$$

Noting that the expected value of $D_{m,n}$ is equal to (16), we first show that the integral in (16) is in fact proportional to the expected value of the maximum of n i.i.d. $\chi^2(2m)$ random variables. To prove that, we assume x_i's for $i = 1, \ldots, n$ are i.i.d. random variables with $\chi^2(2m)$ distribution. We can then write the expected value of the maximum of x_i's as shown in (23) on the next page, where $f_{max}(x)$ and $F_{max}(x)$ are probability distribution and cumulative distribution functions (CDF) of the maximum of x_i's and $F(x)$ is the CDF of x_i. We further know that x_i's are i.i.d. and have $\chi^2(2m)$ distribution and therefore their CDF is the incomplete gamma function and can be written as $F(x) = 1 - S_m(x)e^{-x}$. Therefore, we may write (23) as shown in (24). Therefore to analyze the mean of $D_{m,n}$, we investigate the behavior of the maximum of x_i's. In [32], it is shown that for m fixed, $\max\limits_{1 \le i \le n} x_i$ behaves like with high probability. This would then lead to the result for $E\{D_{m,n}\}$ for large n and fixed m. See [31] for the precise argument.

To obtain the variance, we first note that $D_{m,n} \le mD_{1,n}$ which is clear from the definition of $D_{m,n}$. Now we derive the variance of $D_{1,n}$ and, since m is fixed, the variance of $D_{m,n}$ has the same order. Denote by r_i, for $i = 1, \ldots, n$, the number of transmissions after transmitting at least one packet to $i-1$ users and before i users receive their first packet. Clearly r_i's are independent and have geometric distribution, i.e., $\Pr\{r_i = k\} = \left(\frac{i-1}{n}\right)^{k-1}\left(1 - \frac{i-1}{n}\right)$. The distribution of r_i is obtained by noting that r_i equals k if in the last $k-1$ trials the packet is transmitted to the $i-1$ users that have already been chosen and then in the k'th channel use, one user will be transmitted to from the pool of $n-i+1$ users that have already been chosen.

Using the definition of $D_{1,n}$ and r_i's, it is clear that $D_{1,n} = \sum_{i=1}^{n} r_i$ and therefore the variance of $D_{1,n}$ can be written as,

$$\sigma^2_{D_{1,n}} = n^2 \sum_{i=1}^{n} \frac{1}{i^2} - n \sum_{i=1}^{n} \frac{1}{i}. \tag{25}$$

It is quite straightforward to prove that the first term in the right hand side of (25) behaves like $O(n^2)$ and the second term behaves like $n \log n$. Therefore the variance of $D_{m,n}$ can be written as $\sigma^2_{D_{m,n}} \le m^2 \sigma^2_{D_{1,n}} = O(n^2)$.

In order the prove the other cases, we need to investigate the behavior of the maximum of n i.i.d. $\chi^2(2m)$ random variables when m for large n and when m also grows. We refer the reader to [31] for the proofs. ■

Assuming $m = 1$ and using the result of Theorem 4, we can state that the delay converges to the mean almost surely using Chebychev's inequality as shown in (26) for large n. This implies that the delay hit for sending the first packet successfully to all the users is increased from the minimum of n for the round robin scheduling to $n \log n$ for the opportunistic transmission for large n. So the delay degradation due to exploiting the channel variation and maximizing the throughput of the system is a multiplicative factor of $\log n$. It would be also interesting to investigate the scaling law of the variance of $D_{m,n}$ when m also grow to infinity; this would then imply the type of convergence to the mean for different regions of m and n.

Remark 2: For a homogeneous network, as opportunistic transmission is long term fair (i.e. the probability of choosing all the users is the same), we know that for sufficiently large m, the expected delay should behave like mn. This is confirmed by the fourth part of Theorem 4. Interestingly, Theorem 4 further implies that if m grows faster than $(\log n)^r$ where r is fixed and greater than one the expected delay behaves like mn. This has implications for the time scale after which the system behaves fairly. Moreover, if m grows logarithmically with n, the expected delay is only off by a constant factor of $\alpha = 3.14$, compared to the minimum delay mn. Therefore, our result can be seen as the short term behavior of the delay for any m.

As mentioned, the largest delay hit is when we focus on sending a few packets, i.e. $m = 1$ or m is small. The delay hit gets less when we focus on sending more and more packets (i.e., when m gets larger). Therefore, in the rest of the paper, we mainly focus on the delay for sending the first packet, i.e. $D_{1,n}$.

IV. Delay in Multi Antenna Broadcast Channels

Multiple transmit antennas have been shown to significantly improve the throughput of a broadcast channel. It is shown that dirty-paper coding achieves the sum rate capacity

$$E(D_{m,n}) = \frac{n+1}{1-P_e} E\{\max_{1\leq i\leq n} x_i\} = \frac{n+1}{1-P_e}\int_0^\infty \left(1-(1-S_m(x)e^{-x})^n\right) dx. \tag{24}$$

$$\Pr\left\{|D_{m,n} - \frac{1}{1-P_e} n\log n + O(n\log\log n)| \leq n\sqrt{\log n}\right\} \geq 1-\frac{1}{\log n}, \tag{26}$$

of a Gaussian broadcast channel [33], [34], [35]. However, beamforming has long been proposed as a heuristic method to mitigate the interference in the transmitter and to send multiple beams to different users. Although, beamforming is not optimal in achieving the sum rate capacity, its throughput does scale the same as that of dirty paper coding for a system with many users and has much less complexity than that of dirty paper coding [36], [37].

In this paper, for a system with M transmit antennas, we assume a simple model in which the base station transmits to M different receivers at each channel use. This is certainly a valid model for beamforming or channel inversion, though it does not fit the dirty paper scheduling in which the transmitter sends information to all the users at each time. However, as far as the scaling law of the sum rate throughput is concerned, when M is either fixed or growing logarithmically with n, it can be shown that beamforming, channel inversion, and random beamforming all give the optimal scaling law for the sum rate throughput [32].

For a homogeneous network, our model for the multiple antenna transmitter implies that, at each channel use, the transmitter sends to M *different* users uniformly chosen from the pool of n users (see [32]). In this scheduling the transmitter sends M beams each one is assigned to the user with the best signal-to-noise and interference ratio (SINR) for the corresponding beam. As shown in [32], the best SINR behaves like $\log n$ with high probability for large n. Therefore, we may again assume that each packet carries a fix amount information (roughly about $\log(1+\rho\log n)$).

This scheduling is certainly more balanced compared to the case where we have a single antenna system that works M times faster. This can be justified by noticing the fact that we exclude the possibility of sending to one user twice (or more) in each block of M transmissions and hence the scheduling is more balanced. In particular, assuming that there is no packet dropped as in Theorem 1. Then, we have,

$$D_{m,n}(M) \leq \frac{1}{M} D_{m,n} \tag{27}$$

where $D_{m,n}(M)$ is the delay for sending m packets successfully to n users in an M-transmit antenna system and where $D_{m,n}$ is the delay for a single antenna broadcast system as in Theorem 1.

In fact we can compute exactly the expected delay in transmitting the first packet successfully, i.e. $E\left(D_{1,n}(M)\right)$, for any n and M. Further generalization of the result to $m>1$ is non trivial and we have not been able to do this; however, it is quite easy to show that $D_{m,n}(M) \leq mD_{1,n}(M)$. The next theorem presents the result for $m=1$ and for any n and M.

Theorem 5. *Consider a broadcast channel with M transmit antennas and n users. Assuming that no packet is dropped, we can write the expected delay in sending one packet to all users for any m and n as,*

$$E\left(D_{1,n}(M)\right) = \sum_{k=0}^{\infty}\sum_{r=1}^{n}\sum_{i=0}^{n-r}(-1)^{n-r-i}\frac{\binom{n}{r}}{\binom{n}{M}^k}\binom{n-r}{i}\binom{i}{M}^k. \tag{28}$$

Proof: Similar to the proof of Theorem 3, we first note that the mean of $D_{1,n}(M)$ can be written as,

$$E\left(D_{1,n}(M)\right) = \sum_{k=0}^{\infty}\Pr\left(D_{1,n}(M) > k\right) \tag{29}$$

In order to compute the probability of $D_{1,n} > k$, we define the auxiliary random variable $\mu_n^M(k)$ as the number of users that have received no packets after k channel uses in which the transmitter sends to M different users. From the definition of μ_n^M, it is clear that $\mu_n^M \leq n$ and that $D_{1,n}(M) > k$ is equivalent to $\mu_n^M(k) > 0$. Therefore, Eq. (29) can be written as,

$$\begin{aligned} E\left(D_{1,n}(M)\right) &= \sum_{k=0}^{\infty}\Pr\left(\mu_n^M(k) > 0\right) \\ &= \sum_{k=0}^{\infty}\sum_{r=1}^{n}\Pr\left(\mu_n^M(k) = r\right) \end{aligned} \tag{30}$$

The probability that $\mu_n^M(k) = r$ can be computed as follows. Assuming $\mu_n^M(k) = r$ implies that *only* $n-r$ users have received at least one packet in k channel uses. We then define the event S_i for $i = 0, 1, \ldots, n-r$ as the event that at least $n-r-i$ users have not received any packets among $n-r$ users that are supposed to receive a packet. This implies that there are at most i users that the transmitter sends packets to. It is clear that for $1 \leq i \leq M$ probability of S_i is zero, since the transmitter certainly can transmit to M different users at each channel use. For $i > M$, however we can write the probability of S_i as

$$\Pr\{S_i\} = \binom{n}{r,i}\frac{\binom{i}{M}^k}{\binom{n}{M}^k} = \binom{n}{r}\binom{n-r}{i}\frac{\binom{i}{M}^k}{\binom{n}{M}^k}$$
$$i = 0, 1, \ldots, n-r. \tag{31}$$

where we first chose two sets of users with cardinality r and i from the set of n users and then we distributed packets among i of them k times by choosing M different users at each time.

Considering the definition of $\mu_n^M(k) = r$ and the S_i's, we can use the inclusion-exclusion principle (see chapter 4 of [28]) to obtain (32). Substituting (32) in (30), we can write the expected delay as (33). This completes the proof for the Theorem. ∎

$$\begin{aligned}\Pr\left(\mu_n^M(k)=r\right) &= \Pr(S_{n-r})-\Pr(S_{n-r-1})+\ldots+\Pr(S_0)\\ &= \sum_{i=0}^{n-r}(-1)^{n-r-i}\Pr(S_i)=\frac{\binom{n}{r}}{\binom{n}{M}^k}\sum_{i=M}^{n-r}(-1)^{n-r-i}\binom{n-r}{i}\binom{i}{M}^k \end{aligned}\tag{32}$$

$$\begin{aligned}E\left(D_{1,n}(M)\right) &= \sum_{k=0}^{\infty}\Pr\left(\mu_n^M(k)>0\right)\\ &= \sum_{k=0}^{\infty}\sum_{r=1}^{n}\frac{\binom{n}{r}}{\binom{n}{M}^k}\sum_{i=M}^{n-r}(-1)^{n-r-i}\binom{n-r}{i}\binom{i}{M}^k\end{aligned}\tag{33}$$

Remark 3: It is worth mentioning that we can also obtain the generating function $F(z)$ that would lead to the moments of $D_{1,n}(M)$ for any M and n. In fact, $F(z)$ is equal to

$$\begin{aligned}F(z) &= \sum_{k=0}^{\infty}z^k\Pr(D_{1,n}(M)>k)\\ &= \sum_{k=0}^{\infty}\sum_{r=1}^{n}\sum_{i=0}^{n-r}(-1)^{n-r-i}\frac{z^k\binom{n}{r}}{\binom{n}{M}^k}\binom{n-r}{i}\binom{i}{M}^k\end{aligned}\tag{34}$$

Using (6) and (34), we can easily obtain the variance (and other moments) of $D_{1,n}(M)$.

Although Theorem 5 gives us the exact value of the expected delay for any number of users, it does not make clear how much improvement on the delay we can get in using multi-antenna transmitter over that of the single antenna system. We can in fact asymptotically analyze the expected delay derived in Theorem 5 for large number of users to get a better intuition about this result.

Theorem 6. *Consider the setting of Theorem 5. Then the expected delay in sending at least one packet to all n users using an M-antenna transmitter derived in (28) behaves like*

$$E\left(D_{1,n}(M)\right)=\frac{\sum_{k=1}^{n}\frac{1}{k}}{\sum_{r=0}^{M-1}\frac{1}{n-r}}+O(1).\tag{35}$$

for large n and when M grows no faster than $\log n$.

Proof: The proof is quite involved and we omit it due to lack of space. The interested reader is referred to [31] for the proof. ∎

For the special case of $M=1$, the problem reduces to the coupon collector problem when $m=1$ (one packet). It can be easily shown that the expected delay is equal to $n\sum_{i=1}^{n}\frac{1}{i}\approx n\log n$. Clearly the result of Theorem 5 confirms this result for one transmit antenna, i.e. $M=1$.

Remark 4: As mentioned in (27), using multiple transmit antennas in the transmitter should improve the delay. We may write the improvement on the expected delay by using M transmit antennas over that of single antenna case as shown in (36). Eq. (36) implies that when M is not growing faster than $\log n$, the gain in delay is a factor of M which comes from the fact that we are transmitting packets M times faster. Therefore, multiple transmit antenna systems incur pretty much the same delay as that of a single antenna transmitter that operates M times faster when there is no channel correlation.

Although the gain on delay in using multiple transmit antennas is not that much, multiple transmit antennas can significantly improve the long term fairness in a heterogeneous network. More precisely, in [32], it is proves that if M grows logarithmically with the number of users, the probability of choosing each user become independent of its SNR and approaches to $\frac{1}{n}$. Moreover, when there is channel correlation, multiple antenna systems can significantly reduce the delay by "decorrelating in time" the effective channel through means such as random beamforming [32], [38].

V. Trading Delay with the Throughput: d-Algorithm

Previously, we showed the delay hit in using the optimal throughput scheduling is a $\log n$ fold increase compared to the minimum achievable delay. In this section, we propose an algorithm that can reduce the expected delay for sending the first packet at the price of a little throughput degradation. The goal is to improve the $\log n$ fold degradation in the delay without too much reducing the throughput of the system.

In order to improve the delay, we have to introduce more options to the scheduler at each channel use. For single antenna systems, this can be done by looking at the d best users in terms of capacity and transmit to the user among those d users that has received the least number of packets. We call this scheduling the *d-algorithm.* For a large number of users and fixed d, it is quite easy to show that the capacity of the best user and that of the d'th best user is quite close almost surely. This in fact guarantees that the throughput degradation using our algorithm is not that much. The next Theorem quantifies the performance of the d algorithm precisely.

Theorem 7. *Consider the setting of Theorem 1 and suppose the transmitter uses the d algorithm. We denote the expected delay in sending the first packet by $E(D_{1,n}^d)$. Then, for any d,*

$$E(D_{1,n}^d)=n\int_0^{1-\frac{d}{n}}\frac{1}{1-x^d}dx+O(1)\tag{37}$$

Asymptotically, we can further prove that if d is fixed,

$$\lim_{n\to\infty}\frac{E(D_{1,n}(d))}{E(D_{1,n})}=\lim_{n\to\infty}\frac{E(D_{1,n}^d)}{n\log n}=\frac{1}{d}.\tag{38}$$

Proof: In order to compute the expected delay, we again define the variable r_i as the number of channel uses after

$$\text{Gain on the expected delay with } M \text{ antenna transmitter} = \frac{1}{\sum_{r=0}^{M-1} \frac{n}{n-r}} = M + O\left(\frac{M^2}{n}\right). \tag{36}$$

sending at least one packet to $i-1$ users and before completing the transmission of at least one packet to i users. Clearly r_i has a Geometric distribution as,

$$\Pr(r_i = k) = (1-p_i)^{k-1} p_i \qquad k = 1, 2, \ldots \tag{39}$$

where p_i is the probability that all the d best users have been chosen before, therefore

$$\begin{aligned} p_i &= 0 \qquad 1 \le i \le d-1 \\ p_i &= 1 - \frac{\binom{i}{d}}{\binom{n}{d}}, \qquad d \le i \le n-1 \end{aligned} \tag{40}$$

Noting that $D_{1,n} = \sum_{i=0}^{n-1} r_i$, and also using the fact that the mean value of r_i is $\frac{1}{p_i}$, we can obtain the expected value of $D_{1,n}$ as

$$\begin{aligned} E(D_{1,n}^d) = \sum_{i=d}^{n-1} \frac{1}{p_i} &= \sum_{i=d}^{n-1} \frac{1}{1 - \frac{i(i-1)\ldots(i-d+1)}{n(n-1)\ldots(n-d+1)}} \\ &\le \sum_{i=d}^{n-1} \frac{1}{1 - \left(\frac{i-d+1}{n}\right)^d} \end{aligned} \tag{41}$$

where we used a simple upper bound for $\binom{i}{d}/\binom{n}{d}$. To evaluate the summation in the right hand side of (41), we may take integrals from $x = 1$ to $x = n - d + 1$ from both sides of

$$\frac{1}{1-(x/n)^d} \ge \frac{1}{1-(\lfloor x \rfloor/n)^d} \ge \frac{1}{1-((x-1)/n)^d}, \tag{42}$$

to obtain

$$E(D_{1,n}^d) = n \int_0^{1-d/n} \frac{dx}{1-x^d} + O(1), \tag{43}$$

which completes the proof for the first part of the Theorem. To prove the second part, we define the integral in the right hand side of (43) as $G(n)$. Then it is quite easy to show that when d is fixed, we have

$$\lim_{n\to\infty} \frac{G(n)}{\log n} = \lim_{n\to\infty} \frac{d}{n(1-(1-\frac{d}{n})^d)} = \frac{1}{d}. \tag{44}$$

where we used the L'Hopital's rule in (44). Considering that $E(D_{1,n})$ scales like $n \log n$ as proved in Theorem 4, the second part of the theorem immediately follows from (44). ■

Fig. 3 shows the delay improvement for different values of d and for different number of users. As d increases the delay improves though with less pace. Clearly, we can get most of the improvement by just checking the the best two users ($d = 2$) and further increasing d will not improve the expected delay as much as before.

There is of course a price to pay on the rate for the delay improvement. In order to see the throughput hit, we look into the ergodic throughput of the channel (denoted by $R(d)$) using the d algorithm defined as

$$R(d) = E \log\left(1 + \rho \max_{1\le i\le n}{}^k |h_i|^2\right) \tag{45}$$

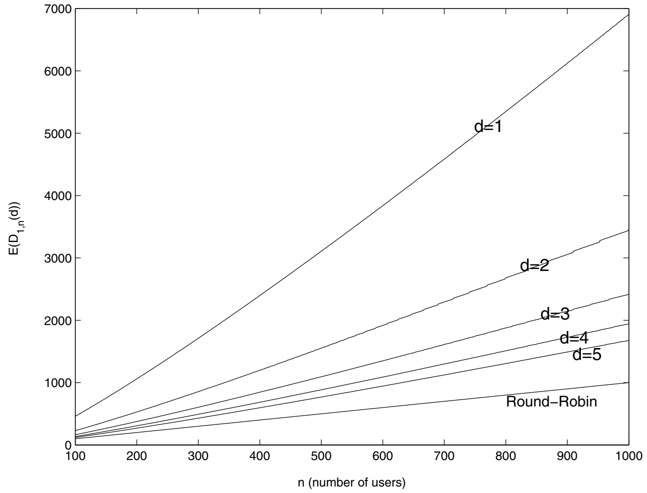

Fig. 3. Expected delay $E(D_{1,n}^d)$ for different values of d and n.

where $\max^k$ denotes the k'th maximum and k is a random variable uniformly distributed between 1 and d. Using results on the extreme value theory, it is quite straightforward to show that,

$$\lim_{n\to\infty} R(d) - R(1) = 0, \tag{46}$$

when d is fixed. The proof is based on the fact that is d is fixed, the first and the d best user both have SNR of about $logn$(see [39], [32]. Eq. (46) implies that in the limit of large n, the difference of the throughput of the d algorithm and the maximum throughput converges to zero.

Remark 5: It is worth mentioning that the transmitter may use a round-robin type scheduling and also exploits the channel. This can be done by sending to the best user among n users at the first channel use, and then sending to the best user among $n-1$ users that have not been chosen and so on. This method can make sure that the worst case delay is equal to n. The ergodic throughput of this scheme can be written as,

$$R_{RR} = E\left\{\frac{1}{n}\sum_{k=1}^{n} \log\left(1 + \rho \max_{1\le i\le k} |h_i|^2\right)\right\} \tag{47}$$

Assuming that the channel is Rayleigh fading, we can show that in the limit the ratio of R_{RR} over $R(1)$ is one. Of course, the convergence in (46) for d-algorithm holds in a stronger sense. Moreover, it is worth mentioning that this scheduling may require packets with different amount of information.

Remark 6: Another approach to trade the delay with throughput is to consider a threshold for the capacity and to send to the user that has received the least number of packets among the users with instantaneous capacity above the threshold value (C_{Th}). In this case, we basically have a random d that has a binomial distribution where the binomial parameter q depends on the threshold value C_{Th}. We can in fact bound the delay for sending one packet to all users using d algorithm as,

$$E(D_{1,n}) = E_d\{E\{D_{1,n}|d\}\} \leq E_d\left\{\sum_{l=1}^{n-d+1} \frac{1}{1-\left(\frac{l-1}{n}\right)^d}\right\}. \tag{48}$$

where d has binomial distribution with parameter $q = \Pr\{\log(1+\rho_i|h_i|^2) \geq C_{Th}\}$.

VI. Conclusion

Providing quality of service (QoS) and also maximizing the throughput in a cellular system are the main challenges that require designing the physical layer and multiple access layer together. In this paper, we consider the downlink of a cellular system (i.e., a broadcast channel) and we also consider a notion of worst case delay which is defined as the delay $D_{m,n}$ incurred in receiving m packets by *all* the n users in the system. Clearly this definition of the delay is stronger than the average delay and represents the worst case delay among the users. In order to maximize the throughput, the transmitter has to send a packet to the user with the best channel condition which increases the delay. The main goal of this paper is to analyze this delay increase.

Assuming a block fading i.i.d. channel and a single antenna broadcast system with n backlogged users, we derive the moment generating function of the delay for any m and n and for a general hetereogeouns network where a packet can be dropped if outage capacity occurs. We further discuss how our results can be extended to the non-backlogged case. Asymptotically, for a homogeneous network where the throughput optimal scheduling is long-term fair (i.e., the probability of choosing users are equal), the result implies that the average delay in sending one packet to all users behaves like $n\log n$ as opposed to n for a round robin scheduling. We also prove that when m grows like $(\log n)^r$, for some $r > 1$, then to the first order the delay scales as mn. This roughly determines the time-scale required for the system to behave fairly. We also look into the delay analysis for a system equipped with multiple transmit antennas. Finally we propose an algorithm that without sacrificing too much on the throughput can significantly improve the delay. The algorithm always considers the first d user with the best channel conditions and transmits to the one that has received the least number of packets.

There are still questions remain to be answered. For example, in the model we considered, all the users always have packets of equal size for transmission, it would be quite interesting to generalize the results to the case where each user have a random rate of arrival or different transmission rates and analyze the behavior of the length of the longest queue among n users.

References

[1] X. Qin and R. Berry, "Exploiting multiuser diversity for medium access control in wireless networks," in *Proc. of INFOCOM 2003*, pp. 1084–1094.

[2] S. Shamai and E. Telatar, "Some information theoretic aspects of decentralized power control in multiple access fading channels," in *Proc. Information Theory and Networking Workshop 1999*.

[3] I. Bettesh and S. Shamai, "Optimal power and rate control for fading channels," in *Proc. Veh. Tech. Conf. 2001*, pp. 1063–1067.

[4] G. Caire, G. Taricco, and E. Biglieri, "Optimum power allocation over fading channels," *IEEE Trans. Inf. Theory*, vol. 45, no. 5, pp. 1468–1489, July 1999.

[5] R. A. Berry and R. G. Gallager, "Communication over fading channels with delay constraints," *IEEE Trans. Inf. Theory*, vol. 48, no. 5, pp. 1135–1149, May 2002.

[6] N. Abramson, "The ALOHA systems-another alternative for computer communications," in *Proc. Fall Joint Comput. Conf. 1970*, pp. 281–285.

[7] R. Gallager, "A perspective on multiaccess channels," *IEEE Trans. Inf. Theory*, vol. 31, no. 3, pp. 124–142, Mar. 1985.

[8] A. Ephremides and B. Hajek, "Information theory and communication networks: an unconsummated union," *IEEE Trans. Inf. Theory*, vol. 44, no. 10, pp. 2416–2434, Oct. 1998.

[9] D. N. Tse and S. V. Hanly, "Multiaccess fading channels. I. polymatroid structure, optimal resource allocation and throughput capacities," *IEEE Trans. Inf. Theory*, vol. 44, no. 7, pp. 2796–2815, Nov. 1998.

[10] L. Tong, V. Naware, and P. Venkitasubramaniam, "Signal processing in random access: a cross layer perspective," *IEEE Signal Processing Mag.*, July 2004.

[11] M. J. Neely and E. Modiano, "Dynamic power allocation and routing of time-varying wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 23, no. 1, Jan. 2005.

[12] A. Ganti, E. Modiano, and J. Tsitsiklis, "Optimal transmission scheduling in symmetric communication models with intermittent connectivity," available at http://web.mit.edu/jnt/www/publ.html, 2004.

[13] A. Eryilmaz and R. Srikant, "Scheduling with Quality of Service Constraint over Rayleigh Fading Channels," in *Proc. IEEE Conference on Decision and Control 2003*, pp. 245–250.

[14] A. Stoylar and K. Ramanan, "Largest weighted delay first scheduling: large deviations and optimality," *Annals Applied Probability*, no. 11, pp. 1–48, Nov. 2001.

[15] S. Borst, "User level performance of channel aware scheduling algorithms in wireless data networks," in *Proc. INFOCOM 2003*.

[16] M. Agrawal and A. Puri, "Base station scheduling of requests with fixed deadlines," in *Proc. INFOCOM 2002*.

[17] S. Kumar and P. R. Kumar, "Performance bounds for queueing networks and scheduling policies," *IEEE Trans. Auto. Control*, vol. 39, no. 9, Aug. 1994.

[18] P. R. Kumar and S. Meyn, "Stability of queueing networks and scheduling policies," *IEEE Trans. Auto. Control*, vol. 40, no. 2, Feb. 1995.

[19] E. Yeh and A. S. Cohen, "Throughput and delay optimal resource allocation in multiaccess fading channels," in *Proc. IEEE ISIT 2003*, pp. 245–245.

[20] J. I. Capetanakis, "Tree algorithms for packet broadcast channels," *IEEE Trans. Inf. Theory*, vol. 25, no. 9, pp. 505–515, Sept. 1979.

[21] L. Tassiulas and A. Ephremides, "Dynamic server allocation to parallel queues with randomly varying connectivity," *IEEE Trans. Inf. Theory*, vol. 39, no. 2, Mar. 1993.

[22] A. Eryilmaz, R. Srikant, and J. Perkins, "Stable scheduling policies for broadcast channels," in *Proc. IEEE Inter. Symp. Info.*, July 2002, p. 382.

[23] M. Andrew, K. Kumaran, K. Ramanan, A. Stoylar, P. Whiting, and R. Vijaykumar, "Providing quality of service over a shared wireless link," *IEEE Commun. Mag.*, vol. 39, no. 2, pp. 246–251, Feb. 2001.

[24] J. F. Kingman, "Inequalities in the theory of queues," *J. Royal Statistical Society: Series B*, vol. 32, no. 1, pp. 102–110, Jan. 1970.

[25] M. J. Ferguson, "On the control. stability, and waiting time in a slotted ALOHA random access system," *IEEE Trans. Commun.*, vol. 23, no. 10, Oct. 1975.

[26] L. H. Ozarow, S. Shamai, and A. D. Wyner, "Information theoretic considerations for cellular mobile radio," *IEEE Trans. Veh. Technol.*, vol. 43, no. 2, pp. 359–378, May 1994.

[27] A. Ephremides and R. Zhu, "Delay analysis of interacting queues with an approximate model," *IEEE Trans. Commun.*, vol. 35, no. 2, Feb. 1987.

[28] W. Feller, *An Introduction to Probability Theory and its Applications*. John Wiley and Sons, Inc., 1967.

[29] N. L. Johnson and S. Kotz, *Urn Models and Their Application*. John Wiley and Sons, Inc., 1977.

[30] D. J. Newman and L. Shepp, "The double dixie cup problem," *Amer. Math. Monthly*, vol. 67, no. 1, pp. 58–61, Jan. 1960.

[31] M. Sharif and B. Hassibi, "Delay analysis of throughput optimal scheduling in broadcast fading channels," Technical Report, California Institute of Technology, available at www.its.caltech.edu/ masoud/delaybc.pdf, 2004.

[32] M. Sharif and B. Hassibi, "On the capacity of MIMO BC channel with partial side information," *IEEE Trans. Inf. Theory*, no. 2, pp. 506–523, Feb. 2005.

[33] P. Viswanath and D. N. Tse, "Sum capacity of the vector Gaussian broadcast channel and downlink-uplink duality," *IEEE Trans. Inf. Theory*, vol. 49, no. 8, pp. 1912–1921, Aug. 2003.

[34] G. Caire and S. Shamai, "On the achievable throughput of a multi-antenna Gaussian broadcast channel," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1691–1706, July 2003.

[35] S. Vishwanath, N. Jindal, and A. Goldsmith, "Duality, achievable rates and sum rate capacity of Gaussian MIMO broadcast channle," submitted to *IEEE Trans. Inf. Theory*, 2002.

[36] M. Sharif and B. Hassibi, "A comparison of time-sharing, DPC, and beamforming for MIMO broadcast channels with many users," in *Proc. International Symp. on Information Theory 2004*.

[37] Y. Xie and C. Georghiades, "Some results on the sum rate capacity of MIMO fading broadcast channel," in *Proc. Inter. Symp. in Advances in Wireless Commun. 2002*.

[38] P. Viswanath, D. N. Tse, and R. Laroia, "Opportunistic beamforming using dump antennas," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1277–1294, June 2002.

[39] M. R. Leadbetter, "Extreme value theory under weak mixing conditions," *Studies in Probability Theory, MAA Studies in MAthematics*, pp. 46–110, 1978.

Masoud Sharif received his Ph.D. in Electrical Engineering (2005) from California Institute of Technology. In 2005, he was a post-doctoral scholar in the EE department at Caltech. Since January 2006, he has been an assistant Professor at Boston University. Dr. Sharif was awarded the C.H. Wilts Prize in 2006 for best doctoral thesis in Electrical Engineering at Caltech. He is a member of the Center for Information and Systems Engineering at Boston University. His research interests include ad-hoc and sensor networks, multiple-user multiple-antenna communication channels, cross-layer design for wireless networks, and multi-user information theory. His recent research has focused on collaborative communication scheme in ad-hoc and sensor networks and the capacity of multiple antenna broadcast channels.

Babak Hassibi was born in Tehran, Iran, in 1967. He received the B.S. degree from the University of Tehran in 1989, and the M.S. and Ph.D. degrees from Stanford University in 1993 and 1996, respectively, all in electrical engineering. From October 1996 to October 1998 he was a research associate at the Information Systems Laboratory, Stanford University, and from November 1998 to December 2000 he was a Member of the Technical Staff in the Mathematical Sciences Research Center at Bell Laboratories, Murray Hill, NJ. Since January 2001 he has been with the department of electrical engineering at the California Institute of Technology, Pasadena, CA., where he is currently an associate professor. He has also held short-tem appointments at Ricoh California Research Center, the Indian Institute of Science, and Linkoping University, Sweden. His research interests include wireless communications, robust estimation and control, adaptive signal processing and linear algebra. He is the coauthor of the books *Indefinite Quadratic Estimation and Control: A Unified Approach to H2 and H1 Theories* (New York: SIAM, 1999) and *Linear Estimation* (Englewood Cliffs, NJ: Prentice Hall, 2000). He is a recipient of an Alborz Foundation Fellowship, the 1999 O. Hugo Schuck best paper award of the American Automatic Control Council, the 2002 National Science Foundation Career Award, the 2002 Okawa Foundation Research Grant for Information and Telecommunications, the 2003 David and Lucille Packard Fellowship for Science and Engineering and the 2003 Presidential Early Career Award for Scientists and Engineers (PECASE). He has been a Guest Editor for the *IEEE Transactions on Information Theory* special issue on "space-time transmission, reception, coding and signal processing," was an Associate Editor for Communications of the *IEEE Transactions on Information Theory* during 2004-2006, and is currently an Editor for the journal *Foundations and Trends in Information and Communication*.