On the Throughput-Delay Tradeoff in Cellular Multicast

Praveen Kumar Gopala and Hesham El Gamal
Dept. of Electrical & Computer Engineering
The Ohio State University
Email: {gopalap,helgamal}@ece.osu.edu

Abstract— **In this paper, we adopt a cross layer design approach for analyzing the throughput-delay tradeoff of the multicast channel in a single cell system. To illustrate the main ideas, we start with the single group case, i.e., pure multicast, where a common information stream is requested by all the users. We consider three classes of scheduling algorithms with progressively increasing complexity. The first class strives for minimum complexity by resorting to a static scheduling strategy along with memoryless decoding. Our analysis for this class of scheduling algorithms reveals the existence of a static scheduling policy that achieves the optimal scaling law of the throughput at the expense of a delay that increases exponentially with the number of users. The second scheduling policy resorts to a higher complexity incremental redundancy encoding/decoding strategy to achieve a superior throughput-delay tradeoff. The third, and most complex, scheduling strategy benefits from the cooperation between the different users to minimize the delay while achieving the optimal scaling law of the throughput. In particular, the proposed cooperative multicast strategy is shown to simultaneously achieve the optimal scaling laws of both throughput and delay. Then, we generalize our scheduling algorithms to exploit the multi-group diversity available when different information streams are requested by different subsets of the user population. Finally, we discuss the potential gains of equipping the base station with multiple transmit antennas and present simulation results that validate our theoretical claims.**

I. Introduction

Traditional information theoretic investigations pay little, if any, attention to the notion of delay. Clearly, this approach is not adequate for many applications, especially those with strict Quality of Service (QoS) constraints. To avoid this shortcoming, there has been a growing interest in cross layer design approaches in recent years. The underlying idea in these approaches is to jointly optimize the physical, data link, and network layers in order to satisfy the QoS constraints with the minimum expenditure of network resources. Recent works on cross layer design have considered multi-user cellular networks [1–3]. These works have enhanced our understanding of the fundamental tradeoffs in this scenario and the structure of optimal resource allocation strategies. In this paper, we take a first step towards generalizing this cross layer approach to the wireless multicast scenario. This scenario is characterized by a strong interaction between the network, medium access, and physical layers. This interaction adds significant complexity to the problem which motivated the adoption of a simplified on-off model for the wireless channel in the recent works on wireless multicast [4]. However, we argue that employing more accurate models for the wireless channel allows for valuable opportunities for exploiting the wireless medium to yield performance gains. More specifically, our work sheds light on the role of the following characteristics of the wireless channel in the design of multicast scheduling strategies: 1) The *multi-user diversity* resulting from the statistically independent channels seen by the different users [5], 2) The *wireless multicast gain* resulting from the fact that any information transmitted over the wireless channel is *overheard* by all the users (with possibly different attenuations), and 3) The *cooperative gain* resulting from antenna sharing between users [6].

To illustrate the main ideas, we first focus on the single group (pure multicast) scenario where the same information stream is transmitted to all users in the network [7]. We propose three classes of scheduling algorithms with progressively increasing complexity and characterize the throughput-delay tradeoff achieved by each class. Thereby, we establish the asymptotic throughput optimality of the median user scheduler and show that the proposed cooperative multicast scheme achieves the optimal scaling laws of both throughput and delay at the expense of a high complexity. Then, we extend our study to the multi-group scenario where independent streams of information are transmitted to different groups of users. Here, we generalize our scheduling algorithms to exploit the multi-group diversity available in such scenarios. Finally, we quantify the potential performance gains allowed by equipping the base station with multiple transmit antennas.

II. System Model

We consider the downlink of a single cell system where a base station serves G symmetric groups of users. The information streams requested by the different groups from the base station are independent of each other. Each group consists of N symmetric users. All the users within a group request the same information from the base station. Unless otherwise stated, the base station is assumed to be equipped with a single transmit antenna, and communicates to the users through a wireless fading channel. Each user is assumed to have only a single receive antenna. The fading is assumed to be Rayleigh distributed with unit mean and is i.i.d. across users. The noise at each user is assumed to be zero-mean, unit variance, circularly symmetric complex AWGN. We consider quasi-static fading wherein the fading coefficients remain constant for a

0-7803-9305-8/05/$20.00 ©2005 IEEE

coherence interval and change independently from one interval to the next. The short term average power (over each coherence interval) used at the base station is constrained to be less than P. Each packet transmitted by the base station is assumed to be of constant size. We assume that rate adaptation is possible at the base station. The proposed scheduling schemes, except the incremental redundancy scheme, assume perfect knowledge of the channel state information (CSI) at both the transmitter and receiver. We further assume the use of capacity achieving codes at the base station.

We compare the proposed scheduling schemes in terms of their throughput and delay performance. The **throughput** of a scheme is defined as the sum of throughputs provided to all the users within all the groups in the system. In our delay analysis, we consider backlogged queues. We define the **delay** of a scheme as the delay between the start of transmission of a packet belonging to a particular group of users, and the instant when the packet is successfully decoded by all the users in that group. Thus our definition of delay includes only the transmission delay and does not account for the queuing delay experienced by the packets. To facilitate analytical tractability, we only focus on evaluating the asymptotic scaling laws of the throughput and delay, and we make an exponential server assumption, i.e., the service rate in any time slot is assumed to follow an exponential distribution with the same mean as that obtained from our problem formulation. We use the set of Knuth's asymptotic notations[1] throughout the paper.

III. Single Group (Pure Multicast) Scenario

In this section, we consider the pure multicast scenario where the same information stream is transmitted to all users in the network. In the non-cooperative scenario, the throughput-optimal scheme is an N-level superposition coding/successive decoding scheme. However, this strategy suffers from excessive complexity, which motivates our work on the throughput-delay tradeoff of low complexity scheduling schemes. Interestingly, we identify a low complexity static scheduling scheme that achieves the optimal scaling law of the throughput. Furthermore, we establish the optimality of the proposed cooperative multicast scheme in terms of the scaling laws of both delay and throughput. The throughput of **any** scheduling scheme for this scenario can be upper bounded by

$$R_{tot} \leq E\left[\sum_{i=1}^{N} \log\left(1+|h_i|^2 P\right)\right] = \Theta(N).$$

A. Static Scheduling With Memoryless Decoding

In this class of scheduling algorithms (referred to as "static schedulers"), we schedule transmission to a fraction of the users with favorable channel conditions. The transmission rate is adjusted such that each transmission by the base station is intended for successful reception by (N/α) users in the system. While the identity of the target users change, based on the channel conditions, the static nature of the algorithm is manifested in the fact that a **fixed** fraction of the users is able to decode every transmitted packet (i.e., α is not a function of time). Hence at any time instant, the base station transmits to the user whose instantaneous SNR occupies the $(N-(N/\alpha)+1)^{th}$ position in the ordered list of instantaneous SNRs of all users. The other $((N/\alpha)-1)$ users with higher channel gains can also decode the transmitted information. The parameter α of the scheme is restricted to be a factor of N and satisfies $\alpha \in \mathcal{Z}^+$ and $1 \leq \alpha \leq N$. The memoryless decoding[2] assumption is imposed to limit the complexity of the encoding/decoding process. As shown later, this class of algorithms exploit both the multi-user diversity and multicast gains, to varying degrees, depending on the parameter α.

The average throughput of this general static scheduling scheme is given by

$$R_{tot} = \left(\frac{N}{\alpha}\right) E\left[\log\left(1+|h_{\pi(N-\frac{N}{\alpha}+1)}|^2 P\right)\right],$$

where $|h_{\pi(N-\frac{N}{\alpha}+1)}|^2$ is the channel power gain of the user whose SNR occupies the $(N-(N/\alpha)+1)^{th}$ position in the ordered list of SNRs of all users. Throughout the paper, the log(.) function refers to the natural logarithm, and hence, the average throughput is expressed in nats.

A critical step in the delay analysis is to identify the queuing model. In our model, the base station maintains $\binom{N}{N/\alpha}$ queues, one for each combination of (N/α) users. These queues can be divided into sets with α *coupled* queues in each set such that the combinations of users served by the α queues within a set are mutually exclusive (to ensure that multiple copies of the same packet are not sent to any of the users) and collectively exhaustive (to ensure that the packet reaches all the users), i.e., every user in the system is served by exactly one of the α queues in each set. For example, with $N = 6$ users and $\alpha = 3$, we have 15 queues divided into 5 sets with three queues in each set. One possible set of coupled queues serve users $\{(1,2),(3,4),(5,6)\}$ and another possible set may serve users $\{(1,4),(2,5),(3,6)\}$. Note that each user occurs once and only once in each set (See Fig. 1). Hence, any packet that arrives at the base station is routed towards one of the sets[3] where it is stored in all the α queues within that set (since it needs to be transmitted to all the users in the system). Thus the delay in transmitting a particular packet to all the users is given by the delay in transmitting that packet from each of the α coupled queues in the corresponding set. Moreover, the base station services only one of the $\binom{N}{N/\alpha}$ queues at any time, which is chosen based on the instantaneous fading coefficients of all the users.

In our analysis, we benefit from the concept of worst case delay proposed in [8] for analyzing the delay in unicast networks. In this work, the authors characterized the worst case delay by restating their problem as the "coupon collector problem" which has been studied extensively in the mathematics literature [9], [10]. In the coupon collector problem,

[1] 1) $f(n) = O(g(n))$ iff there are constants c and n_0 such that $f(n) \leq cg(n)\ \forall n > n_0$. 2) $f(n) = \Omega(g(n))$ iff there are constants c and n_0 such that $f(n) \geq cg(n)\ \forall n > n_0$, and 3) $f(n) = \Theta(g(n))$ iff there are constants c_1, c_2 and n_0 such that $c_1 g(n) \leq f(n) \leq c_2 g(n)\ \forall n > n_0$.

[2] Memoryless decoding refers to the fact that the decoder memory is flushed in case of decoding failure.

[3] Here, we use a probabilistic approach for choosing the set with a uniform distribution.

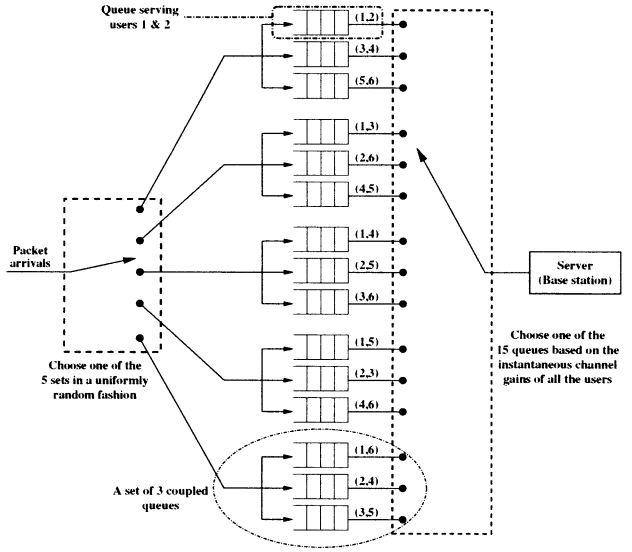

Fig. 1. A queuing model for a system with $N = 6$ users and $\alpha = 3$

the users are assumed to have coupons and the transmitter is the collector that selects one of the users randomly (with uniform distribution) and collects his coupon. The problem is to characterize the average number of trials required to ensure that the collector collects m coupons from all the users. Our queuing problem is analogous to the coupon collector problem with the only *fundamental* difference being that the size of the coupons is time-varying in our problem due to rate adaptation. We now characterize the scaling laws of throughput and delay for the different static scheduling algorithms.

Theorem 1: *The average throughput R_{tot} of the general static scheduling scheme is given by*

$$R_{tot} = \frac{N}{\alpha} \int_0^\infty \log(1 + xP) \mathrm{d}F(x), \tag{1}$$

where

$$F(x) = \sum_{k=(N-\frac{N}{\alpha}+1)}^{N} \binom{N}{k} \left(1 - e^{-x}\right)^k e^{-(N-k)x}, \quad x \geq 0.$$

The average delay of this scheme satisfies

$$D = \max\left\{\Omega\left(\binom{N}{N/\alpha}\frac{\log\alpha}{\log\log N}\right), \Omega\left(\binom{N}{N/\alpha}E[X_{min}]\right)\right\} \tag{2}$$

where $X_{min} = \min_{i=1}^{\alpha} X_i$ and the X_i's are defined as the service times required for transmitting a packet from the i^{th} queue of a set of α queues assuming that the server always services the i^{th} queue.

Proof: The distribution of $|h_{\pi(N-\frac{N}{\alpha}+1)}|^2$ is given by $F(x)$. Hence the throughput is as given in (1). The delay analysis follows the same lines as that in [8] but is modified to incorporate the effect of rate adaptation at the base station. For a detailed proof, refer [11]. ■

We now study three special cases of this general static scheduling scheme in more detail to shed light on the throughput-delay tradeoff achievable by varying α.

1) Worst User Scheduler ($\alpha = 1$): The worst user scheme maximally exploits the multicast gain by always transmitting to the user with the least instantaneous SNR. This enables each transmission to be successfully decoded by all the users in the system. However, the multi-user diversity inherent in the system works against the performance of this scheme and results in a decrease in the individual throughput to any user. The base station maintains a single queue that caters to all the users in the system.

Lemma 1: *The average throughput of the worst user scheme scales as*

$$R_{tot} = \Theta(1) \tag{3}$$

with the number of users N. The average delay scales as

$$D = \Theta(N). \tag{4}$$

Proof: It can be shown that $E\left[|h_{\pi(1)}|^2\right] = \Theta(1/N)$. Hence $R_{tot} = NE\left[\log(1 + |h_{\pi(1)}|^2 P)\right] = \Theta(1)$. Since the rate to any user is $\Theta(1/N)$, the delay can be shown to be $\Theta(N)$. For a detailed proof, refer [11]. ■

2) Best User Scheduler ($\alpha = N$): This scheme maximally exploits the multi-user diversity available in the system. Since the transmission rate is adjusted based on the user with the maximum instantaneous SNR, this scheme fails to exploit any of the multicast gain and any particular packet must be *repeated* N times. The base station maintains N queues, one for each user in the system, and any packet that arrives into the system enters all the N queues.

Lemma 2: *The average throughput of the best user scheme scales as*

$$R_{tot} = \Theta(\log\log N) \tag{5}$$

with the number of users N. The average delay scales as

$$D = \Omega\left(\frac{N\log N}{\log\log N}\right). \tag{6}$$

Proof: The throughput of the best user scheme is shown to be $R_{tot} = \Theta(\log\log N)$ in [8]. The worst case delay for a constant rate is shown to be $\Theta(N\log N)$ in [8]. Since the rate to any user now is $\Theta(\log\log N)$, the delay can be shown to be as given in (6). For a detailed proof, refer [11]. ■

From Lemmas 1 and 2, one can conclude that *maximally* exploiting the multi-user diversity yields higher throughput gains than *maximally* exploiting the multicast gain. This throughput gain, however, is obtained at the expense of a higher delay.

3) Median User Scheduler ($\alpha = 2$): This scheme strikes a balance between exploiting multi-user diversity and multicast gain. The base station always transmits to the user whose instantaneous SNR occupies the median position of the ordered list of SNRs. Each transmission is, therefore, successfully decoded by half the users in the system and the same information needs to be repeated only twice before it reaches all the users. Thus, unlike the best user scheduler, this scheduler benefits from the wireless multicast gain. Moreover, unlike the worst user scheduler, the inherent multi-user diversity does not degrade the performance of this scheduler (since the instantaneous SNR of the median user is not expected to degrade with N). In fact, we show in the following that this scheme achieves the optimal scaling law of the throughput as the number of users N grows to infinity.

Lemma 3: *The proposed median user scheme achieves the optimal scaling law of the throughput. The average throughput of this scheme scales as*

$$R_{tot} = \Theta(N) \tag{7}$$

with the number of users N. The average delay scales as

$$D = \Theta\left(\binom{N}{N/2}\right) = \Theta\left(\frac{2^N}{\sqrt{N}}\right). \tag{8}$$

Proof: Since $E\left[\log(1+|h_{\pi(\frac{N}{2}+1)}|^2 P)\right] = \Theta(1)$, we have $R_{tot} = \Theta(N)$. Since there are $\binom{N}{N/2}$ queues in the system (divided into sets of two coupled queues) and the base station serves only one queue at any time, the delay can be shown to be $\Theta\left(\binom{N}{N/2}\right)$. For a detailed proof, refer [11]. ■

Thus the throughput optimality of the median user scheduler is obtained at the expense of an exponentially increasing delay with the number of users N.

B. Incremental Redundancy Multicast

In this section, we relax the memoryless decoding requirement and propose a hybrid Automatic Repeat reQuest (ARQ) scheme that employs a higher complexity incremental redundancy encoding/decoding strategy to achieve a better throughput-delay tradeoff than the static scheduling schemes. The proposed scheme is an extension of the incremental redundancy scheme in [12]. An information sequence of b bits is encoded into a codeword of length LM, where M refers to the rate constraint. The first L bits of the codeword are transmitted in the first attempt. If a user is unable to successfully decode the transmission, it sends back an ARQ request to the base station. If the base station receives an ARQ request from any of the users, it transmits the next L bits of the same codeword in the next attempt. This process continues until either all N users successfully decode the information or the rate constraint M is violated. Then the codeword corresponding to the next b information bits is transmitted in the same fashion. This scheme does not require the knowledge of perfect CSI at the base station. The base station only needs to know when to stop transmission of the current codeword. Hence the feedback required is minimal. The following result for the unconstrained case ($M \to \infty$) establishes the superior throughput-delay tradeoff achieved by this scheme.

Theorem 2: *The average throughput of the incremental redundancy scheme scales as*

$$R_{tot} = \Theta\left(\frac{N \log\log N}{\log N}\right) \tag{9}$$

with the number of users N. The average delay scales as

$$D = \Theta\left(\frac{\log N}{\log\log N}\right). \tag{10}$$

Proof: Refer [11] for a detailed proof. ■

Thus the incremental redundancy scheme avoids the exponentially growing delay of the median user scheduler at the expense of a minimal penalty in throughput. Moreover, the base station needs to maintain only a single queue that serves all the users in the system. This approach, however, entails added complexity in the incremental redundancy encoding and the storage and joint decoding of all the observations.

C. Cooperative Multicast

In this section, we demonstrate the benefits of user cooperation and quantify the tremendous gains that can be achieved by allowing the users to cooperate with each other. In particular, we propose a cooperation scheme that minimizes the delay while achieving the optimal scaling law of the throughput. This scheme is divided into two stages. In the first half of each time slot, the base station transmits the packet to one half of the users in the system (i.e., the median user scheduler). During the next half of the slot, the base station remains silent. Meanwhile, all the users that successfully decoded the packet in the first half of the slot cooperate with each other and transmit the packet to the other $(N/2)$ users in the system. Through antenna sharing, the $(N/2)$ cooperating users mimic a multi-antenna system with $(N/2)$ transmit antennas. If R_{s1} and R_{s2} are the rates supported in the first and second stage respectively, then the actual transmission rate is chosen to be $\min\{R_{s1}, R_{s2}\}$ in both stages of the cooperation scheme. Note that the rate R_{s2} is chosen such that the information can be successfully decoded even by the worst of the remaining $(N/2)$ users. Here, we note that this scheme requires the base station to know the CSI of the inter-user channels. The scheme, however, does not require the users to have such transmitter CSI (i.e., in the second stage the users cooperate blindly by using i.i.d. random coding). The average throughput of the proposed cooperation scheme is thus given by

$$R_{tot} = \left(\frac{N}{2}\right) E\left[\min\{R_{s1}, R_{s2}\}\right].$$

The following result establishes the optimality of the proposed scheme, in terms of the scaling laws of both the delay and the throughput. Here we assume that the inter-user channels have the same fading statistics as the channels between the base station and users, and the **total** transmitted power is upper bounded by P.

Theorem 3: *The proposed cooperation scheme achieves the optimal scaling laws of both delay and throughput. In particular, the average throughput of this scheme scales as*

$$R_{tot} = \Theta(N) \tag{11}$$

with the number of users N, while the average delay scales as

$$D = \Theta(1). \tag{12}$$

Proof: From Lemma 3, the throughput of the first stage is known to be $\Theta(N)$. It can be shown that the throughput of the second stage is also $\Theta(N)$. Since the information transmitted in the first half of each time slot reaches all the N users at the end of the slot, and the rate to any user is $\Theta(1)$, the delay scales as $\Theta(1)$. For a detailed proof, refer [11]. ■

The price for this optimal performance is the added complexity needed to 1) equip every user with a transmitter, 2) decode/re-encode the information at each cooperating user, and 3) provide perfect CSI of the inter-user channels to the base station.

IV. Multi-Group Diversity

In this section, we generalize the scheduling schemes proposed in Section III to the multi-group scenario where

different information streams are requested by different subsets of the user population. We modify the proposed schemes to exploit the multi-group diversity available in this scenario by always transmitting to the best group. We characterize the asymptotic scaling laws of the throughput and delay of the static schedulers with the number of users per group N and the number of groups G in the following theorem.

Theorem 4: 1) *The average throughput of the best among worst users scheme scales as*[4]

$$R_{tot} = \Theta(\log G) \tag{13}$$

with N and G. The average delay scales as

$$D = \Theta\left(\frac{NG}{\log G}\right). \tag{14}$$

2) *The average throughput of the best among best users scheme scales as*

$$R_{tot} = \Theta(\log\log NG) \tag{15}$$

with N and G. The average delay scales as

$$D = \Omega\left(\frac{NG\log N}{\log\log NG}\right). \tag{16}$$

3) *The average throughput of the best among median users scheme satisfies*

$$\Omega(N) = R_{tot} = O(N\log\log G), \tag{17}$$

while the average delay of this scheme satisfies

$$\Omega\left(\frac{G2^N}{\sqrt{N}\log\log G}\right) = D = O\left(\frac{G2^N}{\sqrt{N}}\right). \tag{18}$$

Proof: Refer [11] for a detailed proof. ■

In the multi-group incremental redundancy scheme, the information bits corresponding to each of the groups are encoded independently. During each time slot, the base station selects that group for which it can send the highest total instantaneous rate to the users who failed to decode up to this point. This selection process makes the scheme "dynamic" in the sense that the outcome of the scheduling process at any particular time slot depends on the outcomes in all previous slots. Unfortunately, this dynamic nature of the proposed scheme adds significant complexity to the problem and, at the moment, we do not have an analytical characterization of the corresponding scaling laws.

In the multi-group cooperation scheme, during each time slot, the base station selects the best group $\hat{g}$ for transmission according to the condition

$$\hat{g} = \arg\max_{g=1,\cdots,G}\left\{\left(\frac{N}{2}\right)\min\{R_{s1}^g, R_{s2}^g\}\right\}. \tag{19}$$

Theorem 5: *The average throughput of the proposed multi-group cooperation scheme satisfies*

$$\Omega(N) = R_{tot} = O\left(N\log\log G\right), \tag{20}$$

while the average delay of this scheme satisfies

$$\Omega\left(\frac{G}{\log\log G}\right) = D = O(G). \tag{21}$$

[4]Note that $\Theta(\log G) \neq 0$ when $G = 1$, since $k + \log G = \Theta(\log G)$, for any constant k.

Proof: Refer [11] for a detailed proof. ■

As expected, the throughput gain resulting from the multi-group diversity entails a corresponding price in the increased delay.

V. Multi-Transmit Antenna Gain

The performance of the proposed static scheduling schemes depends on the spread of the fading distribution. For exploiting significant multi-user diversity gains, the distribution needs to be well-spread out. The lower the spread of the distribution, the lesser the multi-user diversity gain (or loss as shown in the following). To illustrate this point, we consider a scenario where the base station is equipped with L transmit antennas. We assume that the base station has knowledge of only the total effective SNR at any particular user and does not know the individual channel gains from each transmit antenna to that user. Under this assumption, the base station just distributes the available power equally among all the L transmit antennas. Thus the effective fading power gains follow a normalized Chi-square distribution with $2L$ degrees of freedom. Note that the fading power gains are exponentially distributed (Chi-square with 2 degrees of freedom) in the single transmit antenna case. We now characterize the asymptotic scaling laws of the throughput of the proposed static schedulers for this multi-transmit antenna scenario. Here L is assumed to be a constant and does not scale with N.

A. Worst User Scheduler

For the worst user scheme, the average throughput is given by

$$R_{tot} = NE\left[\log\left(1 + |\chi_{min}|^2 P\right)\right],$$

where $|\chi_{min}|^2 = \min_{i=1}^{N} |\chi_i|^2$, and $|\chi_i|^2$ corresponds to the effective fading power gain at the i^{th} user that follows a normalized Chi-square distribution with $2L$ degrees of freedom.

Lemma 4: *When the base station is equipped with L transmit antennas, the average throughput of the worst user scheme scales as*

$$R_{tot} = \Theta\left(N^{\left(\frac{L-1}{L}\right)}\right). \tag{22}$$

Proof: Refer [11] for a detailed proof. ■

Thus the average throughput increases with L. This is expected since the performance of the worst user scheduler is *degraded* by the tail of the fading distribution. Hence, as L increases, the spread of the fading distribution decreases, and consequently, the inherent multi-user diversity has a reduced effect on the performance of the scheduler. This leads to a rise in the average throughput from $\Theta(1)$ for the single transmit antenna case to $\Theta(N)$ for large values of L.

B. Best User Scheduler

For the best user scheme, the average throughput is given by

$$R_{tot} = E\left[\log\left(1 + |\chi_{max}|^2 P\right)\right],$$

where $|\chi_{max}|^2 = \max_{i=1}^{N} |\chi_i|^2$.

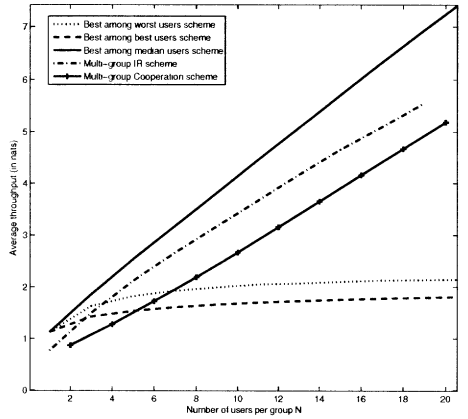

Fig. 2. Comparison of the throughput of the proposed schemes for $G = 5$ groups

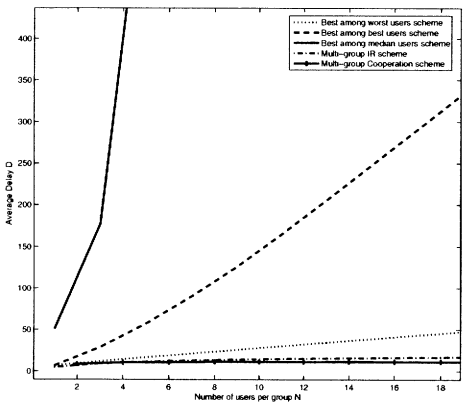

Fig. 3. Comparison of the delay of the proposed schemes for $G = 5$ groups

Lemma 5: *When the base station is equipped with L transmit antennas, the average throughput of the best user scheme scales as*

$$R_{tot} = \Theta\left(\log\left(1 + \frac{\log N + (L-1)\log\log N}{L}\right)\right). \quad (23)$$

Proof: Refer [11] for a detailed proof. ■

Since the best user scheduler leverages multi-user diversity to enhance the throughput, one can see that the throughput of the best user scheme decreases as L increases.

VI. Numerical Results

Here we present simulation results that validate our theoretical claims. These results were obtained through Monte-Carlo simulations and were averaged over at least 5000 iterations. The power constraint P is taken to be unity. For a comparison of the throughput and delay of all the schemes proposed for the pure multicast scenario in Section III, please refer [7]. In Fig. 2, we present a throughput-comparison for the different scheduling schemes proposed in Section IV for the multi-group scenario for increasing values of N with $G = 5$ groups. The corresponding delay-comparison is presented in Fig. 3. Although the best among worst users scheduler performs better than the best among best users scheme, in terms of throughput, for the range of N values shown in the plot, it should be noted that the latter eventually outperforms the former for large values of N ($N > 600$). Except for this case, we see that the simulation results follow the same trends predicted by our asymptotic analysis. Finally, we observe that the utility of our asymptotic analysis is manifested in its accurate predictions even with the relatively small number of users used in our simulations (i.e., in the order of $N = 10$).

VII. Conclusions

In this paper, we have used a cross layer design approach to shed more light on the throughput-delay tradeoff in the cellular multicast channel. Towards this end, we proposed three classes of scheduling algorithms with progressively increasing complexity, and analyzed the throughput-delay tradeoff achieved by each class. We showed that the median user scheduler achieves the optimal scaling law of the throughput at the expense of an exponentially increasing delay with the number of users. We further showed that the proposed cooperative multicast scheme achieves the optimal scaling laws of both throughput and delay at the expense of a high RF and computational complexity. We then generalized our schemes to the multi-group scenario and characterized their ability to exploit the multi-group diversity offered by the wireless channel. Finally, we quantified the performance gains in the multi-transmit antenna scenario with limited feedback and presented simulation results that validate our theoretical claims.

References

[1] E. M. Yeh and A. S. Cohen, "Information theory, queueing, and resource allocation in multi-user fading communications," in *38th Annual Conference on Information Sciences and Systems*, March 2004.

[2] J. Zhang and D. Zheng, "Ad hoc networking over fading channels: Multi-channel diversity, mimo signaling, and opportunistic medium access control," in *41st Allerton Conference on Communications, Control, and Computing*, October 2003.

[3] P. Liu, R. Berry, and M. Honig, "Delay-sensitive packet scheduling in wireless networks," in *IEEE WCNC 2003*, March 2003.

[4] P. Chaporkar and S. Sarkar, "On-line optimal wireless multicast," in *2nd Workshop On Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks*, (Cambridge, England), pp. 282–291, March 2004.

[5] R. Knopp and P. Humblet, "Information capacity and power control in single cell multiuser communications," in *IEEE International Computer Conference (ICC'95)*, (Seattle, WA), June 1995.

[6] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity-part i: System description," *IEEE Transactions on Communications*, vol. 51, pp. 1927–1938, November 2003.

[7] P. K. Gopala and H. E. Gamal, "Opportunistic multicasting," in *Asilomar Conference on Signals, Systems and Computers*, November 2004.

[8] M. Sharif and B. Hassibi, "Delay analysis of throughput optimal scheduling in broadcast fading channels," *Submitted to IEEE Transactions on Information Theory*, 2004.

[9] D. J. Newman and L. Shepp, "The double dixie cup problem," *Amer. Math. Monthly*, vol. 67, pp. 58–61, January 1960.

[10] W. Feller, *An introduction to probability theory and its applications*. John Wiley and Sons, Inc., 1967.

[11] P. K. Gopala and H. E. Gamal, "On the throughput-delay tradeoff in cellular multicast," *Submitted to IEEE Transactions on Information Theory* (Available at www.ece.osu.edu/~gopalap), 2005.

[12] G. Caire and D. Tuninetti, "The throughput of hybrid-arq protocols for the gaussian collision channel," *IEEE Transactions on Information Theory*, vol. 47, July 2001.