# Scheduling for Cellular Multicast: A Cross-Layer Perspective

Praveen Kumar Gopala and Hesham El Gamal

**Abstract**

Recent years have witnessed the importance of cross-layer design in wireless communication systems. The adoption of simplified on/off models for the wireless channel, in an attempt to reduce it to the traditional wireline case, has been shown to be highly sub-optimal. In this paper, we consider more realistic models for the wireless channel and demonstrate the significant performance gains that can be leveraged from the wireless medium through intelligent scheduling strategies. We consider a cellular multicast channel, where a central base station transmits the same information stream to all the users within the system. We design scheduling algorithms for this scenario from a cross-layer perspective, which aids us in exploiting the characteristics of the wireless medium. We consider three classes of scheduling algorithms with progressively increasing complexity and study their asymptotic throughput-delay performance with the number of users in the system. The first class strives for minimum complexity by resorting to static scheduling along with memoryless decoding. Our analysis reveals the existence of a scheduling scheme, within this class, that achieves near-optimal scaling laws of both delay and throughput. 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 and is shown to simultaneously achieve the optimal scaling laws of both delay and throughput. We further study the effect of equipping the base station with multiple transmit antennas on the throughput performance of the proposed low-complexity static schedulers. Finally, we present simulation results that validate our theoretical claims.

## I. INTRODUCTION

Wireless networks are becoming increasingly popular primarily due to their ease of installation and the freedom of mobility that they offer to the users. Traditional data link and network layer protocols designed for wireless networks adopt simplified on/off models for the physical layer, and thereby focus on reducing the system to a wireline scenario. This approach of designing algorithms based on the assumption that the wireless channel behaves like a reliable, time-invariant bit-pipe has been shown to be highly sub-optimal, especially for applications with strict Quality of Service (QoS) constraints. Recent years have witnessed a growing interest in cross-layer design approaches for wireless system design. The underlying idea in these approaches is to jointly optimize the physical, data link, and networking layers in order to satisfy the QoS constraints with the minimum expenditure of network resources. Early investigations on cross-layer design have focused on the single user case [1], [2]. These works have shed light on the fundamental tradeoffs in this scenario and devised efficient power and rate control policies that approach these limits. More recent works have considered multi-user cellular networks [3]–[7]. These studies have enhanced our understanding of the fundamental limits 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 channel, where the same information stream is transmitted by the base station to multiple users within the network. Such multicast scenarios have become highly common in broadband wireless networks, due to their support for streaming video (Mobile TV) applications using the IP Datacasting (IPDC) and Digital Video Broadcasting-Handheld (DVB-H) frameworks. Moreover, these multicast scenarios are 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 several of the recent works on wireless multicast [8]–[10]. In the sequel, 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 [11], 2) The *wireless multicast gain* resulting from the fact that any information transmitted over the wireless channel is *overheard* by all users, possibly with different attenuation factors, and 3) The *cooperative gain* resulting from antenna sharing between users [12].

We focus on the scenario where the same information stream is transmitted to all the users in the network. We consider three classes of scheduling algorithms with progressively increasing complexity. The first class strives for minimum complexity by resorting to static scheduling along with memoryless decoding (i.e., the decoder memory is flushed in case of a decoding failure). Within this class, we study three different schemes that highlight the tradeoff between exploiting the multiuser diversity gain and the wireless multicast gain offered by the wireless medium. We characterize the asymptotic scaling of the throughput and delay achieved by each of these schemes with the number of users in the system. Thereby, we establish the near-optimality of the proposed median user scheduler both in terms of delay and throughput. The second scheduling policy is based on a hybrid Automatic Repeat reQuest (ARQ) strategy, and uses a higher complexity incremental redundancy encoder/decoder to achieve a better throughput-delay tradeoff. The third, and most complex, scheduling strategy allows for cooperation between the different users, thereby exploiting the cooperative gain possible through antenna sharing in wireless channels. We show that this scheme achieves the optimal scaling of both delay and throughput with the user population, thereby striking a nice balance between exploiting the three different opportunities offered by the wireless medium. Finally, we study the effect of equipping the base station with multiple transmit antennas on the throughput performance of the proposed low-complexity static schedulers. In particular, we show that when the number of transmit antennas is large, the wireless multicast gain harnessed by a scheme dominates its throughput performance.

The rest of the paper is organized as follows. In Section II, we introduce the system model along with our notation. In Section III, we propose the three classes of scheduling algorithms for the multicast channel and characterize their asymptotic throughput-delay performance. The potential performance gains allowed by multi-transmit antenna base stations are quantified in Section IV. In Section V, we present numerical results that validate our theoretical claims in certain representative scenarios. Finally, some concluding remarks are offered in Section VI. In order to enhance the flow of the paper, we collect all the proofs in the Appendices.

## II. System Model

We consider the downlink of a single cell system where a base station (BS) serves a group of $N$ users. All the users request the same information from the BS. Unless otherwise stated, the BS is assumed to be equipped with a single transmit antenna. Each user is assumed to have only a single receive antenna. We consider time-slotted transmission wherein the received symbol vector at user $i$ in time slot $k$ is given by

$$\underline{y_i}[k] = h_i[k]\underline{x}[k] + \underline{n_i}[k],$$

where $\underline{x}[k]$ denotes the complex-valued vector of length $m$ transmitted by the BS in slot $k$, $h_i[k]$ represents the complex flat fading coefficient of the channel between the BS and the $i^{th}$ user in time slot $k$, and $\underline{n_i}[k]$ represents the zero-mean unit-variance complex additive white Gaussian noise vector at the $i^{th}$ user in slot $k$. The noise processes are assumed to be circularly symmetric and independent across users. The channel between the BS and each user is assumed to be quasi-static with coherence time $T_c$. Thus the fading coefficients remain constant throughout an interval of length $T_c$ (or $m$ channel uses) and change independently from one interval to the next. The fading coefficients $\{h_i\}$ are assumed to be independent and identically distributed (i.i.d.) across the users and follow a Rayleigh distribution with $\mathbb{E}\left[|h_i[k]|^2\right] = 1,\ \forall i, k$. We restrict our attention to this symmetric scenario, and hence, issues related to fairness are outside the scope of this work. Each packet transmitted by the BS is assumed to be of constant size $S$. We further employ the following short-term average power constraint at the base station

$$\frac{1}{m}\,\mathbb{E}\left[\|\underline{x}[k]\|^2\right] \quad \leq \quad P.$$

Clearly, further performance gain may be reaped through a carefully constructed power allocation policy if this short term power constraint is replaced by a long term one. This line of work, however, is not pursued here and we only rely on rate adaptation and scheduling based on the instantaneous channel state available at the BS. Hence the proposed scheduling strategies, except the incremental redundancy scheme[1], assume perfect knowledge of the channel state information (CSI) at both the transmitter and receiver. In our throughput analysis, we use capacity expressions for the channel transmission rates. Here we implicitly assume that the BS employs coding schemes that approach the channel capacity, which justifies our use of the fundamental information theoretic limit of the channel.

In our delay analysis, we consider backlogged queues, and hence, the only meaningful measure of delay is the transmission delay. This leads to the following definitions for throughput and delay that will be adopted in the sequel.

*Definition 1:* The **throughput** of a scheduling scheme is defined as the sum of the throughputs provided by the base station to each individual user within the system.

*Definition 2:* The **delay** of a scheduling scheme is defined as the delay between the instant representing the start of transmission of a packet, and the instant when that packet is successfully decoded by all the users in the system.

We note here that our notion of delay does not account for the queuing delay experienced by the packets. We adopt this restricted notion to simplify the delay analysis, since significant complexity is added to the queuing delay analysis by the formation of coupled queues[2] in the multicast setting. However, as argued in the sequel, our delay analysis offers a lower bound on

[1]For the incremental redundancy scheme, the BS only needs to know when to stop transmission of the current codeword.

[2]This notion of coupled queues will be made clear in the discussion of the best user scheduler in the next section.

the *total* delay which is very tight in several important special cases. Furthermore, this analysis provides a very useful tool for rank-ordering the different classes of scheduling algorithms and sheds light on their structural properties.

To facilitate analytical tractability, we focus only on evaluating the asymptotic scaling laws of the throughput and delay of the proposed schedulers in the sequel. In this analysis, we use the following asymptotic notations throughout the paper:

a) $f(n) = O(g(n))$ iff there are constants $c$ and $n_0$ such that $f(n) \leq cg(n)$, $\forall n > n_0$.

b) $f(n) = \Omega(g(n))$ iff there are constants $c$ and $n_0$ such that $f(n) \geq cg(n)$, $\forall n > n_0$.

c) $f(n) = \Theta(g(n))$ iff there are constants $c_1$, $c_2$ and $n_0$ such that $c_1 g(n) \leq f(n) \leq c_2 g(n)$, $\forall n > n_0$.

## III. Cross-layer Multicast Schedulers

In a non-cooperative setting, wherein the users cannot help each other to decode the multicast transmissions, the throughput-optimal scheme is an $N$-level superposition coding/successive decoding scheme [13], where $N$ parallel streams (layers) are transmitted simultaneously from the BS (Layered Multicast). While the user with the best channel can successfully decode all the $N$ streams, the user with the worst channel can decode only one of the streams successfully (i.e., with arbitrarily small probability of error). This layered strategy, however, suffers from excessive complexity (especially when there are a large number of users in the system) and might be infeasible to implement in practice, since the mobile nodes have low processing power and tight battery power constraints. This motivates our work where we focus on the throughput and delay achieved by low complexity scheduling schemes, which do not require any superposition coding/successive decoding at the BS and users respectively, and are based only on single stream transmissions by the BS. Interestingly, we identify low complexity schedulers that achieve near-optimal scaling laws of both throughput and delay with the number of users $N$. Furthermore, we establish the asymptotic optimality of the proposed cooperative multicast scheduler in terms of both delay and throughput.

### A. Static Schedulers with Memoryless Decoding

In this class of schedulers, the BS always schedules transmissions to a desired fraction of the users with favorable channel conditions (by adjusting the transmission rate accordingly). While the identity of the users change, based on the instantaneous channel conditions, the fraction of users that are able to decode every transmitted packet always remains the same (and hence the name “static”). The memoryless decoding property dictates that the remaining users, who do not succeed in decoding, *flush* their memories and wait for future re-transmissions of the packet. This assumption is imposed to limit the complexity of the encoding/decoding process. In Section III-B, we relax this memoryless decoding assumption and quantify the gains offered by carefully constructed ARQ schemes.

Under this class of schedulers, we now study three different schemes of scheduling transmissions to the best, worst and median user in detail. These three schemes highlight the tradeoff between exploiting the multi-user diversity and the wireless multicast gain offered by the wireless medium. While the worst user scheduler maximally exploits the wireless multicast gain offered by the system, the best user scheduler, on the other hand, maximally exploits the available multi-user diversity in the system. The proposed median user scheduler then strikes a balance between these two conflicting notions to yield significant gains in the achieved throughput and delay. We now elucidate these three schemes and characterize the scaling laws of the average throughput and delay achieved by each of them with the number of users in the system.

*1) Worst User Scheduler:* This scheme maximally exploits the wireless multicast gain by always transmitting to the user with the least instantaneous SNR (worst channel) in the system. Hence at any time instant, the BS chooses its transmission rate to be the rate supported by the worst user in the system. This enables all the users to successfully decode the transmission (since any user with a higher supported rate can decode the transmitted information). Thus any data transmitted by the BS reaches all the users in a single transmission. However, since the transmission rate supported by the worst user decays as the number of users becomes large, it is clear that 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 average throughput of the worst user scheduler is given by

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

where $|h_{\pi(1)}|^2 = \min_{i=1}^{N} |h_i|^2$ is the minimum channel gain among all the $N$ users in the system, whose distribution and density functions are given by

$$F_{|h_{\pi(1)}|^2}(x) = 1 - e^{-Nx} \quad \text{and} \quad f_{|h_{\pi(1)}|^2}(x) = Ne^{-Nx}, \quad x \geq 0.$$

Throughout the paper, the $\log(.)$ function refers to the natural logarithm, and hence, the average throughput is expressed in nats. Since any transmission by the BS reaches all the users in the system, the BS needs to maintain only a single common queue for all the users to implement this scheme.

*Theorem 3:* The average throughput and the average delay of the worst user scheduler scale as

$$R_{tot} \;=\; \Theta(1) \qquad \text{and} \qquad D \;=\; \Theta(N), \tag{1}$$

with the number of users $N$.

From Theorem 3, it is clear that the average throughput of the worst user scheduler does not scale with the number of users $N$, while the average delay increases linearly with $N$.

*2) Best User Scheduler:* This scheme maximally exploits the multi-user diversity available in the system. At any instant, the BS chooses its transmission rate to be the rate supported by the best user in the system. Since the transmission rate is adjusted based on the user with the maximum instantaneous SNR, none of the other users in the system will be able to successfully decode the transmission. Thus only one user is targeted in every transmission and the scheme fails to exploit any wireless multicast gain. Hence every packet must be *repeated* $N$ times before it reaches all the users. The average throughput of the best user scheduler is given by

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

where $|h_{\pi(N)}|^2 = \max_{i=1}^{N} |h_i|^2$ is the maximum channel gain among all the $N$ users in the system, whose distribution function is given by

$$F_{|h_{\pi(N)}|^2}(x) = \left(1 - e^{-x}\right)^N, \quad x \geq 0.$$

To implement this scheme, the BS needs to maintain a set of $N$ *coupled* queues, one for each user in the system. These queues are coupled in the sense that any packet that needs to be transmitted enters all the $N$ queues simultaneously (to ensure that the packet reaches all the users), and the BS serves only one of these $N$ queues (the queue corresponding to the best user) at any time. Thus the delay in transmitting a particular packet to all the users is given by the delay in transmitting that packet from all of the $N$ queues at the BS. In our analysis, we benefit from the concept of worst case delay proposed in [14] for analyzing the delay in unicast networks. In [14], 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 [15], [16]. In the coupon collector problem, the users are assumed to have coupons and the transmitter is the collector that selects one of the users randomly (with a 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 (the detailed analysis is presented in the proofs). The following theorem establishes the average throughput and delay achieved by the best user scheduler.

*Theorem 4:* The average throughput and the average delay of the best user scheduler scale as

$$R_{tot} = \Theta(\log\log N) \qquad \text{and} \qquad D = \Omega\left(N \log N\right), \tag{2}$$

with the number of users $N$.

From Theorems 3 and 4, one can conclude that *maximally* exploiting the multi-user diversity yields higher throughput gains than *maximally* exploiting the wireless multicast gain, when there are a large number of users in the system. This throughput gain, however, is obtained at the expense of a higher delay. This observation motivates the investigation of other static schedulers that achieve a better throughput-delay tradeoff.

*3) Median User Scheduler:* This scheme strikes a balance between exploiting the multi-user diversity and the wireless multicast gain offered by the system. At any instant, the BS chooses its transmission rate such that the better half of the users in the system can successfully decode each transmission, i.e., the rate is adjusted based on the user whose instantaneous SNR is the median among the SNRs of all users. Thus $(N/2)$ users are targeted in each transmission and therefore, 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 $N$). Once the BS starts transmitting a packet, it keeps on repeating the same packet until it is successfully decoded by all the users in the system. Since the BS has perfect channel knowledge, it can easily keep track whether or not the transmitted packet has been decoded by all the users in the system. Once the current packet reaches all the $N$ users, the BS immediately starts transmitting the next packet in the same fashion. The BS needs to maintain only a single common queue that caters to all the users in the system.

One drawback of this scheme is that some of the users may receive redundant copies of the same packet (since any user who has decoded the packet has to wait until all the other users have decoded that packet). This redundancy leads to a reduction in the effective throughput achieved by this scheme. However, it is interesting to note that the median user scheduler achieves near-optimal scaling laws of both throughput and delay, as shown in the following theorem.

*Theorem 5:* The average throughput and the average delay of the median user scheduler scale as

$$R_{tot} \;=\; \Theta\left(\frac{N}{\log N}\right) \qquad \text{and} \qquad D \;=\; \Theta\left(\log N\right), \tag{3}$$

with the number of users $N$.

By comparing the results in Theorems 3, 4 and 5, it is clear that the proposed median user scheduler achieves a much superior throughput-delay tradeoff than the best and worst user schedulers, by striking a balance between exploiting multi-user diversity and multicast gain.

An upper bound on the average throughput of **any** scheduling scheme is given by

$$R_{tot} \;\leq\; \mathbb{E}\left[\sum_{i=1}^{N} \log\left(1+|h_i|^2 P\right)\right] \;=\; N\mathbb{E}\left[\log\left(1+|h_1|^2 P\right)\right] \;=\; \Theta(N). \tag{4}$$

Moreover, the average delay of **any** scheduling scheme can also be easily lower bounded as

$$D \;=\; \Omega(1). \tag{5}$$

Comparing these bounds with the results in Theorem 5, we find that the proposed median user scheduler achieves near-optimal scaling laws of both throughput and delay. In fact, the loss in both delay and throughput scaling laws, compared to the optimal values, is only a factor of $(\log N)$. However, this scheme requires perfect knowledge of the channel gains of all the users (CSI) at the BS, and hence entails a significant feedback requirement. We next propose a scheduling scheme that does not require perfect CSI at the BS and entails very minimal feedback.

### *B. Incremental Redundancy Multicast*

In this section, we relax the memoryless decoding requirement and propose a scheme that employs a higher complexity incremental redundancy encoding/decoding strategy to achieve a better throughput-delay tradeoff than the median user scheduler. The proposed scheme is an extension of the incremental redundancy scheme given by Caire *et al* in [17]. 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 BS. If the BS receives an ARQ request from any of the users, it transmits the next $L$ bits of the same codeword in the next attempt. After each transmission attempt, the users try to decode the transmitted information sequence using the received sequence in that attempt *jointly* with the received sequences in all previous ARQ attempts. This process continues until either all $N$ users successfully decode the information sequence or the rate constraint $M$ is violated. Then the codeword corresponding to the next $b$ information bits is transmitted in the same fashion.

In this scheme, similar to the median user scheduler proposed earlier, even if some of the users successfully decode the packet in very few attempts, they still have to wait until all the $N$ users successfully receive the packet before any new packet is transmitted to them by the BS. This sub-optimality of the proposed schemes results in significant complexity reduction by avoiding the use of superposition coding and successive decoding. However, unlike the median user scheme, this scheme does not require the knowledge of perfect CSI at the BS. The BS only needs to know when to stop transmission of the current codeword. Hence the feedback required is minimal. The following result establishes the superior throughput-delay tradeoff achieved by this scheme compared with the median user scheduler.

*Theorem 6:* The average throughput and the average delay of the incremental redundancy multicast scheduler scale as

$$R_{tot} = \Theta\left(\frac{N \log\log N}{\log N}\right) \qquad \text{and} \qquad D = \Theta\left(\frac{\log N}{\log\log N}\right), \tag{6}$$

with the number of users $N$.

Thus, from Theorem 6 and the throughput and delay bounds in (4) and (5), it is clear that incremental redundancy multicast achieves near-optimal scaling laws of both throughput and delay. The loss in both delay and throughput scaling laws, compared to the optimal values, is only a factor of $(\log N/\log\log N)$. In this approach, the BS again 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 BS 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 BS 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. This is equivalent to a transmission from a transmitter equipped with $(N/2)$ transmit antennas to the worst user in a group of $(N/2)$ users. If $R_{s1}$ and $R_{s2}$ are the rates supported in the first and second stages 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 BS 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

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

The following theorem establishes the optimality of the proposed scheme, in terms of both delay and throughput scaling laws.

*Theorem 7:* The proposed cooperative multicast scheduler achieves the optimal scaling laws of both delay and throughput. In particular, the average throughput and the average delay of this scheduler scale as

$$R_{tot} \;=\; \Theta(N) \qquad \text{and} \qquad D \;=\; \Theta(1), \tag{7}$$

with the number of users $N$. 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 $P$.

The price for this optimal performance is the added complexity needed to 1) equip every user terminal with a transmitter, 2) decode/re-encode the information at each cooperating user terminal, and 3) inform the BS with perfect CSI of the inter-user channels. In Table I, we provide a comparison of the throughput-delay tradeoffs achieved by each of the proposed multicast schedulers.

## IV. Multi-Transmit Antenna Gain

The performance of the best and worst user schedulers proposed in Section III-A depends on the spread of the fading distribution of the users' channels. For exploiting significant multi-user diversity gains, this distribution needs to be well-spread out [18]. 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 BS is equipped with $L$ transmit antennas. We assume that the BS 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 BS 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 were exponentially distributed (Chi-square with 2 degrees of freedom) in the single transmit antenna case. We now characterize the asymptotic throughput scaling laws of the best and worst user schedulers for this multi-transmit antenna scenario. Note that all the results in this section are derived for the case where $L$ is a constant and does not scale with $N$.

### A. Worst User Scheduler

The average throughput of the worst user scheduler is given by

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

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 and whose distribution function is given by

$$F(x) = 1 - e^{-Lx}\left(\sum_{k=0}^{L-1}\frac{(Lx)^k}{k!}\right), \quad x \geq 0. \tag{8}$$

*Lemma 8:* When the base station is equipped with $L$ transmit antennas, the average throughput of the worst user scheduler scales as

$$R_{tot} \;=\; \Theta\left(N^{\left(\frac{L-1}{L}\right)}\right). \qquad (9)$$

From (9), it is clear that the throughput scaling law of the worst user scheduler improves as $L$ increases. This throughput improvement is expected since the performance of the worst user scheduler is known to be *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 of the worst user scheduler from $\Theta(1)$ for the single transmit antenna case to almost $\Theta(N)$ for large values of $L$. Thus the worst user scheduler achieves a near-optimal throughput scaling for large values of $L$.

### B. Best User Scheduler

The average throughput of the best user scheduler is given by

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

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

*Lemma 9:* When the base station is equipped with $L$ transmit antennas, the average throughput of the best user scheduler scales as

$$R_{tot} \;=\; \Theta\left(\log\left(1 + \frac{\log N + (L-1)\log\log N}{L}\right)\right). \qquad (10)$$

Since the best user scheduler leverages multi-user diversity to enhance the throughput, one can see from (10) that its throughput decreases as $L$ increases.

It is interesting to note that even when the BS is equipped with only 2 transmit antennas ($L = 2$), the throughput of the worst user scheduler is significantly higher than that of the best user scheduler. This is contrary to our conclusion in Section III-A for the single transmit antenna case, where we showed that the best user scheduler outperforms the worst user scheduler. Since the multi-user diversity gain decreases as $L$ increases (due to channel hardening [19]), the wireless multicast gain starts to dominate the achieved throughput for large values of $L$, which accounts for the better performance of the worst user scheduler.

## V. Numerical Results

We now present simulation results that validate our theoretical claims. Our results were obtained through Monte-Carlo simulations and were averaged over 10000 iterations. The power constraint $P$ is taken to be unity. A comparison of the throughput of all the scheduling schemes proposed in Section III is presented in Fig. 1 for increasing values of $N$. Although the incremental redundancy scheme outperforms the cooperative multicast scheme for small values of $N$, it is clear that the latter eventually outperforms the former for large values of $N$ ($N > 45$). The corresponding delay-comparison of the proposed schedulers is presented in Fig. 2. It is clear that the simulation results follow the same trends that were predicted by our asymptotic analysis. In Fig. 3, we present a comparison of the throughput of the best and worst user schedulers for the multi-transmit antenna case discussed in Section IV. As predicted by our analysis, the worst user scheduler outperforms the best user scheduler even for the $L = 2$ case. It is also clear that the throughput scaling of the worst user scheduler is almost linear for large values of $L$ ($L \geq 10$). 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 $N = 10$).

## VI. Conclusions

In this paper, we adopted a cross-layer perspective to study the throughput-delay tradeoff in the cellular multicast channel. Towards this end, we proposed three classes of scheduling algorithms with progressively increasing complexity, and characterized the throughput-delay points achieved by each class. We first considered the class of low-complexity static schedulers with memoryless decoding and showed that the proposed median user scheduler achieves near-optimal scaling laws of both throughput and delay. We then proposed an incremental redundancy multicast scheme that achieves a superior throughput-delay tradeoff, at the expense of increased encoding/decoding complexity. We further proposed a cooperation scheme that achieves the optimal scaling laws of both throughput and delay at the expense of a high RF and computational complexity. We also analyzed the effect of multiple transmit antennas at the base station, and showed that the channel hardening effect enables the worst user scheduler to achieve a near-optimal throughput scaling law for a large number of transmit antennas. Finally, we presented simulation results that establish the accuracy of the predictions of our asymptotic analysis in systems with low to moderate number of users.

## APPENDIX I
## WORST USER SCHEDULER (PROOF OF THEOREM 3)

The average throughput of the worst user scheduler is given by

$$R_{tot} \;=\; N\mathbb{E}\left[\log\left(1+|h_{\pi(1)}|^2P\right)\right].$$

Since the $\{|h_i|^2\}$ are i.i.d. and exponentially distributed, the random variable $|h_{\pi(1)}|^2 = \min_i |h_i|^2$ also follows an exponential distribution, and hence

$$R_{tot} \;=\; N\int_0^{\infty}\log(1+xP)Ne^{-Nx}\mathrm{d}x \;=\; -Ne^{\left(\frac{N}{P}\right)}Ei\left(-\frac{N}{P}\right), \tag{11}$$

where $Ei(x) = \int_{-\infty}^{x}(e^t/t)\mathrm{d}t$. For large values of $x$, we have

$$Ei(-x) \;=\; \int_{-\infty}^{-x}\frac{e^t}{t}\mathrm{d}t \;=\; -\frac{e^{-x}}{x}\left(1+\epsilon\right),$$

where $\epsilon \to 0$ as $x \to \infty$. Using this fact in (11), we get

$$R_{tot} \;=\; P(1+\epsilon) \;=\; \Theta(1).$$

We now calculate the average delay of the worst user scheduler. The BS maintains a single common queue for all the users in the system. We consider each coherence interval of length $T_c$ as a time slot. The service time $X$ is defined as

$$X \;=\; kT_c, \quad \text{when} \quad T_c\left(\sum_{i=1}^{k-1}R_i\right) < \; S \; \leq T_c\left(\sum_{i=1}^{k}R_i\right) \quad (k\in\{1,2,\ldots\}). \tag{12}$$

Here $S$ denotes the size of each packet and $R_i$ represents the service rate in the $i^{th}$ time slot, which is given by $R_i = \log(1+|h^i_{\pi(1)}|^2P)$. We let $C = (S/T_c)$ in the sequel. We consider the sequence of random variables $\{R_i\}$ and define a stopping instant $\tau$ as follows:

$$\tau \;=\; \min\left\{k : \sum_{i=1}^{k}R_i \;\geq\; C\right\}.$$

Using the stopping rule property [20], we get $\mathbb{E}[\tau]\mathbb{E}[R] = \mathbb{E}[\hat{C}] = \mathbb{E}[C+\tilde{C}]$, where $\tilde{C}$ represents the overshoot of the sum of $R_i$'s with respect to the threshold $C$ (Hence $\mathbb{E}[\tilde{C}] \leq \mathbb{E}[R]$). Thus the mean stopping time is given by

$$\frac{C}{\mathbb{E}[R]} \;\leq\; \mathbb{E}[\tau] \;\leq\; 1+\frac{C}{\mathbb{E}[R]} \qquad \Rightarrow \;\; \mathbb{E}[\tau] \;=\; \Theta\left(1+\frac{C}{\mathbb{E}[R]}\right).$$

Thus the average service time is given by

$$\bar{X} \;=\; \mathbb{E}[\tau]T_c \;=\; \Theta\left(T_c + \frac{S}{\mathbb{E}[R]}\right). \tag{13}$$

Since for large values of $N$, the average service rate $\mathbb{E}[R] = (R_{tot}/N) = \Theta(1/N)$, the average delay of the worst user scheduler scales as

$$D \;=\; \bar{X} \;=\; \Theta(T_c + NS) \;=\; \Theta(N).$$

## APPENDIX II

## BEST USER SCHEDULER (PROOF OF THEOREM 4)

The average throughput of the best user scheduler is given by

$$R_{tot} \;=\; \mathbb{E}\left[\log\left(1 + |h_{\pi(N)}|^2 P\right)\right] \;=\; \int_0^\infty \log(1 + xP)\mathrm{d}F(x),$$

where $|h_{\pi(N)}|^2 = \max_i |h_i|^2$ has the distribution $F(x) = \left(1 - e^{-x}\right)^N$, $x \geq 0$. Integrating by parts and simplifying, we get

$$R_{tot} \;=\; \sum_{i=1}^{N} \binom{N}{i} (-1)^i e^{\left(\frac{i}{P}\right)} Ei\left(\frac{-i}{P}\right). \tag{14}$$

It has been shown in [21] that the average throughput in (14) scales as

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

with the number of users $N$.

For calculating the average delay of the best user scheduler, we follow the approach used in [21]. Here the BS maintains $N$ queues, one for each user in the system. These queues are coupled, in the sense that any packet that needs to be transmitted enters all the $N$ queues (since it needs to be transmitted to all the users). Moreover, the BS serves only one of these $N$ queues during any particular time slot. We first calculate the average service time $\bar{X}$ required for transmitting a packet from a queue when the BS *always* serves that particular queue. The average service rate of the best user scheduler is given by $\mathbb{E}[R] = R_{tot}$. Thus, following the argument in the earlier proof and using (15), it can be shown that (refer (13))

$$\bar{X} \;=\; \Theta\left(T_c + \frac{S}{\mathbb{E}[R]}\right) \;=\; \Theta\left(T_c + \frac{S}{\log\log N}\right) \;=\; \Theta(1). \tag{16}$$

We are interested in determining the delay involved in successfully transmitting a particular packet from *all* of the $N$ coupled queues. The actual delay, as defined in Section II, is the time between the start of transmission of a packet and the instant when the packet reaches all the $N$ users in the system. In our analysis, we assume that the packet of interest is at the head of all the $N$ queues during the start of transmission. This assumption thus results in a lower bound on the actual delay.

We characterize the delay based on the observation that our queuing problem is equivalent to the well-known "coupon collector" problem. A similar observation was made earlier in [14], where the authors characterized the delay of the throughput-optimal broadcast scheme. They assumed that the server (BS) offers a constant rate of service, which is independent of the instantaneous channel gains of the users. In our analysis, however, we incorporate the effects of rate adaptation. Let $X_1, X_2, \cdots, X_N$ denote the service times (assuming continuous service) required for transmitting a packet from each of the $N$ queues. Then the delay of the scheduler is directly proportional to the minimum number of trials required to ensure that the first queue is served at least $(X_1/T_c)$ times by the base station, the second queue is served at least $(X_2/T_c)$ times, and so on ...

We lower bound the average delay by calculating the minimum number of trials $N_t$ required to ensure that all the $N$ queues are served at least $(X_{min}/T_c)$ times by the BS, where $X_{min} = \min\{X_1, X_2, \cdots, X_N\}$. We determine the average number of such required trials $\mathbb{E}[N_t|X_{min}]$ using the results derived in [14]. Since the BS serves only one of the $N$ queues at any time and since the fading is symmetric across users, there is an equal probability that the BS serves any one of the queues. Thus the probabilities $\{p_j\}$ of the server choosing the $j^{th}$ queue for service are given by $p_1 = \cdots = p_N = (1/N)$. These probabilities $\{p_j\}$ remain constant through all time slots and are not functions of the instantaneous service rates $\{R_i\}$ provided by the BS. The Moment Generating Function (MGF) of the number of trials required is given by [14]

$$F_{N_t|X_{min}}(z) \;=\; \sum_{i=0}^{\infty} z^i \mathrm{Pr}(N_t > i) \;=\; \sum_{i=0}^{\infty} z^i b_i,$$

where $b_i$ is the probability of failure of sending a packet to all the users in $i$ channel uses. The value of $b_i$ is equal to the polynomial $(x_1 + x_2 + \cdots + x_N)^i/N^i$ evaluated at $x_1 = \cdots = x_N = 1$ after removing all terms that have all $x_i$'s with exponent larger than or equal to $(X_{min}/T_c)$ (denoted by the operator $\{.\}$) [15]. Thus, the MGF of the number of trials required is given by

$$F_{N_t|X_{min}}(z) \;=\; \sum_{i=0}^{\infty} \frac{z^i}{N^i} \left\{ (x_1 + \cdots + x_N)^i \right\}.$$

Using the identities [15]

$$\frac{z^i}{N^i} = \frac{N}{i!z}\int_0^\infty e^{-\frac{Nt}{z}} t^i \mathrm{d}t \qquad \text{and}$$

$$\sum_{i=0}^{\infty} \frac{\{(x_1+\cdots+x_N)^i\}}{i!} = \left\{e^{(x_1+\cdots+x_N)}\right\} = e^{(x_1+\cdots+x_N)} - \prod_{i=1}^{N}\left(e^{x_i} - S_{\left(\frac{X_{min}}{T_c}\right)}(x_i)\right),$$

where $S_m(t) = \sum_{i=0}^{m-1}(t^i/i!)$, we get

$$F_{N_t|X_{min}}(z) = \frac{N}{z}\int_0^\infty e^{-\frac{Nt}{z}}\left(e^{Nt}\left[1-\left(1-S_{\left(\frac{X_{min}}{T_c}\right)}(t)e^{-t}\right)^N\right]\right)\mathrm{d}t.$$

Hence the average number of trials required $\mathbb{E}[N_t|X_{min}]$ is given by [14]

$$\begin{aligned}\mathbb{E}[N_t|X_{min}] &= F_{N_t|X_{min}}(1) = N\int_0^\infty \left[1-\left(1-S_{\left(\frac{X_{min}}{T_c}\right)}(t)e^{-t}\right)^N\right]\mathrm{d}t \\ &= N\mathbb{E}\left[\max_{1\le i\le N} Y_i\right],\end{aligned}$$

where the $Y_i$'s are i.i.d. random variables that follow a Chi-square distribution with $(2X_{min}/T_c)$ degrees of freedom. Using the results in [14], it can be shown that for such a sequence of random variables $\{Y_i\}$,

$$\mathbb{E}\left[\max_{1\le i\le N} Y_i\right] = \max\left\{\Theta(\log N), \Theta\left(\frac{X_{min}}{T_c}\right)\right\}. \tag{17}$$

Thus the average number of trials required is given by

$$\mathbb{E}[N_t|X_{min}] = \max\left\{\Theta\left(N\log N\right), \Theta\left(\frac{NX_{min}}{T_c}\right)\right\}.$$

Hence the average delay of the best user scheduler can be lower bounded by

$$D \ge \mathbb{E}_{X_{min}}\left[\mathbb{E}[N_t|X_{min}]T_c\right] = \mathbb{E}_{X_{min}}\left[\max\left\{\Theta\left(NT_c\log N\right), \Theta\left(NX_{min}\right)\right\}\right].$$

Since $\mathbb{E}\left[\max\{Z_1, Z_2\}\right] \ge \max\left\{\mathbb{E}[Z_1], \mathbb{E}[Z_2]\right\}$, we have

$$D = \max\left\{\Omega\left(NT_c\log N\right), \Omega\left(N\mathbb{E}[X_{min}]\right)\right\}.$$

By observing that $\mathbb{E}[X_{min}] \le \bar{X}$ and using (16), we get

$$D = \Omega\left(NT_c\log N\right) = \Omega\left(N\log N\right). \tag{18}$$

## APPENDIX III
## MEDIAN USER SCHEDULER (PROOF OF THEOREM 5)

In the median user scheduler, the BS keeps on repeating the same packet to the best $(N/2)$ users in each time slot, until all the $N$ users receive it successfully. Due to this repetition, some of the users receive redundant information (multiple copies of the same packet). Hence the average throughput of this scheduler **cannot** be expressed as

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

where $|h_{\pi(\frac{N}{2}+1)}|^2$ is the median of the channel gains among all the $N$ users in the system. However, the average throughput can be easily calculated using the following renewal theory argument. Consider the renewal process wherein the successful reception of a packet of size $S$ by all the $N$ users is taken to be the renewal event. Since the average inter-renewal time is given by the average delay $D$, it is straightforward to show, using the renewal reward theorem, that the average throughput of the scheduler is

$$R_{tot} \;=\; \frac{NS}{D}.$$

Thus, we first need to characterize the average delay $D$ of the median user scheduler. The average service rate provided to any user is given by

$$\mathbb{E}[R] \;=\; \mathbb{E}\left[\log\left(1+|h_{\pi(\frac{N}{2}+1)}|^2P\right)\right].$$

We first characterize the scaling of $\mathbb{E}[R]$ with $N$. Now suppose that the BS does not repeat the same packet after one transmission. Then the average throughput obtained is $T=(N/2)\mathbb{E}[R]$. From the results on central order statistics in [22] (Theorem 8.5.1), we know that the sample median of $N$ i.i.d. exponential random variables converges in distribution to a normal random variable with mean $\theta$ and variance $(1/N)$, where $\theta=\log 2$ is the median of the underlying exponential distribution. Hence

$$\left(|h_{\pi(\frac{N}{2}+1)}|^2-\theta\right)\sqrt{N} \quad\rightarrow\quad W \quad \text{in distribution,} \tag{19}$$

where $W$ is a standard normal random variable. Using Chebyshev's inequality, we get $\forall\epsilon>0$,

$$\begin{aligned}\mathbf{Pr}\left(\left||h_{\pi(\frac{N}{2}+1)}|^2-\theta\right|>\epsilon\right) \;&=\; \mathbf{Pr}\left(\sqrt{N}\left||h_{\pi(\frac{N}{2}+1)}|^2-\theta\right|>\epsilon\sqrt{N}\right)\\ &<\; \frac{\mathbb{E}[W^2]+\delta}{N\epsilon^2}\rightarrow 0 \quad \text{as } N\rightarrow\infty.\end{aligned}$$

Thus $|h_{\pi(\frac{N}{2}+1)}|^2 \to \theta$ in probability. Since the $\log(.)$ function is continuous,

$$\log\left(1+|h_{\pi(\frac{N}{2}+1)}|^2P\right) \quad \to \quad \log(1+\theta P) \quad \text{in probability.} \tag{20}$$

We now derive a lower bound on $T$. We recall the following property of positive random variables. Let $(X_n)$ be a set of positive random variables converging to a constant $A$ in probability. Then $\forall \epsilon > 0$, $\Pr(|X_n - A| \geq \epsilon) < \delta$, for some small $\delta > 0$. Now

$$\mathbb{E}[X_n] \;=\; \int_0^\infty t f_{X_n}(t)\mathrm{d}t \;\;\geq\;\; \int_{A-\epsilon}^{A+\epsilon} t f_{X_n}(t)\mathrm{d}t \;\;\geq\;\; (A-\epsilon)(1-\delta).$$

Taking the limit as $n \to \infty$, we get $\lim_{n\to\infty} \mathbb{E}[X_n] \geq A$. Using this property in (20), we get

$$\lim_{N\to\infty} \mathbb{E}\left[\log\left(1+|h_{\pi(\frac{N}{2}+1)}|^2P\right)\right] \;\geq\; \log(1+\theta P) = \Theta(1) \quad \Rightarrow \quad T \;=\; \Omega(N). \tag{21}$$

Combining this with the upper bound on $T$ in (4), we get

$$T \;=\; \Theta(N) \qquad \Rightarrow \quad \mathbb{E}[R] \;=\; \mathbb{E}\left[\log\left(1+|h_{\pi(\frac{N}{2}+1)}|^2P\right)\right] \;=\; \Theta(1). \tag{22}$$

Thus the average service rate in the median user scheduler does not scale with $N$. We now consider an extension of the coupon collector problem, where the users are assumed to have coupons and the transmitter is the collector that selects $(N/2)$ different users randomly (with a uniform distribution) in each trial, and collects one coupon from each of them. We characterize the average number of trials $\overline{N_t}$ required to ensure that the collector collects at least one coupon from all the $N$ users. An upper bound can be easily derived by considering a *weaker* modified scheme, where in each trial, the $(N/2)$ users are chosen with replacement by the collector from the set of $N$ users. Thus any user can be selected multiple times within the same trial, and hence the average number of trials required for this weaker scheme will be greater than that for the original scheme. It is easy to see that this weaker scheme is in fact the original coupon collector problem [15] with $(N/2)$ independent coupons collected at each instant. Thus

$$\overline{N_t} \;=\; O\left(\frac{N\log N}{(N/2)}\right) \;=\; O(\log N).$$

A lower bound on $\overline{N_t}$ is derived as follows. During the $k^{th}$ trial, the probability that coupon $i$ has not been collected is $(1/2)^k$. The expected number of coupons that have not been collected until the $k^{th}$ trial is given by $E_{N,k} = N(1/2)^k$ [20]. We find the number of trials $k_\delta$ required to ensure that Pr(collecting coupons from all $N$ users within $k_\delta$ trials) $> 1-\delta$, for some small $\delta > 0$. This requires that $E_{N,k} < \epsilon$ for some small $\epsilon > 0$.

$$\Rightarrow \quad \frac{N}{2^{k_\delta}} \; < \; \epsilon \quad \Rightarrow \quad k_\delta \; > \; \log_2\left(\frac{N}{\epsilon}\right) \quad \Rightarrow \quad k_\delta \; = \; \Omega(\log N).$$

Since ensuring that coupons have been collected from all users is stronger than the condition Pr(collecting coupons from all $N$ users within $k_\delta$ trials) $> 1-\delta$, the value $k_\delta$ serves as a lower bound for $\overline{N_t}$. Thus $\overline{N_t} = \Theta(\log N)$. Using (22) and the property (19), it can be shown that the average delay $D$ of the median user scheduler scales as

$$D \; = \; \Theta\left(\overline{N_t}\left(T_c + \frac{S}{\mathbb{E}[R]}\right)\right) \; = \; \Theta(\log N).$$

The average throughput of the median user scheduler is hence given by

$$R_{tot} \; = \; \frac{NS}{D} \; = \; \Theta\left(\frac{N}{\log N}\right).$$

## Appendix IV

## Incremental Redundancy Multicast (Proof of Theorem 6)

Let $A_i$ denote the event that a packet is successfully decoded by all the $N$ users in the system in $i$ transmission attempts. Following the notation in [17], we define

$$q(m) \; = \; \Pr(\overline{A_1}, \ldots, \overline{A_{m-1}}, A_m) \; = \; p(m-1) - p(m),$$

where

$$p(m) \; = \; \Pr(\overline{A_1}, \ldots, \overline{A_{m-1}}, \overline{A_m}) \; = \; 1 - \sum_{l=1}^{m} q(l),$$

with $p(0) = 1$. The rate $\bar{R}$ is defined as $\bar{R} = (b/L)$. We define the random variable $\tau$ to be the number of transmission attempts made between the instant when the codeword is generated and the instant when its transmission is stopped (Transmission is stopped either when the packet is successfully decoded by all the $N$ users or the number of transmission attempts exceeds the rate constraint $M$). The probability distribution of $\tau$ is given by

$$f_\tau(m) \; = \; \begin{cases} 0, \quad m = 0 \\ q(m), \quad 1 \leq m \leq M-1 \\ q(M) + p(M), \quad m = M \end{cases} .$$

We define the random reward $\mathcal{R}$ as follows: $\mathcal{R} = N\bar{R}$ if transmission stops because of successful decoding and $\mathcal{R} = 0$ if transmission stops because of the rate constraint violation. Hence

$$\mathbb{E}[\mathcal{R}] \;=\; N\bar{R}\sum_{m=1}^{M} q(m) \;=\; N\bar{R}[1-p(M)].$$

The mean inter-renewal time is given by

$$\begin{aligned}\mathbb{E}[\tau] \;&=\; \sum_{m=1}^{M} m f_\tau(m) \;=\; \sum_{m=1}^{M} m q(m) + Mp(M)\\ &=\; \sum_{m=1}^{M} m[p(m-1)-p(m)] + Mp(M) \;=\; \sum_{m=0}^{M-1} p(m).\end{aligned}$$

Applying the renewal-reward theorem, we obtain the average throughput of the proposed scheme as $R_{tot} = (\mathbb{E}[\mathcal{R}]/\mathbb{E}[\tau])$ with probability $1$. Hence

$$R_{tot} \;=\; \frac{N\bar{R}\,[1-p(M)]}{1+\sum_{m=1}^{M-1} p(m)}.$$

The average delay $D$ of the scheme is given by the mean inter-renewal time. Hence $D = \mathbb{E}[\tau]$. The unconstrained throughput and delay are obtained by letting $M \to \infty$ and are given by

$$R_{tot} \;=\; \frac{N\bar{R}}{\sum_{m=0}^{\infty} p(m)} \quad \text{and} \quad D \;=\; \sum_{m=0}^{\infty} p(m). \tag{23}$$

From the earlier definitions, we have

$$\begin{aligned}p(m) \;&=\; \mathbf{Pr}(\overline{A_1},\ldots,\overline{A_{m-1}},\overline{A_m}) \;=\; \mathbf{Pr}(\overline{A_m}) \;=\; \mathbf{Pr}\left(\min_{i=1}^{N}\sum_{k=1}^{m} I(X;Y_{ik}) \le \bar{R}\right)\\ &=\; 1-\left[1-\mathbf{Pr}\left(\sum_{k=1}^{m} I(X;Y_{1k}) \le \bar{R}\right)\right]^N. \end{aligned}\tag{24}$$

For a Gaussian input distribution, we have

$$\sum_{k=1}^{m} I(X;Y_{1k}) \;=\; \sum_{k=1}^{m}\log(1+|h_k|^2P).$$

We know that

$$\log\left(1+\sum_{k=1}^{m}|h_k|^2P\right) \;\le\; \sum_{k=1}^{m}\log(1+|h_k|^2P) \;\le\; \sum_{k=1}^{m}|h_k|^2P\ .$$

Hence

$$\mathbf{Pr}\left(\sum_{k=1}^{m}|h_k|^2P \le (e^{\bar{R}}-1)\right) \ge \mathbf{Pr}\left(\sum_{k=1}^{m}\log(1+|h_k|^2P) \le \bar{R}\right) \ge \mathbf{Pr}\left(\sum_{k=1}^{m}|h_k|^2P \le \bar{R}\right).$$

Since both $\bar{R}$ and $(e^{\bar{R}}-1)$ are constants, substituting both the lower and upper bounds in (24) will yield the same scaling with $N$. So we consider only the lower bound on $p(m)$. Let

$$s(m) \;=\; 1-\left[1-\Pr\left(\sum_{k=1}^{m}|h_k|^2 P \leq \bar{R}\right)\right]^N .$$

Hence $\sum_{m=0}^{\infty} p(m) \;=\; \Theta\left(\sum_{m=0}^{\infty} s(m)\right)$ w.r.t $N$. The random variable $\sum_{k=1}^{m}|h_k|^2$ has a $2m$-dimensional Chi-square distribution with the density and distribution functions given by

$$f(x) \;=\; \frac{e^{-x}x^{m-1}}{(m-1)!} \quad \text{and} \quad F(x) \;=\; 1-e^{-x}\left(\sum_{l=0}^{m-1}\frac{x^l}{l!}\right), \quad x \geq 0.$$

Letting $C=(\bar{R}/P)$, we have

$$s(m) \;=\; 1-\left[e^{-C}\left(\sum_{l=0}^{m-1}\frac{C^l}{l!}\right)\right]^N .$$

From Taylor's theorem, we know that (for some $0<\theta<1$)

$$e^C \;=\; \sum_{l=0}^{m-1}\frac{C^l}{l!}+\frac{e^{\theta C}C^m}{m!} \qquad \Rightarrow \quad \sum_{l=0}^{m-1}\frac{C^l}{l!} \;=\; e^C-\frac{e^{\theta C}C^m}{m!}$$

$$\Rightarrow \quad s(m) \;=\; 1-\left(1-\frac{e^{-(1-\theta)C}C^m}{m!}\right)^N .$$

To find the scaling of $\sum_{m=0}^{\infty} s(m)$ w.r.t $N$, we first derive a lower bound by finding the value of $m$ until which $s(m)\to 1$ as $N\to\infty$. Now

$$s(m)\to 1 \quad \Rightarrow \quad \left(1-\frac{e^{-(1-\theta)C}C^m}{m!}\right)^N \to 0 \quad \Rightarrow \quad \frac{e^{-(1-\theta)C}C^m}{m!} \;>\; \Theta\left(\frac{1}{N}\right).$$

Using Stirling's approximation, we have

$$\frac{e^{-(1-\theta)C}C^m}{\sqrt{2\pi m}e^{-m}m^m} \;>\; \frac{k}{N}, \quad \forall \quad \text{constant } k.$$

Taking log on both sides, we get

$$(1-\theta)C-m+m\log\left(\frac{m}{C}\right)+\frac{1}{2}\log(2\pi m) \;<\; \log N-\log k, \quad \forall k.$$

For large $N$, this equation can be reduced to $m\log m<\log N$. This equation is satisfied by all values of $m$ such that

$$m \;<\; \Theta\left(\frac{\log N}{\log\log N}\right).$$

Since $s(m) \to 1$ as $N \to \infty$ for all values of $m$ that satisfy the above equation, the sum of $s(m)$'s can be lower bounded as

$$\sum_{m=0}^{\infty} s(m) \;\geq\; \Theta\left(\frac{\log N}{\log\log N}\right). \tag{25}$$

Similarly an upper bound on $\sum_{m=0}^{\infty} s(m)$ can be derived by finding the value of $m$ from which $s(m) \to 0$ as $N \to \infty$. Following the same procedure as before, we find that $s(m) \to 0$ when $m > \Theta(\log N/\log\log N)$. This yields the following upper bound

$$\sum_{m=0}^{\infty} s(m) \;\leq\; \Theta\left(\frac{\log N}{\log\log N}\right).$$

Combining this with the lower bound in (25), we get

$$\sum_{m=0}^{\infty} s(m) \;=\; \Theta\left(\frac{\log N}{\log\log N}\right).$$

Thus the average delay is given by

$$D \;=\; \sum_{m=0}^{\infty} p(m) \;=\; \Theta\left(\sum_{m=0}^{\infty} s(m)\right) \;=\; \Theta\left(\frac{\log N}{\log\log N}\right).$$

The average throughput of the incremental redundancy scheme is then given by

$$R_{tot} \;=\; \frac{N\bar{R}}{\sum_{m=0}^{\infty} p(m)} \;=\; \frac{N\bar{R}}{D} \;=\; \Theta\left(\frac{N\log\log N}{\log N}\right).$$

## APPENDIX V
## COOPERATIVE MULTICAST (PROOF OF THEOREM 7)

The average throughput in the first stage of the cooperation scheme is given by

$$\left(\frac{N}{2}\right)\mathbb{E}[R_{s1}] \;=\; \left(\frac{N}{2}\right)\mathbb{E}\left[\log\left(1+|h_{\pi(\frac{N}{2}+1)}|^2 P\right)\right].$$

It is shown in (22) that $\mathbb{E}[R_{s1}] = \Theta(1)$. We now characterize the average throughput in the second stage of the cooperation scheme. As noted earlier, the cooperative transmission by the users in the second stage is equivalent to the transmission of packets from a transmitter equipped with $(N/2)$ transmit antennas to the worst user in a group of $(N/2)$ users. Hence the average transmission rate during the cooperative stage is given by

$$\mathbb{E}[R_{s2}] \;=\; \mathbb{E}\left[\min_{i=1,\ldots,(N/2)} \log\left(1+\frac{|h_{1i}|^2+\cdots+|h_{(N/2)i}|^2}{(N/2)}P\right)\right],$$

where the $|h_{ki}|^2$'s represent the inter-user fading power gains that are i.i.d. and exponentially distributed.

$$\Rightarrow \quad \mathbb{E}[R_{s2}] \;=\; \mathbb{E}\left[\log\left(1+\min_{i=1,\ldots,M}\frac{|\chi_{2M}^i|^2}{M}P\right)\right], \tag{26}$$

where $M=(N/2)$ and $|\chi_{2M}^i|^2$'s are Chi-square random variables with $2M$ degrees of freedom whose distribution function is given by

$$F(x) \;=\; 1-e^{-x}\left(\sum_{j=0}^{M-1}\frac{x^j}{j!}\right), \quad x\geq 0.$$

Using the results on extreme order statistics in [22] (Theorems 8.3.2-8.3.6), it can be shown that the random variable $\left(\min_{i=1}^M|\chi_{2M}^i|^2\right)/b_M\to W$ in distribution as $M\to\infty$, where $W$ is a Weibull type random variable and $b_M$ satisfies $F(b_M)=(1/M)$. Now

$$F(b_M) \;=\; \frac{1}{M} \quad\Rightarrow\quad 1-e^{-b_M}\left(\sum_{j=0}^{M-1}\frac{b_M^j}{j!}\right) \;=\; \frac{1}{M}.$$

Using Taylor's theorem, we get for some $0<\beta_M<1$

$$1-e^{-b_M}\left(e^{b_M}-\frac{e^{\beta_M b_M}b_M^M}{M!}\right) \;=\; \frac{1}{M} \quad\Rightarrow\quad \frac{e^{-(1-\beta_M)b_M}b_M^M}{M!} \;=\; \frac{1}{M}.$$

Using Stirling's approximation, we have

$$\frac{e^{-(1-\beta_M)b_M}b_M^M}{\sqrt{2\pi M}M^M e^{-M}} \;=\; \frac{1}{M}.$$

Taking $\log(.)$ on both sides, we get

$$(1-\beta_M)b_M - M\log b_M \;=\; M-\left(M-\frac{1}{2}\right)\log M + C.$$

Since $\beta_M\to 0$ as $M\to\infty$, we get $b_M=\Theta(M)$. Thus $\left(\min_{i=1}^M|\chi_{2M}^i|^2\right)/M\to kW$ in distribution, for some constant $k>0$. Since the $\log(.)$ function is continuous, we have

$$\log\left(1+\frac{\min_{i=1}^M|\chi_{2M}^i|^2}{M}P\right) \;\to\; \log(1+kWP) \quad \text{in distribution, as } M\to\infty.$$

Now, we know

$$\log\left(1+\frac{\min_{i=1}^M|\chi_{2M}^i|^2}{M}P\right) \;\leq\; \frac{\left(\min_{i=1}^M|\chi_{2M}^i|^2\right)P}{M} \;\leq\; \frac{|\chi_{2M}^1|^2P}{M}\;.$$

Since

$$\mathbb{E}\left[\left(\frac{|\chi_{2M}^1|^2P}{M}\right)^2\right] \;=\; \frac{\mathbb{E}[(|\chi_{2M}^1|^2)^2]P^2}{M^2} \;=\; \left(1+\frac{1}{M}\right)P^2 \;\leq\; 2P^2 \;<\; \infty \quad \forall M,$$

the sequence $\{(|\chi_{2M}^1|^2 P)/M \quad ; \quad M \geq 1\}$ is uniformly integrable.

$$\Rightarrow \quad \left\{ \log\left(1 + \frac{\min_{i=1}^{M} |\chi_{2M}^i|^2}{M} P\right) \quad ; \quad M \geq 1 \right\} \quad \text{is uniformly integrable.}$$

It is shown in [23] that if a sequence of random variables $\{X_n\}$ is uniformly integrable and $X_n \to X$ in distribution as $n \to \infty$, then $\mathbb{E}X_n \to \mathbb{E}X$ as $n \to \infty$. Thus

$$\mathbb{E}\left[\log\left(1 + \frac{\min_{i=1}^{M} |\chi_{2M}^i|^2}{M} P\right)\right] \quad \to \quad \mathbb{E}[\log(1 + kWP)] \; = \; \Theta(1).$$

Hence the average transmission rate of the second stage is given by $\mathbb{E}[R_{s2}] = \Theta(1)$ w.r.t $N$. Since both $\mathbb{E}[R_{s1}]$ and $\mathbb{E}[R_{s2}]$ do not scale with $N$, we have $\mathbb{E}\left[\min\{R_{s1}, R_{s2}\}\right] = \Theta(1)$. Thus the average throughput of the cooperation scheme is given by

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

We now determine the average delay of the cooperation scheme. We note that the BS needs to maintain only a single queue that caters to all the $N$ users in the system. The information transmitted by the BS in the first half of each time slot reaches all the $N$ users at the end of that time slot. Hence the average delay is equal to the average service time required for transmitting a packet of size $S$ from the queue. Following the steps in Appendix I, the average delay $D$ for transmitting a packet in the cooperation scheme is given by (refer equation (13))

$$D \; = \; \Theta\left(T_c + \frac{S}{\mathbb{E}[\min\{R_{s1}, R_{s2}\}]}\right) \; = \; \Theta(1).$$

## Appendix VI
## Multi-Transmit Antenna Worst User Scheduler (Proof of Lemma 8)

From the results on extreme order statistics in [22], we know that $(|\chi_{min}|^2/b_N) \to W$ in distribution, where $W$ has a Weibull type distribution and $b_N$ satisfies $F(b_N) = (1/N)$, which implies

$$1 - e^{-Lb_N}\left(\sum_{k=0}^{L-1} \frac{(Lb_N)^k}{k!}\right) \; = \; \frac{1}{N} \; .$$

Using Taylor's theorem, we get for some $0 < \gamma_N < 1$,

$$1 - e^{-Lb_N}\left(e^{Lb_N} - \frac{e^{\gamma_N L b_N}(Lb_N)^L}{L!}\right) \; = \; \frac{1}{N} \quad \Rightarrow \quad \frac{e^{-(1-\gamma_N)Lb_N}(Lb_N)^L}{L!} \; = \; \frac{1}{N} \; .$$

Taking $\log(.)$ on both sides, we get

$$(1-\gamma_N)Lb_N - L\log b_N \;=\; \log N + L\log L - \log(L!)\;.$$

Since $|\chi_{min}|^2 \leq |\chi_1|^2 = \Theta(1)$, we know that $b_N = O(1)$ and hence the $\log b_N$ term dominates the left hand side of the above expression. Thus we have $b_N = \Theta\left(N^{-\left(\frac{1}{L}\right)}\right)$.

$$\Rightarrow \quad N^{\left(\frac{1}{L}\right)}|\chi_{min}|^2 \;\;\rightarrow\;\; kW \quad \text{in distribution, for some constant } k>0.$$

Since $\mathbb{E}\left[|\chi_{min}|^2\right] \leq \mathbb{E}\left[|\chi_1|^2\right] < \infty$, we can use the result in Theorem 2.1 of [24] to conclude that $N^{\left(\frac{1}{L}\right)}\mathbb{E}\left[|\chi_{min}|^2\right] \rightarrow k\mathbb{E}[W] = \Theta(1)$. Thus $\mathbb{E}\left[|\chi_{min}|^2\right] = \Theta\left(N^{-\left(\frac{1}{L}\right)}\right)$. The average throughput of the worst user scheduler can now be upper bounded, using Jensen's inequality, as follows

$$\begin{aligned} R_{tot} \;&=\; N\mathbb{E}\left[\log\left(1+|\chi_{min}|^2P\right)\right] \;\leq\; N\log\left(1+\mathbb{E}\left[|\chi_{min}|^2\right]P\right) \\ \Rightarrow \quad R_{tot} \;&=\; O\left(N^{\left(\frac{L-1}{L}\right)}\right). \end{aligned} \tag{27}$$

We lower bound the average throughput as follows

$$\begin{aligned} R_{tot} \;&=\; N\int_0^\infty \log(1+xP)\mathrm{d}F_{min}(x) \;\geq\; N\int_{b_N}^\infty \log(1+xP)\mathrm{d}F_{min}(x). \\ \Rightarrow \quad R_{tot} \;&\geq\; N\log\left(1+b_NP\right)\left[1-F_{min}(b_N)\right], \end{aligned}$$

where $F_{min}(x) = 1-(1-F(x))^N$. Using the fact that $F(b_N) = (1/N)$, we get

$$F_{min}(b_N) \;=\; 1-\left(1-\frac{1}{N}\right)^N \;=\; 1-e^{N\log\left(1-\frac{1}{N}\right)} \;=\; 1-e^{-1}\left(1+O\left(\frac{1}{N}\right)\right).$$

$$\begin{aligned} \Rightarrow \quad R_{tot} \;&\geq\; N\log\left(1+b_NP\right)\left[e^{-1}+O\left(\frac{1}{N}\right)\right] \\ &=\; \Theta\left(N\log\left(1+N^{-\frac{1}{L}}P\right)\right) \;=\; \Theta\left(N^{\left(\frac{L-1}{L}\right)}\right). \end{aligned}$$

Combining this with the upper bound in (27), we get $R_{tot} = \Theta\left(N^{\left(\frac{L-1}{L}\right)}\right)$.

## APPENDIX VII

## MULTI-TRANSMIT ANTENNA BEST USER SCHEDULER (PROOF OF LEMMA 9)

From the results on extreme order statistics in [22], we know that

$$\left(\frac{|\chi_{max}|^2 - a_N}{b_N}\right) \quad \to \quad W \qquad \text{in distribution,}$$

where $W$ has a Gumbel distribution and $a_N$ and $b_N$ satisfy $F(a_N) = 1 - (1/N)$ and $b_N = (1/Nf(a_N))$, where $f(.)$ denotes the probability density function obtained from (8). Now

$$F(a_N) \;=\; 1 - \frac{1}{N} \quad \Rightarrow \quad \frac{e^{-La_N}(La_N)^{(L-1)}}{(L-1)!}\left(1 + O\left(\frac{1}{a_N}\right)\right) \;=\; \frac{1}{N}.$$

Taking $\log(.)$ on both sides and using Stirling's approximation, we get

$$\begin{aligned} La_N - (L-1)\log a_N \;&=\; \log N + (L-1) - \frac{1}{2}\log(L-1) + K. \\ \Rightarrow a_N \;&=\; \frac{\log N + (L-1)\log\log N}{L} + O(\log\log N). \end{aligned}$$

Since

$$f(a_N) \;=\; \frac{Le^{-La_N}(La_N)^{(L-1)}}{(L-1)!} \;=\; \Theta\left(\frac{1}{N}\right),$$

we have $b_N = C = \Theta(1)$. Thus

$$|\chi_{max}|^2 - \left(\frac{\log N + (L-1)\log\log N}{L} + O(\log\log N)\right) \quad \to \quad CW \quad \text{in distribution.}$$

Using Chebyshev's inequality, it is easy to show that

$$\frac{|\chi_{max}|^2}{\left(\frac{\log N + (L-1)\log\log N}{L}\right)} \quad \to \quad 1 \quad \text{in probability.}$$

Since any Chi-squared random variable with $2L$ degrees of freedom can be expressed as the sum of $L$ exponential i.i.d. random variables, we have

$$\mathbb{E}\left[|\chi_{max}|^2\right] \;=\; \mathbb{E}\left[\max_{i=1}^{N} \left\{\frac{Z_1^i + \cdots + Z_L^i}{L}\right\}\right] \;\leq\; \mathbb{E}\left[\max_{i=1}^{N} Z_1^i\right],$$

where $Z_j^i$'s are exponential random variables with unit mean. Hence

$$\mathbb{E}\left[\frac{|\chi_{max}|^2}{\left(\frac{\log N + (L-1)\log\log N}{L}\right)}\right] \leq \frac{\mathbb{E}\left[\max_{i=1}^{N} Z_1^i\right]}{\left(\frac{\log N + (L-1)\log\log N}{L}\right)} \leq \frac{k\log N}{\left(\frac{\log N + (L-1)\log\log N}{L}\right)} \leq kL < \infty.$$

Thus we can apply the Dominated Convergence Theorem to get

$$\mathbb{E}\left[\frac{|\chi_{max}|^2}{\left(\frac{\log N + (L-1)\log\log N}{L}\right)}\right] \to 1 \qquad \Rightarrow \quad \mathbb{E}\left[|\chi_{max}|^2\right] = \Theta\left(\frac{\log N + (L-1)\log\log N}{L}\right).$$

Using Jensen's inequality, we get

$$\begin{aligned} R_{tot} &= \mathbb{E}\left[\log\left(1+|\chi_{max}|^2 P\right)\right] \;\leq\; \log\left(1+\mathbb{E}\left[|\chi_{max}|^2\right] P\right) \\ \Rightarrow \quad R_{tot} &= O\left(\log\left(1+\frac{\log N + (L-1)\log\log N}{L}\right)\right). \end{aligned} \tag{28}$$

The average throughput of the best user scheduler can be lower bounded as follows

$$\begin{aligned} R_{tot} &= \int_0^\infty \log(1+xP)\mathrm{d}F_{max}(x) \;\geq\; \int_{a_N}^\infty \log(1+xP)\mathrm{d}F_{max}(x). \\ \Rightarrow \quad R_{tot} &\geq \log\left(1+a_N P\right)\left[1-F_{max}(a_N)\right], \end{aligned}$$

where $F_{max}(x) = (F(x))^N$. Using the fact that $F(a_N) = 1 - \frac{1}{N}$, we get

$$F_{max}(a_N) \;=\; (F(a_N))^N \;=\; \left(1-\frac{1}{N}\right)^N \;=\; e^{-1}\left(1+O\left(\frac{1}{N}\right)\right).$$

$$\Rightarrow \quad R_{tot} \;\geq\; \log\left(1+a_N P\right)\left[1-e^{-1}+O\left(\frac{1}{N}\right)\right] \;=\; \Theta\left(\log\left(1+a_N P\right)\right).$$

$$\Rightarrow R_{tot} \;=\; \Omega\left(\log\left(1+\frac{\log N + (L-1)\log\log N}{L}\right)\right).$$

Combining this with the upper bound in (28), we get

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

## References


[1] R. Berry and R. Gallager, "Communication over fading channels with delay constraints," *IEEE Transactions on Information Theory*, vol. 48, pp. 1135–1149, May 2002.

[2] D. Rajan, A. Sabharwal, and B. Aazhang, "Delay bounded packet scheduling of bursty sources over wireless channels," *IEEE Transactions on Information Theory*, vol. 50, pp. 125–144, January 2004.

[3] 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.

[4] 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.

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

[6] S. Shakkottai and A. Stolyar, "Scheduling for multiple flows sharing a time-varying channel: The exponential rule," *American Mathematical Society Translations, Series 2*, vol. 207, 2002.

[7] M. Airy, S. Shakkottai, and R. Heath Jr., "Spatially greedy scheduling in multi-user MIMO wireless systems," in *IEEE Asilomar Conf. on Signals, Systems, and Computers*, November 2003.

[8] 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.

[9] C. Wu and Y. Tay, "Amris: A multicast protocol for ad hoc wireless networks," in *IEEE MILCOM99*, (Atlantic City, NJ), November 1999.

[10] J. Garcia-Luna-Aceves and E. Madruga, "The core-assisted mesh protocol," *IEEE Journal on Selected Areas in Communications*, vol. 17, pp. 1380–1394, August 1999.

[11] 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.

[12] 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.

[13] T. Cover and J. Thomas, *Elements of Information Theory*. New York: John Wiley Sons, Inc., 1991.

[14] M. Sharif and B. Hassibi, "Delay considerations for opportunistic scheduling in broadcast fading channels," *To appear in IEEE Transactions on Wireless Communications*, 2006.

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

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

[17] 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.

[18] P. Viswanath, D. N. C. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Transactions on Information Theory*, vol. 48, pp. 1277–1294, June 2002.

[19] B. M. Hochwald, T. L. Marzetta, and V. Tarokh, "Multiple-antenna channel hardening and its implications for rate feedback and scheduling," *IEEE Transactions on Information Theory*, vol. 50, no. 9, pp. 1893–1909, 2004.

[20] R. G. Gallagher, *Discrete Stochastic Processes*. Kluwer Academic Publishers, 1996.

[21] M. Sharif and B. Hassibi, "On the capacity of MIMO broadcast channels with partial channel state information," *To appear in IEEE Transactions on Information Theory*, vol. 51, no. 2, 2005.

[22] B. C. Arnold, N. Balakrishnan, and H. N. Nagaraja, *A first course in order statistics*. New York: John Wiley Sons, Inc., 1992.

[23] R. Durrett, *Probability: Theory and Examples*. California: Duxbury Press, Inc., 1996.

[24] J. Pickands, "Moment convergence of sample extremes," *The Annals of Mathematical Statistics*, vol. 39, no. 3, pp. 881–889, 1968.

| Scheme | Average Throughput $R_{tot}$ | Average Delay $D$ |
|---|---|---|
| Worst User Scheduler | $\Theta(1)$ | $\Theta(N)$ |
| Best User Scheduler | $\Theta(\log\log N)$ | $\Omega(N\log N)$ |
| Median User Scheduler | $\Theta\left(\frac{N}{\log N}\right)$ | $\Theta(\log N)$ |
| IR Multicast | $\Theta\left(\frac{N\log\log N}{\log N}\right)$ | $\Theta\left(\frac{\log N}{\log\log N}\right)$ |
| Cooperative Multicast | $\Theta(N)$ | $\Theta(1)$ |

TABLE I

COMPARISON OF THE THROUGHPUT-DELAY TRADEOFFS ACHIEVED BY THE PROPOSED MULTICAST SCHEDULERS

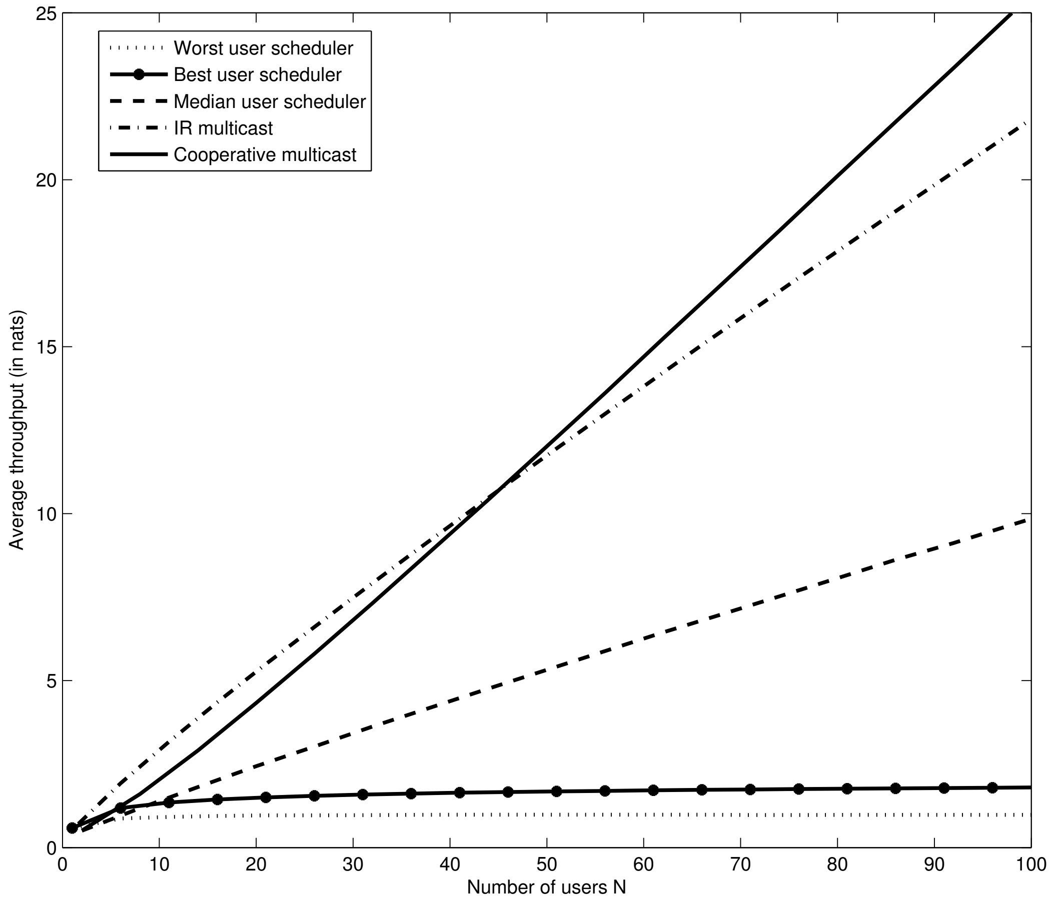


Fig. 1. Comparison of the throughput of the proposed multicast schedulers

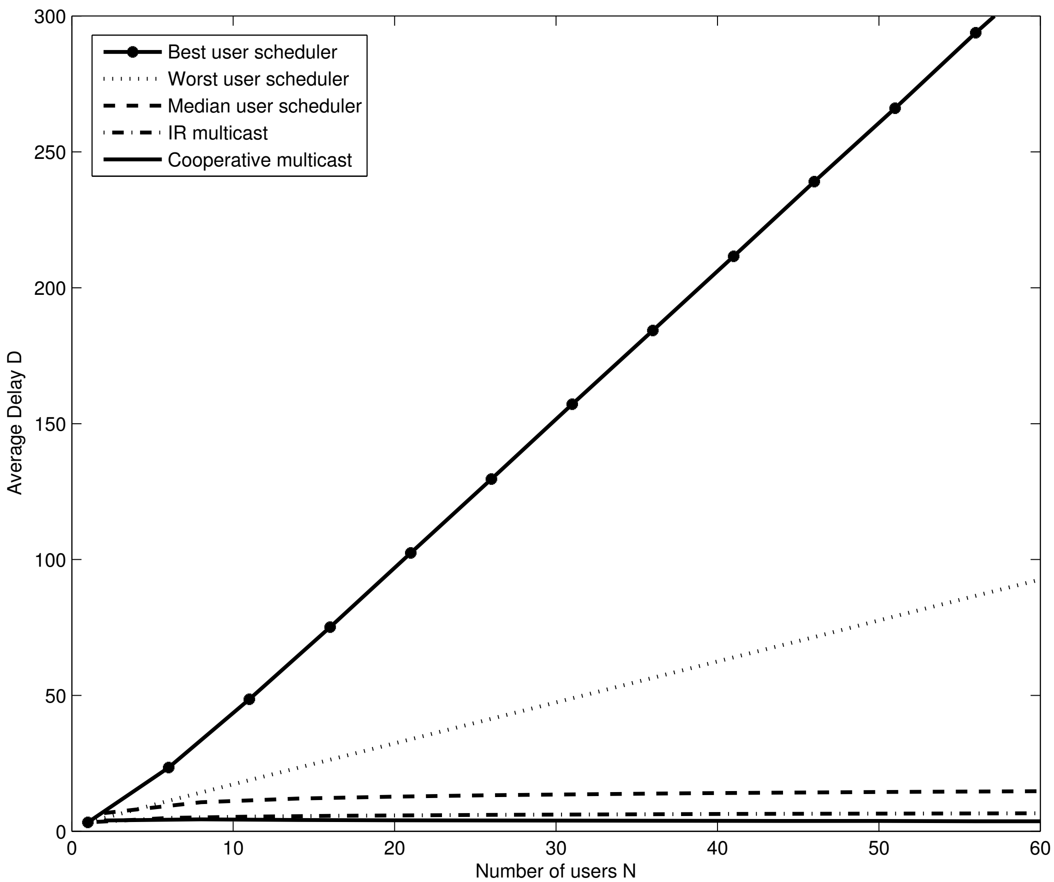


Fig. 2. Comparison of the delay of the proposed multicast schedulers

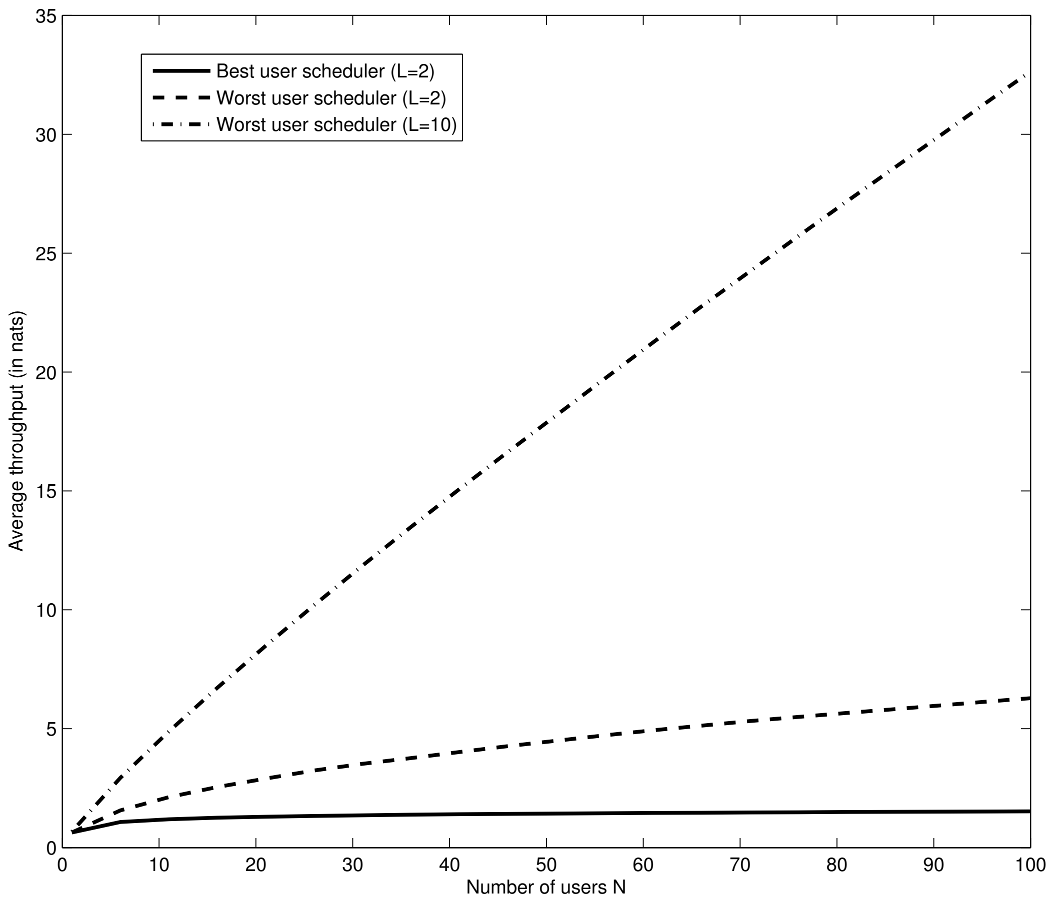


Fig. 3. Comparison of the throughput of the best and worst user schedulers for the multi-transmit antenna scenario (L transmit antennas at the base station)