Opportunistic Multicasting

(Invited Paper)

Praveen Kumar Gopala
Dept. of Electrical & Computer Engineering
The Ohio State University
Columbus, Ohio 43201
Email: gopalap@ece.osu.edu

Hesham El Gamal
Dept. of Electrical & Computer Engineering
The Ohio State University
Columbus, Ohio 43201
Email: helgamal@ece.osu.edu

***Abstract*— In this paper, we develop an information theoretic framework for analyzing the fundamental tradeoffs of the downlink multicasting channel in a single cell system. We consider three classes of scheduling algorithms with varying complexities. The first class strives for minimum complexity by resorting to a static scheduling strategy along with memoryless decoding. Our analysis for the static scheduling algorithms reveals a fundamental throughput-delay tradeoff. In particular, we establish 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 multicasting strategy is shown to achieve the optimal scaling laws of both throughput and delay. Finally, we present simulation results in certain representative scenarios that validate our theoretical claims.**

I. Introduction

It has been observed by several researchers that the joint optimization of the physical, data link, and network layers can allow for significant performance gains in wireless networks. One of the prime examples is the multi-user diversity principle first observed by Knopp and Humblet [1]. Interestingly, through this approach, one can increase the throughput of the wireless system to outperform the *non-faded* additive white Gaussian noise (AWGN) channel [1]. The basic idea in multi-user diversity is to allocate the channel dynamically to the best user, and hence, allow the system to ride the peaks of the fading channel fluctuations. One of the basic assumptions in this approach is that an independent stream of information is requested by every user in the network. In this paper, we consider the multicasting scenario where the same information stream is requested by multiple users. In this scenario, one should attempt to jointly exploit multi-user diversity and the multicasting gain offered by the wireless medium. This multicasting gain results from the fact that a wireless transmission intended to any particular user is naturally *overheard* by all the other users in the network (after passing through possibly different channels).

In this paper, we focus on the **pure** multicasting scenario where the same information stream is transmitted to **all** users in the network. In [2], we extend our study to the multi-group scenario where independent streams of information are transmitted to different groups. In our work, we consider three classes of scheduling algorithms with varying complexities. The first class strives for minimum complexity by resorting to a static scheduling strategy along with memoryless decoding. In this approach, we schedule transmission to a fraction of the users that enjoy favorable channel conditions. We establish the throughput-delay tradeoff allowed by varying the fraction of users targeted in every transmission. To gain more insight into the problem, we study in more detail the three special cases of scheduling transmissions to the best, worst and median user. Here we establish the asymptotic optimality of the throughput attained by the median user scheduler. The second scheduling policy resorts to a higher complexity incremental redundancy encoding/decoding strategy to achieve a better throughput-delay tradeoff. This scheme is based on a hybrid-ARQ strategy and yields a significant reduction in the delay, compared with the median user scheduler, at the expense of a minimal penalty in the throughput. 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. More specifically, we show that the proposed cooperative multicasting strategy achieves the optimal scaling laws of both throughput and delay at the expense of a high complexity.

II. System Model

We consider the downlink of a pure multicasting channel wherein the base station serves N users requesting the same information. The base station has 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. We consider time slotted transmission wherein the signal received by user i in slot k is given by

$$y_i[k] = h_i x[k] + n_i[k],$$

where $x[k]$ denotes the complex-valued signal transmitted in slot k, h_i represents the complex flat fading coefficient of the i^{th} user's channel and $n_i[k]$ represents the complex additive white Gaussian noise (AWGN) at the i^{th} user in slot k. The AWGN is assumed to be circularly symmetric with unit variance and independent across users. The fading is assumed

0-7803-8622-1/04/$20.00 ©2004 IEEE

to be quasi-static with coherence time T_c. Thus the fading coefficients remain constant throughout a slot and change independently from slot to slot. The fading coefficients $\{h_i\}$ are Rayleigh distributed with $E\{|h_i|^2\} = 1$ and i.i.d across users. The users are assumed to be **symmetric** with the same fading statistics. The total transmit power used by the base station is constrained to be less than P, i.e., $E\left[|x[k]|^2\right] \leq P$. We assume perfect CSI at the transmitter (i.e., the downlink channel gains of all users are available at the base station) and that rate adaptation is possible at the base station. We further assume the use of capacity achieving codes and hence appeal to the fundamental information theoretic limits of the channel.

We consider a system of backlogged users (i.e., the base station always has packets to be transmitted to each user in the system). For such a backlogged system, we define the throughput and delay of any scheduling scheme as follows.

Definition 1: The **throughput** of a scheduling scheme is defined as the total system throughput, which is the sum of the throughputs provided by the base station to all the individual users in 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 the packet is successfully decoded by all the users in the system. Hence our definition of delay includes only the transmission delay and does not include queuing delay.

To simplify the delay analysis, we make an exponential server assumption, i.e., the rate of service offered by the server is assumed to follow an exponential distribution with the same mean as that obtained using the original channel rate distibution. For analytical simplicity, we consider scaling laws of the throughput and delay. We use Knuth's asymptotic notations[1] throughout the paper.

III. Static Scheduling Schemes

In a pure multicasting scenario, two specific characteristics of the system can be exploited to yield a significant increase in the throughput. Firstly, the availability of multiple users with independent fading channels can be exploited to yield significant multiuser diversity gains [1], [3]. Secondly, a transmission to any particular user will also be received (with different scalings) by all the other users in the system at no extra cost. Since all the users want the same information, this property of the wireless medium can be exploited to yield significant multicasting gains.

The throughput-optimal scheme for this scenario is an N-level superposition coding/successive decoding scheme. We don't consider this approach in this paper due to its excessive complexity. Instead, we focus on low-complexity static scheduling schemes with memoryless decoding[2] that exploit these two possible gains. In these schemes, we schedule transmissions to a fraction of the users with favorable channel conditions. To gain more insights, we consider two example schemes. The first scheme maximally exploits multiuser diversity by always transmitting to the best user (highest SNR). But the high transmission rate disables the other users from decoding the transmission. Hence the same information needs to be repeated N times before it reaches all the users. Thus this scheme does not exploit any multicasting gains. The second scheme maximally exploits multicasting gains by always transmitting to the worst user (lowest SNR). Due to the low transmission rate, all the users can decode the transmission. However, the inherent multiuser diversity degrades the performance of this scheme. From these two examples, it is clear that there is a tradeoff between exploiting multiuser diversity and multicasting gains. To characterize this tradeoff, we consider a general static scheduling scheme and evaluate the throughput and delay of three special cases of this scheme.

A. General static scheduling scheme

This scheme is such that each transmission by the base station is intended for successful reception by (N/α) users in the system. 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 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 $1 \leq \alpha \leq N$. This scheme is "static" in the sense that the fraction of users targeted in every transmission remains the same (i.e., the parameter α is not a function of time). The average total throughput of this general scheduling scheme is given by

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

where $|h_{(N-\frac{N}{\alpha}+1)}|^2$ 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.

For implementing this scheme, the base station needs to maintain $\binom{N}{N/\alpha}$ 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 and collectively exhaustive (i.e., every user in the system is served by exactly one of the α queues). For example, with $N = 6$ users and $\alpha = 2$, one possible set of coupled queues may serve users $\{1,2,3\}$ & $\{4,5,6\}$. Another possible set may serve users $\{1,2,5\}$ & $\{3,4,6\}$. Hence it is clear that any packet that arrives into a particular set enters 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 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 (i.e., the worst case delay). Moreover, the base station services only one of the $\binom{N}{N/\alpha}$ queues at any time, which is chosen based

[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] In memoryless decoding, the received observations that cannot be successfully decoded by a user are discarded and not used in future decoding attempts

on the instantaneous fading coefficients of all the users. A detailed analysis of the throughput and delay of this general static scheduling scheme is provided in [2].

To characterize the tradeoff between multiuser diversity and multicasting gains, we now consider three special cases of this general static scheduling scheme and determine the scaling of the throughput and delay of each of them with the number of users in the system.

B. Worst user scheme

The worst user scheme corresponds to the case $\alpha = 1$ of the general scheme. This scheme maximally exploits multicasting gains by always transmitting to the worst user in the system. However, the multiuser diversity inherent in the system works against the performance of the scheme and results in a loss in the individual throughput to any user. For implementing this scheme, the base station needs to maintain only a single queue that caters to all the users in the system.

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

$$R_{tot} = \Theta(1) \tag{2}$$

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

$$D = \Theta(N). \tag{3}$$

Sketch of Proof: The average throughput can be calculated from (1) and is given by $R_{tot} = -Ne^{\left(\frac{N}{P}\right)}Ei(-N/P)$, where $Ei(x) = \int_{-\infty}^{x}(e^t/t)\mathrm{d}t$. Using the fact that $Ei(-x) \approx -(e^{-x}/x)$ for large values of x, it can be shown that $R_{tot} = \Theta(1)$. Since $R_{tot} = (N/\alpha)E[R_\alpha]$ and $\alpha = 1$ for this scheme, the average service rate is given by $E[R_1] = \Theta(1/N)$. Hence it can easily be shown that the average delay scales as $D = \Theta(N)$. For a detailed proof, please refer [2]. ■

C. Best user scheme

The best user scheme corresponds to the case $\alpha = N$ of the general scheme. This scheme maximally exploits multiuser diversity by always transmitting to the best user in the system. However, it does not exploit any multicasting gains. Though this scheme is throughput-optimal for the broadcast channel where the users want independent information, it is clearly not throughput-optimal for the multicast channel. For implementing this scheme, the base station needs to maintain N queues, one for each user in the system. At any instant of time, the base station transmits a packet from the queue corresponding to the best user.

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

$$R_{tot} = \Theta(\log\log N) \tag{4}$$

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

$$D = \Omega\left(\frac{N\log N}{\log\log N}\right). \tag{5}$$

Sketch of Proof: It has been shown in [4] that the average throughput of this scheme scales as $R_{tot} = \Theta(\log\log N)$. The average service rate is given by $E[R_N] = R_{tot} = \Theta(\log\log N)$. We first use the exponential server assumption to determine the probability distribution of the required service time when the base station always services the same queue. But at any time instant, the base station serves only the queue corresponding to the best user. Hence any particular user's packet has to wait until the base station services the corresponding queue (i.e., until that user becomes the best user in the system). Moreover since all the users want the same packet, the delay in transmitting the packet is given by the worst case delay. We use the results in [5] and [6] on the "coupon collector" problem to derive a lower bound on this delay. For a detailed proof, please refer [2]. ■

Thus it is clear that exploiting multiuser diversity yields higher throughput gains than exploiting multicasting gains. However, this throughput gain is obtained at the expense of a higher delay and a higher number of required queues at the base station.

D. Median user scheme

The median user scheme corresponds to the case $\alpha = 2$ of the general scheme. This scheme strikes a balance between exploiting multiuser diversity and multicasting gains. The base station always transmits to the median user whose instantaneous SNR is the median of the ordered list of SNRs of all users. Hence each transmission is decoded by half the users in the system and the information needs to be repeated only twice before it reaches all the users. Thus significant multicasting gains are exploited. Moreover, unlike the worst user scheme, the inherent multiuser diversity does not degrade the performance of this scheme. Infact, we show that this scheme achieves the optimal scaling law of the throughput as the number of users N grows to infinity. For implementing this scheme, the base station needs to maintain $\binom{N}{N/2}$ queues, one for each combination of $(N/2)$ users. Hence the number of required queues increases exponentially with N.

Theorem 3: The proposed median user scheme is asymptotically throughput-optimal. The average throughput of this scheme scales as

$$R_{tot} = \Theta(N) \tag{6}$$

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

$$D = \Theta\left(\binom{N}{N/2}\right) = \Theta\left(\frac{2^N}{\sqrt{N}}\right). \tag{7}$$

Sketch of Proof: From (1), the average throughput of this scheme is given by $R_{tot} = \left(\frac{N}{2}\right) E\left[\log\left(1 + |h_{\left(\frac{N}{2}+1\right)}|^2 P\right)\right]$. It is shown in [7] that central order statistics are asymptotically normal under certain conditions. Using this result, it can be shown that the sample median of N i.i.d exponential random variables converges in probability to θ as $N \to \infty$, where θ is the median of the underlying exponential distribution. Thus $R_{tot} = (N/2)\log(1 + \theta P) = \Theta(N)$. An upper bound on the throughput, obtained by considering the ergodic channel case, is given by $R_{tot} \leq R_{erg} = NE\left[\log(1 + |h|^2 P)\right] = \Theta(N)$. This establishes the asymptotic throughput-optimality of the median user scheme. The average service rate is given by $E[R_2] = (2/N)R_{tot} = \Theta(1)$. Since the base station services only one of the $\binom{N}{N/2}$ queues at any time, it can easily be

shown that the average delay scales as $D = \Theta\left(\binom{N}{N/2}\right)$. For a detailed proof, please refer [2]. ■

Hence the throughput optimality of the median user scheme is obtained at the expense of a delay that increases exponentially with N.

IV. Incremental Redundancy Scheme

The schemes proposed earlier employed memoryless decoding. Now we consider a scheme that employs a higher complexity incremental redundancy encoding/decoding strategy to achieve a better throughput-delay tradeoff. The proposed scheme is an extension of the scheme proposed in [8]. 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 are transmitted in the first attempt. If a user cannot decode the transmission, it sends back an ARQ request. If the base station receives an ARQ request from any of the users, it transmits the next L bits of the codeword in the next attempt. This process continues until either all N users successfully decode the information or the rate constraint M is violated. This scheme does not require perfect CSI at the base station[3]. We show that this scheme yields a significant reduction in the delay, compared to the median user scheduler, at the expense of a minimal penalty in the throughput. The unconstrained throughput and delay of the scheme are obtained as $M \to \infty$.

Theorem 4: The average throughput of the incremental redundancy scheme satisfies

$$R_{tot} = \Omega\left(\frac{N \log\log N}{\log N}\right). \tag{8}$$

The average delay D of this scheme satisfies

$$D = O\left(\frac{\log N}{\log\log N}\right). \tag{9}$$

Sketch of Proof: Following the steps in [8], it can be shown that the unconstrained delay and throughput are given by $D = \sum_{m=0}^{\infty} p(m)$ and $R_{tot} = (N\bar{R}/D)$, where $\bar{R} = (b/L)$ and

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

An upper bound on the delay is obtained by finding the values of m beyond which $p(m) \to 0$ as $N \to \infty$. This in turn yields a lower bound on the throughput. For a detailed proof, please refer [2]. ■

Thus the incremental redundancy scheme achieves a better throughput-delay tradeoff than that achieved by the low-complexity static scheduling schemes proposed earlier. However, this scheme uses codewords of much bigger lengths and the decoders need to jointly decode the observations to exploit the incremental redundancy. Hence the better throughput-delay tradeoff is achieved at the expense of a higher complexity.

[3]The base station only needs to know when to stop transmission of the current codeword. Hence the feedback required is minimal.

V. Cooperation Scheme

Here 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. The proposed scheme benefits from the cooperation between the different users to minimize the delay while achieving the optimal scaling law of the throughput. The cooperation scheme is divided into two stages. In the first half of each time slot, the base station transmits at rate R_{s1} to one half of the users in the system. During the next half of the slot, these users cooperate with each other and transmit to the other $(N/2)$ users in the system. The rate of transmission R_{s2} is chosen such that the information can be decoded even by the worst of the remaining $(N/2)$ users. The throughput of the cooperation scheme is given by

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

We show that this scheme achieves the optimal scaling laws of both throughput and delay. Moreover, the base station only needs a single queue to implement this scheme. However to enable cooperation, the users need to decode/re-encode the information and the channel gains of all the users should be known by all other users in the system. Also the base station needs to have perfect knowledge of the inter-user channels. This requires much more feedback from the users. Thus the cooperation scheme has a much higher complexity than all of the schemes proposed earlier.

Theorem 5: The proposed cooperation scheme is asymptotically throughput-optimal and delay-optimal. The throughput of this scheme scales as

$$R_{tot} = \Theta(N) \tag{10}$$

with the number of users N while the delay scales as

$$D = \Theta(1). \tag{11}$$

Sketch of Proof: It can be shown that the rate R_{s2} of the second (cooperative) stage scales as $\Theta(N)$. Using this along with the results obtained for the median user scheme, we get the desired result. For a detailed proof, please refer [2]. ■

Thus the cooperation scheme achieves the optimal scaling laws of both throughput and delay at the expense of a very high complexity.

We have generalized all the scheduling algorithms proposed here to exploit the multi-group diversity available in a general multicast scenario where different information streams are requested by different subsets of the user population. These schemes are not discussed here due to space constraints. For a detailed discussion, please refer [2].

VI. Numerical Results

Here we present simulation results in certain representative scenarios that validate our theoretical claims. These results were obtained through Monte-Carlo simulations and were averaged over atleast 5000 iterations. The power constraint

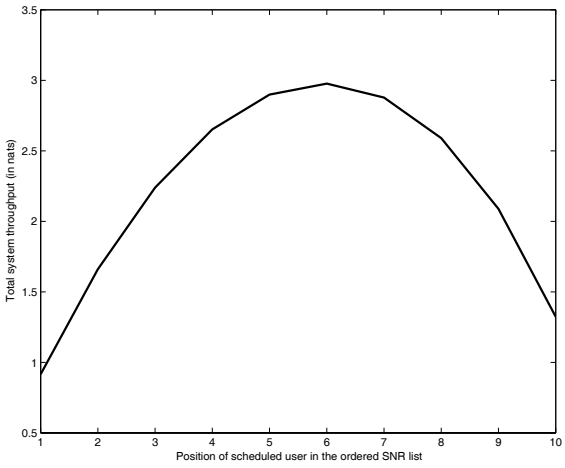

Fig. 1. Throughput of the general static scheduling scheme for different positions of the intended user in the ordered list of SNRs of all users (N=10)

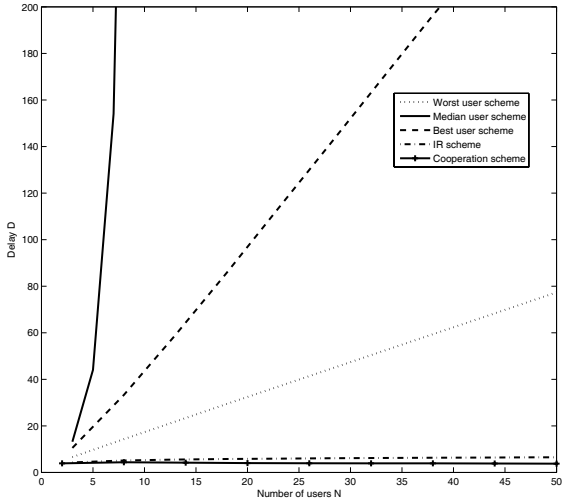

Fig. 3. Comparison of the delay of the proposed schemes

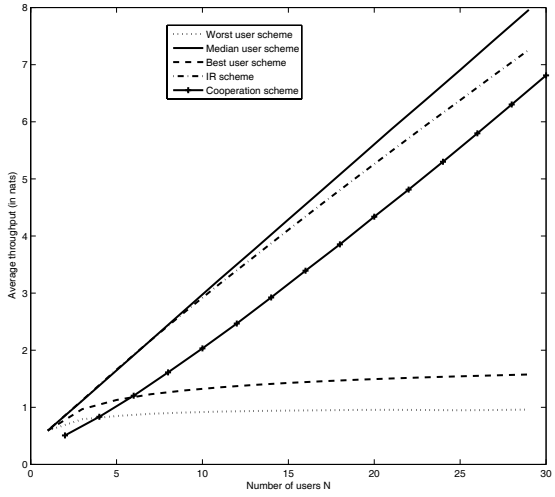

Fig. 2. Comparison of the throughput of the proposed schemes

P is taken to be unity. The throughput of the proposed low-complexity general static scheduling scheme is shown in Fig. 1 for different positions of the intended user in the ordered list of SNRs of all users. It is evident from the figure that, as predicted by the analysis, the throughput of the median user scheme is better than that of the best user scheme, which in turn is better than the throughput of the worst user scheme. In Fig. 2, we present a comparison of the throughputs of all the proposed schemes for increasing values of N. A similar comparison of the delays of all the proposed schemes is presented in Fig 3. It is clear that the presented simulation results follow the same trend as predicted by our theoretical analysis. Morevoer, the simulations show that the trends with small number of users agree with the asymptotic results.

VII. Conclusions

In this paper, we have used an information theoretic framework for characterizing the throughput-delay tradeoffs of the downlink multicasting channel. We first proposed low-complexity static scheduling schemes for the pure multicasting scenario and showed that the proposed median user scheduler achieves the optimal scaling law of the throughput, although at the expense of an exponential delay. We then proposed a more complex incremental redundancy scheme that trades off complexity for delay and achieves a better throughput-delay tradeoff. Finally, we proposed a cooperation scheme that achieves the optimal scaling laws of both throughput and delay at the expense of a high complexity. The validity of our theoretical claims was established through simulations.

References

[1] 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.

[2] Praveen Kumar Gopala and Hesham El Gamal. Opportunistic multicast scheduling: Throughput-delay-complexity tradeoff. *In Preparation*.

[3] David N. C. Tse. Optimal power allocation over parallel gaussian channels. In *International Symposium on Information Theory*, Ulm, Germany, June 1997.

[4] M. Sharif and B. Hassibi. On the capacity of mimo broadcast channel with partial side information. *To appear in IEEE Transactions on Information Theory*, 2004.

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

[6] D. J. Newman and L. Shepp. The double dixie cup problem. *Amer. Math. Monthly*, 67(1):58–61, January 1960.

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

[8] Giuseppe Caire and Daniela Tuninetti. The throughput of hybrid-arq protocols for the gaussian collision channel. *IEEE Transactions on Information Theory*, 47(5), July 2001.