This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE INFOCOM 2007 proceedings.

Resolving Collisions Via Incremental Redundancy: ARQ Diversity

Young-Han Nam, Praveen Kumar Gopala and Hesham El-Gamal
Department of Electrical and Computer Engineering, The Ohio State University,
Email: {namy,gopalap,helgamal}@ece.osu.edu

Abstract— **A cross-layer approach is adopted for the design of finite-user symmetric random access wireless systems. Instead of the traditional collision model, a more realistic physical layer model is adopted. An Incremental Redundancy Automatic Repeat reQuest (IR-ARQ) scheme, tailored to jointly combat the effects of user collisions, multi-path fading, and channel noise, is proposed. The diversity-multiplexing-delay tradeoff of the proposed scheme is analyzed for fully-loaded queues, and compared with that of the Gallager tree algorithm for collision resolution and the network-assisted diversity multiple access (NDMA) protocol of Tsatsanis *et al.*. The fully-loaded queue model is then replaced by one with random arrivals, where the three protocols are compared in terms of the stability region and average delay. Overall, our analytical and numerical results establish the superiority of the proposed IR-ARQ scheme and reveal some important insights. For example, it turns out that the performance is optimized, for a given total throughput, by maximizing the probability that a certain user will send a new packet and minimizing the transmission rate employed by each user.**

I. Background

We consider a random access system with symmetric users who compete to communicate with a common receiver, or a base station. Traditional approaches for analyzing such systems use the simplified collision model ([17] and references therein), which assumes that the base station cannot decode messages when a collision occurs (i.e., when more than one user transmits at the same time), and a message is always received error-free when a single user transmits. Under this model, several protocols, which avoid collisions and rely on single-user transmissions, have been proposed in the literature. Examples of such protocols include the Gallager tree algorithm (GTA) [10] and carrier sense multiple access/collision detection (CSMA/CD). The collision model, however, does not adequately capture some important characteristics of the wireless channel, e.g., multi-path fading, and ignores certain physical layer (PHY) properties like multi-packet reception (MPR) [12]. Recently, researchers have started to focus on cross-layer approaches that exploit the properties of the wireless medium to improve the performance of random access systems. For example, Naware *et al* [12] analyzed the stability and average delay of slotted-ALOHA based random access channels with MPR at the base station. This analysis, however, has abstracted out the physical layer parameters by using MPR reception probabilities. Another example is [3] where Tsatsanis *et al.* have proposed a random access protocol, called network-assisted diversity multiple access (NDMA), which uses the time diversity of **repetition** Automatic Repeat reQuest (ARQ) for collision resolution. As argued in the sequel, this protocol results in a significant loss in throughput owing to repetition coding. In [2], Caire *et al* studied the benefits of using an ARQ protocol with incremental redundancy (IR) transmissions (instead of repetition ARQ), and analyzed the throughput of the IR-ARQ scheme for the Gaussian collision channel [1] with fully-loaded[1] queues and single-user decoders at the base station. By adopting the fully-loaded queuing model, this work ignores the stability issues that arise in practical random access systems with random arrivals. Moreover, the single-user decoders used in this work are sub-optimal and result in considerable throughput losses. To overcome the limitations of these previous works, we adopt a system model that incorporates more realistic physical layer properties, and propose a new random access protocol based on IR-ARQ with joint decoding at the base station, in the next section.

II. ARQ Random Access

In this section, we introduce our system model and briefly review two existing random access schemes; namely, the GTA and the NDMA protocol. To the best of the author's knowledge, these two approaches represent the state of the art in the design of random access networks. More specifically, the GTA belongs to the class of tree collision resolution techniques whereas the NDMA serves as a representative for the class of collision resolution schemes that employ repetition ARQ. We then propose our new IR-ARQ protocol that overcomes the limitations of these existing protocols.

A. System Model

We consider a K-user symmetric random access channel with M antennas at each user and N antennas at the receiver (base station). We assume that all the users' channels experience Rayleigh-flat and long-term static fading where the channel fading coefficients remain constant during all the ARQ rounds corresponding to an information message. We consider individual power constraints on the users, and denote the received SNR of a user's signal by ρ. Time is slotted and a slot is composed of T channel uses. In order to control the number of users colliding, each user selects the slots for transmission according to the probability-p_t rule: in every slot, each user having a packet to transmit transmits a signal burst

[1] Each queue has infinite packets for transmission.

0743-166X/07/$25.00 ©2007 IEEE

with probability p_t and does not transmit with probability $1-p_t$, where $0 < p_t \leq 1$. We assume that the BS can perfectly identify the set of active users (by assigning a different control channel to each user). We initially assume fully-loaded queues in Section III, and then relax this assumption and consider a queuing system with random arrivals in Section IV.

B. Gallager Tree Algorithm (GTA)

This algorithm was proposed by Gallager [10] for the random access channel under the simplified collision model. The extension of this algorithm to our channel model mainly includes the probability-p_t rule and an explicit assumption that the base station does not even try to decode the users' messages in the case of a collision. We describe the extended GTA as follows. The traffic in the channel is interpreted as a flow of collision resolution (CR) epochs. At the beginning of each CR epoch, each user which has packets to transmit, uses the probability-p_t rule to decide whether it should (or should not) transmit in that epoch. If none of the users choose to transmit, the slot remains idle and a new CR epoch starts from the following slot. If only one user chooses to transmit, then it is allowed to transmit in the current slot and a new CR epoch begins from the following slot. But when a collision occurs, i.e., more than one user chooses to transmit in the current slot, the system enters into a CR mode, and only the users that participated in the collision at the beginning of a CR epoch are allowed to transmit until the end of that CR epoch. The colliding users are randomly split into two different groups, say the Left(L) and Right(R) groups. We assume that the base station uses a fair random split, wherein each user has an equal probability of joining either of the groups. The users in L transmit in the next slot. Based on the outcome of this transmission, the algorithm decides whether to include the users in R in the current CR epoch or not. If L has only one user, the users in R will transmit in the following slot. Meanwhile, if L has no users (idle slot), then the users in R are immediately split fairly into two sub-groups, without wasting a slot that is sure to result in a collision (*level skipping*). Finally, if there is a collision (L has multiple users), then the users in R are removed from the current CR epoch and have to wait until the next CR epoch to transmit according to the probability-p_t rule (*tree pruning*). The algorithm continues in the same fashion until all the users who have initiated the CR epoch and not been pruned, get a slot to transmit their packets without collisions. This marks the end of the current CR epoch and a new CR epoch begins from the next slot. This tree algorithm is known to be very efficient and achieves a maximum stable throughput of 0.487 with an infinitely large number of users. A more detailed description of this algorithm is provided in [9], [10], [11].

C. Orthogonal Network-Assisted Diversity Multiple Access (O-NDMA)

The NDMA protocol was proposed by Tsatsanis *et al.* [3] and relies on the use of time diversity through a repetition ARQ scheme to resolve collisions between users. At the beginning of each CR epoch, the transmission of each user will be determined by the probability-p_t rule as in the GTA protocol. If none or only a single user choose to transmit, then the next CR epoch starts from the following slot as before. However, when k (≥ 2) users transmit, then all those users repeat their transmissions in the next $(k-1)$ slots. At the end of k slots, the BS is assumed to be able to reliably decode the k packets, and a new CR epoch begins from the next slot. On the other hand, in [4], Zhang *et al.* proposed a new variant of NDMA which does not rely on time diversity to resolve/detect collisions. This variant, named B-NDMA, relies on a blind signal separation method utilizing a Vandermonde mixing matrix constructed via specially designed user retransmissions. In B-NDMA, the detection and resolution of a k-user collision require $(k+1)$ slots. However, in this paper, we assume the use of separate control channels for collision detection; which allows for a slightly more efficient version than the B-NDMA protocol, named orthogonal NDMA (O-NDMA), which requires only k slots to resolve a k-user collision, without relying on temporal diversity. The behavior of users in O-NDMA is the same as that in NDMA, with the only difference that in case of a k-user collision, user i transmits its symbols scaled by $(w_i)^\ell = (e^{\frac{j2\pi i}{k}})^\ell$, where $i = 1, \cdots, k$ and $j = \sqrt{-1}$, in the ℓ-th slot after the initial collision. At the end of the k^{th} slot, the BS utilizes the orthogonal structure constructed using the w_i's to decompose the joint decoding problem into k single-user problems. For example, suppose that user 1 and user 2 have collided ($k = 2$), and user i's codeword is $\mathbf{x}_i$, for $i = 1, 2$. Then, the BS coordinates the users so that user 1 repeats $\mathbf{x}_1$ whereas user 2 transmits $-\mathbf{x}_2$, in the slot following the collision. To decode user 1, the BS calculates the sum of the received vectors in the two slots, while to decode user 2, it takes the difference (i.e., matched filtering). This way, the multi-user interference is removed, and single-user decoders can be utilized to recover both packets. It is worth noting that O-NDMA requires symbol-level synchronization to facilitate the interference cancellation described above. Hence, our results for O-NDMA can be interpreted as optimistic upper bounds on the performance of repetition based random access protocols.

However, O-NDMA is still sub-optimal for two reasons. First, the BS might be able to decode[2] the messages of k colliding users in less than k time slots. Conversely, it is also possible that k time slots are insufficient for the successful decoding of the k packets. Thus, such a static strategy may result in a throughput loss. Second, O-NDMA is essentially **a repetition based** collision resolution mechanism. Although this results in a low-complexity decoder at the BS, the throughput performance is highly sub-optimal, as shown rigorously in the sequel. A significant improvement in the throughput can be achieved by allowing for IR transmissions

[2] Multiple messages can be jointly decoded in a single transmission block, with an arbitrary small error probability, if a rate-tuple lies within the capacity region of the channel and a sufficiently large block length is used [18].

from the colliding users within the CR epoch, and using joint decoding, across ARQ rounds and users, at the base-station (as discussed next).

D. Proposed Incremental Redundancy ARQ (IR-ARQ) Protocol

To overcome the disadvantages of the existing protocols, we propose a new IR-ARQ random access protocol operating as follows. Each user encodes an information message (packet) of B_T bits using a codebook of length-LT codewords, where L is an integer denoting a deadline constraint (i.e., a constraint on the maximum number of allowed ARQ rounds). Codewords are divided into L sub-blocks of length T. At the beginning of each CR epoch, the users which have packets to transmit, choose to transmit or not based on the probability-p_t rule as before. Once a user chooses to transmit in a particular slot, it transmits its first T symbols during that slot. We adopt an IR-ARQ transmission strategy and use a joint decoder that decodes the received observations both across users and ARQ rounds. If the receiver decodes the transmitted message(s), it feeds back an ACK; otherwise, it returns a NACK. On receiving an ACK, the CR epoch is terminated and a new CR epoch starts from the next slot. Thus a CR epoch can be defined as the time between two successive ACKs from the receiver[3]. On the other hand, if a NACK is received, each of the *colliding* users sends its second sub-block of T codeword symbols in the next slot, while all the other users remain silent. Thus only the users who transmit at the beginning of a CR epoch are allowed to transmit until the end of that epoch. The ACK/NACK rule applies in a similar manner until the L^{th} slot is reached. In this case, the receiver sends an ACK regardless of its decoding result. While decoding, the base-station uses all the observations received up to the current ARQ round. If a user's message is decoded after ℓ ARQ rounds, the effective coding rate becomes R/ℓ bits per channel use (BPCU), where $R = B_T/T$ denotes the rate of the first round. Here we assume that the base station can perfectly identify the set of active users (by assigning a different control channel to each user). Note that the deadline constraint bounds the transmission delay of this protocol by L slots.

III. Diversity-Multiplexing-Delay Tradeoff (DMDT)

In this section, we analyze the DMDT of the proposed IR-ARQ protocol and contrast it with the two benchmark protocols (i.e., GTA and NDMA) under the assumption of fully-loaded queues. The "fully-loaded" assumption allows for analyzing the maximum achievable throughput without focusing on the stability and delay issues, for the moment.

A. Definitions

We borrow the notion of DMDT from [6]. This notion is a generalization of the Zheng-Tse diversity-multiplexing tradeoff (DMT) which characterizes the fundamental tradeoff of fading channels between throughput and reliability in the high SNR regime [7]. In particular, we consider a family of ARQ protocols where the size of the information messages $B_T(\rho)$ depends on the operating SNR ρ. These protocols are based on a family of space time-codes $\{C_\rho\}$ with a first round rate of $R(\rho) = B_T(\rho)/T$ and an overall block length TL. Similar to [6], the delay constraint is on the maximum number of ARQ rounds to be less than or equal to L. For this family of protocols, we define the first round multiplexing gain r as

$$r = \lim_{\rho\to\infty} \frac{R(\rho)}{\log\rho}\ , \tag{1}$$

and the **effective** ARQ multiplexing gain r_e as

$$r_e \triangleq \lim_{\rho\to\infty} \frac{\eta_{FL}(\rho)}{\log\rho}. \tag{2}$$

Here $\eta_{FL}(\rho)$ is the long-term average throughput of the ARQ protocol in the random access channel with fully-loaded (FL) queues, i.e.,

$$\eta_{FL}(\rho) = \lim_{s\to\infty} \frac{b(s)}{sT}, \tag{3}$$

where s is the slot index and $b(s)$ is the total number of message bits transmitted up to slot s. Note that the message bits received in error at the base station are also counted in $b(s)$. The **effective** ARQ diversity gain is defined as

$$d = -\lim_{\rho\to\infty} \frac{\log P_E(\rho)}{\log\rho}, \tag{4}$$

where $P_E(\rho)$ is the system error probability, which is defined as the probability that at least one of the users' messages is not correctly decoded by the base station. In the symmetric random access channel, the diversity gain obtained from (4) is the same as the diversity gain of each individual user, since

$$P_{E^{(i)}}(\rho) \le P_E(\rho) \le \sum_{j=1}^{K} P_{E^{(j)}}(\rho), \quad \forall i\in\{1,\cdots,K\}\ , \tag{5}$$

where $P_{E^{(i)}}(\rho)$ is the error probability of the i^{th} user in the system. In summary, the DMDT of a certain protocol characterizes the set of achievable tuples (d, r_e, L) (here, we observe that our results are information theoretic in the sense that we assume the use of random Gaussian codebooks [7]).

In our analysis, we will make use of the results of Viswanath *et al.* on the diversity-multiplexing tradeoff of **coordinated** multiple access channels [8][4]. In the sequel, we denote the diversity gain of the coordinated multiple access channel with k users as $d_k^{MAC}(r)$, which is given by

$$d_k^{MAC}(r) = \begin{cases} d^{M,N}(r), & r \le \min\{M, \frac{N}{k+1}\} \\ d^{kM,N}(kr), & r \ge \min\{M, \frac{N}{k+1}\} \end{cases}\ , \tag{6}$$

where $d^{M,N}(r)$ is the diversity gain of the point-to-point channel with M transmit and N receive antennas, and multiplexing gain r given in [7].

[3] This definition requires the base station to return an ACK message after an idle slot.

[4] Coordinated multiple access channels differ from our model in the fact that the access mechanism is controlled by the base-station.

In the ARQ setting, we denote the event that a NACK is transmitted in the ℓ^{th} ARQ round, when k users are transmitting simultaneously, by $\bar{\mathcal{A}}_k(\ell)$, for $\ell = 1, \cdots, L-1$, and the error event in the L^{th} round by $\bar{\mathcal{A}}_k(L)$. We also denote the complement of $\bar{\mathcal{A}}_k(\ell)$ by $\mathcal{A}_k(\ell)$. We define

$$\alpha_k(\ell) \triangleq \Pr\left(\bar{\mathcal{A}}_k(1), \cdots, \bar{\mathcal{A}}_k(\ell-1), \mathcal{A}_k(\ell)\right) \tag{7}$$

and

$$\beta_k(\ell) \triangleq \Pr\left(\bar{\mathcal{A}}_k(1), \cdots, \bar{\mathcal{A}}_k(\ell)\right), \text{ for } \ell = 1, \cdots, L, \tag{8}$$

where, by definition, we let $\beta_k(0) = 1$, for $k = 1, \cdots, K$. Note that $\alpha_k(\ell)$ is the probability that the length of a CR epoch is ℓ (slots), given that k users have collided initially. For notational convenience, we denote the pmf of a binomial random variable with population K and probability of success p by,

$$\mathcal{B}(K, k, p) \triangleq \binom{K}{k} p^k (1-p)^{K-k} . \tag{9}$$

B. Main Results

First, we characterize the DMT of GTA [10] (note that we do not have deadlines in this model).

Proposition 1: The DMT for GTA with a given $p_t \in (0, 1]$ is

$$d^{GTA}(r_e) = d_1^{MAC}\left(\frac{\sum_{k=0}^{K} \mathcal{B}(K, k, p_t)\mathcal{X}_k}{\sum_{k=0}^{K} \mathcal{B}(K, k, p_t)J_k} r_e\right), \tag{10}$$

where $\mathcal{X}_k$ and J_k can be found by the following recursions:

$$\begin{aligned} \mathcal{X}_k &= 1 + \mathcal{B}(k, 0, 0.5)\mathcal{X}_k + \mathcal{B}(k, 1, 0.5)(1 + \mathcal{X}_{k-1}) \\ &\quad + \sum_{i=2}^{k} \mathcal{B}(k, i, 0.5)\mathcal{X}_i , \end{aligned} \tag{11}$$

and

$$\begin{aligned} J_k &= \mathcal{B}(k, 0, 0.5)J_k + \mathcal{B}(k, 1, 0.5)(1 + J_{k-1}) \\ &\quad + \sum_{i=2}^{k} \mathcal{B}(k, i, 0.5)J_i , \end{aligned} \tag{12}$$

for $k = 2, 3, \cdots$, with $\mathcal{X}_0 = \mathcal{X}_1 = 1$ and $J_0 = 0$, $J_1 = 1$.

Proof: Since errors occur only when a single user transmits, the diversity gain is given by $d_1^{MAC}(r)$. We observe that CR epochs are renewal intervals, and we apply the renewal-reward theorem [16] to analyze the long-term average throughput:

$$\eta_{FL} = \lim_{s\to\infty} \frac{b(s)}{sT} = \frac{\mathbb{E}[\mathcal{R}]}{\mathbb{E}[\mathcal{X}]}, \tag{13}$$

where $\mathcal{X}$ and $\mathcal{R}$ are random variables representing the number of channel uses and the number of bits transmitted in a renewal interval, respectively. Let J_k denote the average number of users who transmitted in a CR epoch (without being pruned) given that k users have collided initially as, and the average length of the CR epoch conditioned on k as $\mathcal{X}_k$. Then, (2) and (13) give

$$r_e = \frac{\sum_{k=0}^{K} \mathcal{B}(K, k, p_t)J_k}{\sum_{k=0}^{K} \mathcal{B}(K, k, p_t)\mathcal{X}_k} r, \tag{14}$$

and thus we have (10).

Now, we find the recursions for $\mathcal{X}_k$ and J_k similarly as done in [11]. We first consider $\mathcal{X}_k$. It is easy to see that $\mathcal{X}_0 = \mathcal{X}_1 = 1$. For $k \geq 2$, we make the following observations. If L has zero packets, then we have spent two slots for the initial collision and the idle slot; but the level skipping prevents us from an obvious collision in R. On the other hand, if L has one packet, then we have spent two slots for the initial collision and the transmission of the one packet. Finally, if L has i (≥ 2) packets, then we have spent one slot for the initial collision, and $(k - i)$ packets will be pruned (thus it is equivalent to have another initial collision of i packets after the original collision). These observations lead us to the following recursion:

$$\begin{aligned} \mathcal{X}_k &= \mathcal{B}(k, 0, 0.5)(1 + \mathcal{X}_k) + \mathcal{B}(k, 1, 0.5)(2 + \mathcal{X}_{k-1}) \\ &\quad + \sum_{i=2}^{k} \mathcal{B}(k, i, 0.5)(1 + \mathcal{X}_i). \end{aligned} \tag{15}$$

It is easy to see that (15) is equivalent to (11). Next, we consider J_k. It is obvious that $J_0 = 0$ and $J_1 = 1$. For $k \geq 2$, we make the following observations. If L has zero packets, then no packets will be pruned. On the other hand, if L has one packet, then one packet will be transmitted (or $(k - 1)$ packets will remain). Finally, if L has i (≥ 2) packets, then $(k-i)$ packets will be pruned (or i packets will remain). These observations lead us to the recursion (12). ■

Since the GTA protocol is inspired by the simplified collision model, the main idea is to assign a single slot exclusively for transmission of each colliding user (that was not pruned by the algorithm). The resulting DMT, therefore, is given in terms of a single-user performance, i.e., $d_1^{MAC}(.)$ and it is upper-bounded by $d_1^{MAC}(r_e)$. The main drawback of GTA is the relatively large number of slots needed to resolve each collision, which translates into a loss in the effective multiplexing gain, i.e., the argument of $d_1^{MAC}(.)$ in (10). It is now easy to see that GTA cannot achieve the full effective multiplexing gain of the multiple access channel, i.e., $\min\{KM, N\}$. An example highlighting this fact will be provided in the later part of this section. On the other hand, the DMT in (10) reveals the performance dependence on p_t (and r), which implies the possibility of maximizing the diversity gain by choosing the appropriate values, p_t^* and r^* for each $r_e \in [0, \min\{KM, N\})$. At the moment, we do not have a general analytical solution for this optimization problem. However, the solution for the special case of two users is obtained in Section III-C. We also note that the DMT of CSMA/CD is again upper-bounded by the single-user DMT, since CSMA/CD also tries to assign a single user to each slot and thus the throughput cannot exceed 1 packet/slot.

Next, we characterize the optimal DMT for the O-NDMA protocol (Again we do not have a delay parameter in the tradeoff since the number of ARQ rounds is **always** equal to the number of colliding users).

Proposition 2: The *optimal* DMT for O-NDMA is,

$$d^{ONDMA}(r_e) = d_1^{MAC}(r_e) . \tag{16}$$

Proof: The DMT for O-NDMA with a given $p_t \in (0,1]$ and r is found as

$$d^{ONDMA}(r_e) \;=\; d_1^{MAC}(r) \tag{17}$$

where

$$r \;=\; \frac{Kp_t + (1-p_t)^K}{Kp_t} r_e, \tag{18}$$

utilizing the average throughput results in [3], and noting that the average SNR of each single-user decoder is still ρ. Then, it is easy to find that the optimal values $(r^*, p_t^*) = (r_e, 1)$, which yields (16). ■

The matched-filter-like structure utilizing the orthogonality of transmissions over different slots allows the O-NDMA protocol to achieve the single-user performance, as we see from (17). Furthermore, $p_t^* = 1$ ensures that the throughput is maximized, and the optimal DMT is given by (16). By comparing the expressions in (10) and (16), we realize that the O-NDMA protocol achieves a larger diversity gain, as compared with the GTA protocol, for any r_e less than $\min\{M, N\}$.

Finally, the optimal DMDT of the IR-ARQ random access protocol is characterized in the following theorem.

Theorem 3: The *optimal* DMDT for the IR-ARQ protocol is,

$$d^{IR}(r_e, L) \;=\; d_K^{MAC}\left(\frac{r_e}{KL}\right). \tag{19}$$

Proof: First, we assume an asymptotically large block length $T \to \infty$ to allow our error correction (and detection) scheme to operate arbitrarily close to the channel fundamental limits. An application of the renewal-reward theorem [16] gives

$$\eta_{FL} \;=\; \frac{p_t K R}{1 + \sum_{k=1}^{K} \mathcal{B}(K,k,p_t) \sum_{\ell=1}^{L-1} \beta_k(\ell)}\;. \tag{20}$$

In addition, given that joint typical-set decoders [18], which have an inherent ability to detect errors, are used for the channel output over slots 1 to ℓ in ARQ round ℓ, extending the results in [6] and [8], the probability of error P_e is upper-bounded by,

$$P_e \;\leq\; \sum_{k=1}^{K} \mathcal{B}(K,k,p_t)\beta_k(L). \tag{21}$$

Noting that in the high SNR regime $\beta_k(\ell)$ approaches

$$\lim_{\rho\to\infty} \beta_k(\rho,\ell) \;=\; \mathbf{1}\left(r > \min\left\{\ell M, \frac{\ell N}{k}\right\}\right) \tag{22}$$

$$\triangleq\; \begin{cases} 0, & r < \min\{\ell M, \frac{\ell N}{k}\} \\ 1, & r > \min\{\ell M, \frac{\ell N}{k}\} \end{cases}\;, \tag{23}$$

we find the DMDT with a given $p_t \in (0,1]$ as

$$d^{IR}(r_e, L) \;=\; d_K^{MAC}\left(\frac{r}{L}\right), \tag{24}$$

where r can be obtained from r_e using the relation (for $0 \leq r \leq \min\{M, N\}$),

$$r_e \;=\; \frac{p_t K r}{1 + \sum_{k=1}^{K} \mathcal{B}(K,k,p_t) \sum_{\ell=1}^{L-1} \mathbf{1}\left(r > \min\{\ell M, \frac{\ell N}{k}\}\right)}\;, \tag{25}$$

from (20–23), and the results in [6], [8]. Finally, we find the optimal values $(r^*, p_t^*) = (\frac{r_e}{K}, 1)$, which gives (19). A detailed proof will be presented in the journal version of this paper [15]. ■

Two remarks are now in order. First, we elaborate on the intuitive justification for the optimal values $(r^*, p_t^*) = (\frac{r_e}{K}, 1)$ for the IR-ARQ protocol. In the asymptotic case with $\rho \to \infty$, the error probability is dominated by the worst case K-user collision for any $p_t \in (0,1]$, which does not depend on ρ by definition. This implies that choosing $p_t = 1$ will maximize the average throughput, without penalizing the asymptotic behavior of the error probability. Now with $p_t = 1$, choosing $r^* = \frac{r_e}{K} < \min\{M, \frac{N}{K}\}$ will result in an effective multiplexing gain equal to r_e and will minimize the number of rounds needed to decode the colliding messages, since each user is transmitting at a small rate. Furthermore, it is clear that with this choice of r^*, we can achieve any desired effective multiplexing gain less than $\min\{KM, N\}$ (the degrees of freedom in the coordinated multiple access channel). Next, comparing Propositions 1, 2 and Theorem 3, it is straightforward to verify that the DMT of the IR-ARQ protocol is **always** superior to that of the GTA and O-NDMA protocols. This advantage of IR-ARQ is a manifestation of the **ARQ diversity** resulting from the IR transmission and joint decoding. More specifically, the ARQ diversity **scales down** the effective multiplexing in the right hand side of (19), and hence, results in an increased diversity advantage (since $d_K^{MAC}(.)$ is a decreasing function in its argument). The O-NDMA protocol does not allow for **efficiently** exploiting the ARQ diversity due to the sub-optimality of repetition based ARQ.

C. Examples

We numerically illustrate the gain offered by the IR-ARQ protocol, as compared with the GTA and the NDMA protocol for two-user random access channels.

1) Two-User Scalar Random Access Channels: We consider the single-antenna 2-user random access channel, i.e., $M = N = 1$ and $K = 2$. Substituting these parameters in Proposition 1, we obtain the DMT for the GTA protocol as, $d^{GTA}(r_e) = d_1^{MAC}\left(\frac{1+3p_t^2}{2p_t} r_e\right) = 1 - \left(\frac{1+3p_t^2}{2p_t}\right) r_e$. In order to maximize the effective multiplexing gain that achieves nonzero diversity gain, we need to choose $p_t = \frac{1}{\sqrt{3}}$, which yields the optimal DMT for GTA as $d^{GTA}(r_e) = 1 - \sqrt{3} r_e,\; 0 \leq r_e < \frac{1}{\sqrt{3}}$. The optimal DMTs for O-NDMA and IR-ARQ are obtained from Proposition 2 and Theorem 3. Fig. 1 compares the tradeoffs of the three protocols where the IR-ARQ protocol is shown to dominate our two benchmarks, with both $L = 1, 2$. Even though O-NDMA achieves the nominal single-user DMT **without multi-user interference**, i.e., $d(r_e) = 1 - r_e, \forall r_e < 1$, it is still worse than IR-ARQ, since it wastes slots to facilitate single-user decoding and relies on repetition ARQ. In addition, as L increases from 1 to 2, the DMDT of IR-ARQ improves, as expected.

2) Two-User Vector Random Access Channels: We consider a 2-user vector random access channel with $M = 1$ and

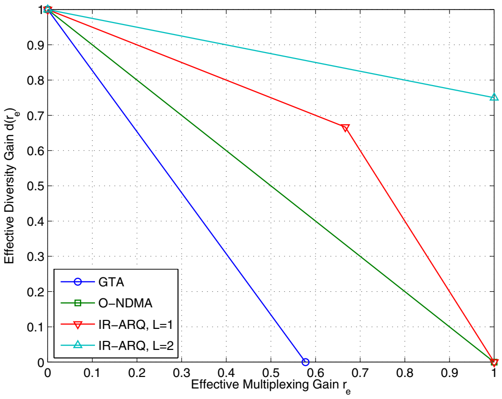

Fig. 1. Diversity-multiplexing tradeoff for various two-user scalar random access systems

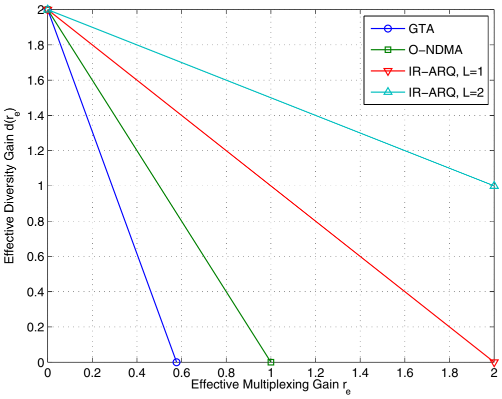

Fig. 2. Diversity-multiplexing tradeoff for various two-user vector random access systems

$N = 2$. By allowing multiple antennas at the BS, the total degrees of freedom of the system is increased by a factor of 2, as compared with the scalar channel in the previous example. The tradeoffs achieved by the three protocols in this scenario are shown in Fig. 2. First, we observe that the three protocols achieve an increased diversity gain, for a given r_e, when compared with the scalar channel in Fig. 1. However, the full effective multiplexing gain, $r_e = 2$, is not achieved by the GTA and O-NDMA protocols, since these two protocols exclude the possibility of first-round decoding when a collision occurs. The IR-ARQ protocol, on the other hand, achieves $r_e = 2$, and the DMDT further improves as L increases.

IV. Random Arrivals

In this section, we relax the assumption of fully-loaded queues adopted in Section III. In our PHY model, depending on the channel conditions, decoding can fail at the base-station even with no collisions. This forces us to allow packets to be dropped from a queue even if it is not successfully decoded at the base station (due to the deadline constraint). Thus, the error probability (or the diversity gain) is also an important performance measure that should be included in the analysis, in addition to the traditional measures of stability region and average delay commonly used in this set-up. In particular, for the proposed IR-ARQ protocol, the choice of the deadline constraint L determines the *tradeoff* between the average delay and the error probability (It will be shown, through simulation results, that an increase in L leads to an increase in the average delay along with a decrease in the error probability).

We consider infinite-length queues at the users, that are fed by randomly-arriving packets of a fixed length of B_A information bits. For simplicity, we assume that $B_T = B_A = B$, i.e., the arrival packet size and the transmission packet size are the same. Thus the first-round transmission rate R is equal to the arrival rate $R_A = (B/T)$. To emphasize that R is a system parameter determined by the arrival packet size, we denote the first-round multiplexing gain r by r_A, and call it *the arrival multiplexing gain*. The arrival rate of user i is $\lambda_i = \lambda/K$ packets/slot, where λ denotes the total arrival rate, and arrivals are assumed to be independent across users.

We first review known results for the stability region and average delay of the GTA and the NDMA protocol. Then, we present the stability region and average delay of the proposed IR-ARQ protocol. Finally, we find the diversity gains achieved by the GTA, the NDMA and the IR-ARQ protocols as simple extensions of our results obtained in the previous section.

A. Stability and Average Delay

We define the notion of stability as follows: Let $\mathbf{g}(m) \triangleq (g_1(m), \cdots, g_K(m))$ be the vector of the backlogs at the beginning of CR epoch m. Then, queue i of the system is stable if [19]

$$\lim_{m\to\infty} \Pr\left(g_i(m) < \bar{g}\right) = F(\bar{g}) \quad \text{and} \quad \lim_{\bar{g}\to\infty} F(\bar{g}) = 1. \tag{26}$$

Furthermore, we say that the system is stable if all the K queues in the system are stable.

The stability region of the GTA can be found using the techniques in [11] as

$$\lambda < \frac{\sum_{k=0}^{K} \mathcal{B}(K,k,p_t) J_k}{\sum_{k=0}^{K} \mathcal{B}(K,k,p_t) \mathcal{X}_k}, \tag{27}$$

and the stability region of the O-NDMA protocol can be found similarly as [5]

$$\lambda < \frac{Kp_t}{Kp_t + (1-p_t)^K}. \tag{28}$$

From the literature ([12] and references therein), we find that there are only limited results on the average delay of the slotted ALOHA system, and it is a non-trivial task to characterize the average delay of random access systems. In this paper, we present only numerical results for the GTA and the O-NDMA protocol, and provide an approximate delay analysis for the proposed IR-ARQ protocol. The average delay of the IR-ARQ scheme can be approximated by using the analysis of the M/G/1 queue with vacations [17], following the approach of Tsatsanis *et al* in [3]. This analysis yields only an approximation of the average delay, since the CR epoch lengths of the IR-ARQ scheme are related to the traffic load (and hence are not independent and identically distributed (i.i.d.) as needed for the result to hold). However, as we will see, in some cases with $\rho \to \infty$ the i.i.d. property holds for these epoch lengths, and hence the result becomes asymptotically accurate. We also note that as K increases, this approximation becomes progressively more accurate [3].

Following the approach of [3], we classify the CR epochs from the viewpoint of a particular user (say user 1) into either *relevant* or *irrelevant* epochs, depending on whether a packet of that user is being transmitted in that CR epoch or not. The *idle* epochs, consisting of only one time slot during which none of the users transmit packets, are a subset of the irrelevant epochs. The lengths of the relevant and the irrelevant epochs of user 1 are random variables, which are denoted by U and V, respectively.

Now, we present the stability region and an approximate average delay for the IR-ARQ protocol in the following theorem.

Theorem 4: Assuming that

$$\exists\ \ell < \infty \text{ with } \ell \in \{1, \cdots, L\}, \text{ such that } \alpha_K(\ell) > 0, \quad (29)$$

the necessary and sufficient condition for the stability of the IR-ARQ protocol is (in packets/slot)

$$\lambda \; < \; \frac{\eta_{FL}}{R} \; = \; \frac{p_t K}{1 + \sum_{k=1}^{K} \mathcal{B}(K,k,p_t) \sum_{\ell=1}^{L-1} \beta_k(\ell)} \; . \quad (30)$$

For Poisson arrivals, when λ satisfies (30), the average delay is *approximately* given by (in slots)

$$D \; = \; \frac{\lambda \left(\mathbb{E}[U^2] + \frac{(2-p_t)(1-p_t)}{p_t^2} \mathbb{E}[V^2] + 2\left(\frac{1}{p_t} - 1\right) \mathbb{E}[U]\mathbb{E}[V] \right)}{2\left(K - \lambda\left(\mathbb{E}[U] + \left(\frac{1}{p_t} - 1\right)\mathbb{E}[V] \right) \right)}$$

$$+\mathbb{E}[U] + \left(\frac{1}{p_t} - 1\right)\mathbb{E}[V] + \frac{\mathbb{E}[V^2]}{2\mathbb{E}[V]} \; , \quad (31)$$

where

$$\mathbb{E}[U] = 1 + \sum_{k=1}^{K} \mathcal{B}(K-1,k-1,p) \sum_{\ell=1}^{L-1} \beta_k(\ell),$$

$$\mathbb{E}[U^2] = 1 + \sum_{k=1}^{K} \mathcal{B}(K-1,k-1,p) \sum_{\ell=1}^{L-1} (2\ell+1)\beta_k(\ell),$$

$$\mathbb{E}[V] = 1 + \sum_{k=1}^{K-1} \mathcal{B}(K-1,k,p) \sum_{\ell=1}^{L-1} \beta_k(\ell),$$

$$\mathbb{E}[V^2] = 1 + \sum_{k=1}^{K-1} \mathcal{B}(K-1,k,p) \sum_{\ell=1}^{L-1} (2\ell+1)\beta_k(\ell),$$

and $p \in (0,1]$ is a solution of the following equation:

$$Kp \; = \; \lambda \left[1 + \sum_{k=1}^{K} \mathcal{B}(K,k,p) \sum_{\ell=1}^{L-1} \beta_k(\ell) \right] . \quad (32)$$

Moreover, the delay expression in (31) holds with probability 1 if U and V are i.i.d. and U and V are mutually independent.

Proof: (sketch) We consider the backlog evolution $\mathbf{g}(m)$ of IR-ARQ, where m is the *epoch* index. We observe that $\mathbf{g}(m)$ is an embedded Markov chain; $g_i(m)$, the backlog evolution of user i is,

$$g_i(m+1) = \begin{cases} (g_i(m)-1)^+ + a_i(m), & \text{with probability } p_t \\ g_i(m) + a_i(m), & \text{with probability } 1-p_t \end{cases} \quad (33)$$

where $a_i(m)$ is the number of packets that arrived at user i's queue during epoch m, and $(x)^+ = x$ if $x \geq 0$, $(x)^+ = 0$ otherwise, for a real number x. We first prove that (30) is the necessary and sufficient condition for the stability of IR-ARQ. We see that under the assumption (29), $\mathbf{g}(m)$ is a homogeneous, irreducible and aperiodic Markov chain, by following the argument in the proof of Proposition 1 in [14]. Given that the Markov chain has these properties, stability of the system is equivalent to the existence of a limiting distribution for the Markov chain, and thus is also equivalent to ergodicity of the Markov chain [13]. Sufficiency and necessity of (30) for the ergodicity can be straightforwardly proved by following the footsteps of the proof of Theorem 1 in [13]. The proof for the average delay closely follows [5]. Detailed proof will be presented in [15]. ■

We note that the assumption (29) always holds when L is finite since the length of any CR epoch is bounded by L. If $L = \infty$, condition (29) requires the existence of a nonzero probability that the length of an epoch is finite. As $\rho \to \infty$, the stability region in (30) approaches

$$\lambda < \frac{p_t K}{1 + \sum_{k=1}^{K} \mathcal{B}(K,k,p_t) \sum_{\ell=1}^{L-1} \mathbf{1}\left(r_A > \min\left\{\ell M, \frac{\ell N}{k}\right\}\right)}. \quad (34)$$

To achieve the maximum stability region, we need to maximize the right hand side of (34) over p_t. At the moment, we do not have a general solution for this problem. Thus, we present results only for one special case: $r_A < \min\{M, \frac{N}{K}\}$. In this case, the stability region is $\lambda < p_t K$, and the maximum stability region is thus given by $\lambda < K$ for the optimal choice of $p_t = 1$. This is a remarkable improvement over the O-NDMA protocol, whose maximum stability region is only $\lambda < 1$, for any r_A.

B. Diversity Gain

As discussed before, the diversity gain should also be included in the comparison of different random access protocols for our PHY model. The diversity gain with random arrivals can be directly obtained from the results in the previous section. One important difference is that in the random arrival case, unlike in the fully-loaded case, one cannot optimize over r_A, since r_A is a system parameter. Based on this observation, we find that the GTA and the O-NDMA achieve a diversity gain

$$d^{GTA}(r_A) \; = \; d^{ONDMA}(r_A) \; = \; d_1^{MAC}(r_A), \quad (35)$$

and the proposed IR-ARQ protocol has diversity gain

$$d^{IR}(r_A) \;=\; d_K^{MAC}\left(\frac{r_A}{L}\right). \qquad (36)$$

C. Examples

1) Two-User Scalar Random Access Channels: Here, we consider the random access channels with $M = N = 1$. For ease of analysis, we assume that $L \geq K = 2$ for the IR-ARQ scheme.

The stability region of the different random access protocols with $\rho \to \infty$ is summarized in Table I. In addition, the error probability, diversity gain and average delay are shown in Fig. 3, Fig. 4 and Fig. 5 respectively. Here, the stability region and diversity gain of the three protocols, and the average delay of the IR-ARQ protocol with $\rho \to \infty$, are obtained analytically. However, the average delay of the GTA and O-NDMA protocols, and the average delay of the IR-ARQ scheme with $\rho < \infty$ are obtained through numerical simulations. In these simulations, we use $R = r_A \log(1+\rho)$ with $r_A = 0.45$, and $p_t = 1$ for the IR-ARQ and the O-NDMA protocols, while $p_t = \frac{1}{\sqrt{3}}$ for the GTA protocol. It is assumed that the transmission results in errors, if and only if the channel is in outage [2]; which is a valid assumption if T is sufficiently large. In addition, for the IR-ARQ protocol, it is assumed that the errors in ℓ^{th} round, where $\ell < L$, are always detected. We also note that when $r_A < 0.5$, the average delay expression for the IR-ARQ scheme, evaluated from Theorem 4, holds with probability 1, and is given by (when $p_t = 1$) $D = 1.5 + \frac{\lambda}{2(2-\lambda)}$. Table I and Fig. 4 shows that both the stability region and diversity gain of the IR-ARQ protocol are the largest. Next, we focus on the delay and the error probability of IR-ARQ with different L's and different ρ's reported in Fig. 3 and Fig. 5. We observe that the delay approaches the asymptotic result with $\rho = \infty$, and the difference of the delay for IR-ARQ with $L = 2$ and with $L = 4$ decreases, as ρ increases, which agrees with the analytical results. Furthermore, Fig. 3 and Fig. 5 reveal an important insight into the relation between the performance of IR-ARQ and the transmission-delay constraint L, i.e., a tradeoff between average delay and error probability emerges. These figures suggest that for certain finite ρ's, a large L achieves a small error probability, at the expense of a large average delay and a small stability region. Therefore, depending on quality-of-service (QoS) requirements, L can be adjusted for achieving the best performance.

2) Two-User Vector Random Access Channels: Here, we consider the 2-user random access protocols with $M = 1$ and $N = 2$ in the high SNR regime ($\rho \to \infty$). We first note that the stability region and delay of the GTA and O-NDMA protocols are not different from the scalar case; only the diversity gain changes with this multiple-antenna setting. For the IR-ARQ protocol, on the other hand, the average delay is given by $D = 1.5 + \frac{\lambda}{2(2-\lambda)}$ for any $r_A \in [0, 1)$, and the stability region is given by, $\lambda < 2, \quad 0 \leq r_A < 1$, with $p_t = 1$. Comparing the stability region of the vector IR-ARQ protocol with that of the scalar IR-ARQ protocol, we find that the vector IR-ARQ achieves a better stability region, especially when $r_A >$

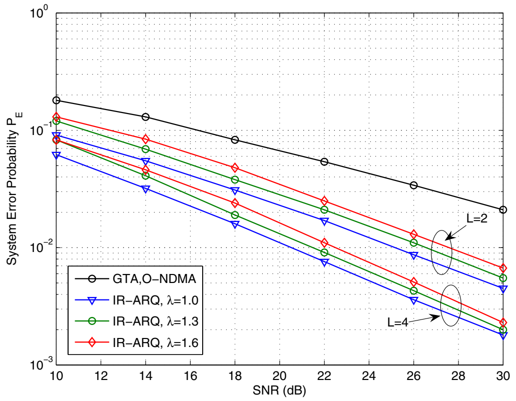

Fig. 3. System error probability versus SNR for various two-user scalar random access systems with random arrivals. Here, $r_A = 0.45$.

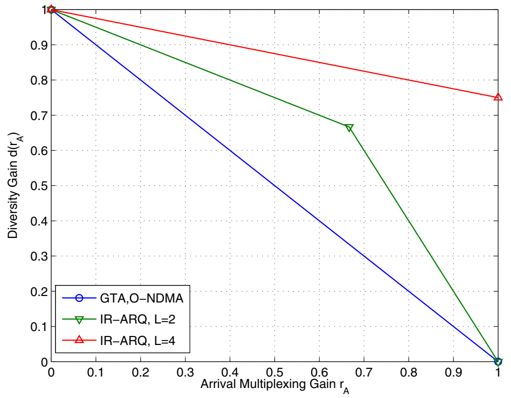

Fig. 4. Diversity gain versus the arrival multiplexing gain r_A for various two-user scalar random access systems with random arrivals.

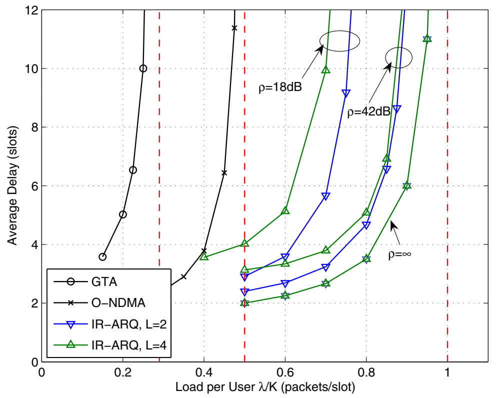

Fig. 5. Load per user versus average delay for various two-user scalar random access systems with random arrivals. Here, $r_A = 0.45$.

TABLE I

STABILITY REGION OF DIFFERENT TWO-USER SCALAR RANDOM ACCESS PROTOCOLS

	Stability Region	Maximum Stability Region
GTA	$\lambda < 2p_t/(1+3p_t^2)$	$\lambda < 1/\sqrt{3}$, with $p_t = 1/\sqrt{3}$
NDMA	$\lambda < 2p_t/(2p_t+(1-p_t)^2)$	$\lambda < 1$, with $p_t = 1$
IR-ARQ	$\begin{cases} \lambda < 2p_t, & r_A < 0.5, \\ \lambda < 2p_t/(1+p_t^2), & r_A > 0.5. \end{cases}$	$\begin{cases} \lambda < 2, & r_A < 0.5, \\ \lambda < 1, & r_A > 0.5. \end{cases}$, with $p_t = 1.$

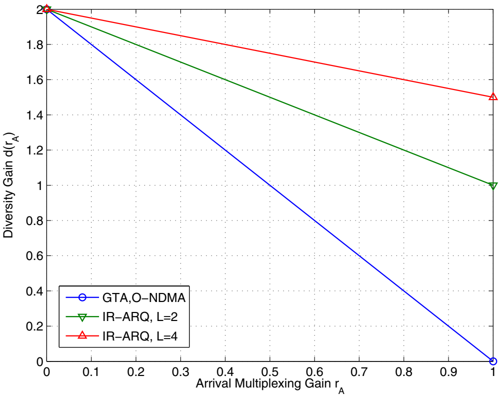

Fig. 6. Diversity gain versus the arrival multiplexing gain r_A for various two-user vector random access systems with random arrivals. Note that the curves for the GTA and the NDMA overlap.

0.5. Finally, Fig. 6 shows the diversity gain achieved with different random access protocols. As expected, the IR-ARQ protocol achieves the best diversity gain, which improves as L increases.

V. CONCLUSIONS

We have proposed a new wireless random access protocol which jointly considers the effects of collisions, multi-path fading, and channel noise. The proposed protocol relies on incremental redundancy transmission and joint decoding to resolve collisions and combat multi-path fading. This approach represents a marked departure from traditional collision resolution algorithms and exhibits significant performance gains, as compared with two benchmarks corresponding to the state of the art in random access protocols; namely GTA and O-NDMA. It is interesting to observe that, in order to fully exploit the benefits of the proposed IR-ARQ protocol, all the users with non-empty queues must transmit with probability one, when given the opportunity, and should use a small transmission rate. Finally, we have identified the tradeoff between average delay and error probability exhibited by the IR-ARQ protocol for certain SNRs, and have shown that this tradeoff can be controlled by adjusting the maximum number of ARQ rounds.

REFERENCES

[1] G. Caire E. Leonardi and E. Viterbo, “Modulation and Coding for the Gaussian Collision Channel,” *IEEE Trans. Inform. Theory*, vol. 46, No. 6, September 2000.

[2] G. Caire and D. Tuninetti, “The Throughput of Hybrid-ARQ Protocols for the Gaussian Collision Channel,” *IEEE Trans. Inform. Theory*, vol. 47, No. 5, July 2001.

[3] M. K. Tsatsanis, R. Zhang and S Banerjee, “Network-Assisted Diversity for Random Access Wireless Networks,” *IEEE Trans. Sig. Proc.*, vol. 48, No. 3, March 2000.

[4] R. Zhang, N. D. Sidiropoulos and M. K. Tsatsanis, “Collision Resolution in Packet Radio Networks Using Rotational Invariance Techniques,” *IEEE Trans. Commun.*, Volume:50, Issue:1, January 2002, Pages:146-155.

[5] G. Dimic, N. D. Sidiropoulos and L Tassiulas, “Wireless Networks with Retransmission Diversity Access Mechanisms: Stable Throughput and Delay Properties,” *IEEE Trans. Sig. Proc.*, vol. 51, No. 8, August 2003.

[6] H. El Gamal, G. Caire and M. O. Damen, “The MIMO ARQ Channel: Diversity-Multiplexing-Delay Tradeoff,” *IEEE Trans. Inform. Theory*, Volume:52, Issue:8, August 2006, Pages:3601-3621.

[7] L. Zheng and D. N. C. Tse, “Diversity and Multiplexing: A Fundamental Tradeoff in Multiple Antenna Channels,” *IEEE Trans. Inform. Theory*, Volume: 49, Issue:5, May 2003 Pages:1073-1096.

[8] D. Tse, P. Viswanath and L. Zheng, “Diversity-Multiplexing Tradeoff in Multiple Access Channels,” *IEEE Trans. Inform. Theory*, Volume: 50, Issue:9, Sept. 2004, Pages:1859-1874

[9] R. G. Gallager, “A Perspective on Multiaccess Channels,” *IEEE Trans. Inform. Theory*, Volume: 31, Issue:2, March 1985, Pages:124-142

[10] R. G. Gallager, “Conflict Resolution in Random Access Broadcast Networks,” *Proc. of the AFOSR Workshop in Communication Theory and Applications,* September 1978, Pages:74-76

[11] M. L. Molle and G. C. Polyzos, “Conflict Resolution Algorithms and their Performance Analysis,” Technical Report, CS93-300, Dept. of Computer Science and Engineering, University of California at San Diego, LaJolla, July 1993

[12] V. Naware and G. Mergen and L.Tong, “Stability and Delay of Finite-User Slotted ALOHA With Multipacket Reception,” *IEEE Trans. Inform. Theory*, Volume: 51, Issue:7, July 2005, Pages:2636-2656

[13] S.Adireddy and L.Tong “Exploiting Decentralized Channel State Information for Random Access,” *IEEE Trans. on Inform. Theory*, Volume: 51, Issue:2, February 2005, Pages:537-561

[14] G. Dimic and N. D. Sidiropoulos, “Stability Analysis of Collision Resolution Protocols with Retransmission Diversity,” *in Proc. of IEEE Int. Conf. Acoust., Speech, Signal Process.*, Orlando FL, May 2002, Pages: III-2133-2136

[15] Y-H. Nam, P. K. Gopala and H. El Gamal, “ARQ Diversity in Fading Random Access Channels,” *submitted to IEEE Trans. Wireless Comm., also available in http://www.ece.osu.edu/~helgamal/*, August 2006.

[16] R. G. Gallager, “Discrete Stochastic Process,” Kluwer Academic Publishers, 1996

[17] D. P. Bertsekas and R. G. Gallager “Data Networks,” Prentice Hall, 1992

[18] T. M. Cover and J. A. Thomas, “Elements of Information Theory,” Wiley-Interscience Publication, 1991

[19] A. S. Tanenbaum, “Computer Networks, 3rd ed.,” Prentice Hall, 1996.