Analyzing Performance of Multi-User Scheduling Jointly with AMC and ARQ

M. Poggioni, L. Rugini, P. Banelli
Dept. of Elect. and Inform. Eng. (D.I.E.I.)
University of Perugia
Perugia, Italy
{mario.poggioni, luca.rugini, paolo.banelli}@diei.unipg.it

***Abstract*—This paper deals with the analytical evaluation of the average delay, the packet-loss rate (PLR) and the throughput of a multi-user (MU) wireless system that capitalizes on a cross-layer design of adaptive modulation and coding (AMC) with automatic repeat request (ARQ). To this end, we propose a heuristic scheduling policy, which has the nice properties to effectively exploit the system resources and to be analytically tractable. Simulation results confirm the effectiveness of the proposed scheduler. Moreover, the excellent match between analytical and simulated performance allows the proposed approach to be used for cross-layer design optimization, avoiding time-consuming simulations.**

Keywords—Adaptive modulation and coding (AMC); automatic repeat request (ARQ); cross-layer design; scheduling.

I. INTRODUCTION

The increasing interest in cross-layer design for wireless systems has highlighted the necessity of an analytical characterization of such systems, capable of accounting for the specific issues arising when more layers are considered together. In this work, we consider the physical layer (PHY-L) and the data link layer (DL-L). More specifically, we focus on cross-layer combining of AMC at PHY-L with ARQ at DL-L [1]-[4]. The combination of AMC and ARQ enhances the spectral efficiency of wireless systems, because the error-correcting capabilities of ARQ allow the use of higher modulation rates at PHY-L. However, the ARQ packet retransmissions introduce a time delay that can become unacceptable for real-time applications. Thus, a main goal is to find a good trade-off between PLR and average delay, depending on the specific application.

Previous work on this subject includes [3], where an interesting combination of queuing with AMC is proposed for single-user (SU) scenarios, taking into account the queuing delay of the packets. Moreover, in [3], the PLR and the throughput are derived analytically, by means of a finite-state Markov chain analysis [5], allowing a cross-layer design of the system. The ARQ protocol, not considered in [3], is taken into account in [4] and [6] using analytical models. In particular, [4] deals with an MU scenario, where multiple (frequency) channels are statically preassigned to multiple users in accordance to their *average* signal-to-noise ratio (SNR) conditions, buffer sizes, packet arrival rates, available transmission modes (TMs), and quality of service (QoS) requirements. Differently, [6] focuses on an SU scenario, with finite buffer length, and proposes a joint design of ARQ and AMC where the user TM is chosen dynamically in accordance with its *instantaneous* SNR. Specifically, in [6], the PLR, the average delay and the system throughput are expressed in closed form, allowing a throughput optimization by means of an exhaustive search in a finite space, bounded with the (aforementioned) QoS constraints. However, the model defined in [6] only considers an SU wireless channel.

In this work, we generalize the approach of [6] to the case of many users that share a single channel. Consequently, we will define and theoretically model a scheduling policy. A heuristic scheduler was proposed in [7] for a similar MU environment, without providing an analytical characterization of its performance, which actually seems to be very challenging. Thus we propose in this paper a different heuristic scheduler that is analytically tractable. We found that the efficiency of the two algorithms is quite similar: however we will not show a simulation comparison between them due to lack of space. Specifically, with respect to [7], we focus on the special case of real-time users, where we refine the policy proposed in [7] by considering the buffer occupancy instead of the time delay. Anyway, this modification alone is not enough to easily cast the derivation of the scheduler performance into the plain extension of [6] to the MU case. Indeed, a rigorous generalization of [6] to the MU scenario, with an arbitrary scheduling policy, would require a complete characterization of the states of all the users, leading to an exponential complexity in the number of users, which is practically unmanageable. To overcome this problem, we introduce some simplifying assumptions that lead to a model with good accuracy and low complexity, which is independent of the number of users.

For the sake of complexity reduction, we consider users with identical statistics and QoS constraints, which allow to define also a simpler scheduling policy. However, it is important to note that our approach can be extended to the case of users with different QoS and traffic characteristics, provided that their states are modeled as proposed in this work. We concentrate on a simpler scenario in order to maintain the description and notation accessible.

II. SYSTEM MODEL AND SCHEDULING

We consider an uplink wireless link between U single-antenna transmitters (users) and a single-antenna receiver (base station). As previously explained, we consider a simple scenario where the users have identical traffic statistics and

1930-529X/07/$25.00 © 2007 IEEE
This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE GLOBECOM 2007 proceedings.

QoS constraints. In particular, perfect power control is assumed, where the average SNR is the same for each user. Each link must support QoS-guaranteed traffic, with a maximum average packet delay δ and a maximum packet loss rate ρ, equal for every user. The TM is chosen by the AMC selector at the receiver end depending on the instantaneous SNR at each decision epoch, and fed back to the transmitting user [6]. We consider the TMs defined in Table 1 of [6], with a Nakagami-*m* fading channel model. The generation of the users' packets is memoryless: specifically, we consider a Poisson packet arrival process. All packets have the same length of N_p bits at DL-L, which are mapped on time slots with different durations at PHY-L, accordingly to the current TM. Each buffer has a maximum packet length B. The basic assumptions of our model, which come from [6], are briefly summarized below:

A1) At PHY-L, time is slotted as in Fig. 1, and each slot contains one packet from DL-L. The overhead is assumed as negligible. We also assume perfect time synchronization among users.

A2) The propagation channel is modeled by a Nakagami-*m* frequency-flat block-fading channel [8], with *coherence time interval* (CTI) of T_f seconds. The channel variation from one CTI to another is captured by a Markov chain model [9].

A3) The TMs are selected at the receiver with perfect CSI knowledge, and they are fed back to the transmitter, without errors and with zero latency.

A4) The error detection at the receiver, by means of CRC, is perfect. The users' packets are dropped either after N_r retransmissions, or when the transmitter buffer is full.

The critical difference with respect to [6] is the presence of several users, which share the same PHY-L resource. This entails the necessity of a scheduling policy in order to have a good throughput, while preserving a certain fairness among users. Thus, we are specifically interested to constrain not only the average delay, but also the maximum delay. On the other hand, an efficient scheduling algorithm should exploit the diversity offered by the high number of users, usually known as multi-user diversity (MUD) [10], which allows to enhance the average system rate with respect to the SU scenario. We also would like to analytically address the system performance in order to avoid extensive simulations and, possibly, gain a deeper insight on the parameters that really affect performance. To this end, we need to define a scheduling policy that is both effective and mathematically tractable. In this view, we propose a heuristic centralized scheduling algorithm which is accomplished at the beginning of each CTI as in [7], with instantaneous decisions, and perfect knowledge of the state of each user. The state $\boldsymbol{\psi}_j^{(i)}$ of the j th user is defined as

$$\boldsymbol{\psi}_j^{(i)} = (c_j^{(i)}, q_j^{(i)}, r_j^{(i)}), \tag{1}$$

where $c_j^{(i)}$ is the channel mode, $q_j^{(i)}$ is the buffer occupancy, and $r_j^{(i)}$ is the actual number of retransmissions at the time instant t_i, which represents the beginning of the i th CTI. We associate to each channel state $c_j^{(i)}$ the channel rate $K_j^{(i)}$ as in Table 1 of [6]. We also consider, as in [7], perfect and instantaneous feedback of the scheduler decision to all the users. Our scheduling policy is aimed at enhancing the rate of the system without introducing an excessive delay. To this end,

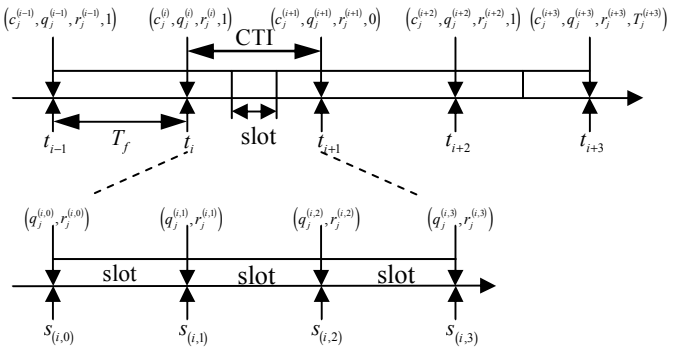

Fig. 1. State and substate transitions.

we define the utility function $\phi_j^{(i)} = K_j^{(i)} q_j^{(i)}$. Nevertheless, in order to reduce the PLR, if some users have full buffer, the priority is assigned to them, irrespectively of $\phi_j^{(i)}$. In the following, we present our scheduling algorithm. We define

$S_{Tx} = \{$ users with full buffer and $TM \neq TM0$ $\}$;

$S_{Tx0} = \{$ users with maximum $\phi_j^{(i)}$ $\}$;

$S_{Tx1} = \{$ users $\in S_{Tx0}$ with maximum $r_j^{(i)}$ $\}$;

$S_{Tx2} = \{$ users $\in S_{Tx}$ with maximum $r_j^{(i)}$ $\}$;

$S_{Tx3} = \{$ users $\in S_{Tx2}$ with maximum $c_j^{(i)}$ $\}$;

the proposed scheduling algorithm can be expressed as:

if $S_{Tx} = \{\}$

then the transmission is randomly assigned to a user $\in S_{Tx1}$;

else the transmission is randomly assigned to a user $\in S_{Tx3}$;

end

It is worth noting that the scheduling algorithm in [7] is not optimized for the minimization of the average delay of real-time users, because it considers different classes of users, and tries to maximize the system rate, while maintaining the delay of real-time users within a suitable bound. Conversely, for equal real-time users, our scheduler adopts a strategy aimed at reducing the average delay of each user. Indeed, from the Little theorem [5], we know that the average delay is proportional to the average buffer occupancy, thus, in the scheduling policy, we consider the buffer occupancy instead of the *instantaneous* delay of the packets, in order to constrain the former and consequently also the *average* delay.

III. Joint AMC and Queuing Analysis

We first briefly summarize the analysis carried out in [6] in SU scenarios for the queuing process induced by the truncated ARQ at DL-L jointly with the AMC scheme at PHY-L. The AMC scheme proposed in [6] partitions the entire SNR range into $N+1$ non-overlapping consecutive intervals, each associated in increasing order to TM0, TM1, …, TMN. Boundary points of each SNR interval are calculated using the same target packet-error rate (PER) for all the TMs (see Eqs. 1-4 in [6]). An exception is made for TM0, where the channel is in a deep fade and by definition PER $= 1$. Thus, the channel evolution is characterized by the $N+1$ probabilities of the channel states, which correspond to the TMs, and are derived accordingly to the PER and the average SNR of the considered

1930-529X/07/$25.00 © 2007 IEEE

Nakagami-*m* model. The transitions between these channel states are modeled as a Markov chain, described by an $(N+1)\times(N+1)$ channel state transition matrix $\mathbf{P}_C$ (see Eq. 5 in [6]), which is derived by means of a level crossing rate analysis [11]. The complete state of the single user is described by $\boldsymbol{\psi}_j^{(i)}$ in (1) (the user index j, not necessary in the SU case [6], will be used later for the MU case). If $c_j^{(i)} = n$, i.e., the AMC selector chooses the n th TM, denoted by TMn, then the CTI is divided in $K_j^{(n)} = K_n$ slots, and a single packet is transmitted in each slot. In addition, the queuing process is described by an embedded Markov chain. Specifically, the transitions of the substate $(q_j^{(i)}, r_j^{(i)})$ in each *slot*, for a given channel state $c_j^{(i)} = n$, are described by the $(B+1)(N_r+1)\times(B+1)(N_r+1)$ transition matrix $\mathbf{T}_n$ (see Eqs. 14-23 in [6]). Consequently, since the system is stable [6], the transitions of the substate $(q_j^{(i)}, r_j^{(i)})$ from a CTI to the next for the TM $c_j^{(i)} = n$ are described by the matrix $\mathbf{T}_n^{K_n}$ (see Fig. 1). The packet arrival process and the channel transitions are assumed independent, thus the transitions of the whole state $\boldsymbol{\psi}_j^{(i)} = (c_j^{(i)}, q_j^{(i)}, r_j^{(i)})$ is described by the $(N+1)(B+1)(N_r+1)\times(N+1)(B+1)(N_r+1)$ matrix (see Eqs. 27-28 in [6])

$$\mathbf{P} = \begin{bmatrix} P_{0,0}\mathbf{T}_0 & \cdots & P_{0,N}\mathbf{T}_0 \\ \vdots & \ddots & \vdots \\ P_{N,0}\mathbf{T}_N^{K_N} & \cdots & P_{N,N}\mathbf{T}_N^{K_N} \end{bmatrix}, \tag{2}$$

where the probabilities $P_{0,0},...,P_{N,N}$ are the elements of $\mathbf{P}_C$. It is worth noting that $K_0 = K_1 = 1$. The stationary probability vector associated to $\boldsymbol{\psi}_j^{(i)}$ is

$$\boldsymbol{\pi} = [\pi_{(0,0,0)}, \pi_{(0,0,1)}, ..., \pi_{(N,B,N_r-1)}, \pi_{(N,B,N_r)}], \tag{3}$$

where $\pi_{(c,q,r)}$ is the probability of the state (c,q,r), and can be classically calculated as (see Eq. 26 in [6])

$$\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}. \tag{4}$$

This is the main result of the embedded Markov chain modeling in [6], from which the packet loss rate (Eqs. 31-37), the system throughput (Eq. 38), and the average delay (Eqs. 39-44) are analytically derived.

The goal of our work is to derive, for *each user* of our MU system, a matrix $\tilde{\mathbf{P}}$ equivalent to $\mathbf{P}$ in (2), and consequently the aforementioned QoS measures accordingly to the approach developed in [6]. In effect, this is theoretically straightforward, since it would be sufficient to define, in the MU scenario, a superstate

$$\mathbf{s}^{(i)} = (c_1^{(i)}, q_1^{(i)}, r_1^{(i)}, c_2^{(i)}, q_2^{(i)}, r_2^{(i)}, ..., c_U^{(i)}, q_U^{(i)}, r_U^{(i)}), \tag{5}$$

which takes into account the state evolution of all the U users. This way, any scheduling policy based on a memoryless utility function

$$u_j^{(i)} = f(c_j^{(i)}, q_j^{(i)}, r_j^{(i)}), \tag{6}$$

could be modeled straightforwardly. Unfortunately, the number of possible superstates would be an exponential function of the number of users. Indeed, the corresponding stationary probability vector $\boldsymbol{\pi}^{(\mathbf{s})}$ associated to $\mathbf{s}^{(i)}$ would have a length equal to the length of the SU case raised to the U th power, as expressed by

$$L_{\boldsymbol{\pi}^{(\mathbf{s})}} = [(N+1)(B+1)(N_r+1)]^U = L_{\boldsymbol{\pi}}{}^U, \tag{7}$$

where $L_{\boldsymbol{\pi}}$ is the size of the vector $\boldsymbol{\pi}$. Thus, due to the eigenvalue decomposition used to solve (4), the computational complexity of the problem is $O(L_{\boldsymbol{\pi}^{(\mathbf{s})}}{}^3)$ [12], and by (7) becomes exponential in the number of users, making this approach impractical. In order to reduce the complexity, we make the key assumption of the *independence* of the users' stationary probability vectors. This assumption allows us to consider a single user as representative of the whole MU system, since we can write

$$\pi_{(c_1,q_1,r_1,...,c_U,q_U,r_U)} = \pi_{(c_1,q_1,r_1)}\pi_{(c_2,q_2,r_2)}...\pi_{(c_U,q_U,r_U)}. \tag{8}$$

The simulation results in Section V will show the good accuracy of this approximation, especially for high values of U. However, in a MU scenario, we must distinguish the cases in which the user is scheduled (Tx) or not (no-Tx). In the Tx case, the state transitions of the representative user are described by

$$\mathbf{P}^{(T)} = \begin{bmatrix} P_{0,0}\tilde{\mathbf{T}}_0 & \cdots & P_{0,N}\tilde{\mathbf{T}}_0 \\ \vdots & \ddots & \vdots \\ P_{N,0}\mathbf{T}_N^{K_N} & \cdots & P_{N,N}\mathbf{T}_N^{K_N} \end{bmatrix}, \tag{9}$$

which is very similar to $\mathbf{P}$. The only difference is represented by the submatrix $\tilde{\mathbf{T}}_0$, which is obtained with the same method of $\mathbf{T}_0$ (see Eqs. 16-23 in [6]), but without the rule that $r_j^{(i)}$ is augmented in the TM0 mode. This rule, which aims at reducing the delay in the SU case, is not convenient in a MU scenario, because this delay reduction is obtained by the scheduling. In the no-Tx case, the state transition matrix is

$$\mathbf{P}^{(N)} = \mathbf{P}_C \otimes \tilde{\mathbf{T}}_0, \tag{10}$$

where $\otimes$ denotes the Kronecker product [12], and accounts only for the channel variations and the packet arrival process. By $\mathbf{P}^{(T)}$ and $\mathbf{P}^{(N)}$, we can describe the state transitions of the representative user, provided that we can determine the probability of being, at the beginning of each CTI, in the Tx state for each different state transition, *based on the scheduling policy*. To this end, we define an extra substate $T_j^{(i)}$ that represents the transmitting condition: $T_j^{(i)} = 1$ when user j is Tx, and $T_j^{(i)} = 0$ when user j is no-Tx. With this approach, the state of *each* single user at t_i can be written as

$$\tilde{\boldsymbol{\psi}}_j^{(i)}(T_j^{(i)}) = (c_j^{(i)}, q_j^{(i)}, r_j^{(i)}, T_j^{(i)}). \tag{11}$$

In the following, we will drop the user index j for notation simplicity. It is worth noting that this way the number of states is $2L_{\boldsymbol{\pi}}$ *for any number of users* U, while in a rigorous analysis would be $L_{\boldsymbol{\pi}}{}^U$. Hence, our approach is scalable with the number of users. Now, we can write the $2L_{\boldsymbol{\pi}}\times 2L_{\boldsymbol{\pi}}$ state transition matrix $\tilde{\mathbf{P}}$ associated to $\tilde{\boldsymbol{\psi}}^{(i)}(T^{(i)})$ as

$$\tilde{\mathbf{P}} = \begin{bmatrix} \mathbf{M}_{(0,0,0)\to(0,0,0)} & \cdots & \mathbf{M}_{(0,0,0)\to(N,B,N_r)} \\ \vdots & \ddots & \vdots \\ \mathbf{M}_{(N,B,N_r)\to(0,0,0)} & \cdots & \mathbf{M}_{(N,B,N_r)\to(N,B,N_r)} \end{bmatrix}, \tag{12}$$

where each transition $\boldsymbol{\psi}^{(i-1)} \to \boldsymbol{\psi}^{(i)}$ is described by the 2×2 substate transition matrix

$$\mathbf{M}_{\boldsymbol{\psi}^{(i-1)}\to\boldsymbol{\psi}^{(i)}} = \begin{bmatrix} p_{\tilde{\boldsymbol{\psi}}^{(i-1)}(0)\to\tilde{\boldsymbol{\psi}}^{(i)}(0)} & p_{\tilde{\boldsymbol{\psi}}^{(i-1)}(0)\to\tilde{\boldsymbol{\psi}}^{(i)}(1)} \\ p_{\tilde{\boldsymbol{\psi}}^{(i-1)}(1)\to\tilde{\boldsymbol{\psi}}^{(i)}(0)} & p_{\tilde{\boldsymbol{\psi}}^{(i-1)}(1)\to\tilde{\boldsymbol{\psi}}^{(i)}(1)} \end{bmatrix}, \tag{13}$$

where we introduced the compact notation $p_E = \Pr\{E\}$. In (13), the elements in the first row represent the substate transitions $\boldsymbol{\psi}^{(i-1)} \to \boldsymbol{\psi}^{(i)}$ starting from the no-Tx case and are described by the matrix $\mathbf{P}^{(N)}$. Thus, by exploiting conditional

1930-529X/07/$25.00 © 2007 IEEE

probability rules, we can derive:

$$p_{\tilde{\psi}^{(i-1)}(0)\to\tilde{\psi}^{(i)}(1)} = [\mathbf{P}^{(N)}]_{k^{(i-1)},k^{(i)}}\, p_{T=0\to T=1|\psi^{(i-1)}\to\psi^{(i)}} ,$$
$$p_{\tilde{\psi}^{(i-1)}(0)\to\tilde{\psi}^{(i)}(0)} = [\mathbf{P}^{(N)}]_{k^{(i-1)},k^{(i)}} - p_{\tilde{\psi}^{(i-1)}(0)\to\tilde{\psi}^{(i)}(1)} , \quad (14)$$

where the index is expressed by $k^{(i)} = c^{(i)}(B+1)(N_r+1) + q^{(i)}(N_r+1) + r^{(i)} + 1$. Analogously, for the Tx case, we have

$$p_{\tilde{\psi}^{(i-1)}(1)\to\tilde{\psi}^{(i)}(1)} = [\mathbf{P}^{(T)}]_{k^{(i-1)},k^{(i)}}\, p_{T=1\to T=1|\psi^{(i-1)}\to\psi^{(i)}} ,$$
$$p_{\tilde{\psi}^{(i-1)}(1)\to\tilde{\psi}^{(i)}(0)} = [\mathbf{P}^{(T)}]_{k^{(i-1)},k^{(i)}} - p_{\tilde{\psi}^{(i-1)}(1)\to\tilde{\psi}^{(i)}(1)} . \quad (15)$$

At this point, to obtain the matrix $\tilde{\mathbf{P}}$, we calculate $p_{T=0\to T=1|\psi^{(i-1)}\to\psi^{(i)}}$ and $p_{T=1\to T=1|\psi^{(i-1)}\to\psi^{(i)}}$ for each transition of the substate $\psi^{(i)}$, based on the scheduling policy adopted.

IV. Stationary State Probability

Even with the simplification of (8), the calculation of the matrix $\tilde{\mathbf{P}}$ by the matrices in (13) is very challenging, mainly because the probabilities in (14)-(15) are related to the probabilities of each user to be in a certain state, which clearly depend on the stationary probability vector

$$\tilde{\boldsymbol{\pi}} = [\tilde{\pi}_{(0,0,0,0)}, \tilde{\pi}_{(0,0,0,1)}, \ldots, \tilde{\pi}_{(N,B,N_r,0)}, \tilde{\pi}_{(N,B,N_r,1)}] . \quad (16)$$

Each $\tilde{\pi}_{(c,q,r,T)}$ in (16) is the probability of the state (c,q,r,T), where it obviously holds true $\pi_{(c,q,r)} = \tilde{\pi}_{(c,q,r,0)} + \tilde{\pi}_{(c,q,r,1)}$. However, $\tilde{\boldsymbol{\pi}}$ is not known and its derivation, which is the goal of this work, would require a direct solution of $\tilde{\boldsymbol{\pi}}\tilde{\mathbf{P}} = \tilde{\boldsymbol{\pi}}$, which is not possible, due to the fact that $\tilde{\mathbf{P}}$ is unknown. Thus, we will resort to an iterative procedure, where, at the n th iteration, $\tilde{\boldsymbol{\pi}}_n$ is computed from the available version $\tilde{\mathbf{P}}_{n-1}$, and is used to derive an updated version $\tilde{\mathbf{P}}_n$, to be employed in the next iteration. As initialization, we set $\tilde{\boldsymbol{\pi}}_0 = \mathbf{1}_{2L_\pi}^T/(2L_\pi)$, where $\mathbf{1}_L$ is the all-ones column vector of size L. Then, the iterative procedure consists of the following three steps.

Step1) We derive the state transition matrix $\tilde{\mathbf{P}}_n$ using (12)-(15). The assumption in (8) greatly simplifies the computation of $p_{T=0\to T=1|\psi^{(i-1)}\to\psi^{(i)}}$ and $p_{T=1\to T=1|\psi^{(i-1)}\to\psi^{(i)}}$, because the state probabilities of each user can be evaluated separately. These transition probabilities appear, by (14) and (15), in the first two rows of the matrix $\tilde{\mathbf{P}}_n$. Moreover, since the scheduling policy is memoryless, i.e., does not depend on $\{\psi_j^{(i-1)}, \forall j\}$, the same values can be used for all the other rows. Accordingly to the scheduling algorithm defined in Section II, we distinguish between two buffer conditions for the j th user (the iteration index n is omitted for simplicity), denoted by C1 and C2.

C1) The buffer is not full, i.e., $q_j^{(i)} < B$. In this case, in order to find the probability that the user j of interest will transmit in the i th CTI, we firstly have to find the probability that no other user has full buffer or product $\phi_h^{(i)} < \phi_j^{(i)}$. Secondly, we have to find the conditional probability that a certain number of users have $\phi_h^{(i)} = \phi_j^{(i)}$, and finally the conditional probability that a certain fraction of those users have $r_h^{(i)} = r_j^{(i)}$. In order to express these probabilities, we define the events L, E, $\overline{T}$, $\overline{N}$, L_r, and E_r, as

$$L = [K_h^{(i)} q_h^{(i)} \le K_j^{(i)} q_j^{(i)}, q_h^{(i)} < B] , \quad \overline{N} = [T_h^{(i-1)} = 0] ,$$
$$E = [K_h^{(i)} q_h^{(i)} = K_j^{(i)} q_j^{(i)}, q_h^{(i)} < B] , \quad \overline{T} = [T_h^{(i-1)} = 1] , \quad (17)$$
$$L_r = [r_h^{(i)} \le r_j^{(i)}] , \quad E_r = [r_h^{(i)} = r_j^{(i)}] .$$

At this point, we can compute $p_{T=0\to T=1|\psi^{(i-1)}\to\psi^{(i)}}$ and $p_{T=1\to T=1|\psi^{(i-1)}\to\psi^{(i)}}$ for the representative user j. However, since these probabilities are *conditioned* to a particular transition of the substate $\psi_j^{(i)}$, we consider the state $\tilde{\psi}_j^{(i)}$ of the user j as fixed, and only the states of the other $U-1$ users as random variables. Moreover, it should be clear that the transmitting state $T_j^{(i-1)}$ of the user j at time t_{i-1} determines the transmitting states of all the other users. Thus, the possible states of the other users have a fixed value of $T_h^{(i-1)}$. Specifically, if $T_j^{(i-1)} = 1$, all the other $U-1$ users are not transmitting, and we can write:

$$p_{T=1\to T=1|\psi^{(i-1)}\to\psi^{(i)}} = \sum_{h=0}^{U-1}\binom{U-1}{h} p_{E|\overline{N}}^h (p_{L|\overline{N}} - p_{E|\overline{N}})^{U-1-h} \times \sum_{u=0}^{h}\binom{h}{u} p_{E_r|(E,\overline{N})}^u \frac{(p_{L_r|(E,\overline{N})} - p_{E_r|(E,\overline{N})})^{h-u}}{u+1} . \quad (18)$$

Conversely, when $T_j^{(i-1)} = 0$, there are other $U-2$ non-transmitting users and one transmitting user, as expressed by

$$p_{T=0\to T=1|\psi^{(i-1)}\to\psi^{(i)}} = \sum_{h=0}^{U-2}\binom{U-2}{h} p_{E|\overline{N}}^h (p_{L|\overline{N}} - p_{E|\overline{N}})^{U-2-h} \sum_{u=0}^{h}\binom{h}{u} p_{E_r|(E,\overline{N})}^u \times (p_{L_r|(E,\overline{N})} - p_{E_r|(E,\overline{N})})^{h-u} \left[\frac{p_{L|\overline{T}} - p_{E|\overline{T}} + p_{E|\overline{T}} p_{L_r|(E,\overline{T})}}{u+1} - \frac{p_{E|\overline{T}} p_{E_r|(E,\overline{T})}}{(u+2)(u+1)}\right].$$

C2) The buffer is full, i.e., $q_j^{(i)} = B$. In this case, we firstly have to find the probability that some other user has a full buffer, secondly the conditional probability that a certain number of users have $r_h^{(i)} = r_j^{(i)}$, and finally the conditional probability that a certain fraction of those users have $K_h^{(i)} = K_j^{(i)}$. We thus define further events $\overline{B}$, L_c, and E_c as

$$\overline{B} = [q_h^{(i)} = B] , \quad L_c = [K_h^{(i)} \le K_j^{(i)}] , \quad E_c = [K_h^{(i)} = K_j^{(i)}] . \quad (19)$$

Exploiting the considerations made in C1, we can write:

$$p_{T=1\to T=1|\psi^{(i-1)}\to\psi^{(i)}} = \sum_{h=0}^{U-1}\binom{U-1}{h} p_{\overline{B}|\overline{N}}^h \left(1 - p_{\overline{B}|\overline{N}}\right)^{U-1-h} \sum_{u=0}^{h}\binom{h}{u} p_{E_r|(\overline{B},\overline{N})}^u \times (p_{L_r|(\overline{B},\overline{N})} - p_{E_r|(\overline{B},\overline{N})})^{h-u} \sum_{v=0}^{u}\binom{u}{v} \frac{p_{E_c|(\overline{B},\overline{N},E_r)}^v (p_{L_c|(\overline{B},\overline{N},E_r)} - p_{E_c|(\overline{B},\overline{N},E_r)})^{u-v}}{v+1} ,$$

$$p_{T=0\to T=1|\psi^{(i-1)}\to\psi^{(i)}} = \sum_{h=0}^{U-2}\binom{U-2}{h} p_{\overline{B}|\overline{N}}^h (1 - p_{\overline{B}|\overline{N}})^{U-2-h} \sum_{u=0}^{h}\binom{h}{u} p_{E_r|(\overline{B},\overline{N})}^u \times (\overline{p}_{L_r|(\overline{B},\overline{N})} - p_{E_r|(\overline{B},\overline{N})})^{h-u} \sum_{v=0}^{u}\binom{u}{v} p_{E_c|(\overline{B},\overline{N},E_r)}^v (\overline{p}_{L_c|(\overline{B},\overline{N},E_r)} - p_{E_c|(\overline{B},\overline{N},E_r)})^{u-v} \times \left(\frac{1 - p_{\overline{B}|\overline{T}}(1 - p_{L_r|(\overline{B},\overline{T})} + p_{E_r|(\overline{B},\overline{T})}) + p_{\overline{B}|\overline{T}} p_{E_r|(\overline{B},\overline{T})} p_{L_c|(\overline{B},\overline{T},E_r)}}{v+1} - \frac{p_{\overline{B}|\overline{T}} p_{E_r|(\overline{B},\overline{T})} p_{E_c|(\overline{B},\overline{T},E_r)}}{(v+1)(v+2)}\right) . \quad (20)$$

Step2) The conditional probabilities in (18)-(20) must be determined by the knowledge of the stationary state probability vectors $\tilde{\boldsymbol{\pi}}$ of all the users $h \neq j$ (which by hypotheses are identical) at iteration $n-1$. However, only the substates $(c_h^{(i-1)}, q_h^{(i-1)}, r_h^{(i-1)})$ can evolve within a fixed CTI, because the scheduling policy, which imposes $T_h^{(i-1)}$, operates only at the beginning of each CTI. Consequently, since we have distinguished in (17)-(20) between Tx and no-Tx users

1930-529X/07/$25.00 © 2007 IEEE
This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE GLOBECOM 2007 proceedings.

(in the $(i-1)$ th CTI), we have to take into account the evolution of the substates $\{\boldsymbol{\psi}_h^{(i-1)}, h \neq j\}$ during the $(i-1)$ th CTI, which depends, for each user h, on the (fixed) substate $T_h^{(i-1)}$. This evolution is captured either by the matrix $\mathbf{P}^{(T)}$ or by $\mathbf{P}^{(N)}$. Consequently, we replace $\tilde{\boldsymbol{\pi}}_{n-1}$ with two different state distribution vectors $\overline{\boldsymbol{\pi}}_{n-1}^{(T)}$ and $\overline{\boldsymbol{\pi}}_{n-1}^{(N)}$, for the Tx and no-Tx cases, respectively, defined as

$$\overline{\boldsymbol{\pi}}_{n-1}^{(T)} = \frac{\tilde{\boldsymbol{\pi}}_{n-1}^{(T)} \mathbf{P}^{(T)}}{\tilde{\boldsymbol{\pi}}_{n-1}^{(T)} \mathbf{1}_{(N+1)(B+1)(N_r+1)}}, \quad [\tilde{\boldsymbol{\pi}}_{n-1}^{(T)}]_i = [\tilde{\boldsymbol{\pi}}_{n-1}]_{2i},$$

$$\overline{\boldsymbol{\pi}}_{n-1}^{(N)} = \frac{\tilde{\boldsymbol{\pi}}_{n-1}^{(N)} \mathbf{P}^{(N)}}{\tilde{\boldsymbol{\pi}}_{n-1}^{(N)} \mathbf{1}_{(N+1)(B+1)(N_r+1)}}, \quad [\tilde{\boldsymbol{\pi}}_{n-1}^{(N)}]_i = [\tilde{\boldsymbol{\pi}}_{n-1}]_{2i-1}. \quad (21)$$

These are the distribution vectors from which we must derive the probabilities used in (18) and (20), for Tx and no-Tx users, respectively. All these probabilities are conditional probabilities, involving specific events defined in (17) and (19), which in turn correspond to suitable subsets of the vectors $\overline{\boldsymbol{\pi}}_{n-1}^{(T)}$ or $\overline{\boldsymbol{\pi}}_{n-1}^{(N)}$. Due to the lack of space we cannot define all these probabilities. As an example, we can express $p_{L_r|(\bar{B},\bar{N}),n} = p_{(L_r,\bar{B},\bar{N}),n} / p_{(\bar{B},\bar{N}),n}$, which becomes:

$$p_{L_r|(\bar{B},\bar{N}),n} = \sum_{c_h=0}^{N} \sum_{r_h=0}^{r_j} \overline{\pi}_{(c_h,B,r_h),n-1}^{(N)} / \sum_{c_h=0}^{N} \sum_{r_h=0}^{N_r} \overline{\pi}_{(c_h,B,r_h),n-1}^{(N)}, \quad (22)$$

where $\overline{\pi}_{(c_h,q_h,r_h),n-1}^{(N)}$ is a generic element of $\overline{\boldsymbol{\pi}}_{n-1}^{(N)}$. This way, the computation of $\tilde{\mathbf{P}}_n$ is maintained affordable, independently of the number of states of each user.

Step3) The state distribution vector $\tilde{\boldsymbol{\pi}}_n$ is derived as

$$\tilde{\boldsymbol{\pi}}_n \tilde{\mathbf{P}}_n = \tilde{\boldsymbol{\pi}}_n. \quad (23)$$

The convergence of the iterative procedure has been investigated and verified by simulations, and a theoretical proof is left for further research. The iterative procedure is stopped when the maximum difference between the elements of $\tilde{\boldsymbol{\pi}}_{n-1}$ and $\tilde{\boldsymbol{\pi}}_n$ is below a suitable threshold. Noteworthy, we can model by the same approach other scheduling policies, by computation of equations analogous to (18)-(23).

V. Simulation Results

We considered a number of users ranging from 2 to 20. The total bandwidth for each link is 1.08M symbols/sec, $N_p = 1080$ bits, and $T_f = 2$ ms. The channel model for each user is based on a Markov chain as detailed in [11], and it is identical to one of those considered in [6], i.e., characterized by Rayleigh fading with SNR = 15 dB and Doppler frequency $f_d = 10$ Hz ($f_d T_f = 0.02$). The channel has six states ($N = 5$), where $K_n \in \{0,1,2,3,6,9\}$ is the average rate, associated with the channel states as in [6]. The average rate of a system with $U = 1$ would be 4.905 packets per CTI (2.64 Mb/sec), while an *ideal system* with infinite buffers' length and no delay constraints could achieve an aggregate *ideal total rate* (ITR)

$$c_{MAX}(U) = \sum_{n=0}^{N} p_{c_{j_{MAX}}^{(i)}=n} K_n = \sum_{n=0}^{N} p_{n,MAX} K_n \quad (24)$$

by scheduling the user j_{MAX} with the maximum rate in each CTI, thus maximally exploiting the MUD. The value of $p_{n,MAX}$ in (24) represent the probability of the best user to be in the TMn state, and is obtained as the stationary probability state associated with the channel matrix $\mathbf{P}_{C,MAX}$, defined similarly to $\mathbf{P}_C$, whose derivation is trivial and omitted for lack of space. We considered a Poisson arrival process, characterized by the conditional probability

$$p_{a|c_j^{(i)}} = \frac{(\lambda L_n)^a}{a!} e^{-\lambda L_n}, \quad a > 0, \quad (25)$$

where λ is the mean arrival rate and a represents the number of packets arriving in the i th CTI to user j during a single *slot*, of duration $L_n = T_f / K_j^{(i)}$. The value of λ is chosen as

$$\lambda = \alpha \frac{c_{MAX}(U)}{U T_f}. \quad (26)$$

The parameter $\alpha \in [0,1[$ quantifies the traffic intensity with respect to the ITR and gives an insight on the load of the scheduling policy. $\alpha = 1$ represents an unreachable limit for any scheduling policy in a practical system with finite buffer length. We considered three values, with $\alpha \in \{0.4, 0.5, 0.66\}$. The buffer length is $B = 10$, while the maximum number of retransmissions is $N_r = 3$. Thus, the size $2L_\pi \times 2L_\pi$ of the matrix $\tilde{\mathbf{P}}$ is 528 × 528. As in [6], we choose the TMs in order to guarantee the same PER = 0.05 for each TM different from TM0. It is meaningful to show only the behavior of the substates $q_j^{(i)}$ and $r_j^{(i)}$, instead of the whole vector $\tilde{\boldsymbol{\pi}}$, because $c_j^{(i)}$ is imposed by the channel matrix $\mathbf{P}_C$.

Fig. 2 shows, for $U = 20$ and $\alpha = 0.66$, the good agreement between the simulated and analytical stationary probability vectors of $q_j^{(i)}$ and $r_j^{(i)}$. Fig. 2 clearly illustrates the benefit for users with full buffer ($q = B = 10$) of being scheduled first, which drastically reduces the probability of overflow and consequently the PLR. Moreover, the probability of reaching the retransmission limit ($r = N_r = 3$) is very low, with negligible effect on the PLR. The PLR is plotted in Fig. 3 for different values of the traffic intensity parameter α as a function of the number U of users. A good agreement between theory and simulation is highlighted, unless for very low PLR values, which would require longer simulations. Fig. 3 also shows that the scheduling policy succeeds in exploiting the MUD of the system, with a significant reduction of the PLR for increasing U even for $\alpha = 0.66$, and an impressive one for lower values of the system load α.

Fig. 4 displays the average delay of the packets. A good agreement with simulations is observed also in this case, especially for higher numbers of users. It is worth noting that, for high U, the increase of the average delay with the number of users is approximately linear, with roughly the same slope. This implies that, for any α and any $U \geq 10$, adding new users causes the same increase in the average delay. In any case, Fig. 4 clearly highlights the benefits of the adopted scheduling policy with respect to the simplest possible one, which randomly schedules the users.

Fig. 5 plots the normalized throughput of each single user for several values of α and U. The normalized throughput is expressed as the ratio between the actual throughput and the ITR $c_{MAX}(U)$. In addition, Fig. 5 exhibits a good agreement between theory and simulations, especially for lower values of α. Moreover, it can be observed that the normalized throughput is largely independent from the number of users for $\alpha = 0.5$ and $\alpha = 0.4$, while for $\alpha = 0.66$ the MUD benefits are highlighted by the increase of the throughput with the number U of users: when $U = 20$, the throughput is close to its maximum theoretical value ($\alpha = 0.66$), while for lower values of U some throughput reduction is experienced. Thus, the good efficiency of our heuristic scheduling policy is confirmed even by Fig. 5.

1930-529X/07/$25.00 © 2007 IEEE
This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE GLOBECOM 2007 proceedings.

VI. CONCLUSIONS

We have proposed a simple scheduling algorithm to balance throughput and delay performance of a multi-user wireless system that exploits AMC and ARQ. The accuracy of the proposed theoretical analysis can be useful to maximize the throughput, similarly to [6], with a significantly reduced computational time with respect to extensive simulations. Further work can extend this approach to users with different QoS and traffic characteristics, as well as to provide performance comparisons with other scheduling strategies.

REFERENCES

[1] L. B. Le, E. Hossain, and A. S. Alfa, "Service differentiation in multirate wireless networks with weighted round-robin scheduling and ARQ-based error control," *IEEE Trans. Commun.*, vol. 54, pp. 208-215, Feb. 2006.

[2] Q. Liu, S. Zhou, and G. B. Giannakis, "Cross-layer combining of adaptive modulation and coding with truncated ARQ over wireless links," *IEEE Trans. Wireless Commun.*, vol. 3, pp. 1746-1755, Sept. 2004.

[3] Q. Liu, S. Zhou, and G. B. Giannakis, "Queuing with adaptive modulation and coding over wireless links: cross-layer analysis and design," *IEEE Trans. Wireless Commun.*, vol. 4, pp. 1142-1153, May 2005.

[4] X. Wang and J. K. Tugnait, "Joint design of channel distribution, truncated ARQ protocol and AMC scheme for multicode CDMA uplink," in *Proc. 38th Conf. Inf. Sciences Syst.*, Princeton Univ., May 2004.

[5] L. Kleinrock, *Queuing Systems*, vol. 1, John Wiley & Sons, NY, 1975.

[6] X. Wang, Q. Liu, and G. B. Giannakis, "Analyzing and optimizing adaptive modulation coding jointly with ARQ for QoS-guaranteed traffic," *IEEE Trans. Veh. Technol.*, vol. 56, pp. 710-720, Mar. 2007.

[7] Q. Liu, X. Wang, and G. B. Giannakis, "A cross-layer scheduling algorithm with QoS support in wireless networks," *IEEE Trans. Veh. Technol.*, vol. 55, pp. 839-847, May 2006.

[8] M. Nakagami, "The *m*-distribution – A general formula of intensity distribution of rapid fading," in *Statistical Methods in Radio Wave Propagation*, Oxford, U. K.: Pergamon, 1960, pp. 3-36.

[9] C. Comaniciu and H. V. Poor, "Jointly optimal power and admission control for delay sensitive traffic in CDMA networks with LMMSE receivers," *IEEE Trans. Signal Process.*, vol. 51, pp. 2031-2042, Aug. 2003.

[10] P. Viswanath, D. N. C. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Trans. Inf. Theory*, vol. 48, pp. 1277-1294, June 2002.

[11] M. D. Yacoub, J. E. Vargas Bautista, and L. Guerra de Rezende Guedes, "On higher order statistics of the Nakagami-*m* distribution," *IEEE Trans. Veh. Technol.*, vol. 48, pp. 790-794, May 1999.

[12] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins Univ. Press, 1996.

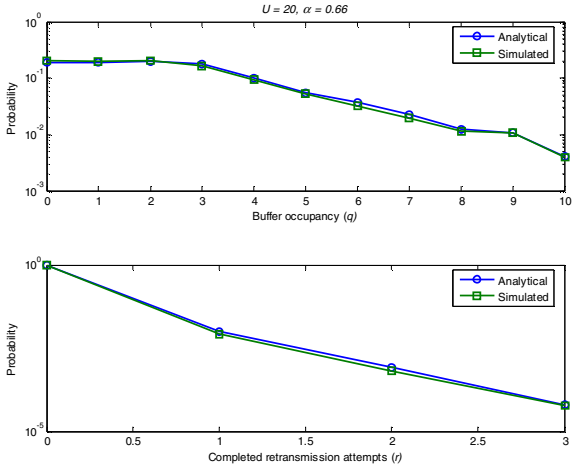

Fig. 2. (q) and (r) state distribution vectors, U =20, α = 0.66

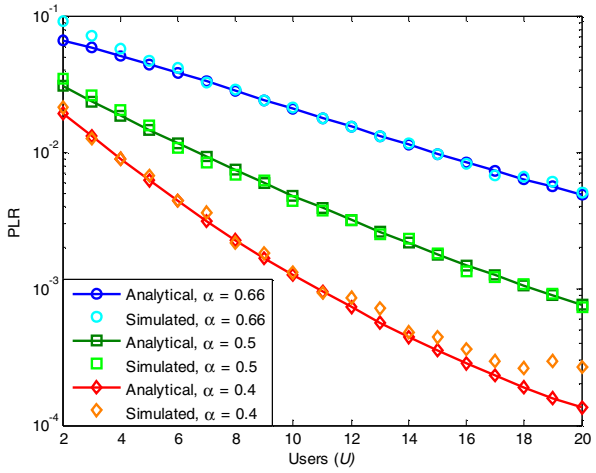

Fig. 3. Packet loss rate for different values of U.

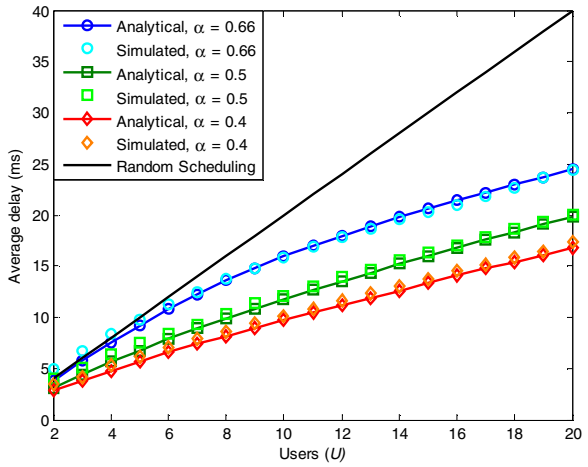

Fig. 4. Average delay for different values of U.

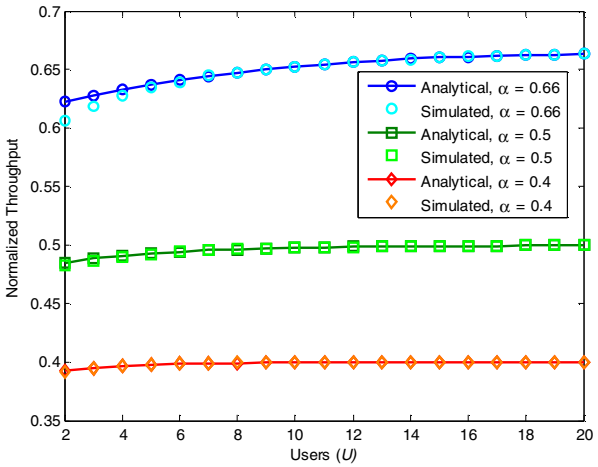

Fig. 5. Normalized throughput for different values of U