EURASIP Journal on Wireless Communications and Networking
a SpringerOpen Journal

RESEARCH Open Access

Channel-aware MAC performance of AMC-ARQ-based wireless systems

Loren Carrasco*, Guillem Femenias and Jaume Ramis

Abstract

This paper proposes a novel framework for the cross-layer design and optimization of wireless networks combining adaptive modulation and coding (AMC) at the physical (PHY) layer with automatic repeat request and channel-aware multiuser scheduling protocols at the data link control (DLC) layer. The proposed framework is based on the use of first-order two-dimensional discrete time Markov chains (DTMCs) jointly modeling the AMC scheme and the amplitude and rate of change of the wireless channel fading envelope. The behavior of the scheduler is embedded into the multidimensional PHY layer Markov model through the use of a service-vacation process. Using this PHY-media access control (MAC) Markov model, the quality of service performance at the DLC layer is discussed considering two different approaches. The first one relies on an analytical framework that is based on the multidimensional DTMC jointly describing the statistical behavior of the arrival process, the queuing system, and the PHY layer. The second one is rooted in the use of the effective bandwidth theory to model the packet arrival process and the effective capacity theory to model the PHY/MAC behavior. Both the DTMC-based and effective bandwidth/capacity-based approaches are analyzed and compared in a cross-layer design aiming at maximizing the average throughput of the system where constraints on the maximum tolerable average packet loss and delay are to be fulfilled.

Keywords: Adaptive modulation and coding; Automatic repeat request; Channel-aware scheduling; Markov models; Effective capacity; Cross-layer design

1 Introduction

Scheduling and automatic repeat request (ARQ) error control protocols at the data link control (DLC) layer and adaptive modulation and coding (AMC) strategies at the physical (PHY) layer are some of the key technologies underpinning state-of-the-art and next-generation wireless communication systems. They are used to optimize resource utilization while providing support to a wide range of multimedia applications with heterogeneous quality of service (QoS) requirements. However, owing to the strong dependencies between DLC and PHY layers in wireless networks, efficiency in system performance may not be warranted using a strictly layered optimization approach. Consequently, cross-layer designs able to jointly optimize the scheduling, ARQ, and AMC functions should be devised.

Although many recent works focus on cross-layer designs that combine AMC schemes with ARQ error control protocols (see, e.g.,[1-12]), proposals also incorporating the multiuser scheduling process at the media access control (MAC) sublayer are much less common (see, e.g.,[13-15]). Liu et al. in [15] presented an opportunistic scheduling scheme to improve the delay performance of secondary users with bursty traffic in cognitive radio (CR) systems. They consider a relay-assisted CR network with a decode-and-forward relaying scheme. Cooperative beamforming is used by the relays to forward packets in either idle or busy time slots without causing interference to primary users. However, in the proposed scheme, although there is a scheduler planning the transmissions of the source and the relays, only one user is considered and the use of AMC is not taken into account. Poggioni et al. in [13] developed a theoretical framework based on a finite-state Markov chain (FSMC) modeling a heterogeneous multiuser scenario where groups of users with different QoS requirements coexist. In this analysis, it

*Correspondence: loren.carrasco@uib.es
Mobile Communication Group, University of the Balearic Islands (UIB), Ctra. de Valldemossa Km. 7,5, Palma 07122, Spain

is assumed that the Markov chain steady-state probabilities of any user can be considered independent from the steady-state probabilities of all the other users in the system. Furthermore, it is assumed that the steady-state probabilities of different users belonging to the same QoS class are identical. These assumptions restrict the possible application scenarios of this approach as they imply that the traffic and channel characteristics are exactly the same for all users belonging to the same QoS class.

The first-order amplitude-based finite state Markov chain (AFSMC) model developed by Le et al. in [5], including both the AMC and ARQ procedures, was extended by the same authors in [14] to incorporate the multiuser scheduling process. The max-rate multiuser scheduler was included in the model through a service-vacation process allowing a manageable number of system states irrespective of the number of users sharing the channel. Nevertheless, the analysis in [14] suffers from an inaccurate modeling of the flat-fading wireless channel caused by the use of a first-order AFSMC (see [9-12] for an in-depth discussion of this and related issues). Moreover, the approach in [14] does not define a cross-layering scheme as a means to optimize the system performance, and on top of this, users are assumed to operate in channels with equal characteristics, thus restricting the usefulness of the presented results.

In this paper, capitalizing on the approach described in [14], a service-vacation process is used to embed the channel-aware scheduling protocol behavior into the AMC/ARQ multidimensional discrete time Markov chain (DTMC) model described in some of our previous contributions [9-12]. Our approach is based on a first-order two-dimensional (2D) Markov model for the wireless flat-fading channel that, as was shown in [9-12], solves most of the drawbacks of the AFSMC model used in [1-8,13,14]. In addition to the max-rate scheduling algorithm discussed in [14], our approach can be extended to the analysis of more sophisticated scheduling algorithms, including the proportional fairness multiuser scheduler. Moreover, it is not constrained by assumptions on the users' traffic and/or channel characteristics. Furthermore, as in [10], two of the principal approaches used in the technical literature to model the DLC layer behavior, namely the DTMC model [4,5] and the effective capacity and effective bandwidth theories [16], are compared in this paper. Both schemes are used to jointly characterize the effects of the multiuser scheduler, the ARQ error control protocols, and the AMC strategies. Finally, another contribution of this paper is the proposal of a cross-layer optimization design that, by tuning selected system parameters such as the average target packet error rate (PER) and/or the average packet arrival rate, is able to coordinate the behavior of AMC, ARQ, and scheduling procedures. The main objective is to optimize the global system performance in terms of average throughput, delay, queue length, and packet loss ratio.

The organization of this paper is as follows: The system model is introduced in Section 2, including subsections describing the AMC scheme, the PHY layer first-order 2D Markov model, and the joint MAC-PHY Markov model. Sections 3 and 4 describe the max-rate and proportional fair schedulers, respectively. Section 5 is devoted to discuss the different approaches that have been used to analyze the interactions between PHY and DLC layers, namely the embedded DTMC approach and the effective bandwidth/capacity theory-based approach. The PHY-MAC cross-layer designs for max-rate and proportional fair schedulers are described in Section 6. In Section 7, analytical and Monte Carlo simulation results are used to validate our model and to establish a fair comparison between DTMC-based and effective bandwidth/capacity-based cross-layer approaches. Finally, the paper concludes in Section 8 with a summary of the main results and contributions.

2 System model and assumptions

A block diagram of the system under consideration is shown in Figure 1. As it can be observed, the downlink scenario of a wireless system with a base station (BS) serving N_s users is considered. At the BS, there are N_s separate radio link level buffers that are used to queue packet arrivals corresponding to every user connected to the BS. These buffers operate in a first-in-first-out (FIFO) fashion and can store up to $\overline{\boldsymbol{Q}} = \{\overline{Q}^1, \ldots, \overline{Q}^{N_s}\}$ packets, where $\overline{Q}^u$ is the queue length of user u. The scheduler, based on channel state information (CSI) collected from the N_s users and using a time division multiplexing scheme, takes scheduling decisions to allocate transmission opportunities to active users. Adaptive transmission is performed by using an ARQ error control scheme at the DLC layer and an AMC strategy at the PHY layer. The processing unit at the DLC layer is a packet and the processing unit at the PHY layer is a frame. The link is assumed to support QoS-guaranteed traffic characterized by a maximum average packet delay $D_{l_{\max}}$ and a target link layer packet loss rate (PLR) $P_{l_{\max}}$.

The AMC scheme is assumed to have a set $\boldsymbol{\mathcal{M}}_p = \{0, \ldots, M_p - 1\}$ of M_p *possible* transmission modes (TMs), each of which corresponding to a particular combination of modulation and coding strategies. It is assumed that when the system uses TM $n \in \boldsymbol{\mathcal{M}}_p$, it transmits $p_n = bR_n$ packets per frame, where R_n denotes the number of information bits per symbol used by TM n and b is a parameter that determines the number of transmitted packets per frame, which is up to the designer's choice. For convenience, we consider that $p_0 < \cdots < p_{M_p-1}$, with $p_0 = 0$ (i.e., TM 0 corresponds to the absence of transmission)

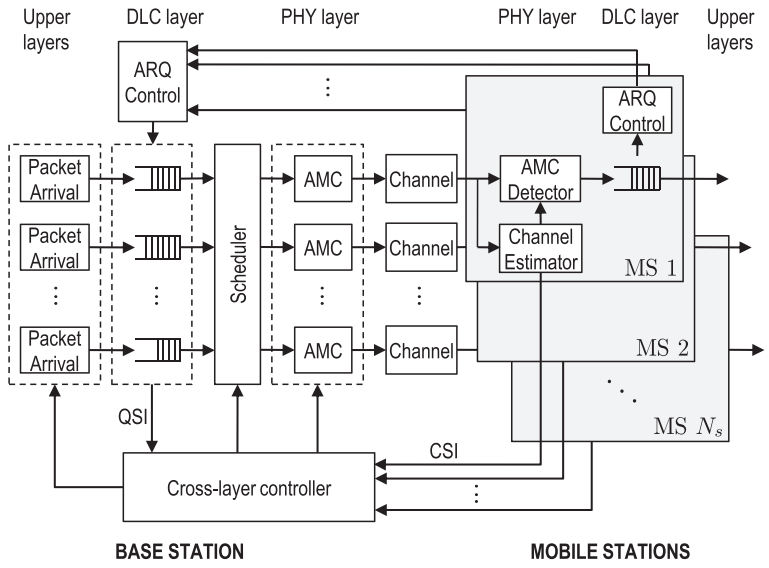

Figure 1 System model.

and $p_{M_p-1} \triangleq \mathcal{C}_p$. As it was shown in [9], depending on the channel conditions and the QoS requirements of the different users, some of these M_p *possible* TMs may be deemed *useless*, and thus, only a set $\boldsymbol{\mathcal{M}} = \{0, \ldots, M^u - 1\}$ of M^u *useful* TMs will be available to the AMC scheme for user u. It will be assumed that when user u is allocated *useful* TM $n \in \boldsymbol{\mathcal{M}}^{\boldsymbol{u}}$, the system transmits c_n packets and, for convenience, we also consider that $c_0 < \cdots < c_{M^u-1}$, with $c_0 \geq 0$ and $c_{M^u-1} = \mathcal{C}^u \leq \mathcal{C}_p$.

A Rayleigh block-fading model [17] is adopted for the propagation channel, according to which the channel is assumed to remain invariant over a time frame interval T_f and is allowed to vary across successive frame intervals[a]. Perfect CSI is assumed to be available at the receiver side, and thus, an ideal frame-by-frame TM selection process is performed at the AMC controller of the receiver. Furthermore, an error-free and instantaneous ARQ feedback channel is assumed.

As in [5,9-12], we assume that the packet generation of user u adheres to a discrete batch Markovian arrival process (D-BMAP). As stated by Blondia in [18], a D-BMAP can be described by substochastic matrices $\boldsymbol{U}_a^u$, $a = 0, 1, 2, \ldots, n$, of the order $\mathcal{A}^u \times \mathcal{A}^u$, with elements $u_a^u(i,j)$ denoting the probability of a transition from phase i to phase j with a batch arrival of size a and $\sum_{a=0}^{\infty} \sum_{j=1}^{\mathcal{A}^u} u_a^u(i,j) = 1$. The transition probability matrix can be obtained as $\boldsymbol{U}^u = \sum_{a=0}^{\infty} \boldsymbol{U}_a^u$.

Owing to the Markovian property of the arrival process, we have $\boldsymbol{\omega}^u = \boldsymbol{\omega}^u \boldsymbol{U}^u$ and $\boldsymbol{\omega}^u \mathbf{1}_{\mathcal{A}^u} = 1$, where $\boldsymbol{\omega}^u$ denotes the D-BMAP steady-phase probability vector and $\mathbf{1}_{\mathcal{A}^u}$ is an all-ones column vector of length $\mathcal{A}^u$. Then, the average arrival rate λ^u can be calculated as

$$\lambda^u = \boldsymbol{\omega}^u \sum_{a=0}^{\mathcal{A}^u - 1} a \, \boldsymbol{U}_a^u \mathbf{1}_{\mathcal{A}^u}. \tag{1}$$

It will be assumed that the average arrival rate to the DLC layer λ^u is a system parameter that can be controlled through a traffic shaping and modeling mechanism in order to comply with the QoS requirements of the system.

2.1 Adaptive modulation and coding

Let γ_ν^u denote the instantaneous received signal-to-noise ratio (SNR) of user u at time instant $t = \nu T_f$. For the assumed Rayleigh block-fading channel model, γ_ν^u can be modeled as an exponentially distributed random variable with mean $\overline{\gamma}^u = E\{\gamma_\nu^u\}$. Given γ_ν^u, the objective of AMC is to select the TM that maximizes the data rate while maintaining an average PER less than or equal to a prescribed value P_0^u. To this end, and according to [3], the entire SNR range is partitioned into a set of nonoverlapping intervals defined by the partition $\boldsymbol{\Gamma}^{u,m} = \left\{0, \gamma_1^{u,m}, \gamma_2^{u,m}, \ldots, \gamma_{M^u-1}^{u,m}, \infty\right\}$ and TM n will be selected when $\gamma_\nu^u \in \left[\gamma_n^{u,m}, \gamma_{n+1}^{u,m}\right]$. In this paper, the partition $\boldsymbol{\Gamma}^{u,m}$ is obtained by using the threshold searching algorithm described in [10]. This searching algorithm has the capability to discriminate between *useful* and *useless* TMs, while guarantying that the average PER fulfills the prescribed constraint. We also assume, without loss of generality, that convolutionally coded M-QAM, adopted from the IEEE 802.11a standard [19], are used in the AMC pool. All possible TMs are listed in ([8], Table one).

2.2 Two-dimensional Markov channel modeling

Let us define the rate of change of the fading as $\delta_\nu^u = \gamma_{\nu-1}^u - \gamma_\nu^u$. Let us also divide the ranges of γ_ν^u and δ_ν^u into sets of nonoverlapping 2D cells defined by the partitions $\mathbf{\Gamma}^{u,c} = \left\{0, \gamma_1^{u,c}, \gamma_2^{u,c}, \ldots, \gamma_{K-1}^{u,c}, \infty\right\}$ and $\mathbf{\Delta} = \{-\infty, 0, \infty\}$, respectively. A first-order 2D Markov channel model can now be defined where each state of the channel corresponds to one of such cells. That is, the Markov chain state of the channel at time instant $t = \nu T_f$ can be denoted as $\boldsymbol{\zeta}_\nu^u = \left(\chi_\nu^u, \Delta_\nu^u\right)$, $\nu = 0, 1, \ldots, \infty$, where $\chi_\nu^u = k$ if and only if $\gamma_k^{u,c} < \gamma_\nu^u \leq \gamma_{k+1}^{u,c}$, and $\Delta_\nu^u = 0$ (or $\Delta_\nu^u = 1$) if and only if $\delta_\nu^u < 0$ (or $\delta_\nu^u \geq 0$).

In our approach the partition $\mathbf{\Gamma}^{u,c}$ is designed assuming that the observable dummy output of our improved first-order 2D Markov model at time instant $t = \nu T_f$ belongs to a codebook of nominal values of SNR $\mathbf{\Psi}^{u,c} = \left\{\Psi_1^{u,c}, \Psi_2^{u,c}, \ldots, \Psi_K^{u,c}\right\}$. The Max-Lloyd algorithm [20,21], developed for the optimum design of nonuniform quantizers, is then used to determine the partition and codebook minimizing the mean square error between γ_ν^u and the quantizer output.

2.3 Physical layer 2D Markov model

Based on the TM selection process used by the AMC scheme (which is defined by the partition $\mathbf{\Gamma}^{u,m}$) and the first-order 2D Markov channel model (which is characterized by the partitions $\mathbf{\Gamma}^{u,c}$ and $\mathbf{\Delta}$), the range of γ_ν^u is partitioned into the set of nonoverlapping intervals defined by $\mathbf{\Gamma}^{u,m,c} = \left\{\left[\gamma_0^{u,m,c}, \gamma_1^{u,m,c}\right) \ldots \left[\gamma_{N_{\text{PHY}}^u\text{-}1}^{u,m,c}, \gamma_{N_{\text{PHY}}^u}^{u,m,c}\right)\right\}$, where N_{PHY}^u denotes the number of PHY states corresponding to user u, and $\left\{\gamma_1^{u,m,c}, \ldots, \gamma_{N_{\text{PHY}}^u\text{-}1}^{u,m,c}\right\} = \text{sort}\left(\left\{\gamma_1^{u,m}, \ldots, \gamma_{M^u\text{-}1}^{u,m}\right\} \cup \left\{\gamma_1^{u,c}, \ldots, \gamma_{K\text{-}1}^{u,c}\right\}\right)$, with $\gamma_0^{u,m,c} = 0$ and $\gamma_{N_{\text{PHY}}^u}^{u,m,c} = \infty$. Each partition interval $\left[\gamma_k^{u,m,c}, \gamma_{k+1}^{u,m,c}\right)$ is characterized by a particular combination of TM and channel state. As in Subsection 2.2, the range of δ_ν^u is also partitioned into the set of nonoverlapping intervals $\mathbf{\Delta} = \{-\infty, 0, \infty\}$.

Using this 2D partitioning, a first-order 2D Markov model for the PHY layer of user u can be defined where each state corresponds to one of such 2D rectangular-shaped cells. Furthermore, the PHY layer Markov chain state at time instant $t = \nu T_f$ is denoted by $\boldsymbol{\varsigma}_\nu^u = \left(\varphi_\nu^u, \Delta_\nu^u\right)$, $\nu = 0, 1, \ldots, \infty$, where $\varphi_\nu^u \in \{0, \ldots, N_{\text{PHY}}^u - 1\}$ represents the combination of TM and channel state in this frame interval and $\Delta_\nu^u \in \{0, 1\}$ is used to denote the *up* or *down*[b] characteristic of the instantaneous SNR over the time frame interval $t = (\nu - 1)T_f$. At any time instant $t = \nu T_f$, the PHY layer state can be univocally identified by an integer number $y_\nu^u = 2\varphi_\nu^u + \Delta_\nu^u$, with $y_\nu^u \in \{0, \ldots, 2N_{\text{PHY}}^u - 1\}$, which can be characterized by an steady-state probability $P_{\text{PHY}}(y_\nu^u)$ and a corresponding conditional average PER $\overline{\text{PER}}_{\text{PHY}}(y_\nu^u)$. Additionally, the PHY layer FSMC will be characterized by a transition probability matrix $\boldsymbol{H}_s^u = \left[H_{i,j}^u\right]_{0 \leq i,j \leq 2N_{\text{PHY}}^u - 1}$, where $H_{i,j}^u = \Pr\{y_{\nu+1}^u = j | y_\nu^u = i\}$. Throughout this paper, the steady-state probabilities, the conditional average PERs, and the state-transition probabilities have all been computed either numerically or by simulation.

2.4 Joint PHY-MAC layer Markov model

Channel-aware-only schedulers can be incorporated to the joint PHY-MAC Markov model by means of a service-vacation process [14]. When a particular user u is selected for transmission in a given time slot, it is said that this user PHY layer is in service; otherwise, it is said to be on vacation. The parameter $z^u \in \{0, 1\}$ is used to denote the service ($z^u = 0$) or vacation ($z^u = 1$) state. The decision wether a user u will be in service or vacation during the next time slot will depend on the possible PHY layer states of all users in the next time slot and on previous scheduling decisions. A D-step memory in the service-vacation process represents the scheduling dependence on D previous decisions and can be used to account for an increased degree of fairness between users.

The joint PHY-MAC layer FSMC state for user u at time instant $t = \nu T_f$ is denoted by the vector of random variables $\boldsymbol{\iota}_\nu^u = (z_\nu^u, z_{\nu-1}^u, \ldots, z_{\nu-D+1}^u, y_\nu^u)$. At any time instant $t = \nu T_f$, the joint PHY-MAC layer state can be univocally identified by an integer number n_ν^u with $n_\nu^u \in \{0, \ldots, N_{\text{PHY-MAC}}^u - 1\}$, where $N_{\text{PHY-MAC}}^u = 2^{D+1} N_{\text{PHY}}^u$. The joint MAC-PHY layer will be in state n_ν^u with a steady-state probability $P_{\text{PHY-MAC}}^u(n_\nu^u)$. Taking into account that user u transmits only when it is in service, the different PHY-MAC states will have a transmission rate (TR), measured in packets per slot, of $\widehat{c}_{n_\nu^u} = c_{y_\nu^u}(1 - z_\nu^u)$, where $c_{y_\nu^u}$ is the TR characterizing PHY layer state y_ν^u. Furthermore, the PHY-MAC layer FSMC will be described by a transition probability matrix $\boldsymbol{P}_s^u = \left[P_{i,j}^u\right]_{0 \leq i,j \leq N_{\text{PHY-MAC}}^u - 1}$, with state transition probabilities

$$P_{i,j}^u = \Pr\left\{n_{\nu+1}^u = j | n_\nu^u = i\right\} \tag{2}$$

that can be analytically calculated for a significant number of scheduling schemes. Without loss of generality, these probabilities are derived in the following sections for the max-rate and proportional fair algorithms, which, in both cases, can be modeled by a service-vacation process with one-step memory ($D = 1$).

3 The max-rate scheduling example

In the max-rate (MR) scheduling rule, the PHY layer states of all active users are assumed to be available at the scheduler without delay. The MR scheduler grants the transmission opportunity to the user that can achieve the highest TR in the current frame. If more than one user

can attain this maximum rate, the scheduler chooses one of them randomly. Although this case was covered in [14], several modifications are included in our analysis in order to adapt it to the 2D PHY layer model and, also, to generalize its application to more realistic scenarios with users experiencing heterogeneous average SNRs.

In one-step memory service-vacation processes, scheduling decisions only rely on the actual system state, and thus, the state transition probabilities in (2) can be simplified to $P^u_{i,j} = \Pr\left\{z^u_{\nu+1}, y^u_{\nu+1} | z^u_\nu, y^u_\nu\right\}$. The transition probability matrix can be expressed as

$$\boldsymbol{P}^u_s = \begin{pmatrix} \boldsymbol{S}^u_{0,0} & \boldsymbol{S}^u_{0,1} \\ \boldsymbol{S}^u_{1,0} & \boldsymbol{S}^u_{1,1} \end{pmatrix},$$

where $\boldsymbol{S}^u_{i,j}$ is a $(2N^u_{\text{PHY}}) \times (2N^u_{\text{PHY}})$ matrix with elements $S^u_{i,j}(k,l) = \Pr\left\{z^u_{\nu+1} = j, y^u_{\nu+1} = l | z^u_\nu = i, y^u_\nu = k\right\}$.

Without loss of generality, user $u = 1$ is considered to be the user of interest and, for notation simplicity, it is assumed that $z^1_\nu = z_\nu$ and $y^1_\nu = y_\nu$. Taking into account that the PHY layer and service-vacation processes are independent, the elements of the $\boldsymbol{S}^1_{i,j}$ matrices can be written as

$$\begin{aligned}\Pr\left\{z_{\nu+1}, y_{\nu+1} | z_\nu, y_\nu\right\} &= \Pr\left\{z_{\nu+1} | z_\nu, y_{\nu+1}, y_\nu\right\} \\ &\quad \times \Pr\left\{y_{\nu+1} | y_\nu\right\},\end{aligned}$$

the latter term being an element of the PHY layer state transition probability matrix $\boldsymbol{H}^1_s$. Moreover, since $z_{\nu+1} \in \{0,1\}$, it holds that

$$\begin{aligned}&\Pr\left\{z_{\nu+1} = 1 | z_\nu = i, y_{\nu+1} = l, y_\nu = k\right\} \\ &= 1 - \Pr\left\{z_{\nu+1} = 0 | z_\nu = i, y_{\nu+1} = l, y_\nu = k\right\},\end{aligned}$$

and therefore, only the case $z_{\nu+1} = 0$ needs to be discussed. Considering now that the service state at time ν depends only on the PHY layer state at time ν, it holds that

$$\begin{aligned}&\Pr\left\{z_{\nu+1} = 0 | z_\nu = i, y_{\nu+1} = l, y_\nu = k\right\} \\ &= \frac{\Pr\left\{z_{\nu+1} = 0, z_\nu = i | y_{\nu+1} = l, y_\nu = k\right\}}{\Pr\left\{z_\nu = i | y_\nu = k\right\}}.\end{aligned} \tag{3}$$

3.1 Calculation of $\Pr\{z_\nu = i | y_\nu = k\}$

Assuming $z_\nu = 0$, the denominator of (3) can be calculated as

$$\begin{aligned}&\Pr\left\{z_\nu = 0 | y_\nu = k\right\} \\ &= \sum_{y^2_\nu=0}^{\widehat{N}^2_{\text{PHY}}} \sum_{y^3_\nu=0}^{\widehat{N}^3_{\text{PHY}}} \cdots \sum_{y^{N_s}_\nu=0}^{\widehat{N}^{N_s}_{\text{PHY}}} \Pr\left\{z_\nu = 0, y^2_\nu, y^3_\nu, \ldots, y^{N_s}_\nu | y_\nu = k\right\},\end{aligned}$$

where $\widehat{N}^u_{\text{PHY}} \triangleq 2N^u_{\text{PHY}} - 1$. At time slot ν, user 1, whose PHY layer is in state k, can only be in service if the rest of users have a PHY layer state with a lower or equal TR. When a users (including user 1) can transmit at maximum TR, then user 1 is chosen for transmission with a probability $1/a$. Thus,

$$\begin{aligned}&\Pr\left\{z_\nu = 0, y^2_\nu, y^3_\nu, \ldots, y^{N_s}_\nu | y_\nu = k\right\} \\ &= \begin{cases} 0, \text{if } \exists u \in \mathcal{U} : c_{y^u_\nu} > c_k \\ \frac{1}{a} \prod_{u=2}^{N_s} P^u_{\text{PHY}}(y^u_\nu), \text{if } \exists \{u_i \in \mathcal{U}\}_{i=1}^{a-1} : c_{y^{u_i}_\nu} = c_k \forall i \end{cases},\end{aligned} \tag{4}$$

where $\mathcal{U} = \{2, \ldots, N_s\}$ is the set of competitor users and c_k is the TR corresponding to $y_\nu = k$.

3.2 Calculation of $\Pr\{z_{\nu+1} = 0, z_\nu = i | y_{\nu+1} = l, y_\nu = k\}$

The numerator of (3) can be written as

$$\begin{aligned}&\Pr\left\{z_\nu = z, z_{\nu+1} = \nu | y_\nu = k, y_{\nu+1} = l\right\} \\ &= \sum_{y^2_{\nu+1}=0}^{\widehat{N}^2_{\text{PHY}}} \cdots \sum_{y^{N_s}_{\nu+1}=0}^{\widehat{N}^{N_s}_{\text{PHY}}} \sum_{y^2_\nu=0}^{\widehat{N}^2_{\text{PHY}}} \cdots \\ &\quad \times \sum_{y^{N_s}_\nu=0}^{\widehat{N}^{N_s}_{\text{PHY}}} \Pr\left\{z_\nu = z, z_{\nu+1} = \nu, y^2_\nu, \ldots, \right. \\ &\quad \left. y^{N_s}_\nu, y^2_{\nu+1}, \ldots, y^{N_s}_{\nu+1} | y_\nu = k, y_{\nu+1} = l\right\}.\end{aligned} \tag{5}$$

In order to obtain the terms inside the summations of this expression, two different cases should be considered:

1. *Case 1* ($z_{\nu+1} = 0, z_\nu = 0$). In this case, user 1 is in service during time slots ν and $\nu+1$ if its PHY states in these time slots are k and l, respectively. This will only happen when the potential TRs of the other $N_s - 1$ users are smaller than or equal to the TRs of user 1 in PHY states k and l during time slots ν and $\nu+1$, respectively. If a and b users (including user 1) can transmit at maximum TR during the ν and $\nu+1$ time slots, respectively, then user 1 will be granted transmission for both time slots with probability $1/(ab)$. Therefore, in this case the probabilities in (5) can be calculated as

$$\begin{aligned}&\Pr\{z_\nu = 0, z_{\nu+1} = 0, y^2_\nu, \ldots, y^{N_s}_\nu, y^2_{\nu+1}, \ldots, y^{N_s}_{\nu+1} | y_\nu = k, y_{\nu+1} = l\} = \\ &\begin{cases} 0, \text{if } \exists u \in \mathcal{U} : c_{y^u_\nu} > c_k \text{ or } c_{y^u_{\nu+1}} > c_l \\ \frac{1}{ab} \prod_{u=2}^{N_s} \Pr\{y^u_\nu, y^u_{\nu+1}\}, \begin{cases} \text{if } \exists \{u_i \in \mathcal{U}\}_{i=1}^{a-1} : c_{y^{u_i}_\nu} = c_k \forall i \\ \text{and } \exists \{u_j \in \mathcal{U}\}_{j=1}^{b-1} : c_{y^{u_j}_{\nu+1}} = c_k \forall j \end{cases} \end{cases},\end{aligned} \tag{6}$$

where $\Pr\{y_\nu = k, y_{\nu+1} = l\} = H^1_{k,l} P^1_{\text{PHY}}(k)$.

2. *Case 2* ($z_{\nu+1} = 0, z_\nu = 1$). In this case, user 1 makes a transition from the vacation state during time slot ν to the service state at $\nu+1$. The service state in time slot $\nu+1$ can occur if b users (including user 1) can

transmit at maximum TR and user 1 is selected for transmission with probability $1/b$. A vacation state during time slot ν can happen as a result of two different situations, either there are users with higher TRs than user 1 or a users (including user 1) can transmit with the maximum TR and user 1 is not selected with probability $(1-\frac{1}{a})$. Then, in case 2, the probabilities in (5) can be obtained using

$$\Pr\{z_\nu=1, z_{\nu+1}=0, y_\nu^2,\ldots,y_\nu^{N_s}, y_{\nu+1}^2,\ldots,y_{\nu+1}^{N_s} | y_\nu=k, y_{\nu+1}=l\} = \begin{cases} 0, \text{if } c_{y_\nu^u} < c_k \forall u \in \mathcal{U} \text{ or } \exists u \in \mathcal{U} : c_{y_{\nu+1}^u} > c_l \\ \frac{1}{b}\prod_{u=2}^{N_s} \Pr\{y_\nu^u, y_{\nu+1}^u\} \begin{cases} \text{if } \exists u \in \mathcal{U} : c_{y_\nu^u} > c_k \\ \text{and } \exists\{u_j \in \mathcal{U}\}_{j=1}^{b-1} : c_{y_{\nu+1}^{u_j}} = c_k \forall j \end{cases} \\ \frac{a-1}{ab}\prod_{u=2}^{N_s} \Pr\{y_\nu^u, y_{\nu+1}^u\} \begin{cases} \text{if } \exists\{u_i \in \mathcal{U}\}_{i=1}^{a-1} : c_{y_\nu^{u_i}} = c_k \forall i \\ \text{and } \exists\{u_j \in \mathcal{U}\}_{j=1}^{b-1} : c_{y_{\nu+1}^{u_j}} = c_k \forall j \end{cases} \end{cases}. \quad (7)$$

4 The proportional fair scheduling example

Originally proposed in the wired network scheduling context, a proportional fair (PF) scheduler promises a tradeoff between the maximization of average throughput and system fairness. At each time instant, the user experiencing the highest instantaneous rate with respect to its average rate is scheduled. That is, user q is selected for transmission during time slot ν if

$$q = \arg \max_{u\in\{1,\ldots,N_s\}} \frac{c_{y_\nu^u}}{T_\nu^u}, \quad (8)$$

where T_ν^u is the average rate of user u. The scheduler defined in (8) maximizes the logarithmic sum of system throughput [22]. The average rate can be computed as a moving average over a time window of length W, that is,

$$T_{\nu+1}^u = \left(1-\frac{1}{W}\right) T_\nu^u + (1-z_\nu^u)\frac{1}{W} c_{y_\nu^u}.$$

We define $\hat{\mathcal{T}}^u \triangleq \lim_{W\to\infty} T^u$ as the stationary throughput of user u and

$$\mathcal{H}^u = \sum_{y=0}^{2N_{\text{PHY}}^u-1} c_y[\,1-\overline{\text{PER}}_{\text{PHY}}(y)]\, P_{\text{PHY}}(y)$$

as the user u effective channel average rate. In this case, using the results presented by Holtzman in [23] and assuming that the fast fading components of all users in the system are identically distributed, it can be shown thatif the rate of user u is a function of its SNR $f(\gamma_\nu^u)$, the fixed point equation described in [23] has a unique solution when the throughputs are proportional to the average rate given by $\overline{f(\gamma_\nu)} = \mathcal{H}$. Thus, given users u and v, $\frac{\hat{\mathcal{T}}^u}{\hat{\mathcal{T}}^v} = \frac{\mathcal{H}^u}{\mathcal{H}^v}$, then the PF weight of user u in time slot ν can be defined as $F_\nu^u = \frac{c_{y_\nu^u}}{\mathcal{H}^u}$. The transition probability matrix can be constructed as in the max-rate example. Expressions (4), (6), and (7) can be rewritten by substituting the TRs $c_{y_\nu^u}$ with the corresponding PF weights F_ν^u. Now a and b will denote the number of users with the maximum PF weights during time slots ν and $\nu+1$, respectively.

5 Queueing model and analysis

Once the PHY layer and MAC sublayer have been properly modeled, the queuing behavior of the DLC layer has to be introduced in the analysis. Two different techniques are proposed.

5.1 Queuing Markov model-based approach

Following the work described in [9-12], the queueing process induced by both the ARQ protocol and the AMC scheme can be formulated in discrete time with one time unit equal to one frame interval. Each user's subsystem states are observed at the beginning of each time unit. Let $\sigma_\nu^u = (q_\nu^u, a_\nu^u, \boldsymbol{\iota}_\nu^u)$ denote the user u subsystem state at time instant $t=\nu T_f$, where $q_\nu^u \in \{0,\ldots,\overline{Q^u}\}$ denotes the queue length at this time instant, $a_\nu^u \in \{0,\ldots,\mathcal{A}^u-1\}$ represents the phase of the D-BMAP, and $\boldsymbol{\iota}_\nu^u \in \{0,\ldots,N_{\text{PHY-MAC}}^u-1\}$ represents the combination of PHY layer state and scheduling decision for user u during this frame interval. Focusing on the set of time instants $t=\nu T_f$, $\nu=0,1,\ldots,\infty$, the transitions between states σ_ν^u are Markovian. Therefore, an embedded Markov chain can be used to describe the underlying queueing process for each user u.

In previous work, we developed the embedded Markov chains describing the underlying queueing process for different AMC schemes, such as the ones described in 802.11 and 802.16 proposals, and different ARQ protocols, including infinitely persistent ARQ [9,10], hybrid ARQ [11], and truncated hybrid ARQ [12] schemes. Using the same technique described in Subsection 2.4, the MAC layer can be incorporated to the models described in these papers and the multiuser case could be also analyzed for those systems. In this paper, as an example and without loss of generality, we have used the model developed in [10] using the transmission mode pool of the IEEE 802.11a system combined with an infinitely persistent selective repeat (SR) ARQ procedure and it has been adapted to the multiuser case. The state space of the user u embedded finite state Markov chain is $\boldsymbol{\mathcal{S}}^u = \{\mathcal{S}_\mu^u\}_{\mu=1}^{N_s^u}$, where $N_s^u = (\overline{Q^u}+1)\mathcal{A}^u N_{\text{PHY-MAC}}^u - 1$ and $\mathcal{S}_\mu^u = (q_\mu^u, a_\mu^u, \boldsymbol{\iota}_\mu^u) \equiv (q_\mu^u, a_\mu^u, n_\mu^u)$.

Taking into account that infinitely persistent SR-ARQ is used at the DLC layer, and assuming that it is conditioned on the instantaneous channel fading, the transmission outcomes (success or failure) of consecutive packets in a frame interval are independent, and the probability that k packets of user u are successfully transmitted (leave the queue) given $c_{n_\mu^u}$ packets are transmitted when the PHY-MAC layer of user u is in state $\mathcal{S}_\mu^u$ can be written as

$$p_{k,c_{n_\mu^u}}^{(n_\mu^u)} = \binom{c_{n_\mu^u}}{k} \left(\overline{\text{PER}}_{n_\mu^u}^{\text{PHY-MAC}}\right)^{c_{n_\mu^u}-k} \left(1-\overline{\text{PER}}_{n_\mu^u}^{\text{PHY-MAC}}\right)^{k}. \quad (9)$$

Thus, the probability that h packets are successfully transmitted given that there are q packets in the queue before transmission and the PHY-MAC layer changes from state n_μ^u to state $n_{\mu'}^u$ can be expressed as

$$t_{h,q}^u\left(n_\mu^u, n_{\mu'}^u\right) = p_{h,\min\left\{c_{n_\mu^u},q\right\}}^{(n_\mu^u)} P_{n_\mu^u,n_{\mu'}^u}^u, \quad (10)$$

for $q \in \{0,\ldots,\overline{Q}^u\}$, $h \in \{0,\ldots,\min\{q,\mathcal{C}^u\}\}$, and n_μ^u, $n_{\mu'}^u \in \left\{0,\ldots,N_{\text{PHY-MAC}}^u - 1\right\}$. The $N_{\text{PHY-MAC}}^u \times N_{\text{PHY-MAC}}^u$ terms in (10) capturing all the cases where h packets are successfully transmitted given that there are q packets in the queue before transmission can be expressed in matrix form as $\boldsymbol{T}_{h,q}^u = \mathcal{D}\left(\boldsymbol{p}_{h,q}^u\right)\boldsymbol{P}_s^u$, for $q \in \{0,\ldots,\overline{Q}^u\}$ and $h \in \{0,\ldots,\min\{q,\mathcal{C}^u\}\}$, where $\boldsymbol{p}_{h,q}^u = \left(p_{h,\min\{\mathcal{C}_0^u,q\}}^{(0)} \cdots p_{h,\min\left\{\mathcal{C}_{N_{\text{PHY-MAC}}^u-1},q\right\}}^{(N_{\text{PHY-MAC}}^u-1)}\right)$ and $\mathcal{D}(\boldsymbol{x})$ is used to denote a diagonal matrix with the elements of the vector $\boldsymbol{x}$ in its main diagonal. Notice that for $q \geq \mathcal{C}_{N_{\text{PHY-MAC}}^u-1} = \mathcal{C}^u$, the probabilities in these matrices do not depend on q and $\boldsymbol{T}_{h,q}^u = \boldsymbol{T}_{h,\mathcal{C}^u}^u$.

Let $q_\mu = q$ and $q_{\mu'} = q + \mathcal{A}^u - l - 1$ be the number of packets in the queue of user u in two consecutive frames[c]. Also, let a and h be the number of arriving packets and the number of packets successfully transmitted in the first one of these frame intervals, respectively. In this case, $q_{\mu'} - q_\mu = \mathcal{A}^u - l - 1 = a - h$ or $h = l - \mathcal{A}^u + a + 1$. Thus, given that $0 \leq l - \mathcal{A}^u + a + 1 \leq \min\{q,\mathcal{C}^u\}$ or $\mathcal{A}^u - l - 1 \leq a \leq \mathcal{A}^u - l + \min\{q,\mathcal{C}^u\} - 1$, and $0 \leq a \leq \mathcal{A}^u - 1$, the probability that the queueing system changes from a generic state $\mathcal{S}_\mu^u = (q, a_\mu, n_\mu) \in \boldsymbol{\mathcal{S}^u}$ to another generic state $\mathcal{S}_{\mu'}^u = (q + \mathcal{A}^u - l - 1, a_{\mu'}, n_{\mu'}) \in \boldsymbol{\mathcal{S}^u}$ can be expressed as

$$A_{q,l}^u\left(\mathcal{S}_\mu^u, \mathcal{S}_{\mu'}^u\right) = \sum_{a=a_{\min}}^{a_{\max}} u_a^u(a_\mu, a_{\mu'})\, t_{l-\mathcal{A}+a+1,q}^u\left(n_\mu, n_{\mu'}\right), \quad (11)$$

where $a_{\min} = \max\{0, \mathcal{A}^u - l - 1\}$ and $a_{\max} = \min\{\mathcal{A}^u - 1, \mathcal{A}^u + \min\{q,\mathcal{C}^u\} - 1 - l\}$, for $q \in \{0,\ldots,\overline{Q}^u\}$, $l \in \{0,\ldots,\mathcal{A}^u + \min\{q,\mathcal{C}^u\} - 1\}$, $a_\mu, a_{\mu'} \in \{0,\ldots,\mathcal{A}^u - 1\}$, and $n_\mu, n_{\mu'} \in \left\{0,\ldots,N_{\text{PHY-MAC}}^u - 1\right\}$. The $\mathcal{A}^u N_{\text{PHY-MAC}}^u \times \mathcal{A}^u N_{\text{PHY-MAC}}^u$ terms capturing all the cases where the queue length changes from q packets in one frame interval to $q + \mathcal{A}^u - l - 1$ packets in the next frame interval can be expressed in matrix form as

$$\boldsymbol{A}_{q,l}^u = \begin{cases} \sum_{a=a_{\min}}^{a_{\max}} \boldsymbol{U}_a^u \otimes \boldsymbol{T}_{l-\mathcal{A}^u+a+1,q}^u, & l \in \{0,\ldots,l_{\max}^u(q)\} \\ \boldsymbol{0}, & \text{otherwise} \end{cases} \quad (12)$$

for $q \in \{0,\ldots,\overline{Q}^u\}$, where $\otimes$ denotes the Kronecker product and $l_{\max}^u(q) = \mathcal{A}^u + \min\{q,\mathcal{C}^u\} - 1$. The resulting transition matrix of the Markov chain can then be written as

$$\boldsymbol{P}^u = \begin{bmatrix} \boldsymbol{A}_{0,\mathcal{A}^u-1}^u & \cdots & \boldsymbol{A}_{0,\mathcal{A}^u-\overline{Q}^u}^u & \overline{\boldsymbol{A}}_{0,\mathcal{A}^u-\overline{Q}^u-1}^u \\ \vdots & & \vdots & \vdots \\ \boldsymbol{A}_{\overline{Q}^u,\mathcal{A}^u+\overline{Q}^u-1}^u & \cdots & \boldsymbol{A}_{\overline{Q}^u,\mathcal{A}^u}^u & \overline{\boldsymbol{A}}_{\overline{Q}^u,\mathcal{A}^u-1}^u \end{bmatrix}, \quad (13)$$

where $\overline{\boldsymbol{A}}_{q,i}^u = \sum_{a=0}^{i} \boldsymbol{A}_{q,a}^u$. Notice that for $q \geq \mathcal{C}^u$, the transition probabilities in these matrix blocks do not depend on q, and therefore, for simplicity this index can be omitted, that is, $\boldsymbol{A}_{q,l}^u = \boldsymbol{A}_l^u$ and $\overline{\boldsymbol{A}}_{q,l}^u = \overline{\boldsymbol{A}}_l^u$ for all $q \geq \mathcal{C}^u$.

To derive the system performance measures, we need to obtain the steady-state probability vectors corresponding to each level of the transition matrix, which can be calculated using the fact that the transition probability matrix $\boldsymbol{P}^u$ and steady-state probability vector $\boldsymbol{\pi}^u = \left[\boldsymbol{\pi}_0^u\ \boldsymbol{\pi}_1^u\ \cdots\ \boldsymbol{\pi}_{\overline{Q}^u}^u\right]$ satisfy $\boldsymbol{\pi}^u\boldsymbol{P}^u = \boldsymbol{\pi}^u$ along with the normalization condition $\sum_{i=0}^{\overline{Q}^u} \boldsymbol{\pi}_i^u \mathbf{1} = 1$, where $\mathbf{1}$ is a column vector of all ones with the appropriate length. To calculate $\boldsymbol{\pi}^u$, the method described by Le et al. in [5] is used to reduce the complexity in solving the matrix $\boldsymbol{P}^u$.

5.2 Performance measures

In our finite buffering ARQ-based error control system with infinite persistence, the PLR of user u, P_l^u (measured in packets per second), is simply equal to the buffer overflow probability. As in [5], we denote by $\boldsymbol{V}_k^u$ the stationary vector describing the probabilities that k packets are lost due to buffer overflow upon arrival of a burst of data packets. Assuming that a batch of a packets arrive at the link layer buffer, if there are $q > \overline{Q}^u - a$ packets in the queue at the end of the previous frame interval and h packets are successfully transmitted, then the number of packets that will be lost due to buffer overflow is $k = a - h - \overline{Q}^u + q$. Therefore, $\boldsymbol{V}_k^u$ can be written as

$$\boldsymbol{V}_k^u = \sum_{a=1}^{\mathcal{A}^u-1} \sum_{q=\max\{0,\overline{Q}^u-a+1\}}^{\overline{Q}^u} \boldsymbol{\pi}_q^u \sum_{\substack{h=0 \\ a-h=\overline{Q}^u-q+k}}^{\mathcal{C}^u} \boldsymbol{U}_a^u \otimes \boldsymbol{T}_{h,q}^u. \quad (14)$$

The PLR of user u can then be calculated as the ratio between the average number of lost packets due to buffer overflow N_l^u and the average number of arriving packets λ^u in one frame interval, that is,

$$P_l^u = N_l^u/\lambda^u = (1/\lambda) \sum_{k=1}^{\mathcal{A}^u-1} k\boldsymbol{V}_k^u\mathbf{1}. \quad (15)$$

Given the PLR, the average throughput (measured in packets per frame) can be calculated as

$$\eta^u = \lambda^u(1 - P_l^u). \tag{16}$$

Then, using the well-known Little's formula [24], the average delay can be calculated as

$$D_l^u = L_q^u/\lambda(1 - P_l^u) = L_q^u/\eta^u, \tag{17}$$

where L_q^u denotes the average number of packets in queue of user u, which can be obtained as $L_q^u = \sum_{q=1}^{\overline{Q}^u} q\,\boldsymbol{\pi}_q^u\mathbf{1}$.

5.3 Effective bandwidth/capacity-based approach

The DLC layer can also be modeled by applying the effective bandwidth/capacity-based approach [16]. The effective capacity and effective bandwidth allow the analysis of the so-called PLR bound probability. The analysis is analogous to the one developed in [10]. The effective bandwidth of the D-BMAP arrival process of user u, characterized by a transition matrix $\boldsymbol{U^u}$, can be calculated as [25] $E_B^u(\psi) = \psi^{-1}\log(\Upsilon_U^u(\psi))$, where Υ_U^u is the Perron-Frobenius eigenvalue of the matrix $\boldsymbol{U}_{\boldsymbol{\varpi}}^u = \mathcal{D}(\boldsymbol{\varpi}^u)\boldsymbol{U}^u$, with $\boldsymbol{\varpi} \triangleq \left(e^{\lambda_0^u\psi}, \ldots, e^{\lambda_{\mathcal{A}^u-1}^u\psi}\right)$, where λ_n^u denotes the number of packets per frame generated when the source of user u is in state n. The effective capacity of the service process that models the behavior of the MAC and PHY layers for the user of interest, which is characterized by a transition probability matrix $\boldsymbol{P}_s^u$, can be obtained as $E_C^u(\psi) = -\psi^{-1}\log(\Upsilon_P^u(-\psi))$. In this expression, Υ_P^u denotes the Perron-Frobenius eigenvalue of the matrix $\boldsymbol{P}_{\upsilon_v}^u = \mathcal{D}(\boldsymbol{\upsilon}_v^u)\boldsymbol{P}_s^u$, with $\boldsymbol{\upsilon}_v^u \triangleq \left(e^{-\check{c}_0^u\psi}, \ldots, e^{-\check{c}_{(4N_{\mathrm{PHY}}^u-1)}^u\psi}\right)$, where $\check{c}_n^u$ denotes the number of packets per frame leaving the queue when the PHY-MAC for user u is in state n, which, for an SR-ARQ infinitely persistent scheme, can be calculated as $\check{c}_{n^u} = \sum_{k=0}^{\widehat{c}_{n^u}} k p_{k,\widehat{c}_{n^u}}^{(n^u)}$ with $p_{k,\widehat{c}_{n^u}}^{(n^u)}$ defined in (9).

The effective bandwidth/capacity-based approach can only provide statistical QoS guarantees. For instance, the target link layer PLR $P_{l\max}$ can only be guaranteed with a small violation probability ϵ, that is, $\Pr\{P_l^u \leq P_{l\max}\} \approx \kappa^u e^{-\psi_u^* P_{l\max}} \leq \epsilon$, where ψ_u^* is the unique real solution of $E_B^u(\psi) - E_C^u(\psi) = 0$ and κ^u is the relation between average arrival rate $\varpi_A^u \triangleq \lim_{\psi\to 0} E_B^u(\psi)$ and average service rate $\varpi_S^u \triangleq \lim_{\psi\to 0} E_C^u(\psi)$, as shown by Tang and Zhang in [26]. It is worth stating at this point that except for low input data rates, the tail probability tends to overestimate the packet loss probability and it can only be used as

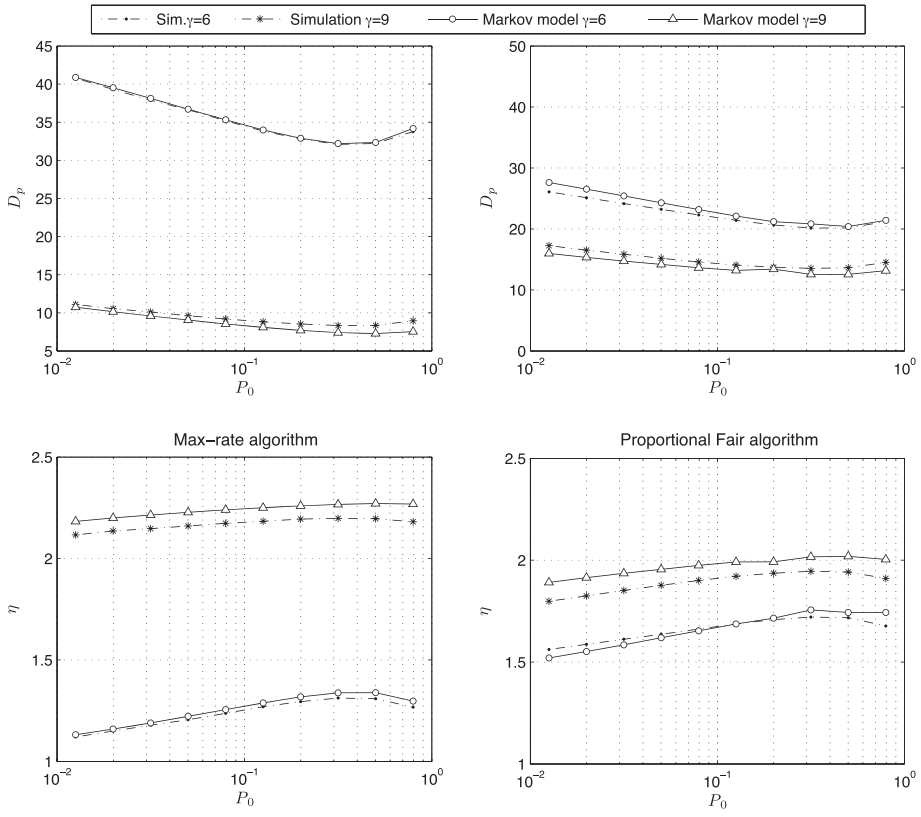

Figure 2 Average delay (D_p) and throughput (η) of the max-rate (left) and PF algorithms (right) vs. target PER.

long as $\varpi_A \leqslant \varpi_S$. Then, the throughput of user u can be calculated as

$$\eta^u = \lambda^u(1 - P_l^u) \approx \lambda^u(1 - \kappa^u e^{-\psi_u^* \overline{Q}^u}).$$

6 Cross-layer design

Given the sets of maximum allowable queue lengths $\overline{\boldsymbol{Q}} = \{\overline{Q}^1, \ldots, \overline{Q}^{N_s}\}$, average SNRs $\bar{\boldsymbol{\gamma}} = \{\bar{\gamma}^1, \ldots, \bar{\gamma}^{N_s}\}$, normalized maximum Doppler frequencies $\boldsymbol{f_d}T_f$ with $\boldsymbol{f_d} = \{f_d^1, \ldots, f_d^{N_s}\}$, and assuming the use of the MR algorithm in the MAC layer, the derived PHY-MAC-DLC model basically depends on both the set of prescribed average PERs $\boldsymbol{P_0} = \{P_0^1, \ldots, P_0^{N_s}\}$, with P_0^u being a real number in the range $\boldsymbol{\Phi} = [0, 1]$, and the set of measured or estimated arrival packet rates $\lambda^u \in \boldsymbol{\Theta}$, where $\boldsymbol{\Theta}$ is the range of feasible arrival rate values controlled by the traffic shaping mechanism. Thus, if the users in the system are to support QoS-guaranteed traffic characterized by a maximum PLR $P_{l\max}$ and a maximum average packet delay $D_{l\max}$, the proposed cross-layer design must aim at determining the prescribed average PER vector $\boldsymbol{P_0}$ and average packet arrival rate vector $\boldsymbol{\lambda} = \{\lambda^1, \ldots, \lambda^{N_s}\}$ solving the constrained optimization problem given by

$$\begin{aligned} \left(\boldsymbol{P_0}^{\text{opt}}, \boldsymbol{\lambda}^{\text{opt}}\right) &= \arg \max_{\boldsymbol{P_0} \in \boldsymbol{\Phi}, \boldsymbol{\lambda} \in \hat{\boldsymbol{\Theta}}} \sum_{u=1}^{N_s} \eta^u(\boldsymbol{P_0}, \boldsymbol{\lambda}) \\ \text{subject to } & P_l^u(\boldsymbol{P_0}, \boldsymbol{\lambda}) \leq P_{l\max}, \quad \forall u \\ & D_l^u(\boldsymbol{P_0}, \boldsymbol{\lambda}) \leq D_{l\max}, \quad \forall u. \end{aligned} \tag{18}$$

A similar cross-layer design can be derived for the PF algorithm, but taking into account that this algorithm maximizes the logarithmic sum of throughput, the optimization function must be designed accordingly as

$$\left(\boldsymbol{P_0}^{\text{opt}}, \boldsymbol{\lambda}^{\text{opt}}\right) = \arg \max_{\boldsymbol{P_0} \in \boldsymbol{\Phi}, \boldsymbol{\lambda} \in \hat{\boldsymbol{\Theta}}} \sum_{u=1}^{N_s} \log(\eta^u(\boldsymbol{P_0}, \boldsymbol{\lambda})). \tag{19}$$

In both cases, the analytical expressions for η^u, P_l^u, and D_l^u do not leave much room for developing efficient algorithms in solving our constrained optimization problem. However, considering that $\boldsymbol{P_0}$ and $\boldsymbol{\lambda}$ lie in a bounded space $\boldsymbol{\Phi}^u \times \boldsymbol{\Theta}^u$, a multidimensional exhaustive search can be used to numerically solve the proposed cross-layer optimization problem.

7 Numerical results

In order to verify the validity of the proposed cross-layer framework, analytical results will be confronted with computer simulation results obtained using Clarke's statistical Rayleigh fading model of the wireless flat-fading channel [27]. Unless otherwise specified, numerical results are presented for the following default parameters: a normalized maximum Doppler frequency $f_d T_f = 0.02$, a queue length $\overline{Q} = 50$, a number of channel states $K = 10$, a parameter $b = 2$, and a D-BMAP parameterized to obtain a truncated Poisson process with a variable average arrival rate λ. These parameters apply to all users in the system.

Figure 2 shows the dependence of the average delay D_l^u and throughput η^u on the target average PER P_0^u of the two users in the system. In this figure, P_0^1 and P_0^2 have been set to a common value P_0 in order to show the analytical and simulation results of both users simultaneously. As it can be observed, in all cases, the behavior of the simulation of the system under consideration with different scheduling algorithms, namely the MR algorithm (left) and the PF algorithm (right), is faithfully reproduced by our analytical PHY-MAC-DLC layer model. In particular, it is interesting to note how the shape and location of the minimum/maximum of the curves obtained by simulation (Clarke's model) coincide with those obtained using the analytical framework, even for a small number of channel states K. The accuracy in determining the location of the maximum of the throughput and the minimum of the average packet loss rate or the average packet delay is particularly important in order to ensure an optimal cross-layer design. Regarding the scheduling performance, it can be observed in Figure 2 that, as expected, PF attains higher fairness at the expense of a global throughput loss.

Figure 3 shows the system sum throughput as a function of the number of active users in the system when using either the MR or the PF scheduling algorithm. Results presented in this figure have been obtained by placing N_s users on the coverage area at distances $R/(N_s + 1), 2R/(N_s + 1), \ldots, N_sR/(N_s + 1)$ from the BS, where R denotes the cell radius. The traffic arrival for each active user in the system has been modeled as a D-BMAP parameterized to obtain a truncated Poisson process with an average arrival rate $\lambda \simeq 3$ packets/frame. Furthermore, the target PER has been set to $P_0^u = 10^{-0.4}$, for all $u \in \{1, \ldots, N_s\}$. As it can be observed, the sum throughput increases with the number of active users in the system.

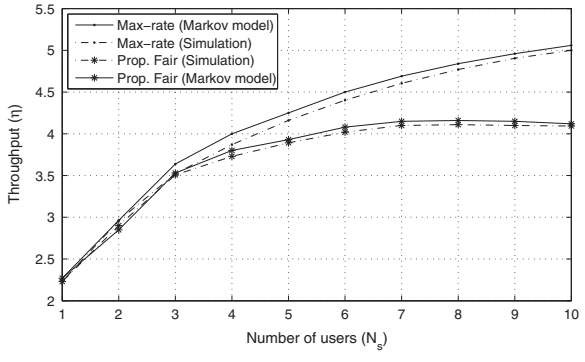

Figure 3 Sum throughput vs. number of users using max-rate and PF algorithms.

However, it can also be seen that the sum throughput gain is subject to the diminishing capacity returns as the number of users increases. Furthermore, as expected, Figure 3 shows that the MR algorithm achieves a higher sum throughput in comparison with the PF strategy, at the cost of unfair treatment of the arriving traffic flows.

Figure 4a,b,c shows the system sum throughput and per-user throughput when applying the cross-layer optimi-

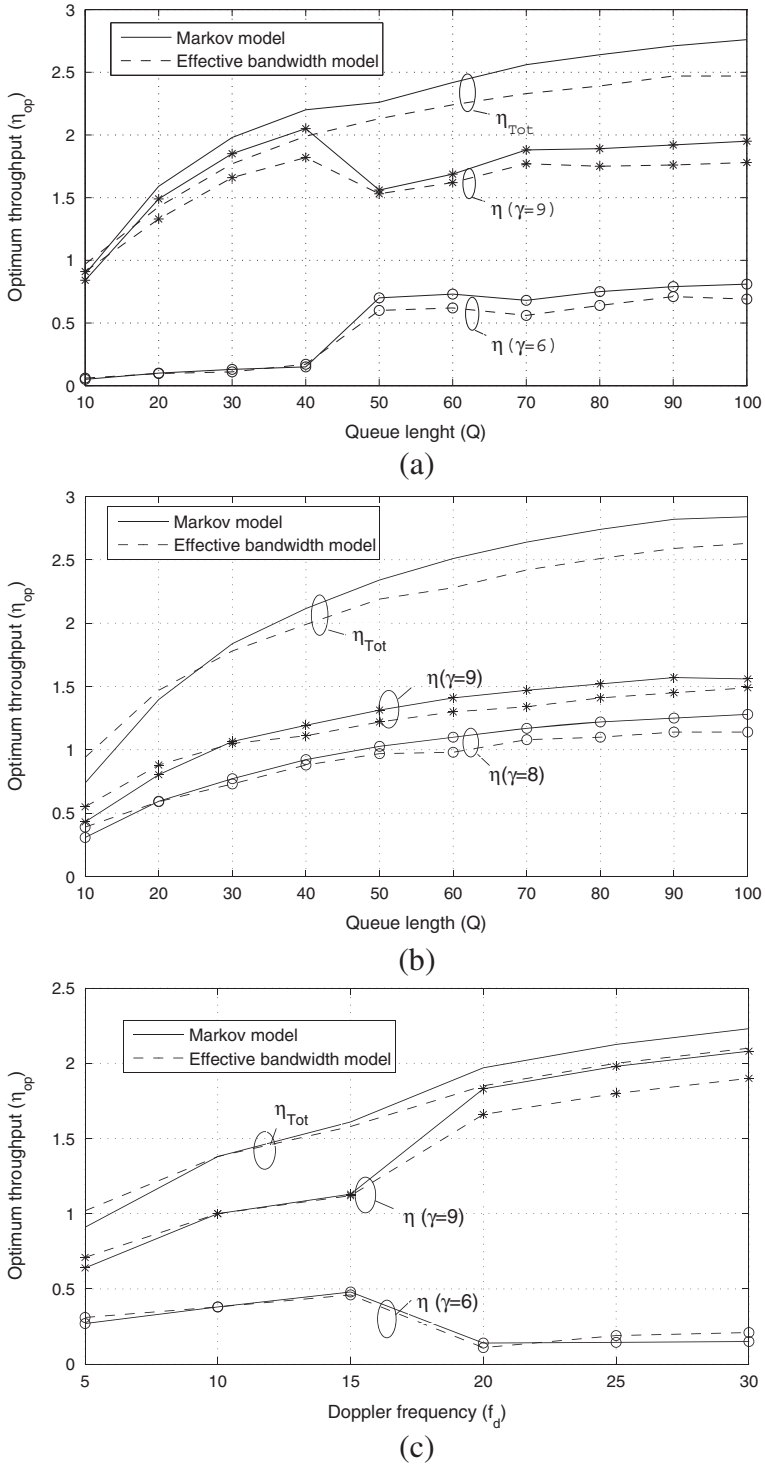

Figure 4 Max-rate algorithm example.

zation defined in (18) for the MR algorithm. These figures have been obtained applying a maximum affordable PLR $P_{l_{\max}} \leq 0.01$ and using either a Markov model or the effective bandwidth/capacity-based approach to model the DLC buffer behavior as described in Section 5. As expected, the effective bandwith approximation tends to overestimate the PLR, thus predicting a lower throughput than the Markov model except for low input data rates. The optimization process further increases the aggregate throughput of the MR policy while maintaining a desired level of QoS in the form of a maximum PLR. This is accomplished by tuning the PHY layers of the users through the P_0^u parameters and shaping the users' average arrival rate λ^u. Figure 4a reveals that for short queue lengths ($\overline{Q} < 50$), the higher sum throughput is obtained by assigning a very low P_0 to the user with a lower average SNR ($\gamma = 6$ dB), which results in a very low throughput for that user. The same behavior is observed in Figure 4c for high Doppler frequencies ($f_d > 20$ Hz). When the queue length increases or the Doppler frequency decreases, the system achieves higher capacity by assigning similar P_0 values to both users and the throughput of the lower SNR user increases accordingly. Logically, when both users are subject to similar average SNR values, as shown in Figure 4b, the maximum sum throughput is always achieved by assigning similar P_0 values.

Figure 5a,b shows the aggregated and per-user throughputs obtained when applying the cross-layer optimization defined by (19) for the PF algorithm. Results presented in Figure 5a have been derived using a maximum packet loss constraint $P_{l_{\max}} \leq 0.01$ and a maximum average delay constraint $D_{l_{\max}} \leq 10$. As expected, for low values of the queue length, $\overline{Q} < 40$, the constraint limiting the throughput is the packet loss, which is mainly caused by the buffer overflow, and therefore, the throughput increases with $\overline{Q}$. For higher queue lengths, $\overline{Q} > 50$, the limiting constraint is the maximum average delay, and in this case, additional increases in the queue length have a negligible effect

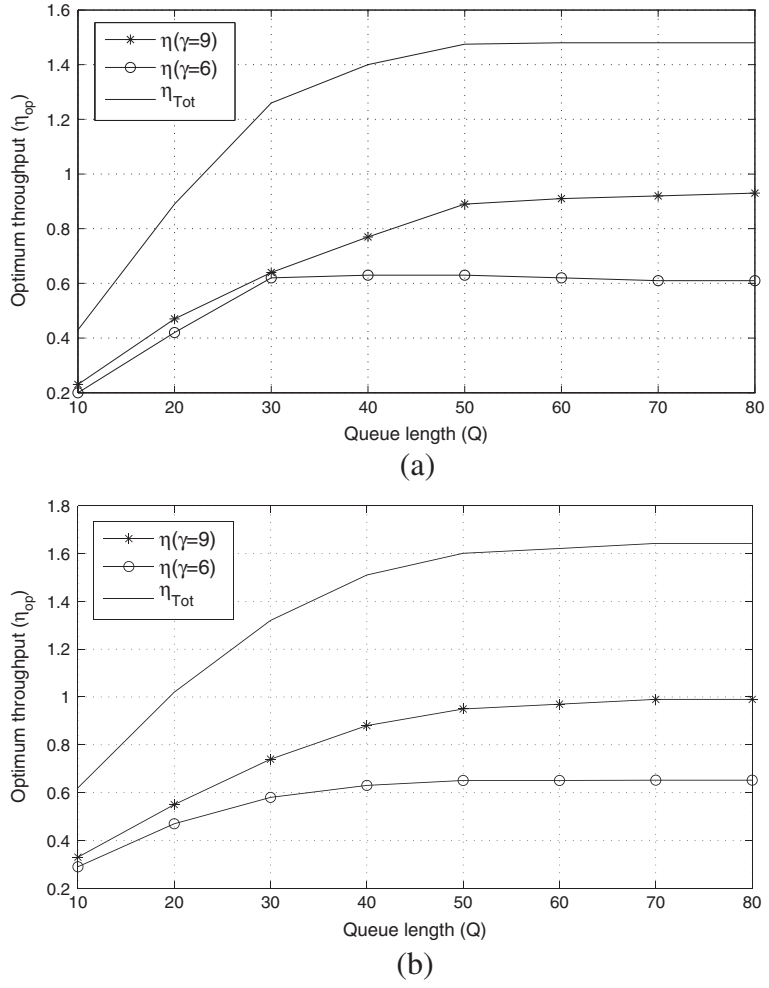

Figure 5 Proportional fair example.

over the throughput. Figure 5b depicts results obtained when using an optimization performed using the effective bandwidth model formulated in (19). In this case, the constraint in (19) has been modified to $\Pr\{D_l^u(\boldsymbol{P_0}, \boldsymbol{\lambda}) \geq D_{l_{\max}}\} \leq \epsilon \quad \forall u$, as the effective bandwidth theory only offers statistical QoS guarantees. The specific values of the constraint used to generate this figure are $\Pr\{D_l^u \geq 50\} \leq 0.01 \quad \forall u$. Results presented in Figure 5b show a similar behavior as those in Figure 5a. For a queue length below $\overline{Q} = 50$, the limiting constraint is the PLR, and therefore, the throughput increases with $\overline{Q}$. In contrast, for $\overline{Q} > 50$, the active constraint is the maximum affordable delay, which does not depend on the queue length, causing the throughput to remain nearly constant with respect to the queue length.

8 Conclusions

This paper extends the analytical framework presented in [10] to incorporate the MAC sublayer in the proposed analytical model that now includes a multiple-user-shared channel scenario. Channel-aware-only schedulers have been embedded in a joint PHY-MAC Markov model by using a service-vacation process to model the scheduling decisions. Two widely used scheduling rules have been considered, the MR and PF algorithms. As in [10], two different approaches have been used to model the DLC level queueing behavior: an analytical Markov model and an approach based on the effective bandwidth theory. Results show that the use of the effective bandwidth approach significantly simplifies the global model and is therefore an interesting technique to use by the resource allocation algorithms. Numerical examples confirm that the derived performance metrics obtained with the PHY-MAC-DLC analytical model faithfully reproduce simulation results. It is important to point out that the multiple user model obtained in this paper can be easily adapted to include truncated or hybrid ARQ techniques in the DLC layer as it was proposed in [11,12]. Finally, a cross-layer design interrelating the PHY, MAC, and DLC layers has been described. The obtained results show the potential of cross-layer resource allocation designs where slot-by-slot decisions are left to simple and efficient channel-aware schedulers, such as the MR or PF algorithm, while QoS control is performed at a higher level by well-selected optimization functions. These optimization functions enhance and complement the scheduling algorithms while maintaining an adequate QoS performance by modifying average parameters of the different layers in the system. The proposed cross-layer approach fits in the radio resource management (RRM) framework proposed for state-of-the-art networks such as LTE that define a division between fast dynamic layer 1 and layer 2 RRM functions working at the transmission time interval level and semi-dynamic layer 3 RRM procedures.

Endnotes

[a]It is assumed in this paper that the frame duration is smaller than the coherence time of the channel.

[b]If $\gamma_\nu^u < \gamma_{\nu-1}^u$, then the instantaneous SNR is descending and it can be tagged as $down(\delta_\nu^u = 1)$; on the contrary, if $\gamma_\nu^u \geq \gamma_{\nu-1}^u$, then the instantaneous SNR is ascending and it can be tagged as $up(\delta_\nu^u = 0)$.

[c]The maximum number of packet arrivals in one frame interval is equal to $\mathcal{A}^u - 1$, and the maximum number of successfully transmitted packets in one frame interval is equal to $\min\{q, \mathcal{C}^u\}$; thus, it is quite obvious that $q_\mu - \min\{q, \mathcal{C}^u\} \leq q_{\mu'} \leq q_\mu + \mathcal{A}^u - 1$ or $0 \leq l \leq \mathcal{A}^u + \min\{q, \mathcal{C}^u\} - 1$.

Competing interests
The authors declare that they have no competing interests.

Acknowledgments
This work has been partially funded by MEC and FEDER through project COSMOS (TEC2008-02422) and project AM3DIO (TEC2011-25446).

Received: 28 January 2013 Accepted: 30 July 2013
Published: 23 August 2013

References

1. Q Liu, S Zhou, GB Giannakis, Cross-layer combining of adaptive modulation and coding with truncated ARQ over wireless links. IEEE Trans. Wireless Commun. **3**(5), 1746—1755 (2004)
2. Q Liu, S Zhou, GB Giannakis, Queuing with adaptive modulation and coding over wireless links: cross-layer analysis and design. IEEE Trans. Wireless Commun. **4**(3), 1142–1153 (2005)
3. Q Liu, S Zhou, GB Giannakis, Cross-layer scheduling with prescribed QoS guarantees in adaptive wireless networks. IEEE J. Selected Areas Commun. **23**(5), 1056–1066 (2005)
4. LB Le, E Hossain, AS Alfa, Service differentiation in multirate wireless networks with weighted round-robin scheduling and ARQ-based error control. IEEE Trans. Commun. **54**(2), 208–215 (2006)
5. LB Le, E Hossain, AS Alfa, Radio link level performance evaluation in wireless networks using multi-rate transmission with ARQ-based error control. IEEE Trans. Wireless Commun. **5**(10), 2647–2653 (2006)
6. F Ishizaki, GU Hwang, Cross-layer design and analysis of wireless networks using the effective bandwidth function. IEEE Trans. Wireless Commun. **6**(9), 3214–3219 (2007)
7. M Poggioni, L Rugini, P Banelli, in *Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM)*. Analyzing performance of multi-user scheduling jointly with AMC and ARQ (Washington, D.C., 26), pp. 3483–3488
8. X Wang, Q Liu, GB Giannakis, Analyzing and optimizing adaptive modulation coding jointly with ARQ for QoS-guaranteed traffic. IEEE Trans. Veh. Technol. **56**(2), 710–720 (2007)
9. J Ramis, L Carrasco, G Femenias, in *Proceedings of the IEEE GLOBECOM*. A two-dimensional Markov model for cross-layer design in AMC/ARQ-based wireless networks (New Orleans, 30 Nov–4 Dec 2008), pp. 4637–4642
10. G Femenias, L Carrasco, J Ramis, Using two-dimensional Markov models and the effective-capacity approach for cross-layer design in AMC/ARQ-based wireless networks. IEEE Trans. Veh. Technol. **58**(8), 4193–4203 (2009)
11. J Ramis, G Femenias, F Riera-Palou, L Carrasco, in *Proceedings of the IEEE GLOBECOM*. Cross-layer optimization of adaptive multi-rate wireless networks using truncated chase combining HARQ (Miami, 6–10 Dec 2010)
12. J Ramis, G Femenias, Cross-layer design of adaptive multirate wireless networks using truncated HARQ. IEEE Trans. Veh. Technol. **60**(3), 944–954 (2011)
13. M Poggioni, L Rugini, P Banelli, QoS analysis of a scheduling policy for heterogeneous users employing AMC jointly with ARQ. IEEE Trans. Commun. **58**, 9 (2010)

14. LB Le, E Hossain, AS Alfa, Delay statistics and throughput performance for multi-rate wireless networks under multiuser diversity. IEEE Trans. Wireless Commun. **5**(11), 3234–3243 (2006)
15. J Liu, W Chen, Z Cao, YJ Zhang, Delay optimal scheduling for cognitive radios with cooperative beamforming: a structured matrix-geometric method. IEEE Trans. Mobile Comput. **11**(8), 1412–1423 (2012)
16. D Wu, R Negi, Effective capacity: a wireless link model for support of quality of service. IEEE Trans. Wireless Commun. **2**(4), 630–643 (2003)
17. E Biglieri, G Caire, G Taricco, Limiting performance of block fading channels with multiple antennas. IEEE Trans. Inf. Theory. **47**(4), 1273–1289 (2001)
18. C Blondia, A discrete time batch Markovian arrival process as B-ISDN traffic model. Belgian J. Oper. Res. Stat. Comput. Sci. **32**(3/4), 3–23 (1993)
19. IEEE, *802.11: Standard for Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*. (IEEE, New York, 1997)
20. J Max, Quantization for minimum distortion. IRE Trans. Inf. Theory. **IT-6**, 7–12 (1960)
21. SP Lloyd, Least squares quantization in PCM. IEEE Trans. Inf. Theory. **IT-28**, 129–137 (1982)
22. KKH Kim, Y Han, in *Proceedings of the IEEE PIMRC*. An efficient scheduling algorithm for QoS provision in wireless packet data transmission (Lisboa, 15–18 Sept 2002), pp. 73–77
23. J Holtzman, in *Proceedings of the IEEE PIMRC*. Asymptotic analysis of the proportional fair algorithm (San Diego, 30 Sept–3 Oct 2001), pp. F33—F37
24. L Kleinrock, *Queuing Systems*, vol. 1. (Wiley, New York, 1975)
25. C Chang (ed.), *Performance Guarantees in Communication Networks*, 1st edn. (Springer, Berlin, 2000)
26. J Tang, X Zhang, Cross-layer modeling for quality of service guarantees over wireless links. IEEE Trans. Wireless Commun. **6**(12), 4504–4512 (2007)
27. RH Clarke, A statistical theory of mobile radio reception. Bell Syst. Tech. J. **47**(6), 957–1000 (1968)

doi:10.1186/1687-1499-2013-213
Cite this article as: Carrasco *et al.*: **Channel-aware MAC performance of AMC-ARQ-based wireless systems.** *EURASIP Journal on Wireless Communications and Networking* 2013 **2013**:213.