Service Differentiation in Multirate Wireless Networks With Weighted Round-Robin Scheduling and ARQ-Based Error Control

Long B. Le, *Student Member, IEEE*, Ekram Hossain, *Member, IEEE*, and Attahiru S. Alfa, *Member, IEEE*

***Abstract*—The radio link-level delay statistics in a wireless network using adaptive modulation and coding (AMC), weighted round-robin (WRR) scheduling, and automatic repeat request-based error control is analyzed in this letter. WRR scheduling can be used for service differentiation similar to that achievable by using the generalized processor sharing scheduling discipline. The analytical framework presented in this letter captures physical and radio link-level aspects of a multirate multiuser wireless network (e.g., general fading model, AMC, scheduling, error control) in a unified way. It can be used for admission control and cross-layer design under statistical delay constraints. The analytical results are validated by simulations. Typical numerical results are presented, and their useful implications on the system performance are discussed.**

***Index Terms*—Adaptive modulation and coding (AMC), automatic repeat request (ARQ) protocol, finite-state Markov channel (FSMC), service differentiation, weighted round-robin (WRR) scheduling.**

I. Introduction

INTEGRATION of data communications services into wireless networks has received increasing interest from the research community in recent years. An efficient analytical tool for performance investigation at the link layer plays a key role in engineering the wireless system and predicting the higher layer protocol performance (e.g., the transmission control protocol (TCP) performance). Analysis for link-layer automatic repeat request (ARQ) protocols was done extensively in the past ([1], [2], and references therein). Most of the previous works in the literature, however, assumed either an independent or two-state Markov channel, which allows only a single packet to be transmitted over the channel in one time slot. In fact, most (if not all) 2.5/3G wireless networks employ multirate transmission by using adaptive modulation and coding (AMC) [3], [4]. Furthermore, the existing analysis for ARQ protocols is mainly for a single-user scenario. Developing a cross-layer (physical and radio link) analytical framework for multirate transmission in a multiuser environment is still an open and interesting research problem.

Implementation of differentiated services with guaranteed quality of service (QoS) in wireless networks has been another research focus recently. The generalized processor sharing (GPS) scheduling discipline [7] (also known as the weighted fair queueing) has been widely studied as an efficient way to implement differentiated services in a multiuser environment. It can guarantee the throughput share proportional to the assigned weight of each user. Unfortunately, the exact delay statistics for this scheduling cannot be found, and only deterministic or statistical delay bounds can be derived [7], [8]. These delay bounds may not be very tight, which may lead to low resource use. It was shown in [7] that compared with the GPS scheme, the weighted round-robin (WRR) scheduling is simpler to implement. However, the performances of these two scheduling schemes are very close to each other.

In this letter, we present a cross-layer (physical and radio link) analytical framework for radio link-level performance evaluation. In the physical layer, AMC is employed where the number of transmitted packets in one time slot varies depending on the channel condition. The channel is modeled by a finite-state Markov channel (FSMC) model [5], [6]. An ARQ protocol is employed in the link layer to counteract the residual error of an error-correction code in the physical layer. The exact queue length and delay distributions are then derived analytically. To show the usefulness of the presented model, admission control for delay-constrained applications and cross-layer design examples are illustrated.

The rest of this letter is organized as follows. System model and assumptions are described in Section II. In Section III, the queueing problem is formulated and solved to obtain the desired performance metrics. The numerical and simulation results are presented in Section IV. Conclusions are stated in Section V.

II. System Model and Assumptions

Suppose that there are μ separate radio link-level queues at the base station (BS) which correspond to μ different mobile users. One common channel for downlink transmission is shared by all users in a time-division multiplexing (TDM) fashion. The transmission time is slotted, and a WRR scheduler is used to schedule transmissions corresponding to the different mobiles. For the sake of simplicity, we assume that there are only two classes of users: high priority (class one) and low priority (class two). High-priority and low-priority users receive two and one service slots in one cycle, respectively. One cycle is defined to be the smallest interval with time slot assignments for all μ users, and it repeats periodically.

The receiver decodes the received packets and sends negative acknowledgments (NACKs) to the transmitter, asking for retransmission of the erroneous packets (if any). In this letter, an error-free and instantaneous feedback channel is assumed, so that the transmitter knows exactly if there is any transmission error at the end of each service time slot. This assumption holds

Paper approved by R. Fantacci, the Editor for Wireless Networks and Systems of the IEEE Communications Society. Manuscript received October 6, 2004; revised May 10, 2005; July 21, 2005; and July 26, 2005. The work of the second author was supported under a grant from the Natural Sciences and Engineering Research Council (NSERC) of Canada. This paper was presented in part at the IEEE Vehicular Technology Conference, Dallas, TX, September, 2005.

The authors are with the Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, MB R3T 5V6, Canada (e-mail: long@ee.umanitoba.ca; ekram@ee.umanitoba.ca; alfa@ee.umanitoba.ca).

Digital Object Identifier 10.1109/TCOMM.2005.863788

in many cases, because the propagation delay and the processing time for the error-detection code can be very small in comparison with the time-slot interval. In fact, the effect of feedback errors can be easily included in the channel model, as in [1]. The delay obtained in this letter, therefore, can be regarded as the lower bound of the delay obtained with these effects.

AMC is employed in the physical layer with K transmission modes corresponding to different channel states. We assume that, when the channel is in state k, the transmitter transmits c_k packets. We further assume that $c_0 = 0$ (i.e., the transmitter does not transmit in channel state zero to avoid the high probability of transmission errors) and $c_K = N$. The channel-state information (CSI) is fed back to the transmitter to choose the suitable transmission mode. The feedback channel, therefore, carries both CSI for AMC and ACK/NACK for the ARQ protocol.

To capture the variations of the multistate Nakagami fading channel, we employ the FSMC model [6]. The received signal-to-noise ratio (SNR) X is partitioned into $K+1$ intervals, each of which corresponds to a particular channel state. Each channel state corresponds to a unique transmission mode of AMC in the physical layer. The average packet-error rate (PER) for mode k ($\overline{\text{PER}}_k$) can be calculated as in [4].

Let $\mathbf{T}_{i,j}$ $(0 \le i,j \le K)$ denote the transition probability from state i to state j of the channel transition matrix $\mathbf{T}$, which has order $(K+1)\times(K+1)$. If the thresholds of the received SNR are determined, $\mathbf{T}_{i,j}$ can be calculated as in [4], where transitions are only allowed among the neighboring states [4]–[6]. Specifically, the channel transition matrix $\mathbf{T}$ has the following form:

$$\mathbf{T} = \begin{bmatrix} \mathbf{T}_{0,0} & \mathbf{T}_{0,1} & \cdots & \cdots & \mathbf{0} \\ \mathbf{T}_{1,0} & \mathbf{T}_{1,1} & \mathbf{T}_{1,2} & \cdots & \mathbf{0} \\ \mathbf{0} & \ddots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{T}_{K-1,K-2} & \mathbf{T}_{K-1,K-1} & \mathbf{T}_{K-1,K} \\ \mathbf{0} & \cdots & \cdots & \mathbf{T}_{K,K-1} & \mathbf{T}_{K,K} \end{bmatrix}. \tag{1}$$

III. Formulation of the Queueing Model

A. Queueing Model and Analysis

The queueing analysis for a target queue can be performed by using a vacation queueing model. While a particular target user is served in his assigned slots, the queue is assumed to be in service; otherwise, it is said to be on vacation. Since the queueing performances for mobile users of each class are statistically the same, we focus on one user from each class only. In the following, we analyze the queueing performance for each user class separately. For convenience, let the service period start from slot one of each cycle for each user class. Assuming that there are μ_1 high-priority users and $\mu_2 = \mu - \mu_1$ low-priority users, one cycle consists of $L = 2\mu_1 + \mu_2$ time slots.

The queueing problem for both classes of users can be modeled in discrete time with one time interval equal to one time slot. Packet arrival is described by a batch Markovian arrival process (BMAP), which is represented by $M+1$ substochastic matrices $\mathbf{U}_m$ $(m = 0,1,2,\ldots,M)$, each of which has order $M_1 \times M_1$ [9]. The elements $\mathbf{U}_m(i,j)$ represent the transition from phase i to phase j with m arriving packets. With this traffic modeling, there are at most M arriving packets in one time slot, and the correlation in the traffic arrival process is captured by the M_1 arrival phases. The buffer is assumed to be finite with a size of Q packets. We assume that packets arriving during time interval $n-1$ cannot be served until time interval n at the earliest. Furthermore, packet transmissions in a time slot are assumed to finish before arriving packets enter the queue. Any arriving packet which sees the full buffer will be lost.

The discrete-time Markov chain (MC) describing the system has state space

$$\{(x_n, a_n, u_n, s_n),\ x_n \ge 0,\ 1 \le a_n \le M_1,\ 1 \le u_n \le L, 0 \le s_n \le K\}$$

where x_n is the number of packets in the queue, a_n is the arrival phase, u_n is the slot in a particular cycle, and s_n is the channel state, all at time n. The number of packets transmitted in the service time slot n is $\min\{x_n, c_{s_n}\}$.

Let $\mathbf{P}$ and (i,j,h,k) denote the transition matrix and a generic system state for this MC, respectively, and let $(i,j,h,k) \to (i',j',h',k')$ denote the transition of this MC from state (i,j,h,k) to state (i',j',h',k'). For fixed i and i', the probabilities corresponding to these state transitions can be written in matrix blocks $\mathbf{A}_{i,k}$, which correspond to transitions in level i of the transition matrix. Thus, level i of the transition matrix represents the system-state transitions where there are i packets in the queue before the transitions.

The transition matrix describing the MC is written in (2) for $M = 3$ and $N = 4$, and is shown at the bottom of the next page. In (2), the system-state transitions $(i,*,*,*) \to (i-k+M,*,*,*)$ are represented by $\mathbf{A}_{i,k}$ for $i < N$ and by $\mathbf{A}_k$ for $i \ge N$. These matrix blocks capture the transitions among the arrival phases, the slot number in a cycle, and the channel states for the target queue. Note that there are at most M arriving packets, and at most N packets can be successfully transmitted in one time slot. Therefore, the transitions can go up by at most M levels, and go down by at most N levels.

As will be seen later, for level $i \ge N$, the state-transition probabilities are independent of level index i. Therefore, for brevity, we omit the level index in the matrix blocks. Since there are M_1 arrival phases, $K+1$ channel states, and a cycle consists of L slots, the order of the matrix blocks $\mathbf{A}_k$ and $\mathbf{A}_{i,k}$ is $N_1 \times N_1$, where $N_1 = M_1 L(K+1)$. The steady-state probability vector $\mathbf{x} = [\mathbf{x}_0\ \mathbf{x}_1\ \mathbf{x}_2\ \cdots \mathbf{x}_Q]$, where $\mathbf{x}_i$ corresponds to level i of the transition matrix, can be found by blocking the transition matrix to obtain a quasi-birth and death (QBD) process as in [9] and [10].

Now, we derive the matrix blocks for the transition matrix in (2). Let $\theta_k = \overline{\text{PER}}_k$ be the probability of transmission failure when the channel is in state k. Assuming that the transmission outcomes of consecutive packets are independent, the probability that i packets are correctly received, given that j packets were transmitted when the channel state is k, can be written as

$$p_{i,j}^{(k)} = \binom{j}{i} \theta_k^{j-i} (1-\theta_k)^i.$$

Let us define the following matrices.

- $\mathbf{W}_{i,k}$ are matrices of order $(K+1)\times(K+1)$ whose elements $(\mathbf{W}_{i,k})(h_1,h_2)$ represent the probability that k packets are successfully transmitted in a particular service slot, given that there were i packets in the queue before transmission and the channel changes from state h_1 to state h_2.
- $\mathbf{H}_{i,k}^{(j)}$ $(j=1,2)$ are matrices of order $(K+1)L\times(K+1)L$ whose elements $(\mathbf{H}_{i,k}^{(j)})(l_1,l_2)$, (h_1,h_2) represent the probability that k packets are successfully transmitted, given that there were i packets in the queue before transmission, and the channel changes from state h_1 to state h_2 during the evolution from slot l_1 to slot l_2 of a cycle for a class-j user .
- $\mathbf{H}_{v}^{(j)}$ $(j=1,2)$ are matrices of order $(K+1)L\times(K+1)L$ which have the same structure as $\mathbf{H}_{i,k}^{(j)}$ and capture the time slot and the channel evolution in the vacation slot(s) for a class-j user.

We can calculate $\mathbf{W}_{i,k}$ as follows:

$$\mathbf{W}_{i,0}=\mathbf{T}_0+\sum_{h=1}^{K}p_{0,b_h}^{(h)}\mathbf{T}_h \tag{3}$$

$$\mathbf{W}_{i,k}=\sum_{h=1}^{K}p_{k,b_h}^{(h)}\mathbf{T}_h,\quad k>0 \tag{4}$$

where $b_h=\min\{c_h,i\}$ and $p_{k,b_h}^{(h)}=0$ if $k>b_h$. The evolution of the channel state and the time slots in a cycle is captured by matrix $\mathbf{C}$ as follows:

$$\mathbf{C}=\begin{bmatrix}\mathbf{0}&\mathbf{T}&\mathbf{0}&\mathbf{0}&\dots&\mathbf{0}\\ \mathbf{0}&\mathbf{0}&\mathbf{T}&\mathbf{0}&\dots&\mathbf{0}\\ \mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{T}&\dots&\mathbf{0}\\ \vdots&\vdots&\vdots&\vdots&\dots&\vdots\\ \mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\dots&\mathbf{T}\\ \mathbf{T}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\dots&\mathbf{0}\end{bmatrix}. \tag{5}$$

We can write $\mathbf{H}_{v}^{(j)}$ as follows:

$$\mathbf{H}_{v}^{(1)}=\begin{bmatrix}\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\dots&\mathbf{0}\\ \mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\dots&\mathbf{0}\\ \mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{T}&\dots&\mathbf{0}\\ \vdots&\vdots&\vdots&\vdots&\dots&\vdots\\ \mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\dots&\mathbf{T}\\ \mathbf{T}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\dots&\mathbf{0}\end{bmatrix} \tag{6}$$

$$\mathbf{H}_{v}^{(2)}=\begin{bmatrix}\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\dots&\mathbf{0}\\ \mathbf{0}&\mathbf{0}&\mathbf{T}&\mathbf{0}&\dots&\mathbf{0}\\ \mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{T}&\dots&\mathbf{0}\\ \vdots&\vdots&\vdots&\vdots&\dots&\vdots\\ \mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\dots&\mathbf{T}\\ \mathbf{T}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\dots&\mathbf{0}\end{bmatrix}. \tag{7}$$

Let $\mathbf{R}_{i,k}^{(j)}$ be defined as

$$\mathbf{R}_{i,k}^{(1)}=\begin{bmatrix}\mathbf{0}&\mathbf{W}_{i,k}&\mathbf{0}&\mathbf{0}&\dots&\mathbf{0}\\ \mathbf{0}&\mathbf{0}&\mathbf{W}_{i,k}&\mathbf{0}&\dots&\mathbf{0}\\ \mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\dots&\mathbf{0}\\ \vdots&\vdots&\vdots&\vdots&\dots&\vdots\\ \mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\dots&\mathbf{0}\\ \mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\dots&\mathbf{0}\end{bmatrix} \tag{8}$$

$$\mathbf{R}_{i,k}^{(2)}=\begin{bmatrix}\mathbf{0}&\mathbf{W}_{i,k}&\mathbf{0}&\mathbf{0}&\dots&\mathbf{0}\\ \mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\dots&\mathbf{0}\\ \vdots&\vdots&\vdots&\vdots&\dots&\vdots\\ \mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\dots&\mathbf{0}\\ \mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\dots&\mathbf{0}\\ \mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\dots&\mathbf{0}\end{bmatrix}. \tag{9}$$

Then, we have

$$\mathbf{H}_{i,k}^{(j)}=\begin{cases}\mathbf{C}, & i=k=0\\ \mathbf{R}_{i,k}^{(j)}, & i,k>0\\ \mathbf{H}_{v}^{(j)}+\mathbf{R}_{i,k}^{(j)}, & i>0 \text{ and } k=0\end{cases} \tag{10}$$

where $j=1,2$. Each of these matrices has L blocks of rows. Each block of rows captures the evolution of the channel states in a particular time slot of a cycle. Now, the matrix blocks of the transition matrix in (2) can be written as follows:

$$\mathbf{A}_{0,k}=\mathbf{U}_{M-k}\otimes\mathbf{C},\quad 0\le k\le M \tag{11}$$

$$\mathbf{A}_{k}=\sum_{h=0}^{M}\mathbf{U}_h\otimes\mathbf{H}_{N,k-M+h}^{(j)},\quad 0\le k\le M+N \tag{12}$$

$$\mathbf{A}_{i,k}=\sum_{h=0}^{M}\mathbf{U}_h\otimes\mathbf{H}_{i,k-M+h}^{(j)},\quad 1\le i<N,\ 0\le k\le i+M \tag{13}$$

where $\otimes$ denotes Kronecker product.

$$\mathbf{P}=\left[\begin{array}{c|cccc|ccccc|ccccc}
\mathbf{A_{0,3}} & \mathbf{A_{0,2}} & \mathbf{A_{0,1}} & \mathbf{A_{0,0}} & & & & & & & & & & \\ \hline
\mathbf{A_{1,4}} & \mathbf{A_{1,3}} & \mathbf{A_{1,2}} & \mathbf{A_{1,1}} & \mathbf{A_{1,0}} & & & & & & & & & \\
\mathbf{A_{2,5}} & \mathbf{A_{2,4}} & \mathbf{A_{2,3}} & \mathbf{A_{2,2}} & \mathbf{A_{2,1}} & \mathbf{A_{2,0}} & & & & & & & & \\
\mathbf{A_{3,6}} & \mathbf{A_{3,5}} & \mathbf{A_{3,4}} & \mathbf{A_{3,3}} & \mathbf{A_{3,2}} & \mathbf{A_{3,1}} & \mathbf{A_{3,0}} & & & & & & & \\
\mathbf{A_7} & \mathbf{A_6} & \mathbf{A_5} & \mathbf{A_4} & \mathbf{A_3} & \mathbf{A_2} & \mathbf{A_1} & \mathbf{A_0} & & & & & & \\ \hline
 & \mathbf{A_7} & \mathbf{A_6} & \mathbf{A_5} & \mathbf{A_4} & \mathbf{A_3} & \mathbf{A_2} & \mathbf{A_1} & \mathbf{A_0} & & & & & \\
 & & \mathbf{A_7} & \mathbf{A_6} & \mathbf{A_5} & \mathbf{A_4} & \mathbf{A_3} & \mathbf{A_2} & \mathbf{A_1} & \mathbf{A_0} & & & & \\
 & & & \mathbf{A_7} & \mathbf{A_6} & \mathbf{A_5} & \mathbf{A_4} & \mathbf{A_3} & \mathbf{A_2} & \mathbf{A_1} & \mathbf{A_0} & & & \\
 & & & & \mathbf{A_7} & \mathbf{A_6} & \mathbf{A_5} & \mathbf{A_4} & \mathbf{A_3} & \mathbf{A_2} & \mathbf{A_1} & \mathbf{A_0} & & \\ \hline
 & & & & & & & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots
\end{array}\right] \tag{2}$$

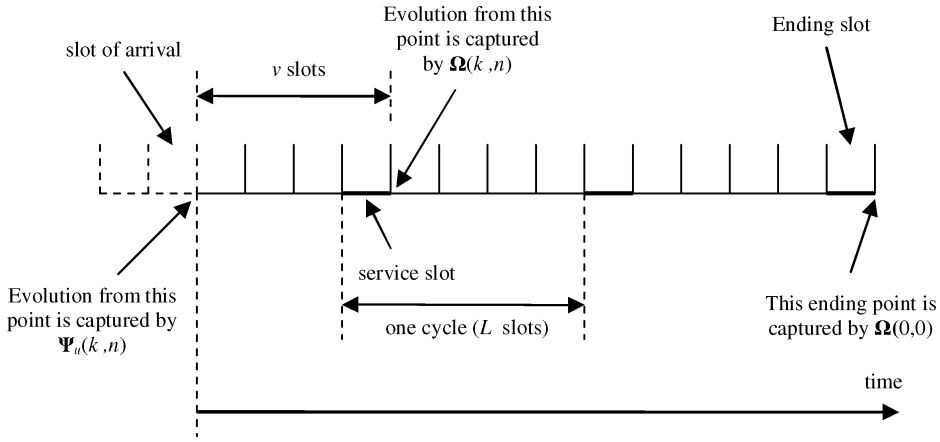

Fig. 1. Delay modeling for a low-priority user.

B. Delay Distribution for a Low-Priority User

In this section, we derive the *delay* distribution for an arriving packet to the queue of a low-priority user. The delay is the time for all packets ahead of the target packet (if any) and itself successfully leaving the queue. Because a low-priority user is assumed to receive service in slot one of a cycle, counting from the end of the arrival slot, the target queue may have to be idle for a while before it is served for the first time. Let the arrival slot be numbered as slot zero, and it is not included in the delay calculation. To avoid confusion, we use "slot u of a cycle" to indicate the uth slot of a particular cycle, and "slot v" to indicate slot v from the arrival slot. Now, if slot one right after the arrival coincides with slot u of a cycle, the target user is in service for the first time in slot v, which satisfies

$$v = \begin{cases} L-u+2 \text{ modulo } L, & L-u+2 \text{ indivisible by } L \\ L, & \text{otherwise.} \end{cases} \tag{14}$$

To calculate the delay for the target packet, we need to keep track of the channel evolution from its arriving slot to the ending slot, where it leaves the queue. The probabilities representing channel transitions and transmission outcomes can be put in the matrix form to facilitate the delay analysis. To this end, let us define the following matrices.

- $\mathbf{\Psi}_u(k,n)$ are matrices of order $(K+1)\times(K+1)$ whose elements $(\mathbf{\Psi}_u(k,n))(i,j)$ describe the probability that an arriving packet spends n slots in the queue, given that it sees k packets waiting in the queue, slot one from the arrival slot coincides with slot u of a cycle, the channel state is i at the beginning of slot one, and is j at the end of slot n.
- $\mathbf{\Omega}(k,n)$ are matrices of order $(K+1)\times(K+1)$ whose elements $(\mathbf{\Omega}(k,n))(i,j)$ represent the probability that k packets are successfully transmitted in n slots counting from the end of the first service slot (slot v), starting in channel state i, and ending in channel state j.
- $\mathbf{S}_{k,l}^{(v)}$ are matrices of order $(K+1)\times(K+1)$ whose elements $(\mathbf{S}_{k,l}^{v})(i,j)$ represent the probability that l packets are successfully transmitted in slot v, given that there were k packets in slot one (there is no transmission from slot one to slot $v-1$), the channel state is i at the beginning of slot one, and is j at the end of slot v.

The modeling of delay for an arriving packet to the queue of a low-priority user is illustrated in Fig. 1. We have the following recursive relations:

$$\mathbf{\Psi}_u(k,n+v) = \sum_{l=0}^{N} \mathbf{S}_{k,l}^{(v)} \mathbf{\Omega}(k-l+1,n) \tag{15}$$

$$\mathbf{\Omega}(k,n) = \sum_{l=0}^{N} \mathbf{S}_{k,l}^{(L)} \mathbf{\Omega}(k-l,n-L) \tag{16}$$

$$\mathbf{\Omega}(0,0) = \mathbf{I}_{K+1}. \tag{17}$$

Equation (15) captures the case where an arriving packet sees k packets waiting in the queue, and the first transmission occurs in slot v with l successfully transmitted packets. Thus, there are $k-l+1$ packets in the queue, including the target packet at the end of slot v if we turn off the arrival source after the target packet enters the queue. These packets will successfully leave the queue in n slots. Equation (16) describes the transmission from the end of slot v, where transmissions occur once in each cycle of L slots. We also have $\mathbf{S}_{k,l}^{(v)} = \mathbf{T}^{v-1}\mathbf{W}_{k,l}$.

Suppose that the delay for an arriving packet is D slots (not including the arrival slot). Recall that in D slots, the first transmission takes place in slot v ($\leq L$) from the arrival, and other J transmissions occur periodically, once every L slots from slot v. We can calculate J as follows:

$$J = \begin{cases} \frac{D}{L}-1, & D \text{ is divisible by } L \\ \left\lfloor \frac{D}{L} \right\rfloor, & \text{otherwise.} \end{cases} \tag{18}$$

Slot v, when the first transmission from the arrival takes place, can be calculated as $v = D - JL$. We can calculate the corresponding u from (14). Let $\mathbf{z}_{i,u}$ be a $(K+1)$-dimensional row vector, whose element $\mathbf{z}_{i,u}(k)$ is the probability that an arbitrary arriving packet sees i packets in the queue, slot one from the arrival coincides with slot u of a cycle, and the channel state is k at the beginning of slot one. If a batch of m ($m = 1,2,\ldots,M$) packets enters the queue, the target packet can be at any position in the arriving batch with probability $1/m$. In fact, if the target packet is at the jth position in the arriving batch, there are $j-1$ packets ahead of it in the arriving batch.

Let $\mathbf{y}_i$ be a N_1-dimensional row vector whose entry $\mathbf{y}_{i,a,u,k}$ is the probability that an arriving packet sees i packets ahead of it, the arrival phase is a, slot one coincides with slot u of a cycle, and the channel state is k at the beginning of slot one. Let

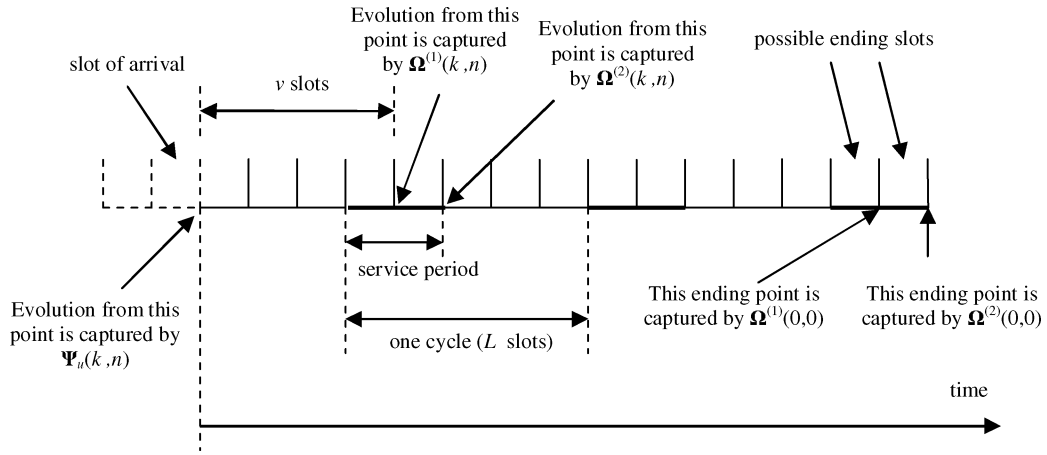

Fig. 2. Delay modeling for a high-priority user.

β denote the probability that there is at least one arriving packet to the queue. Then, $\mathbf{y}_i$ can be calculated as follows:

$$\mathbf{y}_0 = \frac{1}{\beta}\sum_{m=1}^{M}\sum_{l=0}^{N}\frac{1}{m}\mathbf{x}_l\mathbf{U}_m \otimes \mathbf{H}_{l,l}^{(j)} \tag{19}$$

$$\mathbf{y}_i = \frac{1}{\beta}\sum_{m=1}^{M}\sum_{h=1}^{m}\sum_{l=0}^{N}\frac{1}{m}\mathbf{x}_{i+l-h+1}\mathbf{U}_m \otimes \mathbf{H}_{i+l-h+1,l}^{(j)} \tag{20}$$

where $1 \le i \le Q+M-1$ and $j = 1, 2$ (each value of j results in the corresponding vector $\mathbf{y}_i$ for a class-j user). For delay analysis, we are interested only in the arriving packets which are admitted into the queue. The probability that an arriving packet sees the full buffer is $P_f = \sum_{i=Q}^{Q+M-1}\mathbf{y}_i\mathbf{1}_{N_1}$, where $\mathbf{1}_{N_1}$ is a column vector of all ones with dimension N_1. Under the admitting condition of the target packet to the queue, let $\mathbf{y}_i' = \mathbf{y}_i/(1-P_f)$, $(0 \le i \le Q-1)$ be the vector corresponding to $\mathbf{y}_i$.

It is easy to observe that $\mathbf{z}_{i,u}$ are actually the partitions of $\mathbf{y}_i'$ where all arrival phases are lumped together. To obtain $\mathbf{z}_{i,u}$ from $\mathbf{y}_i'$, let $\boldsymbol{\varphi}_u$ $(u = 1, 2, \ldots, L)$ be a matrix of size $N_1 \times (K+1)$ given as

$$\boldsymbol{\varphi}_u = [\underbrace{\mathbf{0}\ \cdots\ \mathbf{I}_{K+1}\ \cdots\ \mathbf{0}}_{L\text{ blocks}}\ \cdots\ \underbrace{\mathbf{0}\ \cdots\ \mathbf{I}_{K+1}\ \cdots\ \mathbf{0}}_{L\text{ blocks}}]^T$$

where there are M_1 groups, each consisting of L blocks as indicated above, and $\mathbf{I}_{K+1}$ is the identity matrix at the uth position in each group. Then, we have $\mathbf{z}_{i,u} = \mathbf{y}_i'\boldsymbol{\varphi}_u$.

The probability that the delay for the target packet is D slots (not including the arrival slot) can be written as

$$P_d^{(2)}(D) = \sum_{i=0}^{W_1}\mathbf{z}_{i,u}\boldsymbol{\Psi}_u(i, D)\mathbf{1}_{N_1} \tag{21}$$

where $W_1 = JN + N - 1$. In (21), the summation is limited by W_1, since at most N packets can be successfully transmitted in one time slot.

C. Delay Distribution for a High-Priority User

In this section, we derive the delay distribution for an arriving packet to the queue of a high-priority user. For a high-priority user, there are two consecutive service slots (assumed to be the first slot and the second slot in each cycle of L slots). If slot one (the first slot right after the arrival slot) coincides with slot u of a cycle, the target user is in service for the first time in slot v, which can be calculated as

$$v = \begin{cases} L - u + 2 \text{ modulo } L, & u \ne 2 \\ 1, & u = 2. \end{cases} \tag{22}$$

Because the target packet and its head-of-line packets can only leave the queue in either slot one or slot two of a cycle, we need to keep track of the transmission outcomes in these two service slots. Also, we have to keep track of the channel evolution during the vacation periods. Again, the channel transition probabilities and transmission outcomes can be captured in matrix forms to ease the analysis. Now, let us define the following matrices.

- $\boldsymbol{\Omega}^{(1)}(k, n)$ are matrices of order $(K+1)\times(K+1)$ whose elements $(\boldsymbol{\Omega}^{(1)}(k, n))(i, j)$ represent the probability that k packets are successfully transmitted in n slots counting from the end of the first slot of a cycle, starting in channel state i in slot one, and ending in channel state j in slot n.
- $\boldsymbol{\Omega}^{(2)}(k, n)$ are matrices of order $(K+1)\times(K+1)$ whose elements $(\boldsymbol{\Omega}^{(2)}(k, n))(i, j)$ represent the probability that k packets are successfully transmitted in n slots counting from the end of the second slot of a cycle, starting in channel state i in slot one, and ending in channel state j in slot n.

The modeling of delay for an arriving packet to the queue of a high-priority user is illustrated in Fig. 2. We have the following recursive relations for these matrices:

$$\boldsymbol{\Psi}_u(k, n+v) = \sum_{l=0}^{N}\mathbf{S}_{k,l}^{(v)}\boldsymbol{\Omega}^{(1)}(k-l+1, n), \quad \text{if } u \ne 2 \tag{23}$$

$$\boldsymbol{\Psi}_u(k, n+1) = \sum_{l=0}^{N}\mathbf{S}_{k,l}^{(1)}\boldsymbol{\Omega}^{(2)}(k-l+1, n), \quad \text{if } u = 2 \tag{24}$$

$$\boldsymbol{\Omega}^{(2)}(k, n) = \sum_{l=0}^{N}\mathbf{S}_{k,l}^{(L-1)}\boldsymbol{\Omega}^{(1)}(k-l, n-L+1) \tag{25}$$

$$\boldsymbol{\Omega}^{(1)}(k, n) = \sum_{l=0}^{N}\mathbf{S}_{k,l}^{(1)}\boldsymbol{\Omega}^{(2)}(k-l, n-1) \tag{26}$$

$$\boldsymbol{\Omega}^{(1)}(0, 0) = \mathbf{I}_{K+1}, \quad \boldsymbol{\Omega}^{(2)}(0, 0) = \mathbf{I}_{K+1}. \tag{27}$$

We can explain the above recursions as follows. Equation (23) describes the case where the first service after the arrival slot occurs in slot v, which coincides with slot one of a cycle. Equation (24) represents the case where the first slot after the arrival slot is slot two of a cycle; therefore, the queue is served in this slot. In both cases, if there are k packets ahead of the target packet and we turn off the arrival source after the target packet enters the queue, and l packets are successfully transmitted in slot v, there will be $k-l+1$ remaining packets which must be transmitted successfully in n slots. Equation (25) captures the fact that counting from the end of the second slot of a cycle, the next service occurs $L-1$ slots after that. Equation (26) describes the fact that from the end of the first slot of a cycle, the queue is still in service in the second slot of that cycle (because a high-priority user receives service in slot one and slot two of a cycle).

We know that the target packet may leave the queue in slot one or slot two of a cycle. Suppose that the first slot after the arrival slot coincides with either slot u_1 or u_2 of a cycle for these two cases, respectively, such that the delay is D time slots. The probability that the delay is D slots (not including the arrival slot) can be written as follows:

$$P_d^{(1)}(D)=\sum_{i=0}^{W_2}\mathbf{z}_{i,u_1}\mathbf{\Psi}_{u_1}(i,D)\mathbf{1}_{N_1}+\sum_{i=0}^{W_3}\mathbf{z}_{i,u_2}\mathbf{\Psi}_{u_2}(i,D)\mathbf{1}_{N_1}. \tag{28}$$

Again, the two sum-terms in (28) are limited by W_2 and W_3, since at most N packets can be successfully transmitted in one service slot. The values of W_2 and W_3 depend on D, but we have $W_2, W_3 \le 2N\lceil D/L\rceil$.

D. Extension for the General-Priority Case

In this section, we extend the previous model by considering a more general scenario with more than two service classes. Suppose there are η user classes and a class-j user receives d_j service slots in one cycle. If there are μ_j class-j users in the system, a cycle consists of $L=\sum_{j=1}^{\eta}\mu_j d_j$ time slots. To analyze the queue corresponding to a class-j user, we can assume that it receives service from slot one to slot d_j of a cycle. The state space of the queue corresponding to the class-j user is still

$$\{(x_n,a_n,u_n,s_n), x_n\ge 0,\ 1\le a_n\le M_1,\ 1\le u_n\le L, 0\le s_n\le K\}.$$

The state transition probabilities can be put in the matrix blocks for each level, and the transition matrix of the MC can still be written as in (2).

The derivations of matrix blocks of the transition matrix for a class-j user can be done in the same way as before, where the matrices $\mathbf{H}_v^{(j)}$ and $\mathbf{H}_{i,k}^{(j)}$ can be written as follows:

$$\mathbf{H}_v^{(j)}=\begin{bmatrix}\mathbf{0}&\mathbf{0}&\mathbf{0}&\dots&\mathbf{0}&\mathbf{0}&\dots&\mathbf{0}\\ \vdots&\vdots&\vdots&\dots&\vdots&\vdots&\dots&\vdots\\ \mathbf{0}&\mathbf{0}&\mathbf{0}&\dots&\mathbf{0}&\mathbf{0}&\dots&\mathbf{0}\\ \mathbf{0}&\mathbf{0}&\mathbf{0}&\dots&\mathbf{T}&\mathbf{0}&\dots&\mathbf{0}\\ \mathbf{0}&\mathbf{0}&\mathbf{0}&\dots&\mathbf{0}&\mathbf{T}&\dots&\mathbf{0}\\ \vdots&\vdots&\vdots&\dots&\vdots&\vdots&\dots&\vdots\\ \mathbf{0}&\mathbf{0}&\mathbf{0}&\dots&\mathbf{0}&\mathbf{0}&\dots&\mathbf{T}\\ \mathbf{T}&\mathbf{0}&\mathbf{0}&\dots&\mathbf{0}&\mathbf{0}&\dots&\mathbf{0}\end{bmatrix} \tag{29}$$

$$\mathbf{H}_{i,k}^{(j)}=\begin{bmatrix}\mathbf{0}&\mathbf{W}_{i,k}&\mathbf{0}&\dots&\mathbf{0}&\mathbf{0}&\dots&\mathbf{0}\\ \vdots&\vdots&\vdots&\dots&\vdots&\vdots&\dots&\vdots\\ \mathbf{0}&\mathbf{0}&\mathbf{0}&\dots&\mathbf{W}_{i,k}&\mathbf{0}&\dots&\mathbf{0}\\ \mathbf{0}&\mathbf{0}&\mathbf{0}&\dots&\mathbf{0}&\mathbf{0}&\dots&\mathbf{0}\\ \mathbf{0}&\mathbf{0}&\mathbf{0}&\dots&\mathbf{0}&\mathbf{0}&\dots&\mathbf{0}\\ \vdots&\vdots&\vdots&\dots&\vdots&\vdots&\dots&\vdots\\ \mathbf{0}&\mathbf{0}&\mathbf{0}&\dots&\mathbf{0}&\mathbf{0}&\dots&\mathbf{0}\\ \mathbf{0}&\mathbf{0}&\mathbf{0}&\dots&\mathbf{0}&\mathbf{0}&\dots&\mathbf{0}\end{bmatrix} \tag{30}$$

where there are L blocks of rows in these matrices, each of them consisting of $K+1$ rows, which captures the channel evolution in the corresponding time slot of a cycle. In $\mathbf{H}_{i,k}^{(j)}$, there are d_j nonzero blocks of rows, which correspond to d_j service slots where k packets successfully leave the queue. In $\mathbf{H}_v^{(j)}$, there are $L-d_j$ nonzero blocks of rows, which capture the channel evolution in the vacation slots. The matrix blocks $\mathbf{A}_{i,k}$ and $\mathbf{A}_k$ can still be calculated as in (11)–(12). To calculate the delay distribution for a class-j user, we have to define $\mathbf{\Omega}^{(h)}(k,n)$, $(h=1,\dots,d_j)$ which captures the system evolution from the end of slot h of a cycle, where k packets are successfully transmitted in n slots. Similar recursive relations as in (23)–(27) can be developed, and delay distribution can be calculated where the target packet leaves the queue in one of the d_j service slots of a cycle.

IV. Validation and Applications of the Queueing Model

We assume that the SNR thresholds for the FSMC model are chosen such that $\overline{\mathrm{PER}}_k=P_0$, and these thresholds are obtained by using the technique outlined in [4]. We assume that $c_k=bS_k$, $b=2$, where $S_k=0.5,1.0,1.5,3.0,4.5$ are the spectral efficiencies of five transmission modes (see [4, Table II]). Let f_d and T_s denote the Doppler shift and the time slot interval, respectively; then f_dT_s represents the normalized fading rate of the wireless channel. A two-state Markovian traffic source, which is a special case of BMAP [9], with the following arrival-state transition matrix used to obtain the numerical results:

$$\mathbf{U}=\begin{bmatrix}0.8&0.2\\0.1&0.9\end{bmatrix}.$$

The complementary cumulative delay distributions for users of both classes obtained by simulation and from the analytical model are shown in Fig. 3. Note that $\Pr\{\text{delay}\ \ge D\}=1-\sum_{i=1}^{D-1}P_d^{(j)}(i)$ for a class-j user. As is evident, the simulation results follow the analytical results very closely. It can also be observed that the higher the Doppler shift f_d and/or Nakagami parameter m, the lower the delay. The complete delay statistics obtained for both user classes enables us to design and engineer the system under statistical delay constraints. Suppose that we are interested in the "95% delay," which refers to the smallest value of D such that $\Pr\{\text{delay}\ < D\}>0.95$. One important design problem is how to choose the SNR thresholds for different transmission modes, such that the delay at the radio link layer is minimized. In Fig. 4, we plot the "95% delay" versus the target PER P_0 for different channel parameters. The optimal point is indicated by a "★." Evidently, using the delay statistics, the mode switching thresholds for AMC can be selected to improve the delay performance.

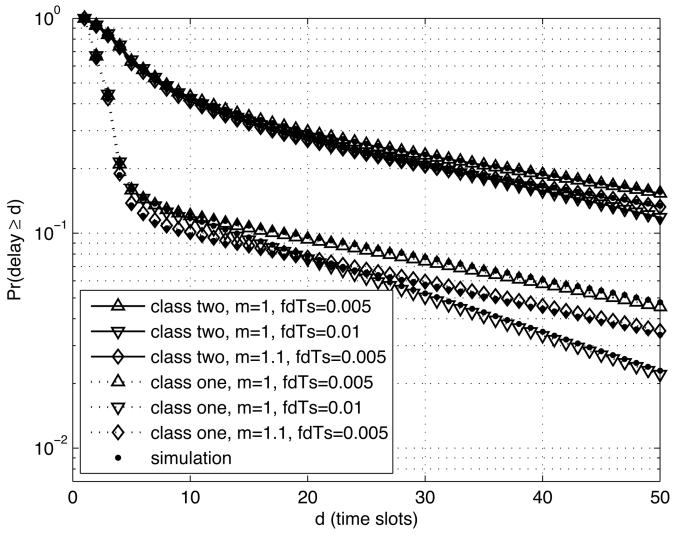

Fig. 3. Complementary cumulative delay distribution (for buffer size $Q = 70$, $L = 4$, average SNR $= 12$ dB, $P_0 = 0.1$, Nakagami parameter $m = 1, 1.1$, $f_d T_s = 0.005, 0.01$).

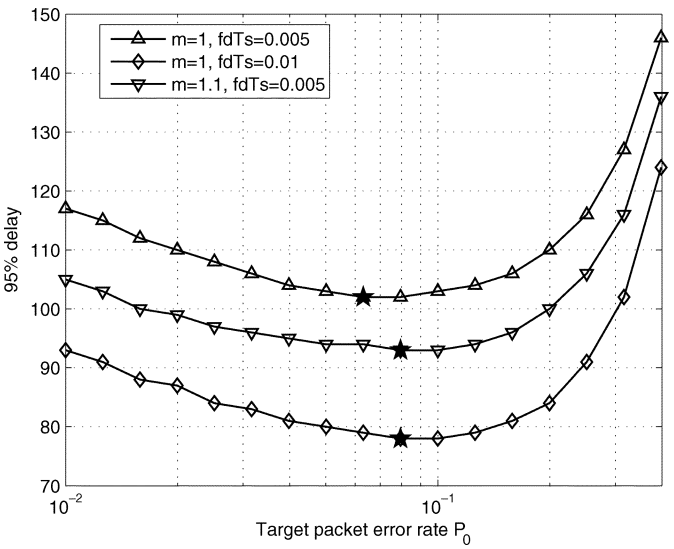

Fig. 4. 95% delay for a class-one user versus target PER P_0 (for buffer size $Q = 70$, $L = 4$, average SNR $= 12$ dB, Nakagami parameter $m = 1, 1.1$, $f_d T_s = 0.005, 0.01$).

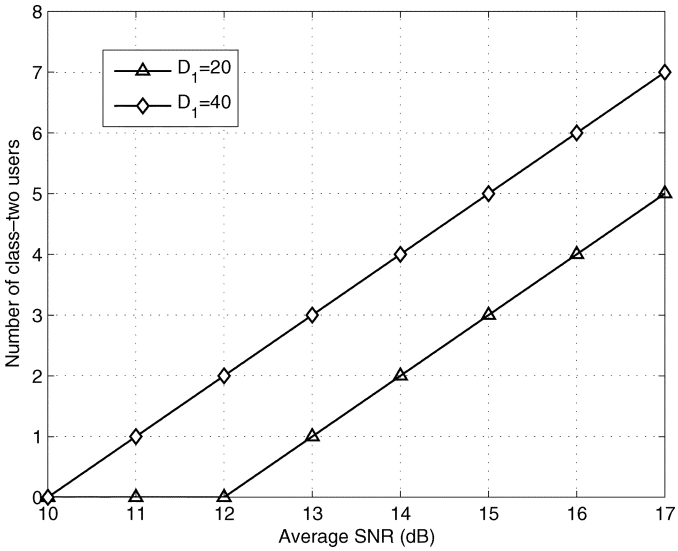

Fig. 5. Maximum admissible number of class-two users versus average SNR (for $D_1 = 20, 40$, $Q = 100$, $P_0 = 0.1$, Nakagami parameter $m = 1$, $f_d T_s = 0.005$).

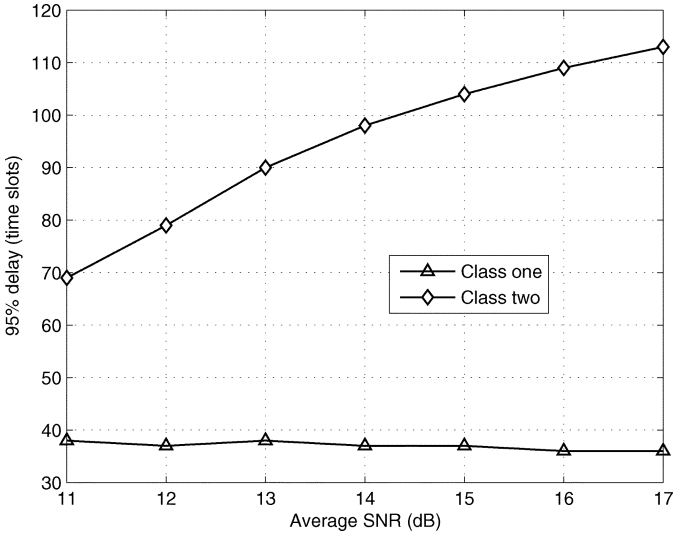

Fig. 6. 95% delay versus average SNR with maximum admissible number of class-two users (for $D_1 = 40$, $Q = 100$, $P_0 = 0.1$, Nakagami parameter $m = 1$, $f_d T_s = 0.005$).

The obtained delay statistics can also be used for admission control under statistical delay constraints. Typical numerical results on admission control are shown in Fig. 5, where one class-one user has already been admitted into the system. Under the constraint $\Pr\{\text{delay} \geq D_1\} \leq 5\%$ for a class-one user, the maximum admissible number of class-two users is determined here. With a very strict delay requirement of $D_1 = 20$, no class-two user can be admitted if the average SNR is less than 12 dB, while for $D_1 = 40$, if the average SNR is greater than 11 dB, class-two users can be admitted into the system. In Fig. 6, we show the variations in "95% delay" for both user classes with the maximum admissible number of class-two users and one class-one user in the system. The achieved "95% delay" for a class-one user is observed to be always smaller than the desired constraint ($D_1 = 40$). Since the delay for a class-two user is not constrained, it is quite large for high average SNR.

V. Conclusions

We have developed an analytical framework for radio link-level performance evaluation of a multirate wireless network with WRR scheduling and ARQ-based error control. We have captured the key physical layer and the radio link-layer features in the analytical model. Traffic arrival has been modeled as a BMAP, which allows correlation in the arrival traffic. The probability distributions for queue length and delay have been obtained analytically, and therefore, the impacts of different channel and system parameters on the system performance can be quantified. Based on the obtained results, we have proposed an admission control policy under the statistical

delay requirement. Also, a cross-layer design example has been shown to highlight the usefulness of the presented model. In summary, the presented analytical framework would be very useful for cross-layer analysis, design, and optimization of multirate wireless systems.

REFERENCES

[1] M. Zorzi and R. R. Rao, "Lateness probability for a retransmission scheme for error control on a two-state Markov channel," *IEEE Trans. Commun.*, vol. 47, no. 10, pp. 1537–1548, Oct. 1999.

[2] J. G. Kim and M. M. Krunz, "Delay analysis of selective repeat ARQ for a Markovian source over wireless channel," *IEEE Trans. Veh. Technol.*, vol. 49, no. 5, pp. 1968–1981, Sep. 2000.

[3] A. Doufexi, S. Armour, M. Butler, A. Nix, D. Bull, J. McGeehan, and P. Karlsson, "A comparison of the HIPERLAN/2 and IEEE 802.11a wireless LAN standards," *IEEE Commun. Mag.*, vol. 40, no. 5, pp. 172–180, May 2002.

[4] Q. Liu, S. Zhou, and G. B. Giannakis, "Queuing with adaptive modulation and coding over wireless link: Cross-layer analysis and design," *IEEE Trans. Wireless Commun.*, vol. 4, no. 5, pp. 1142–1153, May 2005.

[5] H. S. Wang and N. Moayeri, "Finite-state Markov channel—A useful model for radio communication channels," *IEEE Trans. Veh. Technol.*, vol. 44, no. 1, pp. 163–171, Feb. 1995.

[6] Y. L. Guan and L. F. Turner, "Generalized FSMC model for radio channels with correlated fading," *IEE Proc. Commun.*, vol. 146, no. 2, pp. 133–137, Apr. 1999.

[7] A. K. Parekh and R. G. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: The single node case," *IEEE/ACM Trans. Netw.*, vol. 1, no. 3, pp. 344–357, Jun. 1993.

[8] Z. Zhang, D. Towsley, and J. Kurose, "Statistical analysis of the generalized processor sharing scheduling discipline," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 8, pp. 1071–1080, Aug. 1995.

[9] L. B. Le, E. Hossain, and A. S. Alfa, "Radio link level performance evaluation in wireless networks using multi-rate transmission with ARQ-based error control," *IEEE Trans. Wireless Commun.*, to be published.

[10] M. F. Neuts, *Matrix Geometric Solutions in Stochastic Models—An Algorithmic Approach*. Baltimore, MD: John Hopkins Univ. Press, 1981.