Queueing Analysis for GBN and SR ARQ Protocols under Dynamic Radio Link Adaptation with Non-Zero Feedback Delay

Long B. Le, Ekram Hossain, *Senior Member, IEEE*, and Michele Zorzi, *Fellow, IEEE*

***Abstract*— We present a queueing model for performance analysis of go-back-N (GBN) and selective repeat (SR) automatic repeat request (ARQ) protocols in wireless networks using dynamic radio link adaptation with non-instantaneous feedback. Link adaptation technique allows multi-rate transmission which is assumed to be achieved through adaptive modulation and coding. The radio link level queueing models for these two ARQ protocols are formulated in discrete time where the exact queue length and the delay statistics are obtained by using matrix geometric methods under different feedback delay values, channel and system parameters. The link layer delay statistics are useful in many ways, for example, to perform packet level admission control under statistical delay constraints. We validate the analysis by simulation and discuss useful implications of the analytical model on system performance. For dynamic link adaptation, the mode switching thresholds for the received signal-to-noise ratio (SNR) can be chosen to obtain very good link level delay performance. This SNR partitioning is shown to achieve significant cross-layer design gain compared to the case where the mode switching thresholds are chosen to maximize the physical layer throughput.**

***Index Terms*— Automatic repeat request (ARQ) protocols, cross-layer design and analysis, dynamic radio link adaptation, go-back-N (GBN), matrix geometric method (MGM), multi-rate transmission, selective repeat (SR).**

I. Introduction

HIGH-SPEED data transmission will be a key requirement for the next-generation wireless networks. Adaptive modulation and coding (AMC) technique is being used in most of the 2.5/3G wireless networks to increase the transmission rate by exploiting the wireless channel variations [1], [2]. While the design of strong and reliable error correction codes has played a key role in error protection for applications with strict delay requirements (e.g., voice), the deployment of delay-tolerant data services in wireless networks makes automatic repeat request (ARQ)-based error protection very attractive to counteract the residual errors without using costly error correction codes at the physical layer.

Among the three main ARQ protocols (namely, stop-and-wait, go-back-N (GBN-ARQ) and selective repeat (SR-ARQ)), SR-ARQ is the most efficient. GBN-ARQ is less efficient than SR-ARQ but its implementation is simpler than SR-ARQ because packets are always received in order in the receiving buffer. The delay statistics for GBN-ARQ with non-instantaneous feedback delay was derived in [3]. In [4], the approximated average delay for SR-ARQ was obtained for a two-state Markov channel under heavy traffic condition. The exact delay statistics for SR-ARQ over a two-state Markov channel was obtained in [5]. The key weakness of these works, however, is the use of a two-state Markov channel model, which could not capture the multi-rate transmission feature of currently deployed wireless networks.

The analytical model in [5] was extended in [6] for channels with N states. However, the transmission rate was assumed to be constant (i.e., for all channel states one packet is transmitted in one time slot); therefore, the model did not truly take the multi-rate transmission into account. In [7], a link level queueing model with AMC and round-robin scheduling was developed assuming instantaneous feedback (i.e., the transmitter knows the transmission outcomes right at the end of the transmission time slot). Developing an analytical framework for evaluating radio link level queueing performances which can capture multi-rate transmission and non-instantaneous feedback delay is very desirable but challenging.

Note that the availability of radio link-level delay statistics allows wireless network design and engineering under statistical delay constraints of the form $\Pr\{\text{delay} > D_{\max}\} < P_t$ instead of those based on the average delay or delay bounds. Also, it would be very useful in predicting the higher layer protocol (e.g., TCP (Transmission Control Protocol)) performance (e.g., to estimate the round trip time of a TCP flow). So far, investigations on the higher layer protocol performance have been done mostly by simulations [8]. The link model presented in this paper is therefore an important step towards investigating the interaction of TCP with lower layers protocols [9].

A finite-state Markov channel (FSMC) model was proposed for Rayleigh and the more general Nakagami-m fading channels in the literature [10]-[13]. This channel model was validated in [13] and extensively used to analyze system performance at the packet level [14]-[15]. In [15], the authors

Manuscript received February 24, 2006; revised June 26, 2006; accepted August 24, 2006. The editor coordinating the review of this paper and approving it for publication was G. Mandyam. This work was supported in part by the University of Manitoba Graduate Fellowship (UMGF) and in part by a grant from the Natural Sciences and Engineering Research Council (NSERC) of Canada.

L. B. Le is with the Department of Electrical and Computer Engineering, University of Waterloo, Canada (email: longble@engmail.uwaterloo.ca).

E. Hossain is with the Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, MB, Canada R3T 5V6 (e-mail: ekram@ee.umanitoba.ca).

M. Zorzi is with the Department of Information Engineering, University of Padova, Italy (e-mail: zorzi@dei.unipd.it).

Digital Object Identifier 10.1109/TWC.2007.06020038.

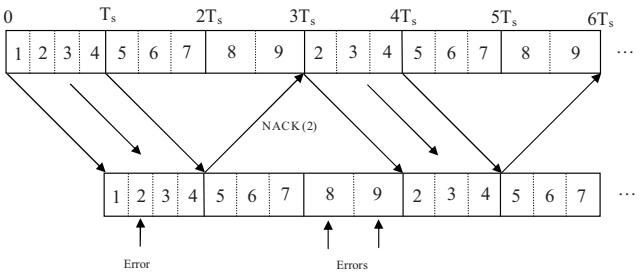

Fig. 1. GBN-ARQ timing diagram.

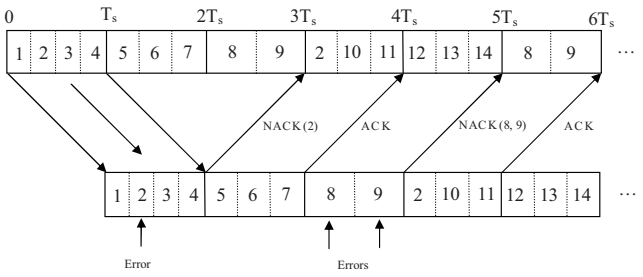

Fig. 2. SR-ARQ timing diagram.

developed a queueing model that takes into account AMC at the physical layer. However, ARQ was not considered at the link layer and the delay statistics cannot be derived from their model. The difficulty of using FSMC model to analyze the performance of multi-rate wireless networks comes from the batch transmission effect where the transmission batch size varies according to the chosen transmission mode at the physical layer. Nevertheless, the interaction among the channel, system and protocol states can be captured in a unified analytical framework which can be used to obtain many performance measures in a multi-rate wireless network. We present such an analytical framework in this paper and obtain the queue length and delay statistics for both GBN-ARQ and SR-ARQ protocols.

We formulate the radio link layer queueing model for these two ARQ protocols in discrete time where the queue length and the delay distributions can be derived by using the matrix geometric method (MGM) [16]. The analytical model enables us to quantify the impacts of physical/radio link and channel parameters on the system performance. Also, it provides interesting insights into the system design. For example, the signal-to-noise ratio (SNR) thresholds for the different transmission modes can be calculated to achieve good link level delay performance. Note that, delay is the primary quality of service (QoS) metric for many data applications and the SNR thresholds which maximize the link level throughput may not be, in general, delay-optimal. Comparison of delay performance of the two ARQ protocols is also highlighted for different values of feedback delay which is necessary to quantify the tradeoff between delay performance and system complexity for link layer protocol design.

The rest of this paper is organized as follows. Section II presents the system model and assumptions used in this paper. We formulate the queueing problem and obtain the performance metrics for GBN-ARQ protocol in Section III. Section IV presents the analysis for SR-ARQ protocol. Model validation and numerical results are presented in Section V. Section VI states the conclusions.

II. System Model and Assumptions

A. System Description

We consider a transmitter node using adaptive modulation and coding at the physical layer and ARQ-based error recovery at the link layer to communicate with a receiver node over a wireless channel. Transmissions occur in fixed-size time slots where the number of packets transmitted during each time slot depends on the chosen transmission mode. The receiver decodes the received packets and sends a feedback packet (i.e., containing acknowledgment (ACK) or negative acknowledgment (NACK) information) to the transmitter. In case of transmission failure of one or more packets transmitted during a time slot, error recovery based on either GBN-ARQ or SR-ARQ protocol is initiated.

For both ARQ protocols, the transmitter continuously transmits packets from the buffer in sequence until it detects a transmission error through NACK in the feedback packet. In case of transmission failure(s), the SR-ARQ protocol only retransmits the erroneous packet(s) while the GBN-ARQ protocol retransmits all the packets starting from the first erroneous packet.

We assume that the feedback packet (i.e., the ACK/NACK information) arrives at the transmitter node n slots after the beginning of the corresponding transmission slot. In this paper, an error-free feedback channel is assumed[1]. In addition to the ACK/NACK information, the feedback channel carries the channel state information (CSI) or the selected transmission mode to be used for dynamic link adaptation. We assume that CSI is available at the transmitter without delay. This assumption is reasonable in slow fading channels where the channel conditions are static over several transmission intervals (or time slots). The maximum number of retransmissions allowed for a packet is assumed to be unbounded. Therefore, the delay obtained in this paper can be considered as an upper bound for the case where finite number of retransmissions are allowed at the link layer.

Examples: The operations of the GBN-ARQ and SR-ARQ protocols are illustrated in Fig. 1 and Fig. 2, respectively, for $n = 3$, where the transmission batch size is denoted by the rate defined in terms of packets/slot. In these two figures, packet 2 in time slot 1 and packets 8, 9 in time slot 3 are assumed to be in error. For the GBN-ARQ protocol, the NACK for packet 2 arrives at the transmitter side at the end of time slot 3 and retransmission of all packets starting from packet 2 begins at time slot 4. Note that, packets $3, 4, \cdots, 7$ are retransmitted even though they were correctly received before. For the SR-ARQ protocol, packet 2 and packets 8, 9 are "selectively"

[1]This is a very standard assumption because in practice a strong error correction code can be used in the feedback channel.

retransmitted in time slots 4 and 6, respectively, together with the new packets when the channel state allows.

The channel is modeled as an FSMC with $K+1$ states $(0, 1, \cdots, K)$ as will be described in Section II-B. When the channel is in state k $(1, 2, \cdots, K)$, h_k packets are transmitted in one time slot. In fact, each channel state corresponds to one transmission mode of the AMC technique as will be modeled in the next section. We further assume that the transmitter does not transmit in channel state 0 to avoid high probability of transmission error.

The radio link level queueing for both ARQ protocols is modeled in discrete time with one time interval equal to one time slot and the system states are observed at the beginning of each time slot. The buffer size is assumed to be infinite. Packet arrivals follow a Bernoulli process with arrival probability λ. We assume that a packet arriving during time interval $t-1$ cannot be transmitted until time interval t at the earliest.

B. Channel Modeling for Adaptive Modulation and Coding

The channel model used in this paper is captured by an FSMC representing the multiple states of a slow Nakagami-m fading channel. In this channel model, the SNR at the receiver is partitioned into a finite number of intervals. Let $X_0(=0) < X_1 < X_2 < \cdots < X_{K+1}(=\infty)$ denote the thresholds of the received SNR for the different channel states. The channel is said to be in state k if $X_k \le X < X_{k+1}$ $(k = 0, 1, 2, \cdots, K)$, where X is the received SNR. As the thresholds of the received SNR are determined, the transition probability matrix for the channel states $\mathbf{T}$, in which element $\mathbf{T}_{k,l}$ is the transition probability from state k to state l can be obtained [10]-[15].

At the physical layer of the considered system, each transmission mode corresponds to a unique modulation and coding scheme. The number of packets transmitted in mode k (equal to h_k) is therefore proportional to the spectral efficiency of mode k. For example, if the spectral efficiencies of five transmission modes in a particular system are 0.5, 1, 1.5, 2.5 and 3.5 (bits/s/Hz) and 1 packet can be transmitted in mode one (with spectral efficiency of 0.5), the number of packets transmitted in one time slot using the other modes is 2, 3, 5, 7, respectively. The maximum number of packets that can be transmitted in one time slot (in channel state K) is denoted by L. Also note that, no transmission is allowed in channel state 0. To calculate the *packet error rate* (PER) for mode k, we use the following approximation as in [15]:

$$\mathrm{PER}_k(X) \approx \begin{cases} 1, & 0 < X < X_{pk} \\ a_k \exp(-g_k X), & X \ge X_{pk} \end{cases} \quad (1)$$

where a_k, g_k and X_{pk} are obtained by curve fitting.

The average PER for mode k can be written as follows:

$$\overline{\mathrm{PER}}_k = \frac{1}{\Pr(k)} \int_{X_k}^{X_{k+1}} a_k \exp(-g_k X) p_X(X) dX$$
$$= \frac{1}{\Pr(k)} \frac{a_k}{\Gamma(m)} \left(\frac{m}{\overline{X}}\right)^m \frac{\Gamma(m, b_k X_k) - \Gamma(m, b_k X_{k+1})}{(b_k)^m} \quad (2)$$

where $m \ge 1/2$ is the Nakagami parameter, $\overline{X}$ is the average SNR, $b_k = m/\overline{X} + g_k$ $(k = 1, \cdots, K)$, $\Gamma(m, x) = \int_x^\infty t^{m-1} \exp(-t) dt$ is the complementary incomplete Gamma function, and $\Pr(k)$ is the probability of channel state k, which can be calculated as in [15].

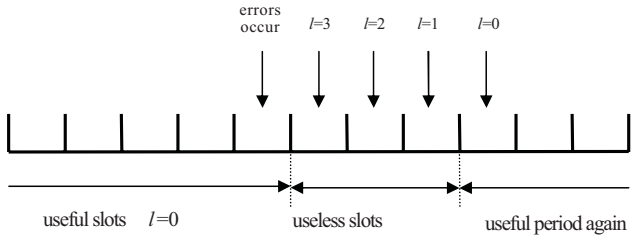

Fig. 3. Modeling of GBN-ARQ for $n = 4$.

III. Queueing Model and Analysis for Go-Back-N ARQ

A. Queueing Model

Since the result of the decoding process for each packet only reaches the transmitter n slots after the beginning of the transmission slot, if a transmitted packet is in error in time slot t, all transmissions from time slot $t+1$ to $t+n-1$ will be discarded. Therefore, we need to keep track not only of the channel state, which determines how many packets can be transmitted in one time slot, but also the *useful* time slot, which is defined as the slot where the transmitted packets, if successfully decoded, will be accepted by the receiver.

Let $x(t) \ge 0$ represent the number of packets in the queue, including packets which will be retransmitted but whose NACKs have not yet been received, $0 \le s(t) \le n-1$ track the useful time slot, and $0 \le c(t) \le K$ represent the channel state. We assign the value for $s(t)$ as follows. If a transmission failure occurs in a useful time slot, $s(t)$ will be equal to $n-1$ in the next time slot. Then, it will be decreased during the subsequent slots until $s(t) = 0$, where a useful transmission period starts. This is illustrated in Fig. 3 for $n = 4$, where the evolution of $s(t)$ is shown. As is evident, the number of *useless* slots following transmission error(s) in a *useful* slot is $n-1$. It can be shown that the random process $X(t) = \{x(t), s(t), c(t)\}$ forms a discrete-time Markov chain (MC). For brevity, we will omit time index t in the related variables if it does not cause confusion.

In order to calculate the steady state probability for the underlying MC, it is important to put its transition probability matrix in a nice form where its specific transition structure can be exploited. Now, let (i, j, k) be the generic system state (i.e., $x = i$, $s = j$, and $c = k$) and $(i, j, k) \to (i', j', k')$ denote the system transition from state (i, j, k) to state (i', j', k'). For fixed i, the probabilities corresponding to system state transitions $(i, *, *) \to (i+1-l, *, *)$ can be written in a matrix block $\mathbf{D}_{i,l}$. We further put the probabilities of state transitions $(i, j, *) \to (i+1-l, j', *)$ into a submatrix $\mathbf{D}_{i,l}(j, j')$ of $\mathbf{D}_{i,l}$. Also, the probability of transition $(i, j, k) \to (i+1-l, j', k')$ is denoted by $\mathbf{D}_{i,l}(j, j')(k, k')$, which is an element of $\mathbf{D}_{i,l}(j, j')$. In fact, the probabilities corresponding to transitions from state (i, j, k) to any other state will be in the $(i(K+1)n + j(K+1) + k)$-th row of the probability transition matrix and they are elements of

$\mathbf{D}_{i,l}$ for some value of l depending on the destination state. The following example clarifies further how the system state transition probabilities are ordered in the matrix form.

Example: For ease of exposition, we consider a very simple case where there are 2 channel states (states 0, 1) and feedback delay is 2 slots (i.e., n = 2). The matrix block $\mathbf{D}_{i,l}$ can be expanded as in (3), shown at the bottom of this page. In this equation, element $\mathbf{D}_{i,l}(1,0)(1,0)$, for example, represents the probability of transition $(i,1,1) \rightarrow (i+1-l,0,0)$.

The resulting transition matrix for $X(t)$ is written in (4) for $L = 3$, as shown at the bottom of this page. Recall that L is the maximum number of packets which can be transmitted in one time slot (i.e., equal to h_K). In this transition matrix, $\mathbf{D}_{i,l}$ contains the probabilities of system transitions where $x = i$ before the transitions. All transitions captured by $\mathbf{D}_{i,l}$ for different l will be called transitions in level i of the transition matrix in the sequel. Note that, in the generic system state (i,j,k), j can have n possibilities and k can have $K+1$ possibilities (i.e., $K+1$ channel states). Thus, the order of $\mathbf{D}_{i,l}$ is $n(K+1)\times n(K+1)$. The derivations of matrix blocks $\mathbf{D}_{i,l}$ and $\mathbf{D}_l$ are detailed in **Appendix I**. As can be seen in Appendix I, for $i \geq L$, $\mathbf{D}_{i,l}$ is independent of the level index i; therefore, for brevity we denote $\mathbf{D}_{i,l}$ by $\mathbf{D}_l$ in (4).

Note that, in (4), there is at most one arriving packet and at most $L = 3$ packets successfully transmitted in one time slot. Therefore, for level $i \geq 3$, the transitions can go up at most one level (represented by $\mathbf{D}_0$) and go down at most three levels (represented by $\mathbf{D}_4$). The transition matrix in (4) describes a GI/M/1 Markov chain, where the solution can be found by the well-established method proposed by Neuts [16]. In fact, the steady-state probability $\mathbf{x} = [\mathbf{x}_0\, \mathbf{x}_1\, \mathbf{x}_2\, \cdots]$ satisfies

$$\mathbf{xP} = \mathbf{x}, \qquad \sum_{i=0}^{\infty} \mathbf{x}_i \mathbf{e} = 1 \tag{5}$$

where $\mathbf{e}$ is a column vector of all ones with the same dimension as $\mathbf{x}_i$ which is $n(K+1)$. We can find $\mathbf{x}_0$, $\mathbf{x}_1$, $\cdots$, $\mathbf{x}_L$ using the boundary and the normalization conditions. Other values of $\mathbf{x}_i$ $(i > L)$ can be calculated from $\mathbf{x}_L$ by using a non-negative matrix $\mathbf{R}$ as follows: $\mathbf{x}_i = \mathbf{x}_L \mathbf{R}^{i-L}$ [16]. Here, the order of matrix $\mathbf{R}$ is $n(K+1) \times n(K+1)$.

B. Delay Analysis

In this section, we derive the delay distribution of a packet arriving at the queue for the GBN ARQ protocol. The delay is the time for all packets ahead of the target packet (if any) and itself successfully leaving the queue. In the following calculation, the delay is considered at the transmitter side. Let the arrival slot be numbered as slot zero and it is not included in the delay calculation.

Now let $\mathbf{\Phi}_{(p,d)}$ be matrices with order $n(K+1)\times n(K+1)$ whose element $\left(\mathbf{\Phi}_{(p,d)}\right)$ $(j,j')(k,k')$ $(0 \leq j,j' \leq n-1,$ $0 \leq k,k' \leq K)$ is the probability of state transition $(p,j,k) \rightarrow (0,j',k')$ in d time slots. In short, $\mathbf{\Phi}_{(p,d)}$ contains system transition probabilities such that p packets are successfully transmitted in d slots. From the definition of $\mathbf{\Phi}_{(p,d)}$, $(\mathbf{\Phi}_{(p,d)})(j,j')$ contains the channel state transition probabilities such that s (in the system state (x,s,c)) evolves from j to j'. Thus, the order of $(\mathbf{\Phi}_{(p,d)})(j,j')$ is $(K+1)\times(K+1)$.

We also define $\mathbf{C}_{h,p}$ to be matrices of order $n(K+1)\times n(K+1)$ with the same structure as $\mathbf{\Phi}_{(p,d)}$ whose elements are the system transition probabilities such that h packets are successfully transmitted in one particular time slot given that there are p packets in the queue at the beginning of the time slot. The derivation of $\mathbf{C}_{h,p}$ is given in **Appendix I**. We have the following recursive relation:

$$\mathbf{\Phi}_{(p,d)} = \sum_{h=0}^{L} \mathbf{C}_{h,p} \mathbf{\Phi}_{(p-h,d-1)}, \quad \text{where} \quad \mathbf{\Phi}_{(0,0)} = \mathbf{I}_{n(K+1)}. \tag{6}$$

Equation (6) can be interpreted as follows. If there are p packets which must be delivered in d time slots (captured by $\mathbf{\Phi}_{(p,d)}$) and h packets are successfully transmitted in the first time slot (captured by $\mathbf{C}_{h,p}$), there are remaining $p-h$ packets to be delivered in $d-1$ slots (captured by $\mathbf{\Phi}_{(p-h,d-1)}$). Here, $\mathbf{\Phi}_{(0,0)}$ simply captures the end point where the target packet leaves the queue.

To calculate the delay statistics, we need to obtain the steady-state vector seen by an arriving packet to the queue. Note that we have assumed a Bernoulli arrival process so that the ASTA (arrivals see time averages) property holds here. Also, packets arriving to the queue after the tagged packet do not affect the delay experienced by the tagged packet so they are ignored in the following derivation. Let $\mathbf{y}_i$ be a

$$\mathbf{D}_{i,l} = \left[\begin{array}{c|c} \mathbf{D}_{i,l}(0,0) & \mathbf{D}_{i,l}(0,1) \\ \hline \mathbf{D}_{i,l}(1,0) & \mathbf{D}_{i,l}(1,1) \end{array}\right] = \left[\begin{array}{cc|cc} \mathbf{D}_{i,l}(0,0)(0,0) & \mathbf{D}_{i,l}(0,0)(0,1) & \mathbf{D}_{i,l}(0,1)(0,0) & \mathbf{D}_{i,l}(0,1)(0,1) \\ \mathbf{D}_{i,l}(0,0)(1,0) & \mathbf{D}_{i,l}(0,0)(1,1) & \mathbf{D}_{i,l}(0,1)(1,0) & \mathbf{D}_{i,l}(0,1)(1,1) \\ \hline \mathbf{D}_{i,l}(1,0)(0,0) & \mathbf{D}_{i,l}(1,0)(0,1) & \mathbf{D}_{i,l}(1,1)(0,0) & \mathbf{D}_{i,l}(1,1)(0,1) \\ \mathbf{D}_{i,l}(1,0)(1,0) & \mathbf{D}_{i,l}(1,0)(1,1) & \mathbf{D}_{i,l}(1,1)(1,0) & \mathbf{D}_{i,l}(1,1)(1,1) \end{array}\right] \tag{3}$$

$$\mathbf{P} = \begin{bmatrix} \mathbf{D}_{0,1} & \mathbf{D}_{0,0} & & & & & \\ \mathbf{D}_{1,2} & \mathbf{D}_{1,2} & \mathbf{D}_{1,0} & & & & \\ \mathbf{D}_{2,3} & \mathbf{D}_{2,2} & \mathbf{D}_{2,1} & \mathbf{D}_{2,0} & & & \\ \mathbf{D}_4 & \mathbf{D}_3 & \mathbf{D}_2 & \mathbf{D}_1 & \mathbf{D}_0 & & \\ & \mathbf{D}_4 & \mathbf{D}_3 & \mathbf{D}_2 & \mathbf{D}_1 & \mathbf{D}_0 & \\ & & & & \ddots & \ddots & \ddots \end{bmatrix}. \tag{4}$$

vector of dimension $n(K+1)$ which represents the system state probabilities where an arriving packet sees i head-of-line (HOL) packets at the end of its arrival time slot. We have

$$\mathbf{y}_i = \sum_{h=0}^{L} \mathbf{x}_{i+h}\mathbf{C}_{h,i+h}. \tag{7}$$

Equation (7) can be interpreted as follows. If there are $i+h$ packets in the queue at the beginning of the arrival time slot (captured by $\mathbf{x}_{i+h}$) and h packets successfully leave the queue in this time slot (captured by $\mathbf{C}_{h,i+h}$), the arriving packet will see exactly i HOL packets (captured in $\mathbf{y}_i$) at the end of this time slot. The probability that the delay is D slots (not including the arrival slot) can therefore be written as follows:

$$P_d(D) = \sum_{h=0}^{DL-1} \mathbf{y}_h \mathbf{\Phi}_{(h+1,D)} \mathbf{e}_{n(K+1)} \tag{8}$$

where $\mathbf{e}_{n(K+1)}$ is a column vector of all ones with dimension $n(K+1)$. The sum in (8) is limited to $DL-1$ since at most L packets can be successfully transmitted in one time slot.

IV. Queueing Model and Analysis for Selective Repeat ARQ

A. Queueing Model

Since the outcome of the decoding process for each packet only reaches the transmitter n slots after the beginning of the corresponding transmission slot and only erroneous packets are "selectively" retransmitted, we need to keep track of the number of erroneous packets in a window of n slots.

Let $\mathbf{b}(t) = [b_1(t), b_2(t), \cdots, b_n(t)]$ be an n-dimensional vector whose elements $b_i(t)$, $(i = 1, \cdots, n)$ represent the number of erroneous packets among those transmitted in the slot which is i slots before the current slot ($0 \le b_i(t) \le L$). Note that, we do not need to differentiate two different cases which can lead to $b_i(t) = 0$: no packet is transmitted because the channel is in state zero and all of the transmitted packets are successfully decoded at the receiver. This is due to the operation of the protocol where only erroneous packets are retransmitted if any.

We can observe that $\mathbf{b}(t+1) = [\beta, b_1(t), b_2(t), \cdots, b_{n-1}(t)]$, where β is the number of packets in error among those transmitted in time slot t. To facilitate the analysis, we represent vector $\mathbf{b}(t)$ by a number $y(t) = \sum_{i=1}^{n} b_i \times (L+1)^{(i-1)}$. Since $\mathbf{b}(t+1) = (\beta, b_1(t), b_2(t), \cdots, b_{n-1}(t))$, for a given $y(t) = k$ there are at most $L+1$ transitions to $y(t+1) = l$ corresponding to different values of β. Also note that there is a unique mapping between $y(t)$ and $\mathbf{b}(t)$; therefore, the use of $y(t)$ instead of $\mathbf{b}(t)$ makes the analysis easier without affecting the semantics of the problem.

Example: Consider the SR-ARQ protocol with feedback delay n = 4 (time slots) and $L = 4$. Suppose that the current vector $\mathbf{b}$ is $\mathbf{b}(t) = (b_1(t), b_2(t), b_3(t), b_4(t)) = (1, 0, 3, 4)$ and two packets among those transmitted in the current time slot are in error. The vector $\mathbf{b}$ in the next time slot will be $\mathbf{b}(t+1) = (2, 1, 0, 3)$. The corresponding transition is $\mathbf{b}(t) \to \mathbf{b}(t+1) \equiv \{y(t) = 576\} \to \{y(t+1) = 382\}$.

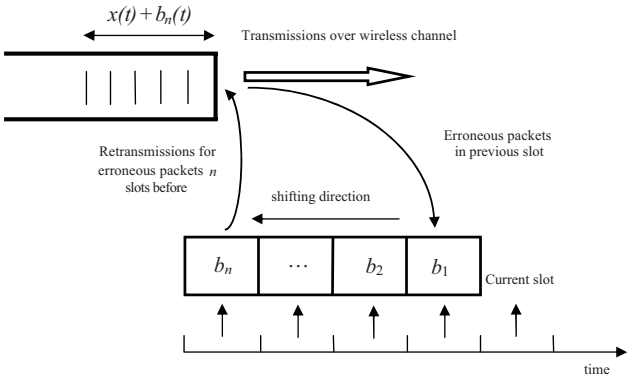

Fig. 4. The SR-ARQ model.

Let $x(t) \ge 0$ represent the number of packets in the queue excluding the packets which were transmitted and the transmitter is waiting for their ACKs/NACKs, $y(t)$ corresponding to vector $\mathbf{b}(t)$ capture the transmission outcomes in the past n time slots, and $0 \le c(t) \le K$ represent the channel state. Then, the random process $Y(t) = \{x(t), y(t), c(t)\}$ forms a discrete-time Markov chain.

At any time slot t, the number of packets available for transmission is $x(t) + b_n(t)$, where $b_n(t)$ is the number of erroneous packets which were transmitted n slots before the current slot and are being retransmitted. The number of packets transmitted at time t is given by $\min\{h_{c(t)}, x(t) + b_n(t)\}$, which is the minimum of the number of available packets in the queue (equal to $x(t) + b_n(t)$) and the transmission capability of channel state $c(t)$ (equal to $h_{c(t)}$). If $a(t)$ denotes the number of arriving packets in time slot t, we have $x(t+1) = x(t) + a(t) + b_n(t) - \min\{h_{c(t)}, x(t) + b_n(t)\}$. The protocol modeling is illustrated in Fig. 4. We will omit time index t in the related variables if it does not cause confusion in the sequel. Note that, element b_i of vector $\mathbf{b}$ stores the number of packets transmitted erroneously in the past regardless of how many packets were really transmitted in the corresponding time slots.

Similar to the model for GBN-ARQ protocol, we put the transition probability matrix of $Y(t)$ in a matrix form. Now, let (i, j, k) be the generic system state and $(i, j, k) \to (i', j', k')$ denote the system transition from state (i, j, k) to state (i', j', k'). For fixed i, we write the probabilities of system transitions $(i, *, *) \to (i + K + 1 - l, *, *)$ in a matrix block $\mathbf{A}_{i,l}$. In the generic system state, j has $(L+1)^n$ possibilities (because each element b_i of vector $\mathbf{b}$ has $L+1$ possibilities) and k has $K+1$ possibilities (i.e., $K+1$ channel states); therefore, the order of $\mathbf{A}_{i,l}$ is $(K+1)(L+1)^n \times (K+1)(L+1)^n$.

The resulting transition matrix for $Y(t)$ is written in (9) for $L = 3$, shown at the bottom of the next page. As before, each block of rows in (9) captures the transitions in one level of the transition matrix. Also, the elements of $\mathbf{A}_{i,l}$ are $(\mathbf{A}_{i,l})(j, j')(k, k')$, which is the probability of system transition $(i, j, k) \to (i+L+1-l, j', k')$. The derivations of matrix blocks $\mathbf{A}_{i,l}$ and $\mathbf{A}_l$ are given in **Appendix II**. Note that, for $i \ge L$, we denote $\mathbf{A}_{i,l}$ by $\mathbf{A}_l$ for brevity because these matrix blocks are independent of the level index. Since we have

$x(t+1)-x(t) = a(t)+b_n(t)-\min\{h_{c(t)}, x(t)+b_n(t)\}$, each level in the transition matrix (9) can go up at most $L+1$ levels (when $a(t)=1$, $b_n(t)=L$, and $\min\{h_{c(t)}, x(t)+b_n(t)\} = 0$) and go down at most L levels (when $a(t)=0$, $b_n(t)=0$, and $\min\{h_{c(t)}, x(t)+b_n(t)\} = L$).

B. Steady-State Solution

We re-block the transition matrix as indicated in (9) to obtain a quasi-birth and death (QBD) process, and the re-blocked transition matrix can be written as follows:

$$\mathbf{P} = \begin{bmatrix} \mathbf{B}_{0,1} & \mathbf{B}_{0,0} & & & & \\ \mathbf{B}_{1,2} & \mathbf{B}_{1,1} & \mathbf{B}_{1,0} & & & \\ & \mathbf{B}_2 & \mathbf{B}_1 & \mathbf{B}_0 & & \\ & & \mathbf{B}_2 & \mathbf{B}_1 & \mathbf{B}_0 & \\ & & & \ddots & \ddots & \ddots \end{bmatrix}. \tag{10}$$

The solution for the QBD process can be found by the well-established method proposed by Neuts [16]. Let $\pi = [\pi_0, \pi_1, \pi_2, \cdots]$ be the steady state probability vector of (10) and $\mathbf{z} = [\mathbf{z}_0, \mathbf{z}_1, \mathbf{z}_2, \cdots]$ be the steady state probability vector of the original transition matrix (9) where $\mathbf{z}_i$ corresponds to level i of the transition matrix (9). Recall that we have combined $L+1$ blocks in each dimension of the transition matrix (9) to obtain (10). Thus, letting $N_1 = (K+1)(L+1)^n$, the dimension of $\mathbf{z}_i$ ($i \geq 0$) is N_1 and the dimension of π_i ($i \geq 1$) is $N_1(L+1)$. For $i \geq 1$, π_i can be written as $\pi_i = [\pi_{i,1}, \pi_{i,2}, \cdots, \pi_{i,L+1}]$. Hence, we have

$$\mathbf{z}_0 = \pi_0, \quad \mathbf{z}_{(i-1)(L+1)+j} = \pi_{i,j} \tag{11}$$

for $i \geq 1$ and $1 \leq j \leq L+1$.

C. Delay Analysis

In this section, we derive the delay distribution of a packet arriving at the queue for the SR-ARQ protocol. Again, the delay is considered at the transmitter side. Let the arrival slot be numbered as slot zero and it is not included in the delay calculation.

Let $N = (L+1)^n - 1$. We define the following matrices:

- $\mathbf{\Omega}_{(p,d)}$ are matrices of order $N_1 \times N_1$ in which element $\left(\mathbf{\Omega}_{(p,d)}\right)(j,j')(k,k')$, $(0 \leq j, j' \leq N, 0 \leq k, k' \leq K)$ is the probability of system transition $(p,j,k) \rightarrow (0,j',k')$ in d time slots. In short, $\mathbf{\Omega}_{(p,d)}$ contains system transition probabilities such that in addition to the p packets captured in $x(t)$ all erroneous packets in the past n time slots are successfully transmitted in d time slots.
- $\mathbf{\Theta}_{(p,h)}$ are matrices of order $N_1 \times N_1$ in which element $\left(\mathbf{\Theta}_{(p,h)}\right)(j,j')(k,k')$, $(0 \leq j, j' \leq N, 0 \leq k, k' \leq K)$ is the probability of system transition $(p,j,k) \rightarrow (p-h,j',k')$ in one time slot. The derivations of $\mathbf{\Theta}_{(p,h)}$ are given in **Appendix III**.

We have the following recursive relation:

$$\mathbf{\Omega}_{(p,d)} = \sum_{h=-L}^{L} \mathbf{\Theta}_{(p,h)} \mathbf{\Omega}_{(p-h,d-1)} \tag{12}$$

where

$$\mathbf{\Omega}_{(0,0)} = \begin{bmatrix} \mathbf{I}_{K+1} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

In (12), $\mathbf{\Omega}_{(p,d)}$ captures the system evolution from time slot d to time slot $d-1$ counting back from the last time slot. Also $\mathbf{\Omega}_{(0,0)}$ simply captures the ending point where all HOL packets (if any), erroneous packets in a window of n slots and the target packet have successfully left the queue. Note that, due to the special structure of $\mathbf{\Omega}_{(0,0)}$, only transitions for which all remaining packets successfully leave the system in the last time slot are allowed to occur. This is a bit different from the corresponding transitions captured in (6) for the GBN-ARQ protocol.

To calculate the delay statistics, we need to obtain the steady-state vector seen by an arriving packet to the queue. Let $\mathbf{w}_i$ be a vector of dimension N_1 which represents the system state probability at the end of the arrival slot where an arriving packet sees i HOL packets in the queue. Then, we have

$$\mathbf{w}_i = \sum_{h=-L}^{L} \mathbf{z}_{i+h} \mathbf{\Theta}_{(i+h,h)} \tag{13}$$

where $\mathbf{z}_{i+h} = \mathbf{0}$, and $\mathbf{\Theta}_{(i+h,h)} = \mathbf{0}$ if $i+h < 0$.

The interpretations of (12) and (13) are similar to those of (6) and (7). The probability that the delay is D time slots (not including the arrival slot) can, therefore, be written as follows:

$$P_d(D) = \sum_{h=0}^{DL-1} \mathbf{w}_h \mathbf{\Omega}_{(h+1,D)} \mathbf{e}_{N_1} \tag{14}$$

$$\mathbf{P} = \left[\begin{array}{ccccc|cccc|cccc} \mathbf{A}_{0,4} & \mathbf{A}_{0,3} & \mathbf{A}_{0,2} & \mathbf{A}_{0,1} & \mathbf{A}_{0,0} & & & & & & & & \\ \mathbf{A}_{1,5} & \mathbf{A}_{1,4} & \mathbf{A}_{1,3} & \mathbf{A}_{1,2} & \mathbf{A}_{1,1} & \mathbf{A}_{1,0} & & & & & & & \\ \mathbf{A}_{2,6} & \mathbf{A}_{2,5} & \mathbf{A}_{2,4} & \mathbf{A}_{2,3} & \mathbf{A}_{2,2} & \mathbf{A}_{2,1} & \mathbf{A}_{2,0} & & & & & & \\ \mathbf{A}_7 & \mathbf{A}_6 & \mathbf{A}_5 & \mathbf{A}_4 & \mathbf{A}_3 & \mathbf{A}_2 & \mathbf{A}_1 & \mathbf{A}_0 & & & & & \\ \mathbf{0} & \mathbf{A}_7 & \mathbf{A}_6 & \mathbf{A}_5 & \mathbf{A}_4 & \mathbf{A}_3 & \mathbf{A}_2 & \mathbf{A}_1 & \mathbf{A}_0 & & & & \\ \hline & & \mathbf{A}_7 & \mathbf{A}_6 & \mathbf{A}_5 & \mathbf{A}_4 & \mathbf{A}_3 & \mathbf{A}_2 & \mathbf{A}_1 & \mathbf{A}_0 & & & \\ & & & \mathbf{A}_7 & \mathbf{A}_6 & \mathbf{A}_5 & \mathbf{A}_4 & \mathbf{A}_3 & \mathbf{A}_2 & \mathbf{A}_1 & \mathbf{A}_0 & & \\ & & & & \mathbf{A}_7 & \mathbf{A}_6 & \mathbf{A}_5 & \mathbf{A}_4 & \mathbf{A}_3 & \mathbf{A}_2 & \mathbf{A}_1 & \mathbf{A}_0 & \\ & & & & & \mathbf{A}_7 & \mathbf{A}_6 & \mathbf{A}_5 & \mathbf{A}_4 & \mathbf{A}_3 & \mathbf{A}_2 & \mathbf{A}_1 & \mathbf{A}_0 \\ \hline & & & & & & & & & \ddots & \ddots & \ddots & \ddots \end{array}\right]. \tag{9}$$

TABLE I
FSMC PARAMETERS (m = 1, f_d = 20 Hz, Average SNR = 10 dB, P_0 = 0.1)

State i	X_{i+1} (dB)	$\mathbf{T}_{i,i-1}$	$\mathbf{T}_{i,i}$	$\mathbf{T}_{i,i+1}$
0	7.1517	0.0000	0.9469	0.0531
1	9.9712	0.0956	0.8221	0.0823
2	15.2475	0.0552	0.9352	0.0096
3	17.5523	0.1015	0.8858	0.0127
4	21.8868	0.1196	0.8804	0.0000
5	∞	0.1970	0.8030	0.0000

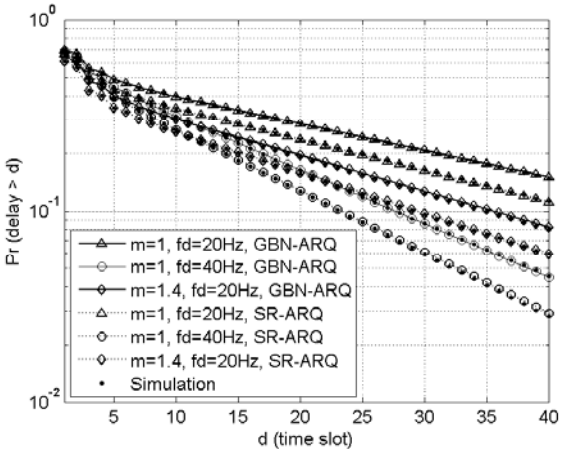

Fig. 5. Complementary cumulative delay distributions (for arrival rate λ = 0.2, feedback delay n = 2, average SNR = 10 dB, P_0 = 0.3, m = 1, 1.4, and Doppler shift f_d = 20, 40 Hz).

where $\mathbf{e}_{N_1}$ is a column vector of all ones with dimension N_1. The summation above contains DL terms, since at most L packets can be successfully transmitted in one time slot.

V. MODEL VALIDATION AND NUMERICAL RESULTS

We use the PER fitting values of a_k and g_k in Table I of [15] to obtain the SNR thresholds of the FSMC model such that $\overline{\text{PER}}_k = P_0$ for all the transmission modes, where P_0 is a certain target packet error rate. A wireless system using adaptive modulation is considered where $h_k = k$ (i.e., the transmitter transmits k packets/slot in channel state k). To save simulation time in validating the analytical model, we consider three transmission modes (i.e., K = 3) in Fig. 5, while for the other results we assume five transmission modes (i.e., K = 5). We assume that the time slot interval T_s = 1 ms. The SNR thresholds and transition probabilities for the FSMC model are summarized in Table I. The complementary delay distributions (i.e., $\Pr(\text{delay} > d) = 1 - \sum_{k=1}^{d} P_d(k)$) obtained from both the simulation and the analytical model for both GBN-ARQ and SR-ARQ are shown in Fig. 5.

The simulations are done by using a discrete-event simulator. For the given channel and system parameters, we calculate the channel transition probability matrix $\mathbf{T}$. In each time slot, the channel state is randomly generated according to the transition matrix $\mathbf{T}$, which determines the maximum number of packets that can be transmitted. For a given transmission error probability P_0, the number of packets reaching the receiver is determined by the corresponding probability and the ARQ protocol rule without introducing any approximation. As can be seen, the analytical results follow the simulation results very closely. We emphasize that the delay statistics for GBN-ARQ and SR-ARQ obtained here is for a very general system model which takes the dynamic radio link adaptation into account, and therefore, is more generally applicable compared to those obtained for a two-state Markov channel.

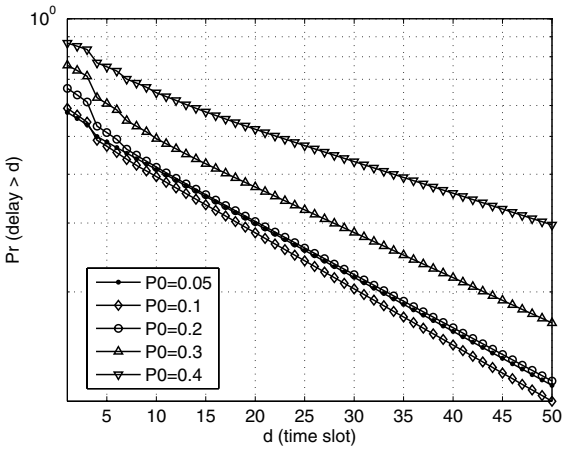

Fig. 6. Complementary cumulative delay distribution under different target packet error rate P_0 (for arrival rate λ = 0.2, feedback delay n = 3, average SNR = 10 dB, m = 1, and Doppler shift f_d = 20 Hz).

We observe that the higher the value of the Nakagami parameter m and/or the Doppler shift f_d, the smaller the delay. The analytical model thus enables us to analyze the impact of channel parameters on the delay performance. In fact, most data applications have certain delay limits such that packets arriving after the delay limit would be useless.

An important issue in designing dynamic link adaptation mechanisms is the selection of the mode switching SNR thresholds for the different transmission modes (i.e., the SNR thresholds X_k, $k = 1, 2, \cdots, K$, presented in Section II-B). Specifically, we have obtained the SNR thresholds such that the average packet error rate for all modes is equal to some target packet error rate P_0 (i.e., $\overline{\text{PER}}_k = P_0$). Different values of P_0 result in different sets of SNR thresholds for the AMC at the physical layer. Variations in the complementary cumulative delay distributions for GBN-ARQ with different values of P_0 are illustrated in Fig. 6. The lowest delay is obtained for $P_0 = 0.1$ in this case. Basically, for higher values of P_0, the average transmission rate increases but the wireless link becomes less reliable. In other words, we are more likely to use high transmission modes by choosing large P_0. However, the high probability of transmission errors may require many retransmissions, which may increase the link level delay. Thus, there exists a value of P_0 for which the link level delay is minimized. Similar trends can also be observed for SR-ARQ under different channel and system parameters. We have observed that the best delay distribution can be obtained when P_0 is in the range of 0.1-0.2, and is quite insensitive to the other system and traffic parameters (e.g., m, f_d and λ).

For dynamic link adaptation, the SNR thresholds for the different transmission modes are usually chosen based on the achieved physical layer throughput taking the packet error rate for each transmission mode into account [1]. In particular, the

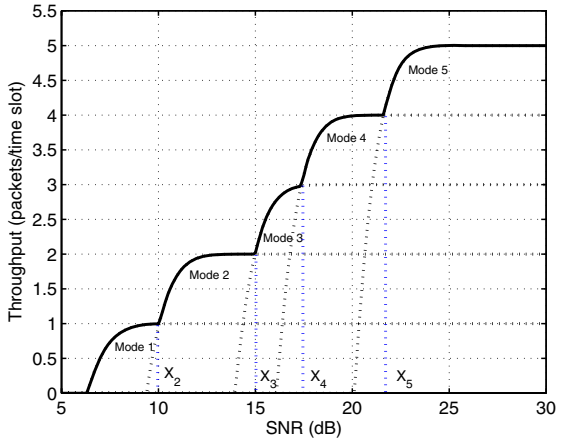

Fig. 7. Throughput for different transmission modes under varying SNR.

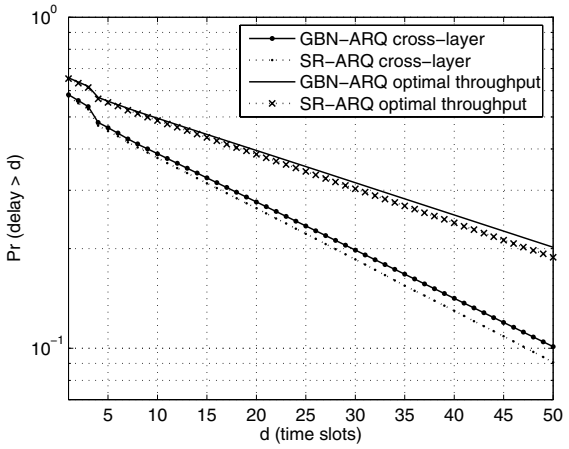

Fig. 8. Complementary cumulative delay distribution under GBN-ARQ and SR-ARQ for two mode switching threshold designs (for arrival rate λ = 0.2, feedback delay n = 3, average SNR = 10 dB, P_0 = 0.1, m = 1, and Doppler shift f_d = 20 Hz).

transmitter transmits h_k packets in one time slot when the channel is in state k. Then the throughput (in packets/slot) is $h_k(1 - \mathrm{PER}_k(X))$ when the received SNR is X and the transmission mode is k. In Fig. 7, we plot the variations in throughput with received SNR for five transmission modes using adaptive modulation without coding. For this traditional SNR threshold design, the transmission mode which achieves the highest throughput will be chosen for each SNR value. For throughput-based link adaptation, the SNR thresholds for the different transmission modes are, therefore, determined by the intersection of these throughput curves as shown in this figure.

We now compare the delay performance obtained by the traditional SNR threshold calculation (as shown in Fig. 7) and our SNR threshold calculation, which is done such that the average packet error rate is equal to a certain target PER P_0 for all transmission modes. We denote the results obtained from the traditional SNR threshold calculation by "optimal throughput" and denote the results obtained from our SNR threshold calculation by "cross-layer" in Fig. 8 for both GBN-ARQ and SR-ARQ. Evidently, the delay obtained from "cross-layer" calculation is significantly lower than that obtained from "optimal throughput" calculation for both the ARQ protocols. This gain in delay performance is achieved simply by choosing the proper SNR thresholds without increasing the system complexity.

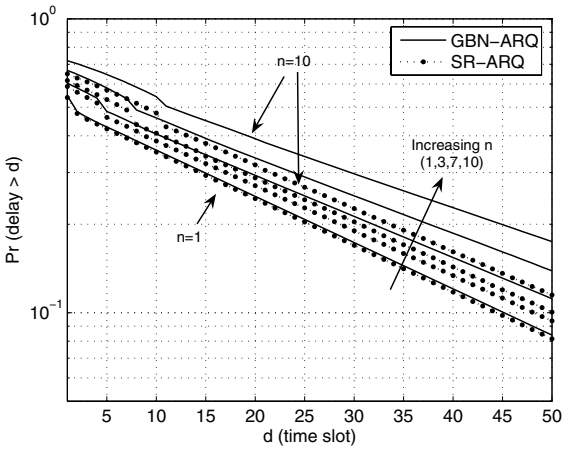

Fig. 9. Complementary cumulative delay distribution under different feedback delay n (for arrival rate λ = 0.2, average SNR = 10 dB, P_0 = 0.1, m = 1, and Doppler shift f_d = 20 Hz).

Typical variations in the complementary cumulative delay distributions for both GBN-ARQ and SR-ARQ are shown in Fig. 9 for different values of the feedback delay n. With the mode switching thresholds chosen such that the average packet error rate is equal to 0.1 for all transmission modes, the delay of GBN-ARQ is significantly larger than that of SR-ARQ only when n is large enough (e.g., $n \geq 10$), which justifies the use of SR-ARQ over GBN-ARQ in this region. Of course, under different choices of the SNR switching thresholds for the AMC, the average PER may be larger than 0.1 and the delay gap between these two protocols may be larger. However, such design is not optimal from the delay point of view as is evident in Fig. 6. Also, the delay-optimal design of the system is important because delay is the ultimate QoS perceived by some of the data applications.

For a particular wireless system, we can quantify the feedback delay, which is the sum of the propagation delay and the processing delay. Therefore, the performance gap between GBN-ARQ and SR-ARQ can be determined exactly under certain system and channel parameter settings. This would enable us to decide which ARQ protocol to implement at the link layer considering the tradeoff between delay performance and system complexity. Also, the obtained delay statistics can be used to perform packet level admission control under statistical delay constraints. For example, for a given maximum delay $D_{\max}$ and delay outage probability P_t, via a simple search, we can find the maximum arrival probability $\lambda_{\max}$ such that the condition $\Pr\{\text{delay} > D_{\max}\} < P_t$ is satisfied.

VI. Conclusion

For a multi-rate wireless network, queueing models for GBN-ARQ and SR-ARQ have been developed under non-instantaneous feedback. The presented models remove the drawback of some of the existing works on analysis of

ARQ protocols which do not capture the multi-rate feature at the physical layer. The radio link layer delay statistics thus obtained for a very general system setup enable radio link protocol design under statistical delay constraints. Also, it is useful in designing dynamic radio link adaptation thresholds based on delay performance (in contrast to the traditional methods based on throughput performance). The obtained delay statistics would be very useful to quantify the tradeoff between delay performance and implementation complexity between these two ARQ protocols.

APPENDIX I
DERIVATIONS OF MATRIX BLOCKS IN (4)

We derive the matrix blocks for the transition matrix (4) in this Appendix. The number of packets transmitted in time slot t is the minimum of the number of packets available in the queue and the transmission capability (i.e., equal to $\min\{x(t), h_{c(t)}\}$). Let $a(t)$ be the number of arriving packets during slot t, and $d(t)$ be the number of packets which will not be retransmitted due to transmissions in slot t (in fact, $d(t) \leq \min\{x(t), h_{c(t)}\}$), we have $x(t+1) = x(t) + a(t) - d(t)$.

To derive the matrix blocks $\mathbf{D}_{i,l}$ and $\mathbf{D}_l$ in (4), we consider the following cases which may occur in each time slot. First, the slot is not useful (i.e., $s(t) \neq 0$), and therefore, no packet can depart. We need to keep track of the channel state evolution only for this case. Second, the slot is useful (i.e., $s(t) = 0$) and all transmitted packets are received correctly. In this case, the number of packets in the queue changes according to the number of successfully transmitted packets and the number of arriving packets in that slot. Also, the next time slot will be a useful one (i.e., $s(t+1) = 0$). Third, the slot is useful (i.e., $s(t) = 0$) and there exists at least one packet among those transmitted in error. The number of packets in the queue at the end of the slot depends on the error pattern and the number of arriving packets in that slot. However, the next time slot will not be useful (in fact, $s(t+1) = n-1$).

Now let us define the following matrices:

- $\mathbf{T}_k$ $(k = 0, 1, \cdots, K)$ are constructed by keeping only the $(k+1)$-st row of the channel transition probability matrix $\mathbf{T}$ and setting all other rows to $\mathbf{0}$. These matrices capture the case the channel is in state k at the beginning of a particular time slot.
- $\mathbf{\Psi}_{i,j}^{(0)}$ are matrices of order $(K+1) \times (K+1)$ in which element $\mathbf{\Psi}_{i,j}^{(0)}(k, k')$ is the probability that all i transmitted packets are received correctly given that there were j packets in the queue before transmission (i.e., $i = \min\{x(t), h_{c(t)}\} = \min\{j, h_{c(t)}\}$), the channel changes from state k to state k' in the transmission slot.
- $\mathbf{\Psi}_{i,j}^{(1)}$ are matrices of order $(K+1) \times (K+1)$ whose element $\mathbf{\Psi}_{i,j}^{(1)}(k, k')$ is the probability that i transmitted packets are received correctly given that there were j packets in the queue before transmission, there are at least one erroneous packet (i.e., $i < \min\{x(t), h_{c(t)}\} = \min\{j, h_{c(t)}\}$), and the channel changes from state k to state k' in the transmission slot.
- $\mathbf{C}_{i,j}^{(k)}$ $(k = 1, 2, 3)$ are matrices of order $n(K+1) \times n(K+1)$ representing the aforementioned three cases, respectively, whose element $\mathbf{C}_{i,j}^{(k)}(l, l')(h, h')$ $(0 \leq l, l' \leq n-1,\ 0 \leq h, h' \leq K)$ represents the probability of system transition $(j, l, h) \to (j-i, l', h')$ (i.e., i packets are successfully transmitted given there were j in the queue before transmissions).

From foregoing definitions, $\mathbf{C}_{i,j}^{(k)}$ can be written as follows:

$$\mathbf{C}_{i,j}^{(1)} = \mathbf{0} \quad \text{if } i \neq 0 \tag{15}$$

$$\mathbf{C}_{0,j}^{(1)} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{T} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{T} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{T} & \mathbf{0} \end{bmatrix} \tag{16}$$

$$\mathbf{C}_{i,j}^{(2)} = \begin{bmatrix} \mathbf{\Psi}_{i,j}^{(0)} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \end{bmatrix} \tag{17}$$

$$\mathbf{C}_{i,j}^{(3)} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{\Psi}_{i,j}^{(1)} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \end{bmatrix}. \tag{18}$$

Equation (16) simply captures the channel state transitions where $s(t) \neq 0$. Note that, $s(t)$ decreases by one in each time slot which explains the structure of $\mathbf{C}_{0,j}^{(1)}$ (i.e., $\mathbf{T}$ is in positions $(s, s-1)$). For $s(t) \neq 0$, no packet can depart; therefore, we have $\mathbf{C}_{i,j}^{(1)} = \mathbf{0}$ for $i \neq 0$. In (17), $\mathbf{C}_{i,j}^{(2)}$ contain the probabilities of transition between two useful slots (i.e., $s(t) = s(t+1) = 0$), and $\mathbf{C}_{i,j}^{(3)}$ contains the probabilities of transition from a useful time slot to a useless one (i.e., $s(t) = 0$ and $s(t+1) = n-1$), where at least one transmission error must have occurred. This explains the position of $\mathbf{\Psi}_{i,j}^{(0)}$ and $\mathbf{\Psi}_{i,j}^{(1)}$ in the matrices $\mathbf{C}_{i,j}^{(2)}$ and $\mathbf{C}_{i,j}^{(3)}$, respectively.

Before we derive the matrix blocks in (4), let $\mathbf{C}_{i,j} = \sum_{k=1}^{k=3} \mathbf{C}_{i,j}^{(k)}$, which contains the probabilities that i packets are successfully transmitted given that there were j packets in the queue before transmission without distinguishing the aforementioned three cases. As we discussed above, if $d(t)$ is the number of packets which will not be retransmitted due to transmissions in slot t, we have $x(t+1) = x(t) + a(t) - d(t)$. It can be easily seen that $\mathbf{D}_{i,l}$ and $\mathbf{D}_l$ contains the probabilities of system state transition where $x(t+1) = x(t) + 1 - l$. Thus, we have $a(t) - d(t) = 1 - l$. As a result, $d(t) = l - 1$ if $a(t) = 0$ (i.e., no arrival) and $d(t) = l$ if $a(t) = 1$ (i.e., one arrival). Therefore, $\mathbf{D}_{i,l}$ can be calculated as follows:

$$\mathbf{D}_{i,l} = (1-\lambda)\mathbf{C}_{l-1,i} + \lambda \mathbf{C}_{l,i}. \tag{19}$$

Similarly, $\mathbf{D}_l$ can be calculated as

$$\mathbf{D}_l = (1-\lambda)\mathbf{C}_{l-1,L} + \lambda \mathbf{C}_{l,L}. \tag{20}$$

It can be easily observed that $\mathbf{C}_{i,j} = \mathbf{C}_{i,L}$ for $j \geq L$.

The remaining task is to determine $\mathbf{\Psi}_{i,j}^{(0)}$ and $\mathbf{\Psi}_{i,j}^{(1)}$, which is pursued now. Let $\theta_k = \overline{\text{PER}}_k$ be the probability of transmission error when the channel is in state k. Let us

assume that the transmission outcomes of different packets are independent and let us define

$$p_i^{(k)} = (1-\theta_k)^i . \tag{21}$$

Then, for $i, j > 0$, $\mathbf{\Psi}_{i,j}^{(0)}$ can be calculated as

$$\mathbf{\Psi}_{i,j}^{(0)} = \begin{cases} p_i^{(k)}\mathbf{T}_k, & i < j,\ i = h_k \\ \mathbf{0}, & i < j,\ \text{if } \exists!\ k \text{ s.t. } i = h_k \\ \sum_{h=k}^{h=K} p_i^{(h)}\mathbf{T}_h, & i = j,\ \ k = \min\{c(t) : h_k \geq j\} \end{cases} . \tag{22}$$

For $i = 0$, $\mathbf{\Psi}_{i,j}^{(0)}$ can be calculated as

$$\mathbf{\Psi}_{0,0}^{(0)} = \mathbf{T}, \quad \mathbf{\Psi}_{0,j}^{(0)} = \mathbf{T}_0, \quad j > 0. \tag{23}$$

Also, $\mathbf{\Psi}_{i,j}^{(1)}$ can be calculated as

$$\mathbf{\Psi}_{i,j}^{(1)} = \sum_{m=k}^{K} p_i^{(m)}\theta_m \mathbf{T}_m \tag{24}$$

where $k = \min\{c(t) : h_k > i\}$.

Equations (22)-(24) can be interpreted as follows. In (22), we must have $i = \min\{j, h_{c(t)}\}$; therefore, $h_{c(t)} = i$ if $j > i$ or $h_{c(t)} \geq i$ if $i = j$. For $\mathbf{\Psi}_{0,0}^{(0)}$, there is no transmission since there is no packet in the queue at the beginning of the slot. Therefore, the channel can be in any state without introducing any transmission error. For $\mathbf{\Psi}_{0,j}^{(0)}$, there are $j > 0$ packets in the queue in this case, and no transmission error occurs only if the channel is in state 0 (since no transmission is allowed in this channel state). The interpretation for (24) is similar but at least one packet among those transmitted must be in error (i.e., $i < \min\{j, h_{c(t)}\}$). Note that in this case the first i packets are received correctly and the $(i+1)$-st packet must be in error.

APPENDIX II
DERIVATIONS OF MATRIX BLOCKS IN (9)

Due to the structure of the system state, we can write $\mathbf{A}_l$ and $\mathbf{A}_{i,l}$ as

$$\mathbf{A}_l = \begin{bmatrix} \mathbf{A}_l(0,0) & \mathbf{A}_l(0,1) & \cdots & \cdots & \mathbf{A}_l(0,N) \\ \mathbf{A}_l(1,0) & \mathbf{A}_l(1,1) & \cdots & \cdots & \mathbf{A}_l(1,N) \\ & \cdots & \cdots & \cdots & \\ \mathbf{A}_l(N,0) & \mathbf{A}_l(N,1) & \cdots & \cdots & \mathbf{A}_l(N,N) \end{bmatrix} \tag{25}$$

$$\mathbf{A}_{i,l} = \begin{bmatrix} \mathbf{A}_{i,l}(0,0) & \mathbf{A}_{i,l}(0,1) & \cdots & \cdots & \mathbf{A}_{i,l}(0,N) \\ \mathbf{A}_{i,l}(1,0) & \mathbf{A}_{i,l}(1,1) & \cdots & \cdots & \mathbf{A}_{i,l}(1,N) \\ & \cdots & \cdots & \cdots & \\ \mathbf{A}_{i,l}(N,0) & \mathbf{A}_{i,l}(N,1) & \cdots & \cdots & \mathbf{A}_{i,l}(N,N) \end{bmatrix} \tag{26}$$

where we have defined $N = (L+1)^n - 1$, sub-matrices $\mathbf{A}_{i,l}(j,j')$ contains the probabilities of system transitions $(i,j,*) \rightarrow (i+K+1-l, j', *)$ and $\mathbf{A}_l(j,j')$ contains the probabilities of the same system transitions for $i \geq L$.

If $\theta_k = \overline{\text{PER}}_k$ is the probability of packet transmission error when the channel is in state k, the probability that i packets are received in error given that j packets were transmitted when the channel is in state k can be written as follows:

$$q_{i,j}^{(k)} = \binom{j}{i} \theta_k^i (1-\theta_k)^{j-i} . \tag{27}$$

Now we show how to calculate $\mathbf{A}_l(j,j')$. It can be checked that $\mathbf{A}_l$ contains the probabilities of system transitions where $x(t+1) - x(t) = L + 1 - l$. Thus we have $x(t+1) - x(t) = L + 1 - l = a(t) + b_n(t) - \min\{h_{c(t)}, x(t) + b_n(t)\} = a(t) + b_n(t) - h_{c(t)}$ since $\min\{h_{c(t)}, x(t) + b_n(t)\} = h_{c(t)}$. Therefore, $h_{c(t)} = a(t) + b_n(t) + l - L - 1$. For given l, j, j', we can find $c(t) = c_1$ for $a(t) = 0$ (no arrival) and $c(t) = c_2$ for $a(t) = 1$ (one arrival) from this relation if they exist. Hence, we have

$$\mathbf{A}_l(j,j') = (1-\lambda) q_{\beta,c_1}^{(c_1)} \mathbf{T}_{c_1} + \lambda q_{\beta,c_2}^{(c_2)} \mathbf{T}_{c_2} \tag{28}$$

where $\mathbf{T}_k$ was defined in **Appendix I**; $b_n(t)$ and β can be found from the mapping $\mathbf{b}(t) \rightarrow \mathbf{b}(t+1) \equiv \{y(t) = j\} \rightarrow \{y(t+1) = j'\}$, with $\mathbf{b}(t) = [b_1(t), b_2(t), \cdots, b_n(t)]$ and $\mathbf{b}(t+1) = [\beta, b_1(t), b_2(t), \cdots, \ b_{n-1}(t)]$.

Similarly, to calculate $\mathbf{A}_{i,l}(j,j')$ we find $c(t)$ from the relation $x(t+1) - x(t) = L + 1 - l = a(t) + b_n(t) - \min\{h_{c(t)}, x(t) + b_n(t)\} = a(t) + b_n(t) - \min\{h_{c(t)}, i + b_n(t)\}$. In this case, we may find more than one $c(t)$ from this relation for the case $a(t) = 0$ (no arrival) and $a(t) = 1$ (one arrival), which are denoted as c_3 and c_4 in the following sums for these two cases, respectively. Hence, we have

$$\mathbf{A}_{i,l}(j,j') = (1-\lambda) \sum_{c_3} q_{\beta,d_1}^{(c_3)} \mathbf{T}_{c_3} + \lambda \sum_{c_4} q_{\beta,d_2}^{(c_4)} \mathbf{T}_{c_4} \tag{29}$$

where $d_1 = \min\{h_{c_3}, i + b_n(t)\}$, $d_2 = \min\{h_{c_4}, i + b_n(t)\}$. Again, $b_n(t)$ and β can be found from the mapping $\mathbf{b}(t) \rightarrow \mathbf{b}(t+1) \equiv \{y(t) = j\} \rightarrow \{y(t+1) = j'\}$.

APPENDIX III
DERIVATIONS OF $\mathbf{\Theta}_{(p,h)}$

We can rewrite $\mathbf{\Theta}_{(p,h)}$ as follows:

$$\mathbf{\Theta}_{(p,h)} = \begin{bmatrix} \mathbf{\Theta}_{(p,h)}(0,0) & \mathbf{\Theta}_{(p,h)}(0,1) & \cdots & \mathbf{\Theta}_{(p,h)}(0,N) \\ \mathbf{\Theta}_{(p,h)}(1,0) & \mathbf{\Theta}_{(p,h)}(1,1) & \cdots & \mathbf{\Theta}_{(p,h)}(1,N) \\ & \cdots & \cdots & \\ \mathbf{\Theta}_{(p,h)}(N,0) & \mathbf{\Theta}_{(p,h)}(N,1) & \cdots & \mathbf{\Theta}_{(p,h)}(N,N) \end{bmatrix} \tag{30}$$

where $\left(\mathbf{\Theta}_{(p,h)}\right)(j,j')$ contains the probabilities of system transitions $(p,j,*) \rightarrow (p-h, j', *)$.

For transitions whose probabilities are captured in $\mathbf{\Theta}_{(p,h)}$, we have $x(t+1) - x(t) = -h = b_n(t) - \min\{h_{c(t)}, x(t) + b_n(t)\}$. Note that, packets arriving at the queue after the target arriving packet do not affect the delay experienced by the target packet; therefore, we let $a(t) = 0$ in this relation. We may find more than one $c(t)$ from this relation which are denoted by c_5 in the following sum. Hence, we have

$$\mathbf{\Theta}_{(p,h)}(j,j') = \sum_{c_5} q_{\beta,d_3}^{(c_5)} \mathbf{T}_{c_5} \tag{31}$$

where $d_3 = \min\{h_{c_5}, p + b_n(t)\}$; $b_n(t)$ and β can be calculated from the mapping $\mathbf{b}(t) \rightarrow \mathbf{b}(t+1) \equiv \{y(t) = j\} \rightarrow \{y(t+1) = j'\}$.

REFERENCES

[1] S. Catreux, V. Erceg, D. Gesbert, and R. W. Heath Jr., "Adaptive modulation and MIMO coding for broadband wireless data networks," *IEEE Commun. Mag.*, vol. 40, no. 6, pp. 108-115, June 2002.

[2] 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.

[3] 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.

[4] 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.

[5] L. Badia, M. Rossi, and M. Zorzi, "SR ARQ packet delay statistics in Markov channels in the presense of variable arrival rate," *IEEE Trans. Wireless Commun.*, vol. 5, no. 7, pp. 1639-1644, July 2006.

[6] M. Rossi, L. Badia, and M. Zorzi, "SR-ARQ delay statistics on N-state Markov channels with finite round trip delay," in *Proc. IEEE GLOBECOM*, Nov.–Dec. 2004, vol. 5, pp. 3032-3036.

[7] L. B. Le, E. Hossain, and A. S. Alfa, "Queuing analysis for radio link level scheduling in a multirate TDMA wireless network," in *Proc. IEEE GLOBECOM*, Nov.–Dec. 2004, vol. 6, pp. 4061-4065.

[8] W. S. Jeon, D. G. Jeong, and B. Kim, "Packet scheduler for mobile internet services using high speed downlink packet access," *IEEE Trans. Wireless Commun.*, vol. 3, no. 5, pp. 1789-1801, Sep. 2004.

[9] E. Cianca, M. De Sanctis, M. Ruggieri, and R. Prasad, "Truncated power control for improving TCP/IP performance over CDMA wireless links," *IEEE Trans. Wireless Commun.*, vol. 4, no. 4, pp. 1704-1714, July 2005.

[10] H. S. Wang and N. Moayeri, "Finite-state Markov channel - A useful model for radio communication channels," *IEEE Trans. Veh. Technol.*, vol. 44, pp. 163-171, Feb. 1995.

[11] Q. Zhang and S. A. Kassam, "Finite-state Markov model for Rayleigh fading channels," *IEEE Trans. Commun.*, vol. 47, no. 11, pp. 1688-1692, Nov. 1999.

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

[13] H. S. Wang and P.-C. Chang, "On verifying the first-order Markovian assumption for a Rayleigh fading channel model," *IEEE Trans. Veh. Technol.*, vol. 45, pp. 353-357, May 1996.

[14] J. Razavilar, K. J. R. Liu, and S. I. Marcus, "Jointly optimized bit-rate/delay control policy for wireless packet networks with fading channels," *IEEE Trans. Commun.*, vol. 50, no. 3, pp. 484-494, Mar. 2002.

[15] 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. 3, pp. 1142-1153, May 2005.

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

Long B. Le received the B.Eng. degree with highest distinction from Ho Chi Minh City University of Technology in 1999, the M.Eng. degree from the Asian Institute of Technology (AIT) in 2002, and the Ph.D. degree from the University of Manitoba in 2007. He is now a postdoc fellow in the Department of Electrical and Computer Engineering at the University of Waterloo. He was awarded the university gold medal in the undergraduate program, Keikyu scholarship, University of Manitoba graduate fellowship, Edward R. Toporeck graduate fellowship in engineering, University of Manitoba students' union scholarship, and IEEE student travel awards for IEEE WCNC 2003 and IEEE ICC 2005. His current research interests include link and transport layer protocol issues, resource allocation, stochastic control, and cross-layer design for wireless communication networks.

Ekram Hossain (S'98-M'01-SM'06) is an Associate Professor in the Department of Electrical and Computer Engineering at the University of Manitoba, Winnipeg, Canada. He received his Ph.D. in Electrical Engineering from the University of Victoria, Canada, in 2000. He was a University of Victoria Fellow and also a recipient of the British Columbia Advanced Systems Institute (ASI) graduate student award. He received his B.Sc. and M.Sc. both in Computer Science and Engineering from Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh, in 1995 and 1997, respectively. Dr. Hossain's research interests include the design, analysis, and optimization of wireless networks and mobile computing. Currently he serves as an Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, *IEEE Wireless Communications*, the *KICS/IEEE Journal of Communications and Networks*, the *Wireless Communications and Mobile Computing* Journal (Wiley InterScience), and the *International Journal of Sensor Networks (IJSNet)* (Inderscience Publishers).

Michele Zorzi (S'89-M'95-SM'98-F'07) was born in Venice, Italy, in 1966. He received the Laurea degree and the Ph.D. degree in Electrical Engineering from the University of Padova, Italy, in 1990 and 1994, respectively. During the Academic Year 1992/93, he was on leave at the University of California, San Diego (UCSD). In 1993, he joined the faculty of the Dipartimento di Elettronica e Informazione, Politecnico di Milano, Italy. After spending three years with the Center for Wireless Communications at UCSD, in 1998 he joined the School of Engineering of the University of Ferrara, Italy, and in 2003 joined the Department of Information Engineering of the University of Padova, Italy, where he is currently a Professor. His present research interests include performance evaluation in mobile communications systems, random access in mobile radio networks, ad hoc and sensor networks, and energy constrained communications protocols. Dr. Zorzi from 2003 to 2005 was the Editor-In-Chief of *IEEE Wireless Communications Magazine*, and currently serves on the Editorial Boards of the IEEE TRANSACTIONS ON COMMUNICATIONS, the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, the *Wiley Journal of Wireless Communications and Mobile Computing* and the *ACM/URSI/Kluwer Journal of Wireless Networks*. He was also guest editor for special issues in *IEEE Personal Communications Magazine* (Energy Management in Personal Communications Systems) and the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS (Multi-media Network Radios).