On the Impact of Correlated Arrivals and Errors on ARQ Delay Terms

Leonardo Badia, *Member, IEEE*

***Abstract*—We analytically investigate the packet delay statistics of the Selective Repeat ARQ scheme with non-instantaneous feedback, with correlation both in the channel errors and the packet arrival process. We highlight interesting trends of the delay terms, which can be extremely useful for multimedia real time services over wireless.**

***Index Terms*—Queueing analysis, automatic repeat request, Markov processes, error analysis.**

I. Introduction

THIS letter discusses the delay statistics for Selective Repeat Automatic Retransmission reQuest (SR ARQ). The way to counteract errors using the SR ARQ technique is to trigger the retransmission of non-acknowledged packets and then to resume transmission from the last packet sent [1]. Non-instantaneous feedback at the transmitter is accounted for, i.e., the round-trip delay is larger than the packet transmission time. For this reason, packets are not always transmitted in numerical increasing order, and this forces the receiver to keep the received packets in a buffer, from where they can be *released* only when all packets with lower identifiers have been acknowledged. Thus, the delay τ_D between the first transmission of a packet and its release from the receiver buffer, which we call *delivery delay*, can not be computed trivially, since it also depends on the outcome of the transmission of all packets with lower id. The analysis discussed in this letter introduces a Markov approach to evaluate the statistics of the delay terms. Importantly, the overall delay experienced by a packet also comprises the time spent in the transmitter's queue [1], which we denote as *queueing delay* (τ_Q). Again, characterizing this term requires a joint analysis of all packets in the system, i.e., ahead in the queue or already transmitted but still pending.

The evaluations of delay performance and other related aspects in such a scenario has been subject of many investigations [2]–[6]. In the 1970s, Towsley and Wolf investigated the statistics of the queueing delay in [2], where they consider a Poisson arrival process and an independent error model for the channel. They use queueing theory to derive the statistics of the time spent in the queue. In [3], Kim and Krunz present an extended analysis, also using queueing theory, for a source characterized by a Markov process and a correlated channel. However, they employed approximations to derive certain terms, e.g., the feedback at the transmitter is assumed to be instantaneous and/or the transmitter's queue is considered to be always full. Seo *et al.*, in [4], derive the delay statistics of a system very similar to the one considered here, with two Markov chains describing arrival and channel error processes, though they focus on Hybrid ARQ. A matrix geometric approach [7] has been used by Le *et al.* [5] to evaluate the performance of ARQ techniques in a radio link with adaptive modulation and coding. To derive the queueing statistics it is observed that the process is Quasi-Birth and Death (QBD), which holds also for the system studied here. Finally, in [6], Luo *et al.* discuss the ARQ delivery delay by focusing on the impact of the link layer ARQ on the performance of upper layers, i.e., the service data unit (SDU) delay. Though their focus is different, they obtain some results by means of simulation, which in what follows will be derived analytically.

Our goal here is to formulate a framework which assembles different aspects which have never been investigated jointly. More precisely, one contribution of the present letter is to present an exact analysis to evaluate the impact of correlation both in the channel errors and in the arrival process on the SR ARQ statistics. This formulation is then used to show and discuss some counter-intuitive results which emerge in the statistics and which, to the best of our knowledge, have never received an analytical characterization.

In particular, in the numerical results' section, we compare the SR ARQ delays with various intensities of the arrival rate and the arrival correlations at the transmitter's queue. Similarly, we investigate the effect of the error correlation in the channel, and we show that the delivery delay may actually decrease for an increasing arrival rate when the channel is moderately correlated. Therefore, in certain cases error correlation may imply a general decrease of the overall delay. These aspects are remarkable to achieve a correct delay estimation in real time multimedia services over wireless.

II. System Assumptions

We consider SR ARQ over a slotted time where each slot corresponds to the transmission of a packet. We assume that the round-trip time equals m slots, with $m > 1$. This is also the value of the ARQ window size. This means that packets are transmitted continuously as long as they are available, but their outcome is known only m slots later, henceforth the system memory consists of the status of the last m transmissions, including the current one. Without loss of generality, we omit the constant propagation delay term, as in [8].

We assume that packets arrive at the transmitter's queuing buffer according to a process described through a Discrete Time Markov chain with two states, referred in the following as "0"=no packet arrival and "1"=packet arrival. Note that multiple arrivals are not possible during the same time slot. As in [4], we denote a transition probability from state i

Paper approved by L. K. Rasmussen, the Editor for Iterative Detection Decoding and ARQ of the IEEE Communications Society. Manuscript received June 7, 2007; revised September 17, 2007.

L. Badia is with the IMT Lucca Institute for Advanced Studies, Piazza S. Ponziano 6, 55100 Lucca, Italy (e-mail: l.badia@imtlucca.it).

Digital Object Identifier 10.1109/TCOMM.2009.0902.070074

to state j with g_{ij}; these values can be collected into the transition matrix $\mathbf{G} = (g_{ij}), i, j \in \{0, 1\}$. The analysis can be easily extended, without changing the rationale and obtaining qualitatively similar results, to a higher number of states. The matrix $\mathbf{G}$ can be fully characterized by means of two independent parameters, the average arrival rate $\lambda = g_{01}/(g_{10}+g_{01})$ and the average arrival burst length $A = 1/g_{10}$. The case where $A = 1/(1-\lambda)$ corresponds to independent identically distributed (*iid*) arrivals with probability λ and will be referred as *iid arrival case*, or as $A = iid$. Newly arrived packets are immediately available for transmission, although retransmissions are prioritized. This means that a packet may be transmitted directly when it arrives, provided the queue is empty and no retransmission takes place.

The data sent from the transmitter's queue arrive at the receiver through a noisy channel. This is modeled through a Discrete Time Markov Channel, which is for example appropriate for fading channels in mobile environment [9] so as to account for error correlation. In general, any channel description through Markov models can be reduced to the introduction of ν "good" states corresponding to error-free transmission and η "bad" states where the packet is always in error. For example, in the Gilbert Elliott model [10] two channel conditions are considered, but they are not error-free or always erroneous. Therefore, there are four possible combinations of channel condition and packet status, which can be represented through a Markov chain with 4 states ($\nu = \eta = 2$) and appropriate transition probabilities. For simplicity, we focus here on a case where $\nu = \eta = 1$, i.e., we have only two channel states, one good (state 0) and one bad (state 1). This two-state Markov channel is similar to the one of [6]. However, the analysis can be easily extended to a higher number of states with qualitatively equivalent results. Transition probabilities p_{ij} from state i to state j are collected in the transition matrix $\mathbf{P} = (p_{ij}), i, j \in \{0, 1\}$. In the following we will describe $\mathbf{P}$ through the following parameters, the steady-state channel error probability $\varepsilon = p_{01}/(p_{10} + p_{01})$ and the average error burst length $B = 1/p_{10}$. Similar to that discussed above, the case where $B = 1/(1-\varepsilon)$ is the *iid error case*, indicated as $B = iid$. In order for the queue to be *stable* [7], the condition $\lambda < 1 - \varepsilon$ must hold.

The receiver answers with positive or negative acknowledgement (ACK/NACK) according to the correct/erroneous reception of these packets, respectively. After a full round-trip time, i.e., after m slots, ACK/NACKs arrive at the transmitter's side, and either a new packet or a retransmission is sent over the channel. We assume error-free ACK/NACKs and unlimited transmitter and receiver buffers. These assumptions are quite standard and do not limit the analysis. On the other hand, by introducing correlation in both the arrival process and the channel error process, we remove simplifications used in many studies, where the focus is on iid arrivals and iid errors only.

III. Markov Model for Selective Repeat ARQ

The mathematical framework can be obtained by generalizing the analysis presented in [8]. Under the assumptions made in the previous section, it is possible to show that the whole SR ARQ system is a Markov chain with a system state comprising the outcome of the last m transmissions, the transmitter's backlog length q, the channel state c and the arrival state r. We denote with b_i a binary variable corresponding to the outcome of the transmission of $m-i$ slots before, where that $b_i = 1$ implies that the transmission is not acknowledged and therefore a retransmission is triggered, 0 otherwise. If $\mathbf{b} = (b_1, b_2, \ldots, b_m)$, the SR ARQ system state is therefore $(q, \mathbf{b}, r, c)$. The system, as well as the variables representing the SR ARQ state, is depicted in Fig. 1.

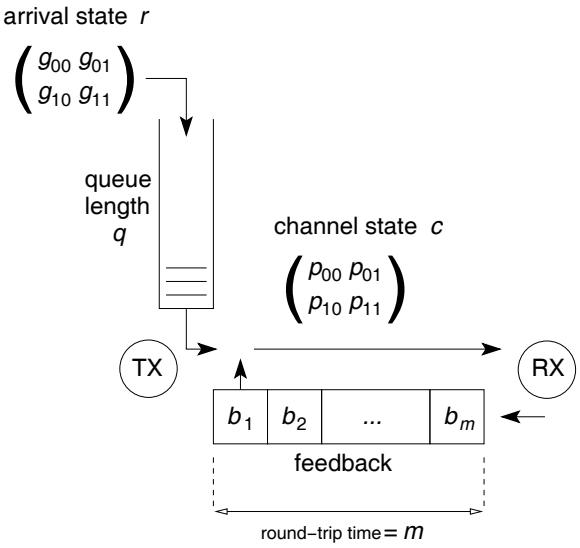

Fig. 1. The SR ARQ system and the whole Markov chain state.

To prove that the system is Markov, observe what follows. The value of q at time t depends on the backlog at time $t-1$ plus possible arrivals (depending on the arrival state r), and possible retransmissions (depending on the entry b_1). All of these terms are part of the system state at time $t-1$. Moreover, the value of b_i at time t for $i = 1, 2, \ldots, m-1$ is equal to b_{i+1} at time $t-1$ by definition. The value of b_m instead depends only on q, b_1 and c, since a retransmission is required if a packet is transmitted (which requires a check whether the transmitter's queue is empty and/or a retransmission is triggered) and the channel is in error. Finally, r and c are clearly Markov, so the whole process is also Markov. Indicating with $\pi(q, b_1, b_2, \ldots, b_m, r, c)$ the stationary probability of state $(q, \mathbf{b}, r, c)$ we can write the following set of balance equations:[1]

$$\text{for } q > 0: \quad \pi(q, b_1, b_2, \ldots, \beta, c, 0, c) \tag{1}$$
$$= \sum_{\alpha=0}^{1} \sum_{x=0}^{1} \pi(q+1-\alpha, \alpha, b_1, b_2, \ldots, \beta, x, \beta) p_{\beta c} g_{x0}$$

$$\pi(0, b_1, b_2, \ldots, \beta, c, 0, c) \tag{2}$$
$$= \sum_{\alpha=0}^{1} \sum_{\mu=\alpha}^{\max(\alpha, 1-c)} \sum_{x=0}^{1} \sum_{y=\beta}^{1} \pi(1-\mu, \alpha, b_1, b_2, \ldots, \beta, x, y) p_{yc} g_{x0}$$

$$\pi(q, b_1, b_2, \ldots, \beta, c, 1, c) \tag{3}$$
$$= \sum_{\alpha=0}^{\min(1,q)} \sum_{x=0}^{1} \sum_{y=\beta}^{1} \pi(q-\alpha, \alpha, b_1, b_2, \ldots, \beta, x, y) p_{yc} g_{x1}$$

[1]The script b_{m-1}, which occurs often, has been replaced by β, to simplify the notation.

$$\pi(0, b_1, b_2, \ldots, \beta, 0, 0, 1) \tag{4}$$
$$= \sum_{x=0}^{1} \sum_{y=\beta}^{1} \pi(0, 0, b_1, b_2, \ldots, \beta, x, y) p_{y1} g_{x0}$$

$$\text{in any other case:} \quad \pi(q, b_1, b_2, \ldots, \beta, b_m, r, c) = 0 \tag{5}$$

The balance equations for such a system can be derived directly from the following observations. The status of the last $m-1$ transmissions is deterministically updated, since their information bit is simply clocked one step in the past. In other words, the binary value referring to the ith previous transmission at time t refers to the $(i+1)$th last transmission at time $t+1$. Analogously, the evolution of the queue length corresponds with the arrival process and the presence or lack thereof of a retransmission indicated by the value of b_1. The values of r and c evolve according to the matrices $\mathbf{G}$ and $\mathbf{P}$, respectively. Moreover, the outcome of the current transmission follows the channel evolution, so that only three combinations of b_m and c are permitted: it is in fact impossible that $b_m = 1$ if $c = 0$. Instead, the condition $b_m = 0$, $c = 1$ is possible, but only if the queue is empty and no retransmission or new arrival take place. In this case, we need to track $c = 1$ to account for the bad channel state, but no retransmission is scheduled since no packet is actually transmitted. This explains why we need to separate (1) from its counterpart for $q = 0$, (2), as in the latter the past channel state y can differ from β. Additionally, (2) considers an additional summation term in μ to account for the possibility where the queue is empty, the channel is good, and no retransmission or new arrival take place. This is the only case where $\mu \neq \alpha$, otherwise this summation includes just one term. For the case with $r = 1$ it is sufficient to consider only (3), where the condition $q = 0$, however, implies that no retransmission occurred: this is the reason for the minimum as the upper limit of the outer sum. Finally, a separate evaluation is required for the case where $q = 0$, $b_m = 0$ and $c = 1$, which is done in (4), whereas in (1)–(3), b_m is always equal to c.

By adding the normalization condition, i.e., that the sum of the πs over all states equals one, we derive the stationary probabilities. Since q can increase or decrease by 1 at most, and thus the whole process is QBD, this can be promptly obtained by putting the equations in a matrix-geometric form and following the approach presented in [7]. Rearranging (1)–(5) one can also write down the SR ARQ system transition matrix $\mathbf{T}(\mathbf{G}, \mathbf{P})$, which is in turn a function of the matrices $\mathbf{G}$ and $\mathbf{P}$.

After the transition matrix is determined, we derive the delay statistics, both for the queueing and the delivery delay. The basic idea is to consider a *tagged* packet entering the queue. When this happens, the system can be in any state but the value of r must be 1 (due to the tagged packet's arrival). At this point, the arrival process can be "turned off," as the delay terms of the tagged packet are not affected by subsequent packet arrivals. Thus, define $\Lambda(q, \mathbf{b}, r, c)$ as the conditional probability of being in state $(q, \mathbf{b}, r, c)$ given that $r = 1$ (trivially, all $\Lambda(q, \mathbf{b}, 0, c)$ are 0). Moreover, consider a matrix $\mathbf{G}^0$ whose entries are set to $g_{00} = g_{10} = 1$, $g_{01} = g_{11} = 0$ and take the Markov chain defined by the transition matrix $\mathbf{T}(\mathbf{G}^0, \mathbf{P})$.

Define $\mathcal{Q} = \{(0, \mathbf{b}, r, c) : \mathbf{b} \in \{0,1\}^m, r, c \in \{0,1\}\}$ and $\mathcal{G} = \{(0, \mathbf{0}, r, c) : r, c \in \{0,1\}\}$, where $\mathbf{0}$ is an m-sized zero vector. Both $\mathcal{Q}$ and $\mathcal{G}$ are absorbing sets for the Markov chain; set $\mathcal{Q}$ is entered when the tagged packet is released from the queueing buffer, where set $\mathcal{G}$ is entered when the tagged packet and all packets transmitted prior to it are acknowledged. We have

$$\mathcal{C}_{\mathcal{Q}}[k] = \mathbf{\Lambda} \cdot [\mathbf{T}(\mathbf{G}^0, \mathbf{P})]^k \cdot \mathbf{e}_{\mathcal{Q}}, \quad k \geq 0\,, \tag{6}$$

where $\mathbf{\Lambda}$ denotes the row vector collecting all $\Lambda(q, \mathbf{b}, r, c)$'s and $\mathbf{e}_Q$ is a column vector of indicator functions of the set $\mathcal{Q}$, i.e., its values are 1 in correspondence with states belonging to $\mathcal{Q}$ and 0 elsewhere. The distribution $\mathcal{C}_{\mathcal{Q}}[t]$ is the probability that the queueing delay is lower than or equal to k slots. Thus, the probability $\text{Prob}\{\tau_Q = k\}$ is determined as:

$$\text{Prob}\{\tau_Q = k\} = \begin{cases} \mathcal{C}_{\mathcal{Q}}[0] & \text{if } k = 0 \\ \mathcal{C}_{\mathcal{Q}}[k] - \mathcal{C}_{\mathcal{Q}}[k-1] & \text{if } k > 0 \end{cases} \tag{7}$$

The statistics of the overall delay τ_G can be evaluated by following the same approach by taking a column vector $\mathbf{e}_G$ containing indicator functions of the set $\mathcal{G}$ instead. Finally, the delivery delay τ_D, corresponding to the delay between the first transmission of the tagged packet over the channel and the resolution of every pending packet, is $\tau_G - \tau_Q$.

IV. Numerical Results

In this section we present some interesting and in certain cases counterintuitive results derived from the Markov approach reported above, in an entire analytical manner. For all of the reported results, m and ε are taken to be equal to 10 and 0.1, respectively, even though other values have been tested and the results agree with the ones shown here. The findings presented are only examples, yet they prove the ability of the analytical model to describe correlated arrivals and errors. Moreover, they show the strong impact of channel and arrival correlation, so as at times the joint effect of particular choices of A and B might lead to unexpected delay behavior (queueing, delivery, or overall).

Fig. 2 shows the queueing delay and the delivery delay as functions of A in the case $B = 3$ (a mildly correlated channel). The delivery delay curves show that the value of τ_D does not significantly change when λ or A varies. The queueing delay, instead, is shown to increase with λ, which may be somehow expected, but also it exhibits a linearly increasing behavior in A. This can be explained by considering that when the arrival process is correlated, packets arrive in bursts and therefore are likely to find many other packets ahead in the queue, which results in a higher τ_Q. This implies that this delay term (and therefore also the overall delay, as the delivery delay is more or less unaffected), may be higher for correlated arrivals with low rate than for iid arrivals with high rate. Thus, the average arrival rate alone is insufficient to determine if delay requirement are met, since correlation can cause delay increases. In the figure, simulation results are also plotted for comparison. They are obtained by averaging the delays experienced by ten million transmitted packets, generated with the same arrival process and transmitted over the same Markov channel, as per Fig. 1. As the analysis is exact, we notice almost perfect agreement between analysis

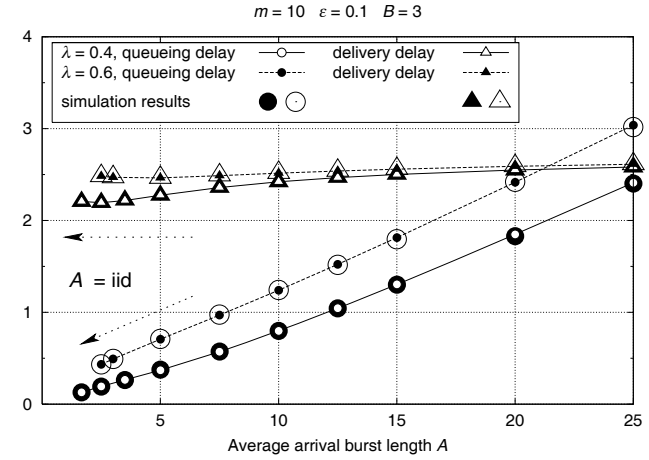

Fig. 2. Average values of the queueing and delivery delay for $m = 10$, $\varepsilon = 0.1$, $B = 3$ as a function of A, for various values of λ.

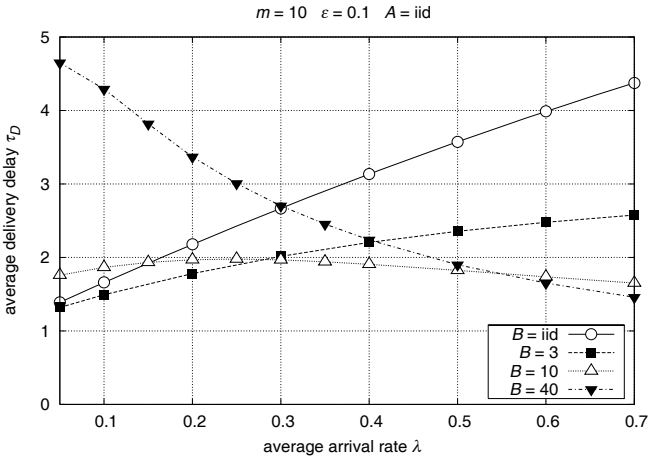

Fig. 3. Average values of the delivery delay for $m = 10$, $\varepsilon = 0.1$, $A = iid$ as a function of λ, for various values of B.

and simulation results. This holds also for the other figures, where, in order to have better distinguishable curves, only the analytical results are shown.

Fig. 3 analyzes instead the delivery delay as a function of λ. We consider $A = iid$, but other values obtain very similar results, since as shown above, A does not affect the delivery delay very much. In this figure, a counterintuitive behavior is emphasized: one might expect that the delay increases with λ, since the system is more heavily loaded. This reasoning is correct for the queuing delay, but not for the delivery delay. Indeed, when the channel is correlated the delivery delay may decrease with increasing λ. This phenomenon can be explained by considering that when the channel is strongly correlated, it is also likely to have long sequences of slots where the channel is in a good state, thus it is easier to solve an entire sequence of packets directly. This behavior is even more acute for large values of B. However, for more realistic cases where the average burst length is lower (e.g., comparable with m), the delivery delay is almost *independent* of the packet arrival rate. These theoretical observations are very important to understand transmission systems based on ARQ techniques. In fact, they imply that the performance of a correlated system

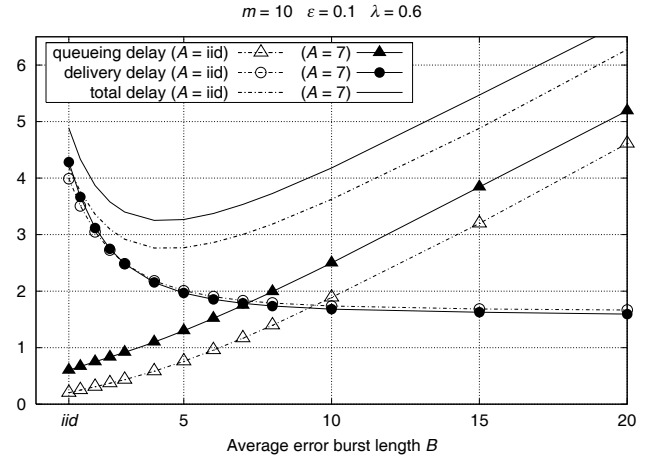

Fig. 4. Comparison between the queueing delay, the delivery delay, and the total delay for $m = 10$, $\varepsilon = 0.1$, $\lambda = 0.6$, as a function of B, for both $A = iid$ and $A = 7$.

is different from the *iid* case [3], though not always worse than it. For the transmission of large amounts of data, like in video-streaming applications, the channel correlation might be helpful, since it reduces delivery delay.

A combination of these considerations can be applied to the results shown in Fig. 4, where the different delay terms are compared for the specific case $\lambda = 0.6$ as a function of B, both in the case of iid and correlated ($A = 7$) arrivals.

It is possible to see that moderate channel burstiness achieves a lower overall delay than the iid channel. Similar curves have been presented in [6], even though they were derived only numerically, and the authors themselves point out the need for an analytical justification. Moreover, in [6] it was observed that the delay decreases at first and then increases linearly. By looking at the figure, we are now able to recognize that this depends on the dominant delay term being either the delivery or the queueing delay. In fact, while τ_D is decreasing when the channel burstiness increases around moderate values, τ_Q is linearly increasing, which becomes the prominent term for high B. In the case of correlated arrivals, the queueing delay increases, as per Fig. 2, but the behavior is similar. As a result, the presence of correlation both in the arrival process and in the channel errors can lead to non-trivial results, which is a fact to consider carefully when designing communication links.

REFERENCES

[1] Z. Rosberg and M. Sidi, "Selective-repeat ARQ: the joint distribution of the transmitter and the receiver resequencing buffer occupancies," *IEEE Trans. Commun.*, vol. 38, no. 9, pp. 1430-1438, 1990.

[2] D. Towsley and J. Wolf, "On the statistical analysis of queue lengths and waiting times for statistical multiplexers with ARQ retransmission schemes," *IEEE Trans. Commun.*, vol. 27, no. 4, pp. 693-702, 1979.

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

[4] J.-B. Seo, Y.-S. Choi, S.-Q. Lee, N.-H. Park, and H.-W. Lee, "Performance analysis of a type-II hybrid-ARQ in a TDMA system with correlated arrival over a non-stationary channel," in *Proc. ISWCS*, Siena, Italy, 2005, pp. 59-63.

[5] 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.*, vol. 5, no. 10, pp. 2647-2653, Oct. 2006.

[6] W. Luo, K. Balachandran, S. Nanda, and K. Chang, "Delay analysis of selective-repeat ARQ with applications to link adaptation in wireless packet data systems," *IEEE Trans. Wireless Commun.*, vol. 4, no. 3, pp. 1017-1029, May 2005.

[7] M. F. Neuts, *Matrix-Geometric Solutions in Stochastic Models*. New York: Dover Publications, Inc., 1981.

[8] M. Rossi, L. Badia, and M. Zorzi, "Exact statistics of ARQ packet delivery delay over Markov channels with finite round-trip delay," *IEEE Trans. Wireless Commun.*, vol. 4, no. 4, pp. 1858-1868, July 2005.

[9] M. Zorzi, R. R. Rao, and L. B. Milstein, "Error statistics in data transmission over fading channels," *IEEE Trans. Commun.*, vol. 46, pp. 1468-1477, 1998.

[10] W. Turin, *Performance Analysis of Digital Transmission Systems*. New York: Computer Science Press, 1990.