

Some Results on the Statistics of Delay Terms in SR ARQ on Markov Channels

Leonardo Badia*, Michele Rossi*, Michele Zorzi*[†]

* Department of Engineering, University of Ferrara, via Saragat 1, 44100 Ferrara, Italy

[†] Department of Information Engineering, University of Padova, Via Gradenigo 6/B, 35131 Padova, Italy

Email: {lbadia, mrossi, zorzi}@ing.unife.it

Abstract—In this paper we explore the packet delay statistics of a Selective Repeat ARQ scheme on a Discrete Time Markov Channel with non-instantaneous round trip delay. In particular, we are interested in obtaining considerations about the queueing delay of the process and also possible comparisons between different delay components. For this reason, we analyze in detail the impact of system parameters, such as the packet arrival rate and the packet error probability, on the terms which constitute the overall delay. Finally, we explore the connection of these numerical evaluations with the QoS requirements connected to delay for multimedia traffic.

Index Terms—Automatic repeat request, Markov processes, error analysis, delay estimation.

I. INTRODUCTION

THE study we present in this paper deals with fine tuning of the system parameters in Selective Repeat Automatic Retransmission reQuest (SR ARQ) systems, in order to better understand the trade-off between the terms which constitute the overall delay experienced by a packet.

SR ARQ [1] is an error control technique, where the retransmission of negatively acknowledged packets is selectively triggered by the receiver, so that after a retransmission the data flow is resumed from the last packet sent so far. In our analysis the time for a packet transmission corresponds to one slot and feedback packets, containing either an acknowledgement (ACK) or a not-acknowledgement (NACK) messages come back at the transmitter after a full round trip time. The main aspect of our investigation, where we choose SR ARQ as a representative also of other ARQ-like techniques, is that we assume that the transmission feedback, expressed by the receiver with ACK/NACK packets, is not instantaneous at the transmitter's side. Between an erroneous transmission and the corresponding retransmission other packets are sent over the channel. Since data packets must be released to higher layers *in-order*, i.e., when all packets with lower identifier have been correctly received, the receiver keeps in a buffer the packet correctly received but not yet released.

Thus, the overall delay term is subdivided into different terms: at the transmitter, a *queueing delay* is experienced by the packets, which are also delayed by retransmissions, which have higher priority. Also, at the receiver packets are released after the correct reception of all packets with lower identifier. To the *transmission delay* of a packet an additional term must then be

added, which takes into account the transmission of previous packets which are still to be acknowledged. The delay spent by a packet in the receiver buffer waiting for the final delivery is referred to as *re-sequencing delay*. The sum of the transmission and re-sequencing delay is also called *delivery delay*. These subdivisions are quite standard in the literature [2] and reflect different aspects of the system.

The transmission of multimedia data, which often is performed by means of ARQ-like techniques, is generally considered to be subject to constraints related to the Quality of Service (QoS). Henceforth, a precise understanding of the delay performance is required. In fact, delay terms can be directly connected with other specific QoS issues. In particular, the queueing delay can be related to the transmission buffer occupancy, whereas the delivery delay has the same role for the receiver buffer [3]. Moreover, for Next Generation services it is expected to have not only throughput requirements (which is the case for non real time data traffic) but also real time constraints, and not only on the average delay but also on the jitter [4], [5]. Thus, a better understanding of delay statistics with analytical instruments is of interest, since it is able to capture in an exact way parameters which are otherwise difficult to characterize. In particular, we characterize also, as clearly allowed by our model, quantities related to second order moments, which somehow relate to the delay jitter.

Another important aspect which is often neglected when performing analysis of the delay performance is that not only do wireless channels have a generally higher error rate than wireline links, but there is an additional factor which heavily affects the performance, i.e., error burstiness. Hence, we are interested in considering error correlation in our study, since it has been proven that the performance of ARQ protocols can change dramatically [6], [7]. A suitable way, which makes analytical investigations possible, is to employ a Markov chain to represent the underlying channel [8]. In fact, channel burstiness can be easily introduced by appropriately defining the elements of the transition probability matrix.

This is important since, as we will show quantitatively, the impact of channel errors can not be taken into account by looking at the average packet error rate only. Moreover, channel burstiness has a heavy impact on the performance, being in particular the queueing delay heavily underestimated when an *iid* error process is considered while neglecting

error correlation. Also, the standard deviation of the queuing delay increases even more rapidly when the channel errors are strongly correlated.

These conclusions are obtained by exploiting an analytical model which directly stems from our previous papers presented in [9], [10]. The contribution here is to extend the analysis in order to show numerical results and describe in detail the impact of system parameters. In this way, several practical conclusions can be drawn for the setup of ARQ-like systems, which we believe are difficult to estimate precisely without an exact analysis.

The rest of this paper is organized as follows: in Section II we discuss other research contributions on topics related to the delay statistics of ARQ systems. In Section III we discuss the system model, with particular focus on the key system parameters. In Section IV we present numerical results with a detailed discussion about the higher order statistics of the queuing delay and the comparison of different delay terms. Finally, Section V concludes the work.

II. RELATED WORK

Several papers can be related with the delay performance of SR ARQ. In [11], an analytical model for the packet delay and buffer occupancy in a SR ARQ system has been proposed for a static channel (fixed error probability), where [12] and [3] analyzed in detail the re-sequencing terms, again in the independent error case. The impact of channel correlation for different ARQ techniques (i.e., also including Stop-and-Wait, Go-Back-N, adaptive SR ARQ and ideal SR ARQ) was considered in [5]–[8], [13]. The separate analysis and comparison of different delay terms for a time varying channel have been considered in [2], but only quantifying approximate mean values. In [9], [10] we investigated instead Markov techniques to derive delay statistics (delivery and queuing delay, respectively) in an exact manner. In particular, both in [10] and in the present paper we adopted a Bernoulli arrival process to characterize the packet generation. This was studied for the first time in [14]. Instead, in [4] a more general packet arrival process is considered. This last contribution and a very recent paper [15] also investigate the queuing delay in detail.

III. SYSTEM MODEL

We consider a pair of communicating entities (a transmitter and a receiver) that exchange packets through a noisy wireless fading link and use an SR ARQ transmission technique with unlimited re-transmission attempts. The time for a packet transmission corresponds to one slot and feedback packets, containing either an acknowledgement (ACK) or a not-acknowledgement (NACK) messages, come back at the transmitter after a full round trip time, which equates m slots. We focus on the more interesting case of non-instantaneous feedback, i.e., $m > 1$. Fig. 1 shows the way in which packets are stored and transmitted, and the consequent delays.

Note that the transmitter continuously transmits new packets in increasing numerical order as long as ACKs are received.

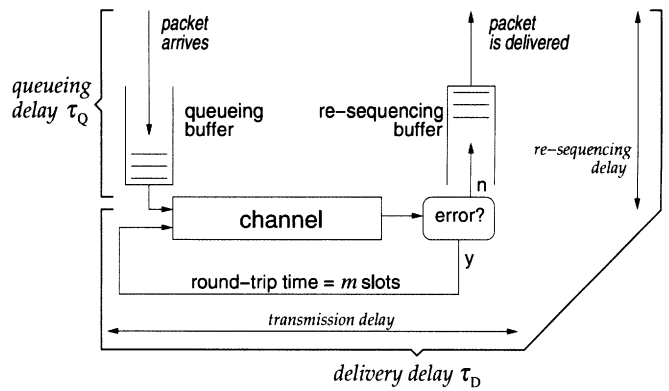


Fig. 1. The model of SR ARQ and its delays

In case of NACK, the packet must be retransmitted, which happens with higher priority with respect to the packets in the queue. However, as ACKs/NACKs refer to the transmission of m slots before, the numerical increasing order of the identifier of the transmitted packets might be broken. This explains why the delay experienced by a packet is related also to the outcome of other packets, due to out-of-order transmissions and re-sequencing.

As will be exploited in the following this implies that, since in every case the bottleneck of previous packets still pending and henceforth blocking the delivery of the ones with higher id can not exceed m slots, we need to keep track of the last m transmissions.

For the purpose of analysis, we assume to have unlimited buffer capacity both for the queue at the transmitter's side and for the re-sequencing buffer at the receiver's side. We also focus on the case of error-free ACK/NACKs. These are quite common assumptions, considered only for the sake of clarity, but they do not substantially change the analysis.

We also adopt a Markov model to represent the channel state [16]. In this way, we are able to describe the transmission with more variables than the average packet error probability alone. The Markov channel we adopt here is Two-State, i.e. state 0 corresponds to an error-free channel condition, where state 1 is always erroneous. The channel transition probability matrix is then:

$$\mathbf{P} = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix}, \quad (1)$$

and the average channel error probability and the average burst length are therefore $\varepsilon = p_{01}/(p_{10} + p_{01})$ and $B = 1/p_{10}$, respectively. If $B = \varepsilon^{-1}$ channel errors are i.i.d., whereas if $B > \varepsilon^{-1}$ there is error correlation. In the following we will refer to these parameters, which completely describe the channel transition matrix, due to their better clarity from the description point of view. Even this simple model accounts for the channel correlation with the average error burst length B , whereas on the other hand offers the advantage of describing correlation with only a single parameter. Of course, more complicated Markov channels can be used as well if necessary [17].

For the arrival process, a Bernoulli model is considered,

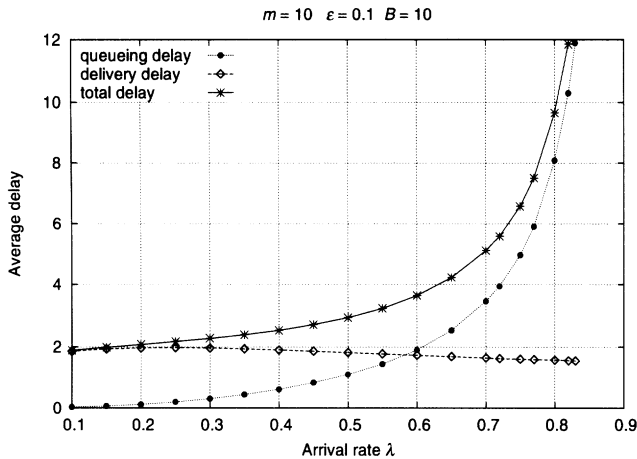


Fig. 2. Average values of the queueing and delivery delay for fixed average error probability and average burst length as a function of the arrival rate.

where a packet arrival occurs with probability λ during a slot. We have chosen this mechanism since it allows a simple representation of different load conditions with a single parameter. Of course a more complicated model, e.g., with correlated arrivals also, would be more realistic, but preliminary investigations have shown that for the kind of insight we discuss in the following, this model is entirely adequate. The reader is referred to [2] for some interesting results on the arrival burstiness.

Based on these assumptions, it can be shown that the system evolution can be modeled by means of a Markov chain. For this chain to admit a steady state distribution, λ must be between 0 and $1 - \varepsilon$, that is the average channel service rate. When $\lambda > 1 - \varepsilon$, τ_Q becomes infinite, since the queueing buffer is saturated. This condition, called in the literature *Heavy Traffic* [2], would anyway allow for an analysis of the delivery delay only, as in [9]. Indeed, as we will show in the following, the Heavy Traffic condition probably suffices for most practical purposes when studying the delivery delay, since the sensitivity of the delivery delay to the arrival intensity is rather weak. However, the more interesting investigations performed in this paper, which involve the queueing delay and/or the overall delay, can be performed only if a variable packet arrival rate is accounted for.

This model can be solved for example by following a Matrix-Geometric approach [18], as we already performed in [10]. For this reason, we only present the model here in the form of a brief summary. Interested readers might follow the details of the analysis, with also proof of correctness and comparison with simulation results, in the aforementioned paper. Note that the contribution here is instead to show how this model can be actually applied to derive considerations about the system dimensioning. In particular, in the following section we will use it to compare different delay terms and also to quantify the queueing delay effect, which, thanks to the analytical evaluation, can be done not only through average values but also with higher order moments.

The model resolution can be directly derived from the following observations. First of all, the system memory comprises the current queueing buffer occupancy and the current channel state. The information about past events can be instead fully characterized by the outcome of the last m transmissions. It is in fact true that the transmissions and the deliveries might be blocked by an error occurred in the transmission arbitrarily far in the past. However, the corresponding packet has to be re-transmitted at most within $m - 1$ slots before the packet currently on the channel.

This implies that an appropriate Markov chain can be defined, where part of the state relates to the channel evolution, hence, it follows the channel DTMC. Another term in the system state derives from the queueing buffer state and evolves according to the following alternatives: if a packet arrives and a retransmission is scheduled, the queue size increases by one; if no packet arrives and no retransmission is scheduled, the buffer occupancy decreases by one, provided it is not zero already; in every other case it stays the same. All these conditions are easily mapped through an appropriate transition matrix of the Markov chain. Finally, the system state comprises also the information about the last m transmissions, which evolves deterministically: for this reason, it is a simple matter to determine the transition probabilities between any pair of states. Thus, after some calculus, one can determine the steady-state values of the system distribution and hence evaluate the delay statistics, as explained in [10].

IV. RESULTS

The first results we show in this paper focus on the comparison of delay terms. An example is shown in Fig. 2. Here, the two main parts of the overall delay are compared, as a function of the arrival rate. This figure refers to a correlated channel with average burst length comparable with the total round-trip-time, since both m and B are taken equal to 10 slots. The average error probability ε has been set to 0.1.

The figure emphasizes that the queueing delay rapidly increases with λ , so that the queueing and the delivery delays are more or less of the same order of magnitude when λ is between 0.5 and 0.7, whereas the queueing delay is as a matter of fact negligible for $\lambda < 0.5$ and explodes for values higher than 0.7.

Therefore, the arrival process mostly affects the queueing term; indeed, the delivery delay is also impacted, but in a more complicated manner that will not be discussed here in detail. A more thorough description of this behavior can be found in [19]. However, for the considered values the average delivery delay remains almost constant. This last property has been verified to hold every time the channel burstiness is comparable with m .

Another similar comparison is reported in Fig. 3, where the average delays are evaluated as a function of the average error probability ε . In this case, we choose λ equal to 0.6, whereas m and B are both still equal to 10. Also this figure shows that, as the channel conditions become more restrictive, the

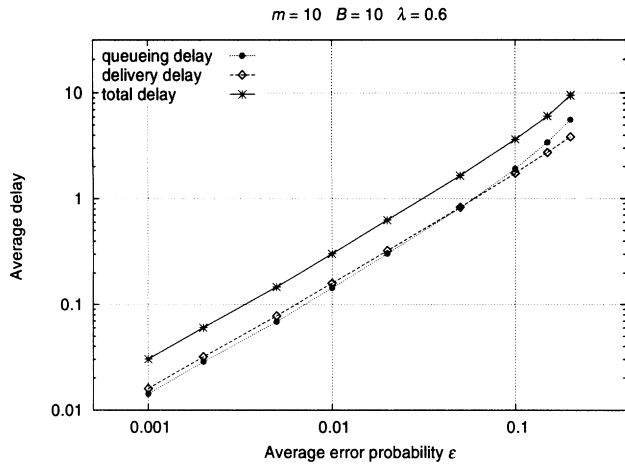


Fig. 3. Average values of the queuing and delivery delay for fixed average burst length and arrival rate as a function of the average error probability.

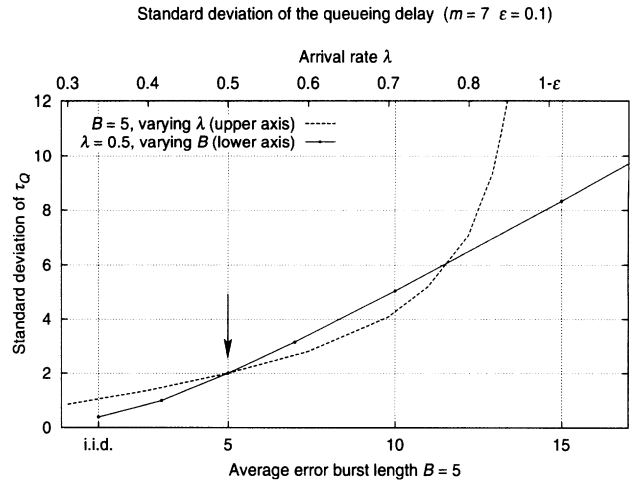


Fig. 5. Standard deviation of the queuing delay as a function of the error probability for fixed $B = 5$ and variable ϵ (upper axis and dashed line) or fixed $\epsilon = 0.1$ and variable B (lower axis and solid line), for $m = 7$, $\lambda = 0.5$.

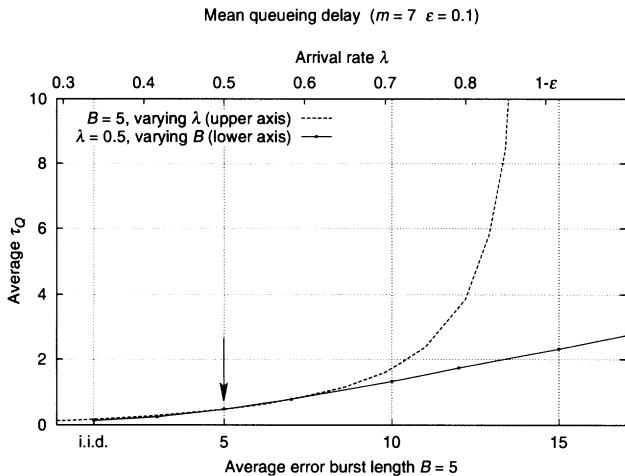


Fig. 4. Average value of the queuing delay as a function of the error probability for fixed $B = 5$ and variable λ (upper axis and dashed line) or fixed $\lambda = 0.5$ and variable B (lower axis and solid line), for $m = 7$, $\epsilon = 0.1$.

queuing delay impact becomes predominant. In fact, for low error probabilities, τ_Q and τ_D are more or less comparable, and therefore the overall delay is roughly twice each of them. However, while the delivery delay increases more or less linearly with ϵ , the queuing delay follows a similar behavior but explodes when the channel error probability is so high that the Heavy Traffic condition is approached. In other words, apart from the cases when the arrival intensity is so high that the queuing buffer is stuck, all delay terms increases more or less linearly with ϵ .

For the delivery delay this proportional increase keeps holding even for very high arrival rates. This confirms again that the delivery delay can be appropriately studied even under the Heavy Traffic assumption, whereas the queuing delay obviously can not. For this reason, in the following results we mainly focus on the queuing delay and its relationship with the arrival process.

Another important aspect to be investigated is the channel

burstiness. While in [9] we already proved that the evaluation of the delivery delay statistics can be accurate only if the channel correlation is properly taken into account, in this paper we quantitatively support a similar statement for the queuing delay also. Even though bursty and independent channels might exhibit a similar qualitative behavior, since of course as shown before τ_Q is strongly dependent on the arrival and departure rate of the system, an i.i.d. channel is definitely not a good model for correlated wireless channels from the point of view of evaluating the delay statistics. For example, the tails of the distributions for bursty channels are heavier, which means that long bursts of errors might cause a significant increase of the queuing delay term.

This can be seen from Fig. 4, where the mean τ_Q is evaluated for different values of the average error burst length B (solid line). In this figure, $m = 7$ and $\epsilon = 0.1$ are considered. The lower value of m with respect to Fig. 2 is simply motivated by the faster computational evaluation. For comparison, the dependence on the arrival rate λ is also considered, with the dashed line, which follows the same behavior of the queuing delay curve in Fig. 2. This figure should be read by keeping in mind that the middle value of both curves (indicated by the arrow) refers to the same case, and moving along the solid or dashed curve means to change B or λ , respectively. In this way, it is possible to see that the queuing delay is clearly larger in the correlated channel than in the case of i.i.d. errors. The increase of the queuing delay can be considered linear in B , whereas it explodes as λ increases and approaches $1 - \epsilon$.

Since we took an analytical approach, we are also able to investigate, e.g., second order moments of the statistics, which have a non negligible impact on the subjective perception of the QoS due to delay jitters. The standard deviation of the queuing delay is plotted in Fig. 5, where we adopt the same way of representing the data as in Fig. 4 in order to focus on and compare the dependence of this value on both λ and B . What is interesting to observe, is that the standard deviation of τ_Q is

also increasing linearly in B . This holds for all values except close to the i.i.d. case, where the behavior is slightly different. This again confirms that considering the channel correlation is unavoidable if a correct evaluation of the statistics is required.

Moreover, note that for what concerns the delay jitter the impact of B is even more relevant than for the average value. In fact the increase in the standard deviation of τ_Q when B is changed from 5 to 10 is comparable with the variation which the same value would have observed if λ were instead increased to almost 0.8, which is a significantly higher value. In other words, Fig. 5 shows that the delay jitter of a highly bursty channel is the same as that of a system with lower B but higher arrival rate. Even though this can be intuitively explained by considering that a higher B leads to more frequent retransmissions (which is eventually similar to a higher arrival rate), Fig. 5 shows that this phenomenon has a severe impact on the delay. This conclusion can be directly connected to the QoS requirements in terms of queueing buffer and delay jitter for real-time traffic, indicating that the negative impact of the error correlation is more significant than what one might expect by looking at the average value only.

V. CONCLUSIONS

In this paper we have studied the delay performance of a Selective Repeat ARQ scheme over a Markov Channel. The statistics obtained with an analytical model have been used to evaluate all the delay terms and their main characteristics are compared as a function of channel error probability and error correlation.

The following interesting conclusions can be inferred from the numerical evaluations. First of all, the queueing delay is very sensitive to the arrival process intensity, so that it generally increases as the packets arrive more frequently, until it reaches instability, which happens when the channel saturation is approached. This might appear as a trivial conclusion, as it is somewhat expected that a higher packet arrival rate will cause an increase of both queue length and queueing delay. However, we found that channel correlation (long bursts) has an unexpectedly strong impact on the queueing delay terms. In particular, the second-order analysis reveals that this effect is even more relevant for the delay jitters, so that avoiding heavily correlated channel errors appears as a key point in order to meet real-time traffic QoS constraints.

REFERENCES

- [1] S. Lin, D. J. Costello, and M. J. Miller, "Automatic-Repeat-reQuest error control schemes," *IEEE Commun. Mag.*, vol. 22, no. 12, pp. 5–17, 1984.
- [2] 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.
- [3] 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.
- [4] M. Yoshimoto, T. Takine, Y. Takahashi, and T. Hasegawa, "Waiting time and queue length distributions for go-back-n and selective-repeat arq protocols," *IEEE Trans. Commun.*, vol. 41, no. 11, pp. 1687–1693, 1993.
- [5] J. Chang and T. Yang, "End-to-end delay of an adaptive Selective Repeat ARQ protocol," *IEEE Trans. Commun.*, vol. 42, pp. 2926–2928, 1994.
- [6] D. L. Lu and J. F. Chang, "Performance of ARQ Protocols in Nonindependent Channel Errors," *IEEE Trans. Commun.*, vol. 41, no. 5, pp. 721–730, May 1993.
- [7] D. Towsley, "A statistical analysis of arq protocols operating in a nonindependent error environment," *IEEE Trans. Commun.*, vol. 29, pp. 971–981, 1981.
- [8] C. Leung, Y. Kikumoto, and S. A. Sorensen, "The throughput efficiency of the Go-back-N ARQ under Markov and related error structures," *IEEE Trans. Commun.*, vol. 36, pp. 231–234, 1988.
- [9] M. Rossi, L. Badia, and M. Zorzi, "Exact statistics of ARQ packet delivery delay over Markov channels with finite round-trip delay," *Proc. IEEE Globecom 2003*, vol. 6, pp. 3356–3360, An extended version is to appear on *IEEE Trans. Wirel. Commun.*, 2005.
- [10] L. Badia, M. Rossi, and M. Zorzi, "SR ARQ packet delay statistics on Markov channels in presence of variable arrival rate," *IEEE Trans. Wirel. Comm.*, accepted for publication.
- [11] A. G. Konheim, "A queueing analysis of two ARQ protocols," *IEEE Trans. Commun.*, vol. 28, pp. 1004–1014, 1980.
- [12] Z. Rosberg and M. Shacham, "Resequencing delay and buffer occupancy under the Selective Repeat ARQ," *IEEE Trans. on Inf. Theory*, vol. 35, no. 1, pp. 166–173, 1989.
- [13] R. Fantacci, "Queueing analysis of the Selective Repeat Automatic Repeat reQuest protocol for wireless packet networks," *IEEE Trans. Veh. Technol.*, vol. 45, no. 2, pp. 258–264, 1996.
- [14] M. E. Anagnostou and E. N. Protonotarios, "Performance analysis of the Selective-Repeat ARQ protocol," *IEEE Trans. Commun.*, vol. 34, no. 2, pp. 127–135, 1986.
- [15] B. L. Long, E. Hossain, and A. S. Alfa, "Queueing analysis and admission control for multi-rate wireless networks with opportunistic scheduling and ARQ-based error control," in *IEEE ICC 2005*, May 2005.
- [16] A. J. Goldsmith and P. P. Varaiya, "Capacity, mutual information, and coding for finite-state Markov channels," *IEEE Trans. Inform. Theory*, vol. 42, no. 3, pp. 868–886, May 1996.
- [17] M. Rossi, L. Badia, and M. Zorzi, "Sr-arq delay statistics on n-state markov channels with finite round trip delay," in *IEEE GLOBECOM 2004*, vol. 5, Nov. 2004, pp. 3032–3036.
- [18] M. F. Neuts, *Matrix-Geometric Solutions in Stochastic Models*. New York: Dover Publications, INC., 1981.
- [19] L. Badia, M. Rossi, and M. Zorzi, "On the effect of packet arrival rates on SR ARQ packet delays in Markov channels," *IEEE Comm. Lett.*, submitted for publication.