Receiver Buffer Behavior for the Selective-Repeat ARQ Protocol

Herwig BRUNEEL, Johan DE VRIENDT and Carla YSEBAERT
Laboratory for Communications Engineering, Ghent State University, Sint-Pietersnieuwstraat 41, B-9000 Gent, Belgium

Abstract. It is well known that the Selective-Repeat ARQ protocol has, among all error detection and retransmission schemes, the best throughput efficiency. However, this protocol has the disadvantage that data blocks may be received out of order, such that some reordering is required at the receiver side, before the blocks can be delivered to their destination. In order to accomplish this reordering, the receiver must dispose of a theoretically infinite amount of buffer space for the storage of correctly received blocks which cannot be delivered yet. The purpose of the paper is to study the statistical behavior of this receiver buffer, under the assumption of independent transmission errors. The study results in an explicit expression for the probability generating function of the number of blocks stored in the receiver buffer (observed at the discrete time epochs when blocks arrive at the receiver side), under the (worst-case) assumption that data blocks are available without interruption at the transmitter side. From this expression several interesting characteristics, such as the mean and variance of the buffer occupancy and the overflow probability of the buffer (if finite) can be derived with arbitrary accuracy.

Keywords. ARQ, selective-repeat protocol, performance evaluation, buffer design.

Herwig Bruneel was born in Zottegem, Belgium, in 1954. He received the M.S. degree in Electrical Engineering, the degree of Licentiate in Computer Science, and the Ph.D. degree in Computer Science in 1978, 1979 and 1984 respectively, all from Ghent State University, Ghent, Belgium. Since 1979, he has been a researcher for the Belgian National Fund for Scientific Research (N.F.W.O.) at Ghent State University. He is also a Professor in the Faculty of Applied Sciences at the same university. His main research interests include statistical modeling of computer communications systems, discrete-time queuing theory, and the study of ARQ protocols.

Johan A. De Vriendt was born in 1964. He received the M.S. degree in Electrical Engineering and the Licentiate degree in Computer Science from Ghent State University, Ghent, Belgium, in 1987 and 1989, respectively. Currently, he is working towards a Ph.D. in Electrical Engineering at the Laboratory for Communications Engineering at the same university. He is supported by the Belgian National Fund for Scientific Research (N.F.W.O.) as an “aspirant”. His present research interests are in the area of computer vision, motion analysis and pattern recognition.

Carla Ysebaert was born in 1964. In 1987, she received the M.S. degree in Electrical Engineering from Ghent State University, Ghent, Belgium, where she has been working on Spread Spectrum communication systems at the Laboratory for Communications Engineering for two years. Recently, she left the laboratory to join the production planning department of Sidmar, Ghent, where she is working on interactive planning systems.

* This work was supported by the Belgian National Fund for Scientific Research (N.F.W.O.).

North-Holland
Computer Networks and ISDN Systems 19 (1990) 129–142

0169-7552/90/$03.50 © 1990 – Elsevier Science Publishers B.V. (North-Holland)
1. Introduction

Every communication system in which digital data blocks are transferred from a sending unit (transmitter) to a receiving unit (receiver) may have to deal with the occurrence of transmission errors. One common technique to diminish the influence of such errors on the overall reliability of the system is the use of the principle of error detection and retransmission, usually referred to by the acronym ARQ (Automatic Repeat reQuest). Simply stated, this amounts to including in each data block a number of redundant (parity) bits, which allow the receiver to decide upon the correctness of this data block. Whenever the receiver finds a data block to be correct, it sends an acknowledgement signal (ACK) to the transmitter, which may then release this block from its buffer storage. However, if the receiver detects an erroneous data block, it sends a non-acknowledgement signal (NACK) to the transmitter, which will then retransmit the block automatically. This procedure continues for each block until a correct copy of this block has eventually reached the receiver.

Many different forms of this principle have been proposed and implemented, each resulting in a different ARQ protocol. A wide variety of ARQ protocols have been described and studied in the literature. Most of these schemes are variations of one of the three basic ARQ strategies, i.e., Stop-and-Wait, Go-Back-N and Selective-Repeat [1–3]. The focus in this paper is on the Selective-Repeat protocol and its descendants [4–11]; some protocols related to the two other basic strategies are discussed, for instance, in [12–22].

In the pure Selective-Repeat ARQ scheme data blocks are sent continuously from the transmitter to the receiver (as in the Go-back-N protocol) and are retransmitted only if they are erroneous (as in the Stop-and-Wait protocol). The nature of the Selective-Repeat strategy implies that the order in which the receiver obtains correct copies for the blocks sent by the transmitter may be different from the arrival order of these blocks at the sending side of the system. It is therefore necessary for the receiver to use a buffer for the storage of correctly received data blocks which cannot be delivered yet to their destination because some data block has not been received correctly yet.

The question then arises as to what amount of buffer storage is required for reordering purposes at the receiver side. In theory, the same data block may (with some finite probability) be received in error any number of times, which implies that no finite amount of buffer space is sufficient to reassure that the reordering can be accomplished in all cases. However, from a practical point of view only a finite amount of buffer space can be provided. It follows that either the ARQ protocol must be modified, such that a finite buffer is sufficient [4], or a limited loss of data blocks must be accepted.

If the first option is taken, some precautions must be made in order to avoid buffer overflow, i.e., whenever the buffer threatens to get filled completely, actions must be taken in order to prohibit the loss of data blocks. Several authors have proposed and investigated such modified Selective-Repeat ARQ protocols with finite receiver buffer [5–8]. These protocols differ in the nature of the procedures used to avoid buffer overflow. Roughly speaking, one could say that in [5–7] the loss of data blocks at the receiver side is prohibited by changing the selective-repeat mode into another mode of operation whereby the receiver buffer occupancy cannot increase any further, whenever a theoretical chance exists that, in the worst of all cases, more data blocks are to be stored than the buffer capacity allows. Thus, precautions against buffer overflow are activated in many cases when the actual buffer occupancy is not even close to the buffer capacity. This is not the case for the scheme proposed in [8], where another mode of operation is started only if the buffer is actually (close to) full. Anyhow, the approach taken in [5–8] is to fix a finite value for the receiver buffer capacity, to adapt the basic Selective-Repeat scheme in order to avoid buffer overflow, and to evaluate (and, occasionally, optimize) the resulting throughput efficiency. No results are obtained with regard to the actual occupancy of the receiver buffer.

The approach taken in the present paper is different in that the main objective of our study is exactly to derive the statistical behavior of the receiver buffer contents of the basic Selective-Repeat scheme (with infinite capacity for the receiver buffer). Apart from leading to a better understanding of the mechanisms according to which the receiver buffer gets filled, this should also give good indications concerning the minimum amount of buffer space required, either to design (new) modified ARQ schemes (without buffer
overflow) whose behavior approaches that of pure Selective-Repeat, or to keep the actual overflow probability (loss of data blocks) below a given limit, if the normal Selective-Repeat procedure is maintained in the presence of a finite-capacity buffer.

The organization of the paper is as follows. In Section 2, the Selective-Repeat protocol is briefly summarized, the main assumptions of the study are stated and some notational conventions are given. Then, in Section 3, some important parameters, needed to characterize the state of the receiver buffer, are introduced; these parameters are used in Section 4 to derive the steady-state probability generating function of the number of blocks stored in the buffer, assuming a block error probability strictly less than 1. The results are discussed and several useful formulas for the most important performance measures of the receiver buffer are derived in Section 5. Finally, in Section 6, the limiting case where the block error probability approaches 1, which is the worst case in terms of buffer requirements, is investigated separately.

It should be mentioned here, as one of the reviewers pointed out, that the receiver-buffer occupancy of a Selective-Repeat ARQ scheme, as well as the waiting time experienced by the data blocks in this buffer, have been studied before in a conference paper by Rosberg and Schacham [23] for a point-to-point connection, and by Schacham and Towsley [24] for a point-to-multipoint system. Although the present study and the one in [23] are independent, and the notations and terminology are quite different, it appears that Rosberg and Schacham—surprisingly enough—have essentially followed the same line of reasoning as we have for the derivation of the generating function of the buffer contents, in the case where the block error probability is strictly less than one. The method is explained in [23], however, in much less detail. Also, the approach taken in [23] to obtain explicit results for the mean, variance and, especially, the quantiles, of the buffer occupancy, is different from ours. In particular, the present paper contains a closed-form expression for the variance and an accurate upper bound for the overflow probability (for small values of the block error probability) which do not appear in [23]. Finally, the derivation of explicit closed-form results for the limiting case where the block error probability approaches 1, by means of an argument which is applicable in this case only, is also new.

2. Problem Formulation

We consider a point-to-point digital communication system in which a sending unit (the transmitter) transfers digital data via some communication channel to a receiving unit (the receiver). The transmitter organizes the digital data, obtained from some source, into fixed size blocks including a number of parity bits, and transmits these blocks at a constant rate via the channel. The receiver checks the blocks it receives for transmission errors and sends an acknowledgement signal to the transmitter for each block it receives, via a so-called return channel: a positive acknowledgement (ACK) for correct blocks and a negative acknowledgement (NACK) for erroneous blocks. A Selective-Repeat ARQ protocol is used. This means that a long as it has blocks to send, the transmitter continuously puts these on the channel, without waiting for the acknowledgement signals corresponding to prior transmissions. As long as positive acknowledgements arrive at the transmitter, new blocks are sent in the order of increasing sequence numbers, i.e., in the order they are generated by the source. Whenever a negative acknowledgement reaches in the transmitter, the normal order of transmitting is temporarily interrupted to allow a retransmission of the erroneous block which caused the NACK signal. This procedure implies that, at the receiver side, correct copies of the consecutive data blocks may arrive out of order. Therefore, the receiver has to reorder the (correctly received) blocks such that they can be delivered to their final destination (henceforth referred to as the user) in the correct order, i.e., according to increasing sequence numbers. In order to do so, the receiver disposes of a buffer (the receiver buffer), in which correct blocks can be stored until they can be delivered to the user.

In this paper, the receiver buffer is assumed to operate as follows. As long as the receiver has obtained correct copies of all the blocks received so far, possibly after several (re)transmissions for some of these blocks, the receiver buffer remains empty. Any new blocks, received without error, are passed immediately,
i.e. without temporarily storing them in the receiver buffer, to the user. If, however, at a given moment, an erroneous block arrives at the receiver, say block $B$, any blocks correctly received after the (first) arrival of $B$ are stored in the receiver buffer, because they cannot be delivered to the user before block $B$. This mode of operation terminates when, after one or more retransmissions, a correct copy of block $B$ finally reaches the receiver. At that time instant, block $B$ and all the correct blocks received after block $B$ but before the next erroneous block (if any) are delivered to the user instantaneously, i.e., the receiver buffer is freed of all blocks which were blocked by block $B$. The only blocks left in the receiver buffer are then the blocks which have been received correctly and whose sequence number is higher than the sequence number of the first block after block $B$ which has caused a NACK signal, say block $C$, if such block exists; otherwise, the receiver buffer becomes empty. In the first case, block $C$ then takes the role of block $B$ previously, i.e., it becomes the eldest erroneous block (E.E.B.); in the latter case, the receiver is back in the state where it was before the first arrival of block $B$. In general then, it is clear that the receiver buffer, at any time, can only contain data blocks whose sequence number is higher than the sequence number of the current E.E.B. The notion of E.E.B. will therefore be crucial in the derivation of the statistical behavior of the receiver buffer in the next sections.

To terminate this section, let us state the main assumptions under which the analysis of the receiver buffer is carried out.

(a) The transmitter sends blocks of constant length equal to $n$ bits at a constant rate of $R$ bits per second.

(b) The round-trip propagation delay of the channel, defined as the time period between the end of the transmission of a block and the subsequent arrival of the corresponding acknowledgement signal at the sending side of the system, is denoted by the symbol $T$, expressed in seconds. It follows that the number of data blocks which can be sent during a round-trip delay is given by

$$s = N - 1 = \text{Int}\left(\frac{RT}{n}\right),$$

where $N$ has the same meaning as in the basic Go-Back-N scheme. The notation $\text{Int}(x)$ indicates the smallest integer greater than or equal to $x$.

(c) We assume that the transmitter has an unlimited supply of data blocks to send, either as first transmissions or as retransmissions, which is a worst-case situation corresponding to the usual assumptions in throughput calculations. These blocks are transmitted synchronously, i.e., the transmissions can only start or end at a set of equally spaced time points along the time axis. The constant "distance" between two consecutive time points is equal to $n/R$ seconds, i.e., the time period required for transmission of one data block. This amount of time will be referred to as a slot length. It is clear that, owing to the constant (forward) propagation delay of the channel, the arrival epochs of the consecutive data blocks at the receiver side are equally spaced as well, with a constant distance of one slot length. In this paper, the term slot will be reserved to indicate an interval of time which begins at the arrival instant of one block and terminates at the arrival instant of the subsequent block at the receiver.

(d) We assume that transmission errors occur independently from block to block with a constant probability equal to $p$. This implies that whenever a block arrives at the receiver side, an ACK signal is issued with probability $1-p$ and a NACK signal occurs with probability $p$, independently of the outcomes for previously received data blocks. We assume here that errors do not occur in the block headers, so that the receiver always knows which blocks to NACK. Also, we assume that the transmitter treats lost blocks (for which no ACK or NACK arrives within a specified time-out period) as if a NACK had arrived. Finally, the return channel of the system is assumed noiseless, i.e., all acknowledgement signals are received error-free by the transmitter.

3. State Description of the Receiver Buffer

Let $S$ denote an arbitrary slot. The aim of this section is to define a number of parameters which describe the state of the receiver buffer at the end point $t$ of slot $S$ in sufficient detail, such that the buffer
occupancy, i.e., the number of blocks present in the buffer at time $t$ can be derived from them. From the previous section it is clear that the location of the slots used for the consecutive transmissions (and retransmissions) of the eldest erroneous block (E.E.B.), if any, will play an important role in this respect.

We first observe that the latest copy of the E.E.B. must have been received during one of the $N$ slots immediately before $t$ (including slot $S$), say during slot $E$ (see Fig. 1). Let $J$ denote the distance between the beginning of slot $E$ and the time point $t$, i.e., $J = 1$ if slots $E$ and $S$ coincide, $J = 2$ if slot $S$ is the slot right after slot $E$, ..., $J = j$ if slot $S$ comes $j - 1$ slots after slot $E$, for $1 \leq j \leq N$.

Next, we associate with slot $S$ or, equivalently, with the time point $t$, a frame structure for the slots before $t$. More specifically, a frame is defined as a group of $N$ contiguous slots, which may commence at the beginning of slot $E$ or a slot at distance $k \cdot N$ of slot $E$, where $k$ is any integer number. Frames are numbered consecutively along the time axis in reversed time order such that slot $E$ (and slot $S$, for that matter) belongs to frame 1; the slots within a frame are numbered from 1 to $m$ in the usual chronological order. Note that, with these conventions, slot $S$ is the $J$th slot of frame 1. Also, if we denote the total number of copies received for the E.E.B. up to time $t$ by the symbol $I$, then the slots used for the consecutive transmissions of the E.E.B. are the first slots of frames 1, 2, ..., $I$. It should be clear then that the parameters $I$ and $J$ describe the “history” of the E.E.B. at time $t$.

Our next step is to derive an expression for the buffer occupancy at time $t$. In order to do so, we observe that the definition of the E.E.B. implies that, at time $t$, the receiver buffer cannot contain any other blocks than the ones which were correctly received during frames 1 to $I$, because any blocks received before frame $I$ have a lower sequence number than the E.E.B. However, some of the blocks received correctly during frames 1 to $I$, may also have a lower sequence number than the E.E.B. and, therefore, do not contribute to the buffer occupancy at time $t$; these blocks are characterized by the fact that they must have been received in error during frame $I + 1$ (and possibly, even before that). On the other hand, blocks which were received correctly during frame $I + 1$ play no role in the buffer occupancy at time $t$. We therefore conclude that the buffer occupancy at time $t$ can be derived unambiguously from a counting argument of the numbers of ACKs and NACKs issued for blocks received during frames 1 to $I + 1$, before $t$.

In order to carry out this counting procedure, let us group the slots of frames 1 to $I + 1$ according to their slot number within the frames they belong to. Let “group $n$” indicate the slots corresponding to slot number $n$, for $1 \leq n \leq N$, and let $a(n)$ denote the number of ACKs corresponding to group $n$. Then clearly $a(1) = 1$ because NACKs are associated to each of the copies of the E.E.B. received during frames 1 to $I$, whereas an ACK must have been sent after the first slot of frame $I + 1$, by the very definitions of the E.E.B. and the parameter $I$. Furthermore, $a(n) \geq 1$ for all $n$ between 2 and $N$, because otherwise the block first received during slot 1 of frame $I$ could not be the E.E.B.

The quantities $I$, $J$ and $a(n)$ for $2 \leq n \leq N$ can be considered as a sufficient state description of the receiver buffer at time $t$ to derive the receiver buffer occupancy. Indeed, let $u$ denote the total number of data blocks stored in the receiver buffer just after slot $S$, i.e., at time $t$, then $u$ can be expressed as

$$u = \sum_{n=2}^{N} [a(n) - 1].$$

(1)
The reasoning behind this formula is that in each group \( n \) (for \( 2 \leq n \leq N \)) there is exactly one slot for which the corresponding ACK signal does not contribute to the buffer occupancy at time \( t \): either the \( n \)th slot of one of the frames 1 to \( I \) (if slot \( n \) of frame \( I + 1 \) yields a NACK) or the \( n \)th slot of frame \( I + 1 \), if it corresponds to an ACK.

4. Probability Generating Function of the Buffer Occupancy

In this section we derive probability distributions for the random variables defined in Section 3. In order to do so we first define \( N \) additional auxiliary random variables as follows:

\[
I_n = \text{the number of consecutive NACKs issued for slots of group } n, \text{ starting at time } t, \text{ in reversed time order, i.e., according to increasing frame numbers; } 1 \leq n \leq N.
\]

It is clear then that the \( I_n \)'s are i.i.d. random variables with common mass function

\[
\Pr[I_n = k] = (1 - p)^{k}p^k, \quad k \geq 0,
\]

where \( p \), the block error probability, is assumed strictly less than 1.

The random variable \( I \) can now be expressed in terms of the \( I_n \)'s as

\[
I = \max(I_1, I_2, \ldots, I_N),
\]

which implies that the mass function of \( I \) is given by

\[
\Pr[I = i] = (1 - p^{i+1})^N - (1 - p^i)^N, \quad i \geq 0.
\]

Further, the joint distribution of \( I \) and \( J \) can be expressed as

\[
\Pr[I = i, J = j] = \Pr[I_i = i \text{ and } I_2, I_3, \ldots, I_j \leq i \text{ and } I_{j+1}, I_{j+2}, \ldots, I_N \leq i - 1],
\]

\[
i \geq 1, 1 \leq j \leq N.
\]

In words: \( I = i \geq 1 \) and \( J = j \) if and only if the highest number of consecutive NACKs (in reversed time, starting at time \( t \)) is found in group 1, i.e., \( j - 1 \) slot positions before our arbitrary slot \( S \), and is equal to \( i \). Furthermore, if the number of consecutive NACKs in any other group is also equal to \( i \), this group can only correspond to the slot positions 2, 3, \ldots, \( j \), i.e., the \( j - 1 \) slot positions just before time \( t \). Note that, if \( I = 0 \), there is no E.E.B. and the random variable \( J \) is not defined. Combining (2) and (5) we then easily obtain

\[
\Pr[I = i, J = j] = (1 - p)^i(1 - p^{i+1})^{j-1}(1 - p^i)^{N-j}, \quad i \geq 1, 1 \leq j \leq N.
\]

Our next step is to derive, for given values of \( I \) and \( J \), the conditional probability generating function of the buffer occupancy \( u \), given in equation (1),

\[
U(z \mid i, j) \triangleq E[z^u \mid I = i, J = j] = \prod_{n=2}^{N} \frac{A_n(z \mid i, j)}{z}, \quad i \geq 1, 1 \leq j \leq N,
\]

where the product form is due to the mutual independence of the \( a(n) \)'s for given values of \( I \) and \( J \), and

\[
A_n(z \mid i, j) \triangleq E[z^{a(n)} \mid I = i, J = j]
\]

is the conditional generating function of \( a(n) \), given that \( I = i \) and \( J = j \). Now, if \( I = i \) and \( J = j \), it is clear that the random variables \( a(n) \) for \( 2 \leq n \leq j \) are equal to the number of ACKs during \( i + 1 \) slots, whereas each of the random variables \( a(n) \) for \( j + 1 \leq n \leq N \) counts the number of ACKs during \( i \) slots, the difference being due to the fact that only that part of frame 1 which comes before time \( t \) plays a role in the present discussion. Furthermore, \( i \geq 1 \) implies that each of the \( a(n) \)'s is at least equal to 1, as discussed in Section 3. It follows that

\[
A_n(z \mid i, j) = E[z^{b(n \mid i, j)} \mid b(n \mid i, j) \geq 1]
\]
where
\[
b(n | i, j) = \begin{cases} 
\sum_{l=1}^{i+1} c_l(n) & \text{if } 2 \leq n \leq j, \\
\sum_{l=1}^{i} c_l(n) & \text{if } j + 1 \leq n \leq N, 
\end{cases}
\] (9)

the \( c_l(n) \)'s being i.i.d. discrete random variables with values 0 or 1 with probabilities \( p \) (no ACK) or \( 1 - p \) (an ACK) respectively. If we denote the generating function of the random variable \( b(n | i, j) \) as \( B_n(z | i, j) \), we then have
\[
A_n(z | i, j) = \frac{B_n(z | i, j) - B_n(0 | i, j)}{1 - B_n(0 | i, j)}
\] (10)

from equation (8), whereas equation (9) yields
\[
B_n(z | i, j) = \begin{cases} 
[p + (1 - p)z]^{i+1} & \text{if } 2 \leq n \leq j, \\
[p + (1 - p)z]^i & \text{if } j + 1 \leq n \leq N.
\end{cases}
\] (11)

Combining (7), (10) and (11) we finally obtain
\[
U(z | i, j) = \left[ \frac{[p + (1 - p)z]^{i+1} - p^{i+1}}{z[1 - p^{i+1}]} \right]^{j-1} \left[ \frac{[p + (1 - p)z]^i - p^i}{z[1 - p^i]} \right]^{N-j}, \quad i \geq 1, 1 \leq j \leq N.
\] (12)

On the other hand, if \( I = 0 \), the conditional generating function of the buffer occupancy \( u \) is given by
\[
U(z | 0) \triangleq E[z^n | I = 0] = 1,
\] (13)

because in these circumstances no E.E.B. exists and the receiver buffer is empty.

The final step of the analysis now consists of deriving the unconditional generating function \( U(z) \) of the buffer occupancy, by averaging the conditional generating functions over all feasible values of \( I \) and \( J \). This lead to
\[
U(z) = \Pr[I = 0]U(z | 0) + \sum_{i=1}^\infty \sum_{j=1}^N \Pr[I = i, J = j]U(z | i, j)
\]
\[
= (1 - p)^N + z^{-N+1} \sum_{i=1}^\infty (1 - p)^i \sum_{j=1}^N \left( [p + (1 - p)z]^{i+1} - p^{i+1} \right)^{j-1}
\]
\[
\times \left( [p + (1 - p)z]^i - p^i \right)^{N-j},
\] (14)

where we have used equations (4), (6), (12) and (13). Another somewhat simpler expression is obtained by carrying out the summation over \( j \); this yields
\[
U(z) = (1 - p)^N + z^{-N+1} \sum_{i=1}^\infty p^i \left( [p + (1 - p)z]^{i+1} - p^{i+1} \right)^N / p^i (z - 1) [p + (1 - p)z]^i.
\] (15)

Equations (14) and (15) express the generating function \( U(z) \) explicitly in terms of known quantities. All kinds of characteristics of the receiver buffer can be derived from this, for the case where the block error probability \( p \) is strictly less than 1, as will be discussed in the next section.
5. Specific Results

In this section we derive and discuss several useful formulas for the most important performance measures of the receiver buffer.

First, we note that the mean buffer occupancy \( E[u] \) can be obtained from \( U(z) \) as

\[
E[u] = \left. \frac{dU(z)}{dz} \right|_{z=1},
\]

which can be reduced to the explicit expression

\[
E[u] = (1 - p) \sum_{i=2}^{N} C_{N}^{i} (-1)^{i} \frac{p^{i-1}}{1 - p^{i-1}}.
\]

Here \( C_{n}^{k} \) denoted the binomial coefficient

\[
C_{n}^{k} = \frac{n!}{k!(n-k)!}.
\]

Similarly, the variance of the receiver buffer occupancy can be obtained from

\[
\sigma_{u}^{2} = \left. \frac{dU(z)}{dz} \right|_{z=1} + \left. \frac{d^{2}U(z)}{dz^{2}} \right|_{z=1} - \left[ \left. \frac{dU(z)}{dz} \right|_{z=1} \right]^{2},
\]

which turns out to be equal to

\[
\sigma_{u}^{2} = E[u][1 - E[u]] + N(N - 1)p^{2} + 2N(1 - p) \sum_{j=2}^{N-1} C_{j-1}^{j-1} (-1)^{j} \frac{p^{j-1} - 2p^{j} + p^{2j-1}}{(1 - p^{j-1})^{2}}
\]

\[
+ 2 \sum_{j=2}^{N} C_{j-2}^{j} (-1)^{j} p^{j-1} \frac{1 - p - (N - 1)(1 - p^{j-1})}{(1 - p^{j-1})^{2}}
\]

\[
- 2 \sum_{j=3}^{N} C_{j-2}^{j} (-1)^{j} p^{j-2} \frac{1 - p^{j}}{1 - p^{j-2}}.
\]

It should be noted, that although equations (17) and (19) represent explicit formulas for \( E[u] \) and \( \sigma_{u}^{2} \), the evaluation of these quantities for specific values of \( N \) and \( p \) is not always trivial. In particular, for large values of \( N \), the number of terms in the summations may become quite substantial. Furthermore, these terms have alternating signs and large absolute values, so that the numerical evaluation (by means of a computer) may lead to serious rounding errors. In these circumstances, it is advantageous to derive \( E[u] \) and \( \sigma_{u}^{2} \) directly from expression (15), which results in an infinite sum over \( i \), from which, in turn, arbitrarily accurate approximations can be obtained by replacing the infinite sum be a finite one, up to some maximum value of \( i \). This approach is numerically stable, because all the terms in the sum are positive (each term in the sum basically represents the contribution of a possible value \( i \) of the random variable \( I \)). Some results are presented in Figs. 2–4.

Figure 2 shows the mean buffer occupancy versus the block error probability \( p \), for a number of (relatively low) values of \( N \). The figure reveals that the mean buffer occupancy increases as the errors occur more frequently. The same behavior is observed for higher values of \( N \). This result is not as obvious as one could think at first glance, in view of the two opposite effects explained below:

(i) If \( p \) increases, then the number of frames contributing to the buffer occupancy, as characterized by the random variable \( I \), typically gets larger;

(ii) the number of (correct) data blocks entering the buffer during each frame typically decreases. The fact that the overall influence of these two effects is an increasing mean buffer occupancy thus means that the first effect is stronger than the second one.
Fig. 2. Mean occupancy of the receiver buffer versus block error probability $p$, for various values of the round-trip delay $N$.

Fig. 3. Mean and standard deviation of receiver-buffer occupancy versus block error probability $p$, for round-trip delay $N = 4$.

Fig. 4. Mean and standard deviation of receiver-buffer occupancy versus block error probability $p$, for round-trip delay $N = 32$. 

In Figs. 3–4 we compare the standard deviation \( \sigma_p \) and the mean value \( E[u] \) for \( N = 4 \) and \( N = 32 \) respectively. These figures are somewhat typical in that they shows \( \sigma_p > E[u] \) for “small” values of \( p \), and the other way round for “large” values of \( p \). It was observed that the range of \( p \)-values of which the standard deviation is smaller than the mean value, i.e., for which the coefficient of variation of the buffer occupancy is smaller than unity, gets larger as the parameter \( N \) increases. Practically speaking, for channels with moderate to long round trip delay, the probability distribution of the buffer occupancy is rather narrow, i.e., concentrated around the mean value, irrespective of the block error probability \( p \).

In view of the complicated nature of the general formulas for \( E[u] \) and \( \sigma_p^2 \), we have approximated these formulas by means of series expansions in the variable \( p \), around the value \( p = 0 \). As a result of neglecting all terms in \( p^k \) for \( k \geq 4 \), we have obtained the following third-order approximations for \( E[u] \) and \( \sigma_p^2 \):

\[
E[u] = C_N^2 p - C_N^3 p^2 + C_{N+1}^4 p^3, \tag{20}
\]

\[
\sigma_p^2 \approx \left( 2C_N^3 + C_N^2 \right) p + \left[ 6C_N^4 + 2C_N^5 + 9C_N^6 - 2NC_N^3 - \left( C_N^2 \right)^2 \right] p^2
+ \left[ 12C_N^5 + 4C_N^4 + 2C_N^3C_{N+2}^2 + 2C_N^3C_N^2 - 22C_N^4 - 6(N + 2)C_N^3 \right] p^3 \tag{21}
\]

These approximations can be used for any value of \( N \), provided the block error probability \( p \) is sufficiently close to zero. More specifically, it was observed, by comparison with the values obtained by direct numerical evaluation, that equation (20) yields results within 5% of the actual results as long as the product \( N \cdot p \) remains lower than 1.5. Equation (21) yields results with the same degree of the accuracy for \( N \cdot p \)-values lower than about 0.7.

To terminate this section, we now concern ourselves with another quantity of interest, namely the probability of buffer overflow. Actual buffer overflow, of course, can occur only if the receiver buffer has a finite capacity. However, we can derive from our infinite capacity model an approximation for the overflow probability \( q(k) \) of a buffer with room for \( ks \) data blocks, as follows:

\[
q(k) = \Pr[u > ks],
\]

where, as before, \( u \) denotes the number of blocks stored in the receiver buffer, and \( s = N - 1 \). We are then faced with the problem of deriving the tail of the distribution of the buffer occupancy \( u \), which is by no means a simple matter. The problem can be overcome, however, by using the meaning of the random variables \( I \) and \( J \), defined in Section 3. It is clear that if \( I = k \), then the number of blocks in the receiver buffer can be no higher than \( ks \). The same is true for \( I = k + 1 \) and \( J = 1 \). It follows that in order for \( u \) to be larger than \( ks \), the random variable \( I \) must be larger than \( k \), but if \( I = k + 1 \) then \( J \) must be different from 1. However, the inverse is not necessarily true: if \( I > k \) (with \( J \neq 1 \), for \( I = k + 1 \)), then \( u \) may also be smaller than or equal to \( ks \). Putting all together, we thus obtain

\[
\Pr[u > ks] < \Pr[I > k] - \Pr[I = k + 1, J = 1],
\]

or, in view of equations (4) and (6),

\[
\Pr[u > ks] < 1 - \left( 1 - p^{k+1} \right)^{N-1} \left( 1 - p^{k+2} \right), \tag{22}
\]

which gives us a fairly simple upper bound for the overflow probability.

It is easily seen that this upper bound is quite close for “small” values of \( p \), because in these circumstances almost all data blocks (except the copies of the E.E.B.) are received correctly, so that \( u > ks \) almost surely if \( I > k \) (with \( J \neq 1 \), for \( I = k + 1 \)). In fact, we have determined the actual overflow probability \( q(k) \) for several values of \( k \), \( N \) and \( p \), by direct numerical inversion of the generating function \( U(z) \), and compared the results with the predictions given by formula (22). We have observed that the relative deviations, apart from being extremely small, are nearly independent of the value of \( N \), are proportional to the (low) value of \( p \) and linearly increasing with \( k \), as shown in Table 1. Actually, it is easily seen from this table that the relative error associated with the upper bound (22) is approximately given by

\[
\text{relative error on } q(k) = (k + 1) p.
\]

For low error rates up to, say, 0.01, the agreement between the actual value of the overflow probability and
Table 1
Relative deviations between the overflow probability \( q(k) \) and its upper bound in equation (22)

<table>
<thead>
<tr>
<th>( p )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0001</td>
<td>0.02%</td>
<td>0.03%</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>0.001</td>
<td>0.2%</td>
<td>0.3%</td>
<td>0.4%</td>
<td>–</td>
</tr>
<tr>
<td>0.01</td>
<td>–</td>
<td>3%</td>
<td>4%</td>
<td>–</td>
</tr>
<tr>
<td>0.1</td>
<td>–</td>
<td>–</td>
<td>40%</td>
<td>50%</td>
</tr>
</tbody>
</table>

its upper bound (22) is extremely good; for error rates up to 0.1, it remains acceptable, as illustrated in Fig. 
5, in which the exact and approximate values are compared for \( p = 0.1 \), for \( N = 32 \) and \( N = 128 \), as 
functions of the buffer length, characterized by the parameter \( k \). In applications such as ATM (in 
broadband ISDN) the block error probability \( p \) is typically very low; in this situation, the overflow 
probability can be very well approximated by the simple upper bound (22). An (exact) numerical 
evaluation, on the contrary, is very time-consuming (and memory-consuming), especially for high values of 
\( N \). Some results for \( p = 0.001 \) are shown in Fig. 6.

Fig. 5. Overflow probability \( q(k) \) and its upper bound, versus \( k \), for block error probability \( p = 0.1 \) and round trip delays 
\( N = 32 \) and 128.

Fig. 6. Overflow probability \( q(k) \) versus \( k \), for block error probability \( p = 0.001 \) and various values of the round-trip 
delay \( N \).
6. The Case \( p = 1 \)

It was observed in the previous section that the most severe buffer requirements occur as the block error probability \( p \) approaches 1. Moreover, the analysis presented in the previous sections assumed \( p \neq 1 \). We therefore treat the case \( p = 1 \) separately in the present section.

If the block error probability is equal to 1, then it can be shown that the group numbers (between 1 and \( N \)) corresponding to two consecutive ACKs are statistically independent; more specifically,

\[
\Pr[\text{next ACK in group } i \mid \text{previous ACK in group } j] = \lim_{p \to 1} \sum_{k=0}^{\infty} p^{(i+N-j-1) \mod N} (1-p)^{kN} \quad p \to 1 \quad \frac{1-p}{1-p^N} p^{(i+N-j-1) \mod N} \quad \frac{1}{N}.
\]

(23)

Now, in order to determine the buffer occupancy at a given time \( t \) (see Fig. 1), let us count, starting at time \( t \) and in reverse time order, the total number of ACKs issued by the receiver, until an ACK has been found for each group, between 1 and \( N \) (the order in which the \( N \) groups are encountered being irrelevant). It is clear that in this operation, the last ACK we will find is the one which was issued exactly \( N \) slots before the first reception of the current E.E.B. Hence, it suffices to subtract \( N \) units, one for each group, from the result of our counting procedure to obtain the number of data blocks stored in the receiver buffer at time \( t \).

Formally, we then obtain

\[
u = \sum_{i=1}^{N} X_i - N \quad (24)
\]

where \( X_i \) is a random variable, equal to the number of ACKs issued between the first appearance (in reverse time order) of an ACK of a given group and the first appearance of an ACK of the next group for which no ACKs had been found yet, when \( i - 1 \) different groups have occurred so far. Clearly, equation (23) implies that each of the \( X_i \)'s is a geometrically distributed random variable, whose parameters \( \alpha_i \) is given by the probability of finding an ACK for an “old” group, given that \( i - 1 \) groups have been encountered so far, i.e.,

\[
\Pr[ X_i = n ] = (1 - \alpha_i) \alpha_i^{n-1}, \quad n \geq 1,
\]

where

\[
\alpha_i = \frac{i-1}{N}, \quad 1 \leq i \leq N.
\]

Using well-known properties of the geometric distribution and the fact that the \( X_i \)'s are independent random variables, we can easily derive the probability generating function \( U(z) \) for the case \( p = 1 \) from (24):

\[
U(z) = \prod_{i=1}^{N} \frac{N-i+1}{N-(i-1)z}.
\]

(25)

The corresponding mean value and variance can be obtained by using (25) in (16) and (18); the results are

\[
E[u] = N \sum_{i=2}^{N} \frac{1}{i} \quad (26)
\]

and

\[
\sigma_u^2 = N \sum_{i=2}^{N} \frac{i-1}{(N-i+1)^2} \quad (27)
\]
In fact, the whole mass function of the buffer occupancy can be easily obtained in this case, by inverse \( z \)-transformation of equation (25); the result is

\[
\Pr[u = j] = \sum_{i=1}^{N-1} C_{N-1}^i (-1)^{N-i-1} \left( \frac{i}{N} \right)^{N-1+j},
\]

i.e., the distribution of \( u \) is a linear combination of geometrics. The overflow probability \( q(k) \), defined in the previous section, can be derived from (28) as

\[
q(k) = \sum_{j=ks+1}^{\infty} \Pr[u = j],
\]

which leads to the following explicit formula:

\[
q(k) = \sum_{i=1}^{N-1} C_{N-1}^i (-1)^{N-i-1} \frac{N}{N-i} \left( \frac{i}{N} \right)^{N+ks}
\]

Some results for the case \( p = 1 \) are presented in Figs. 7–8. In Fig. 7 the mean and the standard deviation of the buffer occupancy are plotted versus the round-trip delay of the channel, characterized by the parameter \( N \), in accordance with equations (26) and (27). Figure 8 shows the overflow probability \( q(k) \) as a function of the buffer size in terms of the parameter \( k \). These graphs can be used to predict the worst-case behavior of the receiver buffer.

7. Conclusion

We have studied in this paper the queuing behavior of the receiver buffer for the basic Selective-Repeat ARQ scheme, under the assumption of uncorrelated block errors with constant probability \( p \), and an
unlimited supply of data blocks at the transmitter side. We have distinguished the relevant parameters needed to describe the buffer, and from this, theoretically derived the probability distribution of the buffer occupancy, in the form of the associated generating function, both for the cases \( p < 1 \) and \( p = 1 \). Finally, some explicit analytical and numerical results as well as some bounds and approximations have been discussed.

References