Calculation of Message Delays and Message Waiting Times in Switching Elements with Slow Access Lines

Herwig Bruneel

Abstract—In this paper we study the delay performance of an ATDM switching element, to which messages composed of a variable number of fixed-length packets arrive via "low-speed" access lines, at the rate of one packet per time slot. More specifically, we analyze the message waiting times and the message delays encountered in this switching element, and compare the results with the case of "fast" access lines, in which all the packets of a message arrive in the same slot. The study is an extension/continuation of previous work, which was mainly concerned with the derivation of buffer occupancies and packet delays.

I. INTRODUCTION

IN digital communication networks of the store-and-forward type, the nodes of the network receive packets from a multitude of sources destined for a multitude of destinations. Hence, switching elements are necessary to route the incoming packets to their correct destinations via an appropriate output link of the node. As the incoming packets may, in general, arrive quite irregularly, the switching element must dispose of buffer space for the temporary storage of packets awaiting transmission to their destination. In principle, a separate "output buffer" is needed for each of the output links of the switching element, i.e., for each of the possible destinations of the packets.

In this paper we concentrate on the queuing performance evaluation of such a switching element, in terms of the waiting times and delays the messages experience in the switch. Since packets having different destinations do not influence each other's delay performance, it suffices to focus on one (arbitrary) output buffer. We make the assumption that each output buffer operates as an Asynchronous Time Division Multiplexing (ATDM) system, i.e., packets are transmitted from the output buffer asynchronously with respect to their sources, in their order of arrival in the buffer, as long as the latter is nonempty [1]. Also we assume that packets are generated by the sources in the form of variable-length entities ("messages" or "sessions") consisting of a (variable) number of packets that belong together. Time is assumed divided into fixed-length intervals, referred to as (time) slots, such that one slot suffices to transmit exactly one packet from the output buffer; packets are assumed to leave the buffer at the end of a slot.

As to the transmission of messages from their source to the output buffer under consideration, it is assumed that messages are sent via "slow" access lines, at the rate of one packet per slot. Notice that packets may arrive in the buffer anywhere during a slot length, as there may well be a phase difference between the clocks at the input and at the output of the switch. The numbers of newly generated messages during the consecutive slots (of the output link at hand) by the totality of all sources together are modeled as a sequence of i.i.d. random variables with common probability generating function (pgf) \( B(z) \). Each message is composed of a random number of packets which, as is customary, is assumed geometrically distributed:

\[
\text{Prob}[\text{message contains } n \text{ packets}] = (1 - \sigma)\sigma^{n-1}, \quad n \geq 1.
\]

Notice that, in terms of queuing theory, we are confronted here with a queueing system with a correlated packet arrival process, owing to the fact that one session (or message) may cause packet arrivals in the buffer during several consecutive time slots. Clearly under the more customary assumption of "fast" access lines, all packets of a message arrive during the same slot and this correlation effect does not appear. Although the majority of the pertinent literature assumes uncorrelated arrivals, queuing systems with different forms of dependence in the arrival stream have been considered before; see, e.g., [2-7]. In fact, this study elaborates on the two previous papers [6,7] in which the \textit{buffer occupancy}, i.e., the number of packets in the buffer, and the \textit{packet delay}, i.e., the elapsed time between the arrival and the departure of an arbitrary packet in the output buffer, were investigated.

In the present paper, however, we derive (closed-form) results for the \textit{message waiting time}, i.e., the time between the generation of (the first packet of) a message (or session) and the epoch when the transmission of this packet is about to take place, and for the \textit{message delay}, i.e., the time between the generation of a message and the completion of its transmission from the buffer. To the best of the author's knowledge, no such studies have been reported before for the case of "slow" access lines, whereby the arrival of a message requires multiple slots. A comparison is also made with the case of "fast" access lines. The results of the study are applicable to a wide variety of time-synchronous networks in which statistical multiplexing occurs, such as ATM networks, multiprocessors and local area networks on the one hand, and cellular, terrestrial and satellite packet networks on the other hand.

Paper approved by Edward G. Sable, the Editor for Communication Switching of the IEEE Communications Society. Manuscript received April 30, 1990; revised August 1, 1991 and March 13, 1992. This work was supported by the Belgian National Fund for Scientific Research (N.F.W.O.). This paper was presented at the IEEE Global Telecommunications Conference, GLOBECOM'90, December 2-5, 1990, San Diego, CA.

The author is with the Laboratory for Communications Engineering, University of Ghent, Belgium.

IEEE Log Number 9400905.

0090-6778/94$04.00 © 1994 IEEE
II. PRELIMINARY RESULTS

Let us concentrate on one output buffer of the switching element. We define the random variables \( a_k, b_k \) and \( u_k \) as follows:

\[
\begin{align*}
\alpha_k & \triangleq \text{total number of packets entering the buffer during slot } k; \\
\beta_k & \triangleq \text{total number of messages generated during slot } k; \\
u_k & \triangleq \text{buffer occupancy, i.e., total number of packets stored in the buffer, at the beginning of slot } k+1, \text{i.e., just after slot } k.
\end{align*}
\]

Let us also define the joint pgf of the random variables \( a_k, b_k \) and \( u_k \) as

\[
F_k(x, y, z) \triangleq E[z^{a_k}y^{b_k}z^{u_k}],
\]

and let \( F \) denote the \( k \to \infty \) limit of \( F_k \), assuming the condition for the existence of a steady state fulfilled, i.e., assuming \( B'(1) < 1 - \sigma \). Using similar methods as in [6,7], the following functional equation for \( F \) can then be established:

\[
F(x, y, z) = \frac{B(xy)z}{z - ((z-1)p_0 + F(1 - \sigma + \sigma x z, 1, z))},
\]

(1)

where \( p_0 \) denotes the steady-state probability of an empty buffer, given by \([6,7] \ p_0 = 1 - B'(1)/(1 - \sigma)\). Note that \( F \) is the joint pgf of the number of packet arrivals and the number of newly generated messages during an arbitrary slot in the steady state together with the buffer occupancy just after this slot. Hence

\[
F(x, y, z) = \sum_{i=0}^{\infty} \sum_{n=0}^{\infty} \sum_{j=0}^{\infty} x^i y^j z^n f(i, j, n),
\]

(2)

where \( f(i, j, n) \) is defined as

\[
f(i, j, n) \triangleq \lim_{k \to \infty} \text{Prob}[a_k = i, b_k = j, u_k = n].
\]

(3)

III. MESSAGE WAITING TIMES

The waiting time of a message is defined as the time period between the end of the slot during which the message is generated, i.e., the slot during which the first packet of this message enters the output buffer, and the start of the slot during which this first packet will be transmitted. Here it is assumed that packets are transmitted from the buffer in their order of arrival, regardless of the message (or session) they belong to. The message waiting time can be viewed as the time it takes a message to "get through". With the definition given it always consists of an integral number of slots, so that it can be considered as a discrete random variable.

Let \( M(\text{arb}) \) indicate an arbitrary message, which is generated during a slot referred to as slot \( J \). Next, let \( \hat{a} \) and \( \hat{b} \) denote the total number of packet arrivals and the total number of newly generated messages during slot \( J \), and \( \hat{u} \) the buffer occupancy just after slot \( J \). Let \( \hat{f}(i, j, n) \) denote the joint mass function of \( \hat{a}, \hat{b} \) and \( \hat{u} \):

\[
\hat{f}(i, j, n) \triangleq \text{Prob}[\hat{a} = i, \hat{b} = j, \hat{u} = n].
\]

(4)

Notice the difference between the quantities \( f(i, j, n) \) and \( \hat{f}(i, j, n) \): the first corresponds to the fraction of slots with \( i \) packet arrivals, \( j \) new messages generated and a buffer occupancy of \( n \) packets just after the slot, whereas the latter describes the fraction of messages which are generated during such a slot, since \( M(\text{arb}) \) is an arbitrary message. It can be shown, using similar methods as in [7], that \( \hat{f}(i, j, n) \) is given by

\[
\hat{f}(i, j, n) = f(i, j, n) \frac{j^{B'(1)}}{B'(1)}
\]

(5)

proportional to \( j \). The joint pgf of \( \hat{a} \) and \( \hat{u} \) can be easily obtained from this as

\[
G(z, x) \triangleq E[z^{\hat{a}}z^{\hat{u}}] = \sum_{i=0}^{\infty} \sum_{n=0}^{\infty} \sum_{j=0}^{\infty} \hat{f}(i, j, n)
\]

(6)

\[
= \frac{1}{B'(1)} \frac{\partial F}{\partial y}(x, 1, z),
\]

where we have used (5) and (2). This result can be further developed by means of the functional equation (1) as follows:

\[
G(z, x) = \frac{z x B'(xz)}{B'(1) B(xz)} F(x, 1, z).
\]

(7)

Now let the random variable \( s \) indicate the number of packets entering the buffer during slot \( J \), before the first packet of our tagged message \( M(\text{arb}) \) does. Due to the first-come-first-served queuing rule, the message waiting time of \( M(\text{arb}) \) is then given by \( w = \hat{u} - \hat{a} + s \). The corresponding pgf \( W(z) \) can be obtained as

\[
W(z) = E[z^{\hat{a} - \hat{u}} E[z^s | \hat{a}, \hat{u}]].
\]

(8)

Now since the first packet of our tagged message \( M(\text{arb}) \) is just an arbitrary packet with respect to the order of arrival within slot \( J \), the random variable \( s \) can take the values 0, 1, ..., \( \hat{a} - 1 \) with equal probabilities, if a total of \( \hat{a} \) packets are known to arrive during slot \( J \). Therefore, we find

\[
E[z^s | \hat{a}, \hat{u}] = \frac{1}{\hat{a}} \sum_{i=0}^{\hat{a} - 1} z^i = \frac{z^{\hat{a} - 1}}{\hat{a}(z - 1)}
\]

so that, in view of (6), \( W(z) \) can be expressed as

\[
W(z) = \frac{1}{z - 1} \int_{1/z}^{1} G(z, x) dx.
\]

(8)

Equation (8) provides us with an explicit expression for the pgf of the message waiting times in terms of the function \( G \), which, in turn, is given in (7) in terms of the function \( F \), for which we have derived a functional equation in (1). Although this not a very transparent way of characterizing the waiting-time distribution of the messages,
explicit closed-form expressions for the various moments of the message waiting time can be easily obtained from this, because only derivatives of \( W(z) \) at \( z = 1 \) are involved there. We illustrate the technique for the case of the mean value \( E[w] \) of the message waiting time. \( E[w] \) can be obtained from (8) as

\[
E[w] = \frac{dW}{dz}(1) = \frac{\partial G}{\partial z}(1, 1) - \frac{1}{2} \frac{\partial G}{\partial z}(1, 1). \tag{9}
\]

The partial derivatives of the \( G \)-function can be calculated by means of (7) in terms of the partial derivatives of the \( P \)-function, which, in turn, can be obtained from the functional equation (1). Putting all together, we finally obtain the following explicit formula for \( E[w] \):

\[
E[w] = \frac{2B'(1)[1 - B'(1)] + B''(1)}{2(1 - \sigma)(1 - \sigma - B'(1))} \left( \frac{\sigma B'(1)}{(1 - \sigma)^2} + \frac{(1 - \sigma)B''(1) - (2 - \sigma)B'(1)^2}{2(1 - \sigma)B'(1)} \right). \tag{10}
\]

Higher-order moments of the message waiting time can be derived similarly, by evaluating higher-order derivatives of \( W(z) \). The integral in expression (8) does not represent any difficulties, because the remaining integrals (of the integrand and of its derivatives with respect to \( z \)) in the eventual expression are equal to zero at \( z = 1 \).

IV. MESSAGE DELAYS

We define the delay of a message as the time period between the (end of the) arrival slot of the first packet of this message in the output buffer, and the epoch when the last packet of the message leaves the buffer. The message delay can thus be viewed as the time required for “deliver” the full message to its destination. Just as the message waiting time, the message delay can be considered as a discrete random variable. However, it turns out that a derivation of the whole probability distribution (for instance, in the form of the pgf) is virtually infeasible, as it leads to very complicated and unmeaningful formulas. Nevertheless, it appears possible to obtain a closed-form expression for the mean message delay. The method is explained below.

Consider again the tagged message \( M(\text{arb}) \), entering the output buffer during slot \( J \). Let \( L \) denote the length of \( M(\text{arb}) \), expressed in packets, and let us indicate the last packet of \( M(\text{arb}) \) as packet \( P \). The message delay \( c \) of \( M(\text{arb}) \) is equal to the time required for the transmission of all those packets that arrive in the output buffer no later than packet \( P \) (packet \( P \) is included), except for those that have left the buffer by the end of slot \( J \). Hence, if we define \( a(J + n) \) as the total number of packet arrivals in the buffer during the \( n \)-th slot following slot \( J \) (referred to as slot “\( J + n \)” in the sequel), and \( \tilde{a}(J + L - 1) \) as the total number of packets arriving during slot \( J + L - 1 \), no later than packet \( P \) does, the message delay \( c \) can be expressed as

\[
c = \tilde{u} - \tilde{a}(J), \quad \text{if } L = 1; \tag{11}
\]

\[
c = \tilde{u} + \sum_{n=1}^{L-2} a(J + n) + \tilde{a}(J + L - 1), \quad \text{if } L \geq 2, \tag{12}
\]

where the summation is equal to zero for \( L = 2 \), by convention.

The mean message delay \( E[c] \) is given by

\[
E[c] = \sum_{l=1}^{\infty} (1 - \sigma)^{l-1} E[c|L = l]. \tag{13}
\]

In order to obtain \( E[c] \), it thus remains for us to find expressions for the quantities \( E[c|L = l] \). We distinguish between \( L = 1 \) and \( L = l \geq 2 \). First, we observe that

\[
E[c|L = 1] = E[w] + 1, \tag{14}
\]

where \( E[w] \) is given in (10). Next, suppose that \( M(\text{arb}) \) contains \( l \geq 2 \) packets, i.e., \( L = l \). From (12) we then have

\[
E[c|L = l] = E[\tilde{u}] + \sum_{n=1}^{l-2} E[a(J + n)] + E[\tilde{a}(J + l - 1)]. \tag{15}
\]

The quantities in the right hand side of this equation can be derived as follows. First, according to (6), \( E[\tilde{u}] \) is nothing else than \( E[\tilde{u}] = \frac{\sigma G}{\sigma^2}(1, 1) \), which can be obtained from (7) and (1). Next, in order to determine the sum in (15), we note that \( a(J + n) \) can be expressed as

\[
a(J + n) = b(J + n) + 1 + \sum_{i=1}^{a(J+n-1)-1} c(i, n), \tag{16}
\]

where \( b(J + n) \) denotes the number of new messages generated during slot \( J + n \), the term \( 1 \) accounts for the packet of \( M(\text{arb}) \) which enters the buffer during this slot, and the sum represents the packets, entering the buffer during slot \( J + n \), belonging to messages other than \( M(\text{arb}) \), whose arrival was already in progress one slot earlier; the \( c(i, n) \)'s are i.i.d. random variables equal to 1 or 0 with probability \( \sigma \) or \( 1 - \sigma \) respectively, in accordance with the geometric message-length distribution. Taking expected values on both sides of (16), and applying the result recursively, \( E[a(J + n)] \) can be expressed in terms of \( E[a(J)] \), which can be derived from (6) as \( E[a(J)] = \frac{\sigma G}{\sigma^2}(1, 1) \), which, in turn, can be obtained from (7) and (1). The last term in (15) can be shown to be given by

\[
E[\tilde{a}(J + l - 1)] = \frac{E[a(J + l - 1)] + 1}{2},
\]

in view of the fact that packet \( P \) can be any of the \( a(J + l - 1) \) arrivals during slot \( J + l - 1 \) with equal probability, and can thus also be obtained explicitly. Consequently, we can derive a closed-form expression for \( E[c|L = l] \), which, in turn, can be used in (13), along with (14), to derive the mean message delay; the final result is
\[ E[c] = \frac{2B'(1)[1 - B'(1)] + B''(1)}{2B'(1)[1 - \sigma - B'(1)]}. \]  

(17)

Comparison of this result with the mean packet delay \( E[d] \), obtained by a different analytic approach in [7], reveals that

\[ E[c] = E[d] + \frac{\sigma}{1 - \sigma} = E[d] + \frac{1}{1 - \sigma} - 1. \]  

(18)

In words: the mean message delay is equal to the sum of the mean packet delay and the mean message length \( 1/(1 - \sigma) \), diminished by 1. It can be shown that this is due to the geometric nature of the message-length distribution.

V. DISCUSSION OF RESULTS

First, let us compare the mean waiting times of an arbitrary message and an arbitrary packet. The first is given by the quantity \( E[w] \), while the latter is

\[ E[v] = E[d] - 1. \]  

(19)

From (10), (17), (18) and (19) it then follows that

\[ E[v] - E[w] = \frac{\sigma}{2(1 - \sigma)} \left[ \frac{1 + \sigma}{1 - \sigma} B'(1) + \frac{B''(1)}{B'(1)} \right], \]  

(20)

which shows that, on the average, an arbitrary packet has to wait longer to "get through" than the first packet of an arbitrary message has to. Next, let us compare the results obtained here with the corresponding results in case of fast access lines, whereby all the packets and of a message enter the buffer during the same slot. In this case, the number of messages in the buffer and the message delay can be investigated by means of a discrete-time queuing model with independent message arrivals according to pgf \( B(z) \) and geometrically distributed message service times with parameter \( \sigma \). Such a model has been studied, for instance, in [8]. Comparing the results for fast access lines with the ones derived in this paper, we have found that the mean message waiting time is always higher in a system with fast access lines, whereas the mean message delay is exactly the same in both cases. These results can be explained intuitively as follows. In a system with fast access lines, all the packets of a message enter the system during the slot when the message is generated, whereas these packets enter one by one, at the rate of one packet per slot, in the case of slow access lines. Hence, the packet arrival process is less bursty in the slow-access-lines case, which results in lower buffer occupancies and, hence, lower packet delays and also lower message waiting times, because these quantities are fully determined by the packets present in the buffer upon arrival of a packet or a message. The message delay, however, is also influenced by packets which arrive later than the first packet of this message (and earlier than the last packet of this message), which is not true in the fast-access-lines case. This phenomenon makes the total transmission time of a message higher than for fast access lines, which explains why the mean message delay in the case of slow access lines can be as high as for fast access lines. The fact that the two results are, in fact, identical, can be shown to be a consequence of the geometric (and, hence, memoryless) nature of the message-length distribution. It should be noted, however, that, in general, the whole distributions of the message delays are not identical. One way to understand this is to bear in mind that the message transmission times are quite different in the two cases: for fast access lines, the transmission time of a message, in slots, is equal to its length in packets, whereas it typically is much higher in the case of slow access lines, i.e., for a given mean value, the message delay takes values in a larger range. Deriving more characteristics of the message-delay distribution in the case of slow access lines remains for further study.

To conclude, let us consider a numerical example. Suppose the messages are generated according to a Poisson distribution with mean \( q \), i.e., \( B(z) = e^{q(1-z)} \), and that the mean message length is 100, i.e., \( \sigma = 0.99 \). The mean packet arrival rate per slot is then equal to \( 67 \) \( \lambda = q/(1 - \sigma) = 100q \). Fig. 1 shows plots of the mean packet delay \( E[d] \), the mean message waiting time \( E[w] \) and the mean message delay \( E[c] \), versus \( \lambda \), for this particular case. The curves make clear that \( E[w] \leq E[v] \leq E[d] \leq E[c] \). Equations (18)-(20) show that the above inequalities are true regardless of the generating function \( B(z) \), i.e., regardless of the distribution of the number of messages generated per slot.

![Fig. 1. Mean message waiting time \( E[w] \), mean packet delay \( E[d] \) and mean message delay \( E[c] \) versus mean arrival rate \( \lambda \) (packets per slot), for a Poisson message generation process and a mean message length of 100.](image)

REFERENCES


