Queue Length and Delay for Statistical Multiplexers with Variable-Length Messages

Sabine Wittevrongel and Herwig Bruneel
Laboratory for Communications Engineering, University of Ghent, Sint-Pietersnieuwstraat 41, B-9000 Gent, Belgium.

ABSTRACT — In this paper we study the performance of a statistical multiplexer, to which messages composed of a variable number of fixed—length packets arrive at the rate of one packet per slot ('train arrivals'), resulting in a correlated packet arrival stream. The distribution of the message lengths is general. The system is analyzed by using an infinite—dimensional state description and a generating—functions approach. An explicit expression is obtained for the probability generating function (pgf) of the queue length and the packet delay. Also, closed—form expressions are derived for the mean and the tail distribution of the queue length and the packet delay. The mean message waiting time is obtained as well. The buffer behavior with train arrivals is compared to that with batch arrivals, where all the packets of a message enter the buffer during the same slot. By means of some numerical examples, the impact of the message—length distribution on the multiplexer performance is investigated.

1. Introduction

The Asynchronous Transfer Mode (ATM) has been widely accepted as the transfer technique for future Broadband Integrated Services Digital Networks (B-ISDN), because of its efficiency and flexibility. It is well known that many of the traffic sources that ATM is expected to support are of a bursty nature. One possible reason for the presence of correlation in cell arrival streams, however, which has not received much attention so far, is the segmentation, at the edge of the ATM network, of large (external) data frames into small—size cells. Such format conversions are necessary if the ATM network is to be used for currently existing data services.

In this paper, we therefore consider a statistical multiplexer, to which variable—length messages composed of multiple fixed—length packets arrive. The generic term message is used in the model for the external data format, whereas the term packet denotes the internal format. Time is divided into fixed—length intervals, called slots, such that one slot suffices to transmit exactly one packet from the multiplexer. Messages arrive at the rate of one packet per slot (so called 'train arrivals'). Hence, a message of length \( n \) contributes to the packet arrival stream into the multiplexer during \( n \) consecutive slots. As mentioned above, this introduces a certain degree of correlation into the packet arrival stream. The distribution of the message lengths is assumed general here. Similar discrete—time models have been investigated in [1]—[4], where either a geometric message—length distribution ([1]—[3]) or constant—length messages ([4]) were assumed. The purpose of this paper, however, is specifically to quantify the effect of the message—length distribution on the multiplexer performance. The paper is related to [5]—[8], where different forms of correlated arrival processes have been studied.

The outline of the paper is as follows. In Section 2, the analytical model under study is described. An explicit expression for the pgf of the queue length is obtained in Section 3. Also a closed—form formula for the mean buffer contents is given there, and the numerical calculation of the tail distribution of the queue length is discussed. Section 4 is concerned with the distribution of the packet delay and the mean message waiting time, for a FCFS queueing rule for packets. In Section 5, train arrivals are compared to batch arrivals. Some numerical examples are given in Section 6.

2. Analytical model description

We consider a discrete—time single—server queueing system with infinite storage capacity, to which variable—length messages consisting of multiple fixed—length packets arrive at the rate of one packet per slot. The message lengths (in terms of packets) are assumed to be i.i.d. random variables with pgf \( A(z) \) and probability mass function (pmf) \( a(i) \). The numbers of new messages generated during a slot are assumed to be i.i.d. with pgf \( B(z) \). Let \( b_k \) denote the number of new messages generated during slot \( k \). Let \( d_{n,k} \) represent the number of users that generate the \( n \)th packet of a message during slot \( k \). Then

\[
d_{1,k} = b_k ; \quad d_{n-1,k-1} = \sum_{i=1}^{n-1} c_{n-1,i}, \quad n > 1 .
\]

Here the \( c_{n-1,i} \)'s are independent sets of i.i.d. random variables with pgf's \( C_{n-1}(z) \), \( n > 1 \), given by

\[
C_{n-1}(z) = \left[ a(n-1) + \sum_{i=1}^{n-1} a(i) \right] \left[ 1 - \sum_{i=1}^{n-2} a(i) \right]^{-1} .
\]

Also, \( c_k \), the number of packet arrivals during slot \( k \), can be expressed as

\[
e_k = \sum_{n=1}^{\infty} d_{n,k} .
\]
Let \( s_k \) denote the system contents (i.e., the number of packets in the queueing system, including the possible packet in transmission) at the beginning of slot \( k \). The following system equation holds:

\[
s_{k+1} = (s_k - 1) + e_k,
\]

(5)

where \((\cdot)^*\) denotes the maximum over \(\cdot\) and 0. In the next section, we will use a generating functions approach to obtain the steady-state pgf's of the arrival process and the buffer occupancy.

3. Arrival process and buffer occupancy in the steady state

Let us define the joint pgf of the random variables \( d_{n,k+1} (n \geq 1) \) and \( s_k \) as

\[
P_k(x_1, x_2, \ldots, z) = E \left[ \prod_{n=1}^{\infty} x_n^{d_{n,k+1}} z^{s_k} \right],
\]

(6)

where \( E[\cdot] \) denotes the expected value of the argument between the square brackets. Then, using (1)–(5), we find

\[
P_{k+1}(x_1, x_2, \ldots, z) = B(x_1 z) E \left[ \prod_{n=1}^{\infty} \left[ C_n(x_{n+1}z) \right]^{d_{n,k+1}} z^{(s_k - 1)^*} \right].
\]

If the system is empty at the beginning of a slot, no cells have arrived in the system during the previous slot, i.e., \( s_k = 0 \) implies \( e_k = 0 \) and hence also \( d_{n,k+1} = 0 \) (n \geq 1). Using this property, and assuming that the queueing system can reach a steady state, we obtain the following functional equation for the steady-state joint pgf \( P(x_1, x_2, \ldots, z) \):

\[
P(x_1, x_2, \ldots, z) = \frac{B(x_1 z)}{z} \left( P(C_1(x_2z), C_2(x_3z), \ldots, z) + (z-1)p_0 \right),
\]

(7)

where \( p_0 \) is the probability of having an empty buffer at the beginning of an arbitrary slot in the steady state.

3.1. The arrival process

Let \( d_n \) denote the steady-state version of \( d_{n,k} \). The joint pgf \( D(x_1, x_2, \ldots) \) of \( d_n \) (n \geq 1) is then given by

\[
P(x_1, x_2, \ldots, z) = B(x_1 z) P(C_1(x_2z), C_2(x_3z), \ldots).
\]

(8)

Successive applications of eq. (8), then lead to

\[
D(x_1, x_2, \ldots, z) = \prod_{n=0}^{\infty} B \left[ \prod_{i=1}^{n} a(i) (1-x_{n+1}z) + x_{n+1}z \right].
\]

(9)

Here we have used the following relationships:

\[
C_1(C_2(\ldots C_n(x_{n+1}))) = \prod_{i=1}^{n} a(i) (1-x_{n+1}z) + x_{n+1}z ;
\]

\[
i \geq 1,
\]

\[
\lim_{n \to \infty} C_1(C_2(\ldots C_n(x_{n+1}))) = 1,
\]

which can be derived from (3). The marginal pgf \( D_n(x_n) \) of \( d_n \) can be obtained by putting \( x_{j+1} = 1 \) (i \geq 1, i \neq n) in eq. (9). The average number of users that generate the \( n \)th packet of a message during an arbitrary slot is then given by

\[
E[d_n] = D_n'(1) = B'(1) \left[ 1 - \sum_{i=1}^{n-1} a(i) \right],
\]

(10)

i.e., the mean number of new messages generated during a slot times the probability of having a message length of at least \( n \) packets, which is intuitively clear.

Let us now denote by \( e \) the number of arrivals during an arbitrary slot in the steady state. Then the pgf \( E(z) \) of \( e \) can be derived from (9) as \( E(z) = D(z, \ldots, z) \). The mean number of packet arrivals per slot is then obtained as

\[
E[e] = E'(1) = A'(1)B'(1).
\]

(11)

Hence, we can express the equilibrium condition as

\[
A'(1)B'(1) < 1.
\]

(12)

3.2. Buffer occupancy

Let \( s \) be the system contents at the beginning of a slot in the steady state. In order to derive from (7) an expression for the pgf \( S(z) \) of \( s \), first, we consider in (7) only those values of \( x_n \) (n \geq 1) for which the arguments of the \( P \)-functions on both sides of (7) are equal, i.e.,

\[
x_n = C_n(x_{n+1}z), \quad n \geq 1.
\]

(13)

From (3) and (13), \( x_n \) (n \geq 1) can be solved in terms of \( z \). Denoting this solution by \( \chi_n(z) \), we get

\[
\chi_n(z) = \left[ \sum_{i=n}^{\infty} a(i) z^{i-n} \right] \left[ 1 - \sum_{i=1}^{n-1} a(i) \right]^{-1}, \quad n \geq 1.
\]

(14)

Note especially that \( \chi_1(z) = A(z)/z \) and \( \chi_n(1) = \chi_n(z)|_{z=1} = 1 \), n \geq 1. Choosing \( x_n/\chi_n(z) \) in (7), we then obtain

\[
P(\chi_1(z), \chi_2(z), \ldots, z) = \left( \frac{z-1}{z-B[A(z)]} \right) p_0 B[A(z)],
\]

(15)

where the quantity \( p_0 \) can be calculated from the normalization condition \( P(\chi_1(z), \chi_2(z), \ldots, z)|_{z=1} = 1 \), as

\[
p_0 = 1 - B'(1)A'(1) = 1 - E'(1).
\]

(16)

Next, the pgf \( S(z) \) of \( s \) is given by \( P(1, \ldots, 1, z) \). Successive applications of (7), then allow to express \( S(z) \) in terms of the function \( P(\chi_1(z), \chi_2(z), \ldots, z) \), given in (15). As a result, we obtain

\[
S(z) = (z-1)p_0 \sum_{n=1}^{\infty} \left[ \prod_{i=1}^{n} B[zC_1(zC_2(\ldots zC_{i-1}(z)))]/z \right] P(\chi_1(z), \chi_2(z), \ldots, z),
\]

(17)

where we have used the property that

\[
\lim_{j \to \infty} C_1(zC_2(\ldots zC_{j-1}(z))) = \chi_j(z), \quad i \geq 1,
\]

as can be shown from (14). By using (3), we can calculate the arguments of the \( B \)-functions in eq. (17) as

\[
zC_1(zC_2(\ldots zC_{i-1}(z))) = z^i + \sum_{j=1}^{i-1} a(j) (z^{i-j-1}).
\]

(18)
Hence, from (15), (17) and (18), the following explicit expression can be derived for $S(z)$:

$$S(z) = \frac{(z-1) \ p_0 \ H(z)}{z - B[A(z)]},$$

(19)

Here $H(z)$ is given by

$$H(z) = \sum_{n=1}^{\infty} \left[ \prod_{i=1}^{n} g_i(z) \right] \left[ z - B[A(z)] \right] + \sum_{i=1}^{\infty} g_i(z) B[A(z)],$$

(20)

where $g_i(z)$ is defined as

$$g_i(z) = B \left[ i + \sum_{j=1}^{i-1} a(j) (z^{-j} - z^{-i}) \right].$$

(21)

In principle, the whole pmf $s(n)$ of $s$ can be calculated by taking the inverse $z$-transform of (19). However, in practice, one is particularly interested in the moments and the tail distribution of the buffer occupancy. Therefore, in the following, we will describe a technique to calculate these quantities, that doesn’t involve the calculation of the whole distribution of $s$ and is not time-consuming.

An expression for the mean system contents $E[s]$ can be derived by evaluating the first derivative of eq. (15) with respect to $z$ at $z=1$. It follows that

$$E[s] = B'(1) \left[ A'(1) - \frac{A'^{(2)}}{2} \right] + B''(1) \left[ A''(1) \right] + B'(1) A'(1).$$

If we now denote by $\mu_a$ and $\sigma_a^2$ the mean and the variance of the message lengths respectively, and by $\mu_b$ and $\sigma_b^2$ the mean and the variance respectively of the number of new messages generated during a slot, then the expression for $E[s]$ can be further transformed into

$$E[s] = \frac{\mu_a \mu_b}{2(1-\mu_a \mu_b)} \sigma_a^2 + \frac{\mu_a^2}{2(1-\mu_a \mu_b)} \sigma_b^2 + \mu_a \mu_b \left[ 1 - \frac{\mu_a}{2} \right].$$

(22)

The above formula clearly indicates that the multiplexer performance is not only determined by the mean message length, but also depends strongly on the actual message—length distribution. For given values of $\mu_a$ and $\mu_b$, the mean system contents is a linearly increasing function of the variances of both the message length and the number of new message arrivals per slot, i.e., $E[s]$ linearly increases with the uncertainty in the arrival process to the queueing system. Furthermore, for a given value of $\mu_b$, the impact of $\sigma_a^2$ on $E[s]$ increases with increasing values of $\mu_a$ and hence with increasing total load $\mu_a \mu_b$. Also, for a given value of the total load $\mu_a \mu_b$, the influence of $\sigma_b^2$ on $E[s]$ increases as $\mu_a$ decreases. In other words, the variance of the message length has a larger impact on $E[s]$ as the buffer is loaded with more shorter messages.

Higher order moments of $s$ can be derived in a similar way. Specifically, one has to evaluate the consecutive partial derivatives of $P(x_1, x_2, ..., x_i)$ with respect to $z$ for $x_i=1$, $i \geq 1$ and $z=1$, by using (7) and (15), and then express the moments of $s$ as a function of these partial derivatives.

In order to derive an expression for the tail distribution of the system contents, we will use an approximation technique described in [9]. It was shown in [4] that this method gives extremely accurate results. Specifically, from the inversion formula for $z$-transforms it follows that the pmf $s(n)$ of $s$ can be expressed as a weighted sum of negative powers of the poles of $S(z)$. As the modulus of all these poles is larger than one, it is clear that $s(n)$ will be dominated by the contribution of the pole having the smallest modulus. Denoting this dominant pole by $z_0$, it is shown in [9] that $z_0$ must necessarily be real and positive in order to ensure that the tail distribution is nonnegative anywhere. Therefore, for $n$ sufficiently large, $s(n)$ can be approximated as

$$s(n) \cong -\theta z_0^{-n-1},$$

(23)

where $\theta$ is the residue of $S(z)$ at the point $z=z_0$. From (19), it follows that $z_0$ is a real root of $z - B(A(z)) = 0$. The residue $\theta$ can be calculated from (19) as

$$\theta = \frac{(z_0 - 1) \ p_0 \ H(z_0)}{1 - B'(A(z_0)) A'(z_0)},$$

(24)

where $H(z)$ is given in (20). In order to calculate $H(z_0)$ numerically, let us introduce the following notation:

$$R_n(z_0) = \prod_{i=1}^{n} g_i(z_0), \quad n \geq 1.$$  

(25)

From (20), it then follows that $H(z_0)$ is given by

$$H(z_0) = R_{\infty}(z_0) B[A(z_0)].$$

(26)

From eq. (25), it is clear that $R_n(z_0)$ must satisfy the following recursive equation:

$$R_n(z_0) = R_{n-1}(z_0) g_n(z_0)/z_0, \quad n \geq 1,$$

(27)

where $R_0(z_0) = 1$. Further, from (21), it can be shown that $g_1(z)$ goes to $B[A(z)]$, if $i$ goes to infinity, and hence, since $B[A(z_0)] = z_0$, it follows that $g_1(z_0)/z_0$ goes to 1, as $i$ goes to infinity. As a consequence, we have, for $n$ sufficiently large, that $R_n(z_0) \cong R_{n-1}(z_0)$. We may therefore conclude that $H(z_0)$ can be calculated as

$$H(z_0) \cong R_n(z_0) z_0, \quad n \text{ sufficiently large},$$

(28)

where $R_n(z_0)$ can be calculated using (27).

4. Packet delay and message waiting time

4.1. Packet delay

In this section, we assume a first—come first—served queueing discipline for packets, i.e., packets leave the buffer in the order of their arrival, whereby packets that arrive in the buffer at the same time are transmitted in random order. We define the packet delay as the period between the end of the packet's arrival slot and the end of the slot during which the packet is transmitted and leaves the buffer. All the moments and the tail distribution of the packet delay can be expressed in terms of the previously derived moments and tail distribution of the system contents, by using the general relationship between packet delay and system contents reported, for instance, in [4] and [10].

1082
4.2. Message waiting time

The message waiting time is defined as the time period between the end of the slot during which the first packet of a message was generated and the time instant at which the transmission of this packet is about to start. Let \( w \) denote the message waiting time. By using a similar method as described in [3], we can derive the pgf \( W(z) \) of \( w \) as

\[
W(z) = \frac{1}{z} \int_0^1 \frac{G(xz)}{x} \frac{1}{z} \; dx,
\]

where \( G(x,z) \) is given by

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

By evaluating the first derivative of (29) with respect to \( z \) at \( z=1 \), the mean message waiting time \( E[w] \) can then be expressed in terms of the first partial derivatives of the \( G \)-function for \( x=z=1 \). These can in turn be calculated from (30), in terms of the following partial derivatives:

\[
\frac{\partial}{\partial x} P(x, \ldots, x, x) \bigg|_{x=z=1} = B'(1)A'(1)
\]

and

\[
\frac{\partial}{\partial z} P(x, \ldots, x, x) \bigg|_{x=z=1} = E[s].
\]

Finally, we obtain the following explicit formula for \( E[w] \):

\[
E[w] = E[s] + \frac{B'(1) - B'(1)^2}{B'(1)} + \frac{A'(1)}{2}.
\]

Hence, for given values of \( \sigma_\alpha \) and \( \eta_\beta \), \( E[w] \) also linearly increases with \( \sigma_\alpha^2 \) and \( \sigma_\beta^2 \).

5. Comparison of train arrivals and batch arrivals

In this section, we will compare the buffer behavior with train arrivals, where messages enter the buffer at the rate of one packet per slot, to that with batch arrivals, where all the packets of a message enter the buffer during the same slot. In the batch arrivals case, let \( v_k \) be the system contents at the beginning of slot \( k \), and let \( \ell_{1:k} \) be the length of the \( k \)th message generated during slot \( k \). We then have

\[
v_{k+1} = (v_k - 1) + \sum_{i=1}^{\ell_{1:k}} \end{equation}

where the pgf of \( \ell_{1:k} \) is given by \( A(z) \). From this, using standard \( x \)-transform techniques, the pgf \( V(z) \) of the buffer contents in the steady state can be derived as

\[
V(z) = \frac{[1-B'(1)A'(1)]}{z} - B[A(z)]
\]

The mean buffer contents in case of batch arrivals \( E[v] \) can be calculated from this. It follows that the difference in mean for the train and batch arrival processes is given by

\[
E[v] - E[s] = \frac{B'(1)A''(1)}{2} \geq 0.
\]

i.e., the mean buffer contents cannot be less with batch arrivals than with train arrivals. Moreover, the two buffer contents are equal only if all messages have a constant length of one packet, which is expected intuitively.

A comparison of (19) and (33) furthermore shows that \( S(z) \) and \( V(z) \) have the same dominant pole \( z_0 \). Therefore, for the batch arrivals case, the pmf \( v(n) \) of the queue length can, for \( n \) sufficiently large, be approximated as

\[
v(n) \sim \frac{\theta_\beta}{z_0} \frac{z_0^n}{n!} - n^{-1},
\]

where \( \theta_\beta \) is given by

\[
\theta_\beta = \frac{(z_0-1)p_0 z_0}{1 - B'[A(z_0)]A'(z_0)}.
\]

Therefore, for \( n \) sufficiently large, we have

\[
s(n) \sim \frac{\theta_\beta}{z_0} = \frac{H(z_0)}{z_0},
\]

i.e., the difference between the two pmf's is constant on a semi-logarithmic scale, for \( n \) sufficiently large. Furthermore, using the expression for \( H(z_0) \), it is possible to show that \( H(z_0) \geq z_0 \), and hence, \( s(n) \leq v(n) \), sufficiently large, where the equality only holds if all the messages have a constant length of one packet.

6. Numerical examples

Let us assume now that users generate new messages according to a Poisson distribution with average value of \( q \) messages per slot, i.e.,

\[
B(z) = \exp[q(z-1)].
\]

Until now, we have assumed an infinite storage capacity for the queueing system. For a buffer of finite size, however, a number of cells will get lost because of buffer overflow. We will approximate the cell loss probability for a finite buffer with a waiting room of size \( S \), by the probability that the buffer contents in the infinite buffer exceeds \( S \), i.e.,

\[
\text{Prob}[s > S] = \frac{\theta_\beta}{z_0 - 1} z_0^{-S} - 1.
\]

In order to investigate the impact of the message-length variance on the overflow probability, we consider the following examples for the pgf \( A(z) \) of the message lengths:

\[
A_1(z) = z^m; \quad A_2(z) = \frac{(1-\sigma)^2 z}{(1-\alpha z)^2}; \quad A_3(z) = \frac{(1-\sigma)^2 z}{1-\sigma z},
\]

i.e., constant-length messages, a negative binomial distribution and a geometric distribution respectively, and we choose the parameters of these distributions such that the mean message lengths in each case equal \( m \). It can be shown that this corresponds to choosing

\[
\sigma = (m-1)/m \quad \text{and} \quad \alpha = (m-1)/(m+1).
\]

The corresponding message-length variances are then

\[
\text{var}_1 = 0; \quad \text{var}_2 = (m-1)(m+1)/2; \quad \text{var}_3 = m(m-1).
\]
In Fig. 1, Prob[\(s>S\)] is shown as a function of \(S\), for Poisson train and Poisson batch arrival processes, for \(q=0.1\), \(A'(1)=8\), and for the above message-length distributions. The variances of the message lengths are \(\text{var}_1 = 0\), \(\text{var}_2 = 31.5\) and \(\text{var}_3 = 56\) respectively. We observe that for a given value of the mean message length, and a given distribution of the number of new messages in a slot, the overflow probability increases with increasing variance of the message length. In Figs. 2 and 3, Prob[\(s>S\)] is plotted versus \(S\), for Poisson train and Poisson batch arrival processes, geometrically distributed message lengths, and various values of the mean message length \(A'(1)\) and the total input load \(qA'(1)\). As expected intuitively, the overflow probability is an increasing function of \(qA'(1)\) for a given value of \(A'(1)\), and an increasing function of \(A'(1)\) for a given value of \(qA'(1)\), i.e., for a given total load, a smaller number of long messages gives rise to higher buffer occupancies. From Figs. 1–3, it is also clear that the overflow probability can not be less in the batch arrivals case compared to the train arrivals case. The difference between the two arrival processes gets larger as the mean message length gets larger and/or the load gets lower.

Acknowledgement

The first and second author are supported by the Belgian Government (D.W.T.C.) and the Belgian National Fund for Scientific Research (N.F.W.O.), respectively.

References