On queueing delays of dispersed messages

Israel Cidon [1], Asad Khamisy and Moshe Sidi

Department of Electrical Engineering, Technion – Israel Institute of Technology, Haifa 32000, Israel

Received 4 September 1992; revised 20 January 1993

We study the message queueing delays in a node of a communication system, where a message consists of a block of consecutive packets. The message delay is defined as the time elapsing between the arrival epoch of the first packet of the message to the system until after the transmission of the last packet of that message is completed. We distinguish between two types of message generation processes. The message can be generated as a *batch* or it can be *dispersed* over time. In this paper we focus on the dispersed generation model. The main difficulty in the analysis is due to the correlation between the system states observed by different packets of the same message. This paper introduces a new technique to analyze the message delay in such systems for different arrival models and different number of sessions. For an $M/M/1$ system with variable size messages and for the bursty traffic model, we obtain an explicit expression for the Laplace–Stieltjes transform (LST) of the message delay. Derivations are also provided for an $M/G/1$ system, for multiple session systems and for fixed message sizes. We show that the correlation has a strong effect on the performance of the system, and that the commonly used *independence assumption*, i.e., the assumption that the delays of packets are independent from packet to packet, can lead to wrong conclusions.

Keywords: Message delay; dispersed messages; *M/M/1*; *M/G/1*; bursty traffic.

1. Introduction

This paper is concerned with the study of message queueing delays in a node of a communication system. A message consists of a block of consecutive packets (where in our terms a packet is the integral transmitted quantity), and it corresponds to a higher layer protocol data unit. The message delay is defined as the time elapsing between the arrival epoch of the first packet of the message to the system until after the transmission (service) of the last packet of that message is completed. We distinguish between two types of message generation processes. The message can be generated as a *batch*, i.e., all the packets that compose the message arrive to

[1] Also at: IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA.

the system at a single instant of time (which corresponds to the well known *batch arrival* model) or it can be *dispersed*, i.e., the packets that compose the message arrive to the system at different times. In this paper we focus on the latter generation model where the message arrival process is spaced over time.

In many systems the message delay, and not the packet delay, is the measure of interest for the network designer. This is due to the fact that packets are data units which are only meaningful at lower layers and are created because of the network data unit size limitations. The ATM (CCITT [3]), TCP/IP (Comer [4]) and TDMA based systems (Rom and Sidi [8]) are examples of such systems, where the application message is segmented into bounded size packets (cells) which are then transmitted through the network. At the receiving end, the transport protocol (or the adaptation layer) reassembles these cells back into a message before the delivery to higher layers. In some applications message delay is not the result of segmentation at the network layer but of the nature of data partitioning in the storage. A file can be composed of multiple records which are stored at different locations in the disk. These records are read individually and may be transmitted as separate packets. However, the entire file transfer delay is the measure of interest for the file transfer application.

Understanding the message delay behavior is important for the proper design of timeout mechanisms for data applications, such as end-to-end protocols for reliable transmission in ATM networks where the retransmitted quantity is the message and not the individual cells. Another design example is the time-out for message retransmissions in data link protocols such as the go-back-N protocol, in which the whole window (or, message) is retransmitted (see, e.g., Bertsekas and Gallager [2]). Individual packet delay distributions are usually not sufficient for proper understanding of the system behavior. In general, the delays of two consecutive packets are strongly correlated, i.e., the delay of the second packet conditioned on the event that the first packet delay is large (small) is larger (smaller) than the delay of an arbitrary packet. We shall show that evaluating the timeout using an independence assumption (that the delays of packets are independent from packet to packet) is quite pessimistic.

The message delay distribution for TDMA systems with a generalized arrival process was presented in Rom and Sidi [8]. The analysis of the message delay was associated with batch arrival processes in Halfin [6] and in Rom and Sidi [8], i.e., each batch corresponds to a message. In this case, the message delay coincides with the delay of the last packet of the message (batch). This fact facilitates the analysis of the message delay distribution. However, in packet switched networks, packets which belong to the same message may arrive at different instants of times (be dispersed), and may be interleaved (due to statistical multiplexing) by packets which belong to other messages. The difficulty that arises in the analysis of the message delay distribution for the dispersed generation model is that there is a correlation between the system states seen by different packets of the same message. The effect of the correlation between successive arrivals to the system on the average packet delays was studied in Sohraby [9] for Poisson cluster processes (PCP). Here, messages arrive to the system according to a Poisson process, but unlike the batch Poisson arrival process, where all the packets of the batch (message) arrive at the same time, the members of a cluster are separated by a random variable. In Sohraby [9], the average delay of packets was approximated for a $PCP/D/1$ system.

This paper introduces a new technique to analyze the message delay in such systems, and shows that this correlation has a strong effect on the performance of the system. This technique is similar to the one introduced recently in Cidon et al. [5]. The model we use in this paper for ascertaining the correlation in the packet delay process consists of a source that generates packets and sends them through a single server with an infinite number of buffers, which represents the communication system. We present an exact analysis of the message delay. In particular, we introduce an efficient recursive procedure to obtain the LST of the message delay for different arrival models and different number of sessions. For the $M/M/1$ system with variable size messages and for the bursty traffic model, we obtain an explicit expression for the LST of the message delay. As was discussed above, the use of an *independence assumption*, i.e., the assumption that the delays of packets are independent from packet to packet, can lead to wrong conclusions. We demonstrate this by comparing the exact variance of the message delay with the variance of the message delay as obtained from the independence assumption. Numerical examples are provided to show that the variance of the message delay may be overestimated by the independence assumption for a wide range of message sizes. In addition, our results demonstrate the "negative feedback" effect that governs the message delay process. If the message's packet arrivals happen to concentrate over a short time interval, then, the message arrival time becomes short. On the other hand this causes a larger queue to be built up, resulting in a larger queueing delay for the last packet of the message. Similarly, if the message's packet arrivals happen to be more dispersed, then, the queueing delay of the last packet tends to become shorter. Thus, the message delay distribution in the dispersed generation model tends to concentrate around the average much more than can be expected using the independence assumption.

The paper is structured as follows: In sections 2 and 3 we focus on continuous time systems and a fixed block size (counted in packets). The continuous time model is suitable for the analysis of variable size packet systems. The analysis is for a single session with Poisson arrivals with exponentially distributed service times in section 2 and with generally distributed service times in section 3. We also discuss the numerical results for some examples. In section 4 we extend the results of section 2 to the case of multiple session multiplexing and obtain the distribution of the message delay for a given message that belongs to a particular session. In section 5 we analyze the single session model of section 2 for the case of variable block size. Here, we obtain an explicit expression for the LST of the message delay. In section 6 we analyze a single session system with binary Markov (bursty) arrival process, and obtain an explicit expression for the LST of the message delay.

2. Single session systems: fixed message size - $M/M/1$

In this section we consider systems with variable length packets. The packets are stored in a queue that can accommodate infinitely many packets and are transmitted according to the first-in-first-out (FIFO) rule. The packets are grouped into messages of an arbitrary size. We consider systems with Poisson arrival process with rate λ. We assume that the transmission time of packets is exponentially distributed with parameter μ. The average load ρ is defined as $\rho \overset{\Delta}{=} \lambda/\mu$, and we assume that $\rho < 1$. In the next section we consider systems with generally distributed transmission time.

Consider an arbitrary message of size n arriving to the system (i.e., the arrival of the first packet of the message) in steady-state. This tagged message will be termed the t-message. The packets which belong to this message will be called the t-packets. The first and the last t-packets will be called the t-header and the t-tail, respectively. Let $d^h_{i,n}, n \geqslant 1, 1 \leqslant i \leqslant n$, ($h$ stands for *h*eader), be a random variable (r.v.) of the time delay from the arrival epoch of the ith t-packet to the system to the departure epoch of the t-tail from the system, given that the t-header is present in the system at that arrival epoch. Denote by $\mathcal{D}^h_{i,n}(s)$ the LST of the r.v. $d^h_{i,n}$. Note that, the message delay for a message of length $n, n \geqslant 1$, is given by the r.v. $d^h_{1,n}$, and we are interested in its LST $\mathcal{D}^h_{1,n}(s)$.

We recall that, for the $M/M/1$ system, the LST of the time spent in the system by an arbitrary packet is exponentially distributed with parameter $\mu - \lambda$ (see, e.g., Kleinrock [7, p. 202]). Since the first packet in a message is arbitrary, then its delay is exponentially distributed with parameter $\mu - \lambda$, and is independent of the inter-arrival and service times of the subsequent packets. Let $d^a_{i,m}, m \geqslant 1, i \geqslant 0$, ($a$ stands for *a*rbitrary), be a r.v. of the time delay from an arbitrary epoch to the departure epoch of the last packet of the next m packets that leave the system (transmitted), given that i packets are present in the system at that arbitrary epoch. Denote by $\mathcal{D}^a_{i,m}(s)$ the LST of the r.v. $d^a_{i,m}$. Consider the ith t-packet that arrives to the system while the t-header is present in the system. Conditioning on the next event; an arrival of a t-packet before the departure of the t-header or a departure of the t-header before the next arrival of a t-packet, and using the well known properties of independent and exponentially distributed r.v.s we have (for $n \geqslant 1$)

$$\mathcal{D}^h_{i,n}(s) = \frac{\mu}{s+\mu}\left(\frac{\lambda}{\mu}\mathcal{D}^h_{i+1,n}(s) + \frac{\mu-\lambda}{\mu}\mathcal{D}^a_{i-1,n-1}(s)\right) \quad 1 \leqslant i \leqslant n-1\,, \tag{1}$$

and for $i = n$ we have

$$\mathcal{D}^h_{n,n}(s) = \left(\frac{\mu-\lambda}{s+\mu-\lambda}\right)\left(\frac{\mu}{s+\mu}\right)^{n-1}, \tag{2}$$

where, the first product term in eq. (1) is the LST of the minimum of two independent r.v.s which are exponentially distributed with rates $\mu - \lambda$ and λ, respectively.

The LST of the message delay $\mathcal{D}^h_{1,n}(s), n \geqslant 1$, is obtained from the set of difference equations (in the parameter i) defined in (1) and (2), and we have

$$\mathcal{D}^h_{1,n}(s) = \frac{(\mu - \lambda)(\mu\lambda)^{n-1}}{(s + \mu - \lambda)(s + \mu)^{2n-2}} + \frac{\mu - \lambda}{s + \mu} \sum_{j=0}^{n-2} \left(\frac{\lambda}{s + \mu}\right)^j \mathcal{D}^a_{j,n-1}(s) . \qquad (3)$$

In order to uniquely determine the LST of the message delay $\mathcal{D}^h_{1,n}(s)$, we need to compute the LSTs $\mathcal{D}^a_{i,n-1}(s), 0 \leqslant i \leqslant n-2$. In what follows, we introduce a recursive procedure similar to the one introduced recently in Cidon et al. [5] for the computation of the LSTs $\mathcal{D}^a_{i,m}(s), 1 \leqslant m \leqslant n-1, 0 \leqslant i \leqslant m-1$. The recursion is initiated at $m = 1$ with the following obvious relation:

$$\mathcal{D}^a_{0,1}(s) = \frac{\lambda\mu}{(s+\lambda)(s+\mu)} . \qquad (4)$$

Conditioning on the next event: an arrival of a t-packet or a departure of a t-packet, we have for $2 \leqslant m \leqslant n-1$ the following recursive equations:

$$\mathcal{D}^a_{0,m}(s) = \frac{\lambda}{s+\lambda} \mathcal{D}^a_{1,m}(s) ,$$

$$\mathcal{D}^a_{i,m}(s) = \frac{\lambda+\mu}{s+\lambda+\mu} \left(\frac{\lambda}{\lambda+\mu} \mathcal{D}^a_{i+1,m}(s) + \frac{\mu}{\lambda+\mu} \mathcal{D}^a_{i-1,m-1}(s) \right) \quad 1 \leqslant i \leqslant m-1 ,$$

$$\mathcal{D}^a_{m,m}(s) = \left(\frac{\mu}{s+\mu}\right)^m . \qquad (5)$$

From the set of difference equations (in the parameter i) defined in (5), we have

$$\mathcal{D}^a_{0,m}(s) = \frac{\lambda}{s+\lambda} \mathcal{D}^a_{1,m}(s) ,$$

$$\mathcal{D}^a_{i,m}(s) = \left(\frac{\lambda}{s+\lambda+\mu}\right)^{m-i} \left(\frac{\mu}{s+\mu}\right)^m + \frac{\mu}{s+\lambda+\mu} \sum_{j=0}^{m-i-1} \left(\frac{\lambda}{s+\lambda+\mu}\right)^j \mathcal{D}^a_{i+j-1,m-1}(s) \quad 1 \leqslant i \leqslant m-1 . \qquad (6)$$

The LSTs $\mathcal{D}^a_{i,m}(s), 2 \leqslant m \leqslant n-1, 0 \leqslant i \leqslant m-1$, can be obtained recursively from (4) and (6). Then, the LST of the message delay is obtained from (3). From eq. (4) and (6), a set of recursive equations for the computation of any moment of the r.v.s $d^a_{i,m}, 2 \leqslant m \leqslant n-1, 0 \leqslant i \leqslant m-1$, can be obtained. The number of simple operations (additions and multiplications) needed for the computation of any such moment is of the order of $O(n^2)$. Then, any moment of the message delay can be obtained from (3).

Yet, for the average message delay we can obtain a simple closed form expression as follows. We note that the message delay is composed of two components. The first is the time elapsing between the arrival epoch of the first packet of the message to the system until the arrival epoch of the last packet of that message to the system (Erlang distribution with parameters $n-1, \lambda$). The second is the time delay of an arbitrary packet (stands for the last packet of that message) in the system which is exponentially distributed with rate $\mu - \lambda$. These two components are of course dependent r.v.s. However, the average message delay can be obtained directly from the sum of the averages of these two components, i.e., $(n-1)/\lambda + 1/(\mu - \lambda)$.

A simple and common way to approximate the message delay is to assume that these two components are independent r.v.s. In the following example we show the relative error of such an approximation.

NUMERICAL EXAMPLE

Using the above independence approximation, the variance of the message delay becomes $(n-1)/\lambda^2 + 1/(\mu-\lambda)^2$. The relative variance error of the message delay, defined as 100∗ [(*approximated variance*)/(*exact variance*)−1] is plotted in fig. 1 versus the message size n for $\mu = 1$ and for different values of λ ($\lambda = 0.5$, 0.8, 0.9). For all cases observe that the approximated variance of the message delay is much larger than the exact one. Observe also that the approximation becomes worse for heavy loads in a wide range of message sizes.

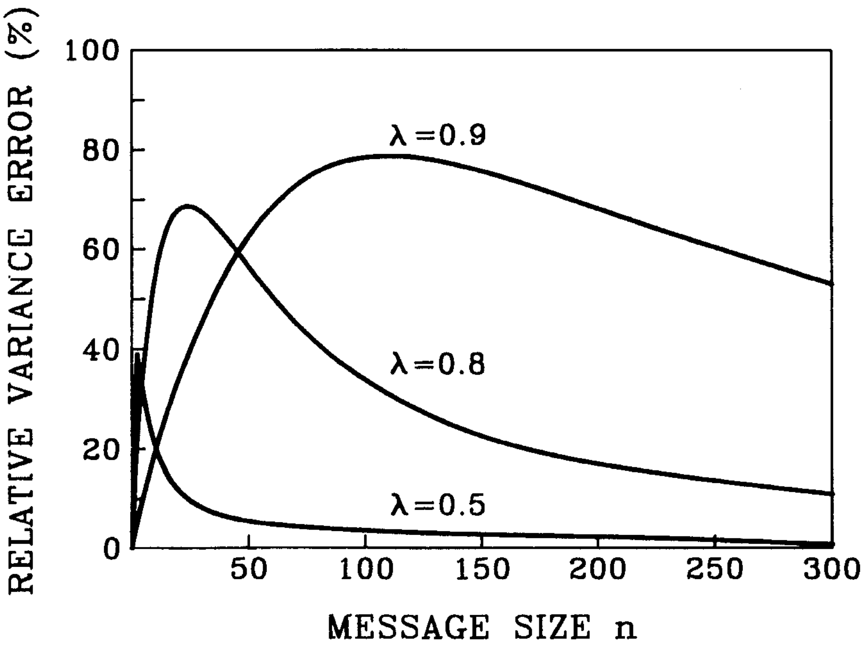

Fig. 1. The relative variance error of the message delay versus the message size *n*.

3. Single session systems: Fixed message size – $M/G/1$

In this section we assume that the transmission time of packets is generally distributed. Let b be the r.v. of the transmission time of a packet. Denote by $\mathcal{A}(s) \triangleq \lambda/(s+\lambda)$ and $\mathcal{B}(s) \triangleq E[\mathrm{e}^{-sb}]$ the LSTs of the inter-arrival time and the transmission time of packets, respectively. The average load ρ is defined as $\rho \triangleq \lambda E[b]$, and we assume that $\rho < 1$.

Let $d_n, n \geqslant 1$, be a r.v. of the message delay for a message of length n. Our purpose in this section is to compute the LST $\mathcal{D}_n(s) \triangleq E[\mathrm{e}^{-sd_n}], n \geqslant 1$, of the message delay d_n. We carry the computation by conditioning on the number of packets in the system (queue) just after the departure of the first packet of the message from the system. To that end, we define $d_{k,m}, m \geqslant 1, k \geqslant 0$, to be the time elapsing between the departure epoch of a packet from the system until after the transmission of the last packet of the next m packets that leave the system (transmitted), given that there are k packets in the system (queue) just after that departure epoch. Denote by $\mathcal{D}_{k,m}(s) \triangleq E[\mathrm{e}^{-sd_{k,m}}]$ the LST of the r.v. $d_{k,m}$. In what follows, we shall need several additional definitions. Let $P_k, k \geqslant 0$, be the probability of k packets in the system (queue) at the departure epoch of the first packet of a message from the system. The z-transform of the probability distribution P_k is given by the Pollaczek–Khinchin transform (PKT) equation for the number of packets in the system (see, e.g., Kleinrock [7, p. 194])

$$Q(z) = \mathcal{B}(\lambda - \lambda z)\frac{(1-\rho)(1-z)}{\mathcal{B}(\lambda - \lambda z) - z}\,.$$

Let $\eta_k, k \geqslant 0$, be the time delay of the first packet of a message in the system (from its arrival to the system to its departure from the system), given that k packets are present in the system (queue) just after the departure epoch of that packet from the system. Since the first packet of a message is arbitrary we have that

$$\mathcal{T}_k(s) \triangleq E[\mathrm{e}^{-s\eta_k}] = \frac{1}{P_k}E\left[\frac{(\lambda\eta)^k \mathrm{e}^{-(\lambda+s)\eta}}{k!}\right] \quad k \geqslant 0\,, \tag{7}$$

where η is the time delay of an arbitrary packet in the system, and its LST is given by the PKT equation for the time delay (see, e.g., Kleinrock [7, p. 200])

$$\mathcal{T}(s) \triangleq E[\mathrm{e}^{-s\eta}] = \frac{s(1-\rho)\mathcal{B}(s)}{s - \lambda + \lambda\mathcal{B}(s)}\,.$$

In the derivation of (7) we used the solution of the following auxiliary problem:

Let γ be a non-negative r.v. and denote its probability density function by $f_\gamma(x), x \geqslant 0$. Assume that packets arrive according to a Poisson process with rate λ. Let v be a r.v. of the number of packets that arrive during a time interval γ. Then we have

$$f(k,x) \triangleq \lim_{dx\to 0} \frac{Pr(v=k, x<\gamma\leqslant x+dx)}{dx} = \lim_{dx\to 0} Pr(v=k|x<\gamma\leqslant x+dx)$$

$$\times \lim_{dx\to 0} \frac{P(x<\gamma\leqslant x+dx)}{dx} = \frac{(\lambda x)^k \mathrm{e}^{-\lambda x}}{k!} f_\gamma(x) \quad x\geqslant 0, k\geqslant 0\,. \tag{8}$$

Let $\gamma(k), k\geqslant 0$, be the time duration of the r.v. γ *when* k packets arrive during γ. Then, using (8) we have

$$E[\mathrm{e}^{-s\gamma(k)}] = \int_0^\infty \mathrm{e}^{-sx} f(k,x)\,dx$$

$$= \int_0^\infty \mathrm{e}^{-sx} \frac{(\lambda x)^k \mathrm{e}^{-\lambda x}}{k!} f_\gamma(x) dx = E\left[\frac{(\lambda\gamma)^k \mathrm{e}^{-(\lambda+s)\gamma}}{k!}\right]. \tag{9}$$

We are now ready to introduce the computation of the LST of the message delay. Since the first packet of a message is arbitrary, we have that $d_1 \overset{d}{=} \eta$, and for $n\geqslant 2$,

$$\mathcal{D}_n(s) \triangleq E[\mathrm{e}^{-sd_n}] = \sum_{k=0}^{\infty} P_k E[\mathrm{e}^{-s(\eta_k+d_{k,n-1})}] = \sum_{k=0}^{\infty} P_k E[\mathrm{e}^{-s\eta_k}] E[\mathrm{e}^{-sd_{k,n-1}}]$$

$$= \sum_{k=0}^{\infty} E\left[\frac{(\lambda\eta)^k \mathrm{e}^{-(\lambda+s)\eta}}{k!}\right] \mathcal{D}_{k,n-1}(s)\,, \tag{10}$$

where the third equality follows from the fact that the time elapsing from the departure epoch of the first packet from the system until after the next $n-1$ packets leave the system and the time delay of the first packet in the system, given that there are k packets in the system at that departure epoch, are independent r.v.s. The last equality follows directly from (7).

To complete the computation we need to compute the LST $\mathcal{D}_{k,n-1}(s), n\geqslant 2$, $k\geqslant 0$. In what follows we introduce a recursive procedure for the computation of those LSTs. The recursion is initiated at $m=1$ with the following obvious relations:

$$\mathcal{D}_{k,1}(s) = \begin{cases} \mathcal{A}(s)\mathcal{B}(s) & k=0\,, \\ \mathcal{B}(s) & k\geqslant 1\,. \end{cases} \tag{11}$$

For $m\geqslant 2$, we have

$$\mathcal{D}_{0,m}(s) = \mathcal{A}(s)\left(\sum_{j=0}^{m-2} E\left[\frac{(\lambda b)^j \mathrm{e}^{-(\lambda+s)b}}{j!}\right]\mathcal{D}_{j,m-1}(s) + \sum_{j=m-1}^{\infty} E\left[\frac{(\lambda b)^j \mathrm{e}^{-(\lambda+s)b}}{j!}\right](\mathcal{B}(s))^{m-1}\right).$$

$$= \mathcal{A}(s)\sum_{j=0}^{m-2} E\left[\frac{(\lambda b)^j \mathrm{e}^{-(\lambda+s)b}}{j!}\right]\left(\mathcal{D}_{j,m-1}(s) - (\mathcal{B}(s))^{m-1}\right) + \mathcal{A}(s)(\mathcal{B}(s))^m\,, \tag{12}$$

$$\mathcal{D}_{k,m}(s) = \sum_{j=0}^{m-k-1} E\left[\frac{(\lambda b)^j e^{-(\lambda+s)b}}{j!}\right]\left(\mathcal{D}_{k+j-1,m-1}(s) - (\mathcal{B}(s))^{m-1}\right) + (\mathcal{B}(s))^m$$
$$1 \leqslant k \leqslant m-1\,, \tag{13}$$

$$\mathcal{D}_{k,m}(s) = (\mathcal{B}(s))^m \quad k \geqslant m\,, \tag{14}$$

where in eqs. (12) and (13) we used the fact that $\sum_{k=0}^{\infty} E[(\lambda b)^k e^{-(\lambda+s)b}/k!] = \mathcal{B}(s)$. The recursions in (12) and (13) are obtained by conditioning on the number of packets in the system (queue) at consecutive departure epochs from the system, and by using the result for the auxiliary problem described above.

From (10) and (14), we have that

$$\mathcal{D}_n(s) = (\mathcal{B}(s))^{n-1}\mathcal{T}(s) + \sum_{k=0}^{n-2} E\left[\frac{(\lambda\eta)^k e^{-(\lambda+s)\eta}}{k!}\right]\left(\mathcal{D}_{k,n-1}(s) - (\mathcal{B}(s))^{n-1}\right). \tag{15}$$

From eqs. (11)–(14), a set of recursive equations for any moment of the r.v.s $d_{k,m}, k \geqslant 0$, for any $m \geqslant 1$ can be obtained by taking the appropriate derivatives at $s = 0$. These derivatives yield expressions of the form $E[b^j e^{-\lambda b}], j \geqslant 1$, that have to be computed in order to obtain the moments. These expressions can be computed by noting that $E[b^j e^{-\lambda b}] = (-1)^j d^j \mathcal{B}(s)/ds^j|_{s=\lambda}, j \geqslant 1$. Then, any moment of the message delay can be obtained from eq. (15). The computation complexity of this procedure is of the order of $O(n^2)$.

Note that the moments of the message delay depends on the LST of the service time $\mathcal{B}(s)$ and its derivatives at $s = \lambda$. That is, it depends on the whole distribution of the service time and not only on its moments as obtained by using the independence assumption.

NUMERICAL EXAMPLE

Consider an $M/E_2/1$ system with $\mathcal{B}(s) = (\mu/(\mu+s))^2$. Denote the arrival rate by λ and define $\rho \triangleq 2\lambda/\mu$. The average message delay can be obtained directly as in the $M/M/1$ system and is equal to $(n-1)/\lambda + (1-\rho)(2\mu-\lambda)/(\mu-2\lambda)^2$. The variance of the message delay can be approximated using the independence assumption, and is given by $(n-1)/\lambda^2 + \mathcal{T}''(0) - (\mathcal{T}'(0))^2$. The relative variance error of the message delay, is plotted in fig. 2 versus the message size n for $\mu = 1$ and for different values of λ ($\lambda = 0.25, 0.4, 0.45$). For all cases observe that the approximated variance of the message delay is much larger than the exact one. Observe also that the approximation becomes worst for heavy loads in a wide range of message sizes.

4. Multiple session systems: Fixed message size

Here we assume that packets arrive to the system from S independent sources,

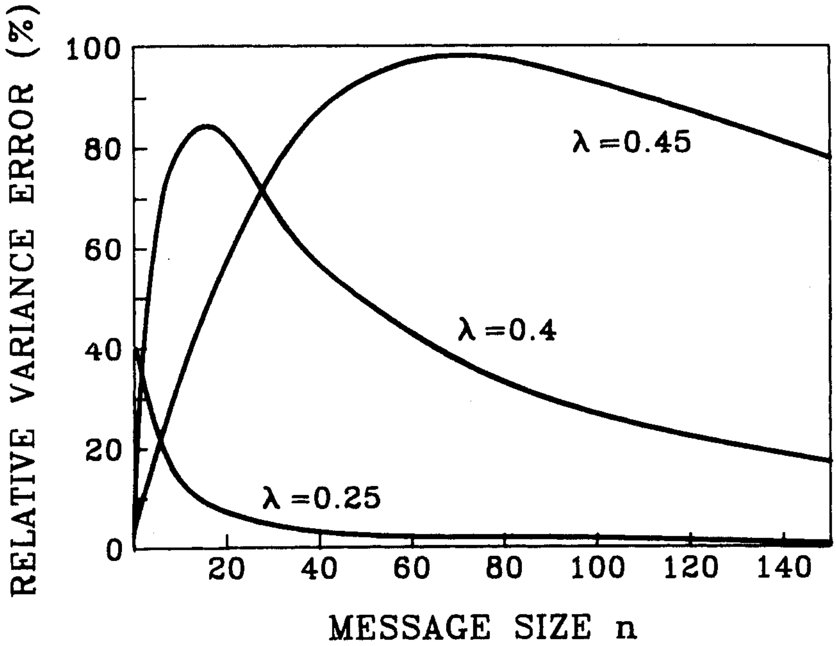

Fig. 2. The relative variance error of the message delay versus the message size n.

that is, the interarrival times and the transmission times of packets from each source are mutually independent. The arrival process from source $g, 1 \leqslant g \leqslant S$, is assumed to be Poisson with rate λ_g. The overall arrival process to the system is then Poisson with rate $\lambda \triangleq \sum_{g=1}^{S} \lambda_g$. For the system with Poisson rate λ and exponential transmission rate μ, the probability of $i, i \geqslant 0$, packets in the system in steady state, is given by, $\Pi(i) = (1-\rho)\rho^i$. Denote the overall arrival rate of all sources other than source g by $\lambda_{\bar{g}} \triangleq \lambda - \lambda_g$.

Consider an arbitrary message of size n arriving to the system from source $g, 1 \leqslant g \leqslant S$, in steady-state. Let $d_n, n \geqslant 1$, be a r.v. of the message delay and denote by $\mathcal{D}_n(s)$ its LST. Let $d_{i,n}, i \geqslant 0, n \geqslant 1$, be a r.v. of the time delay from an arbitrary epoch to the departure epoch of the last packet of the next n packets that *arrive* to the system from source g, given that i packets are present in the system at that arbitrary epoch. Denote by $\mathcal{D}_{i,n}(s)$ its LST, and define

$$\mathcal{D}_{i,0}(s) \triangleq \left(\frac{\mu}{s+\mu}\right)^i \quad i \geqslant 0 . \tag{16}$$

Since the first packet of a message is arbitrary, we have that

$$\mathcal{D}_n(s) = \sum_{i=0}^{\infty} \Pi(i)\mathcal{D}_{i+1,n-1}(s) = \frac{1-\rho}{\rho}\sum_{i=1}^{\infty} \rho^i \mathcal{D}_{i,n-1}(s) , \tag{17}$$

where in (17) we conditioned on the number of packets found in the system at the arrival epoch of the first packet of the message, and used the fact that in the $M/M/1$ system the epoch just after an arrival is equivalent to an arbitrary epoch when we consider the evolution of the system from that epoch on.

We proceed to obtain the quantity $\sum_{i=1}^{\infty}\rho^i\mathcal{D}_{i,n-1}(s)$. In order to do that, we define the power series in the complex variable z, $G_n(s,z) \overset{\Delta}{=} \sum_{i=0}^{\infty}\mathcal{D}_{i,n}(s)z^i$ for $|z|<1$ and for all $s, \mathrm{Re}(s)\geqslant 0$. Since $|\mathcal{D}_{i,n}(s)|\leqslant 1, i\geqslant 0, \mathrm{Re}(s)\geqslant 0$, then using Abel's theorem (see, e.g., Ahlfors [1, p. 38]) it follows that the power series $G_n(s,z)$ for every $s, \mathrm{Re}(s)\geqslant 0$, converges absolutely and is an analytic function in the complex variable z inside the unit disk $|z|<1$. Now, using this definition it follows directly from (17) that

$$\mathcal{D}_n(s) = \frac{1-\rho}{\rho}\left(G_{n-1}(s,\rho) - G_{n-1}(s,0)\right). \tag{18}$$

We proceed to obtain an expression for the power series $G_n(s,z), n\geqslant 0, |z|<1$. In order to do that, we first obtain a set of recursive equations for the computation of the LSTs $\mathcal{D}_{i,n}(s), n\geqslant 0, i\geqslant 0$. Consider an arbitrary epoch in which i packets are present in the system. Condition on the next event; an arrival of a packet from source g (which belongs to the message) before the departure of the next packet from the system, or an arrival of a packet from source other than source g (which doesn't belong to the message) before the departure of the next packet from the system, or a departure of a packet from the system before the next arrival to the system. Then using the well-known properties of independent and exponentially distributed r.v.s (as in section 2), we have that (for $n\geqslant 1$)

$$\mathcal{D}_{0,n}(s) = \frac{\lambda_g}{s+\lambda}\mathcal{D}_{1,n-1}(s) + \frac{\lambda_{\bar{g}}}{s+\lambda}\mathcal{D}_{1,n}(s),$$

$$\begin{aligned}\mathcal{D}_{i,n}(s) = &\frac{\lambda_g}{s+\lambda+\mu}\mathcal{D}_{i+1,n-1}(s) + \frac{\lambda_{\bar{g}}}{s+\lambda+\mu}\mathcal{D}_{i+1,n}(s) + \frac{\mu}{s+\lambda+\mu}\mathcal{D}_{i-1,n}(s)\\ &i\geqslant 1.\end{aligned} \tag{19}$$

Now, substituting the LSTs $\mathcal{D}_{i,n}(s), i\geqslant 0$, from eqs. (16) and (19) in the power series $G_n(s,z)$ and using simple algebra, we have that

$$G_0(s,z) = \frac{s+\mu}{s+\mu-\mu z},$$

$$\begin{aligned}G_n(s,z) = &\frac{1}{s+\lambda+\mu}\left\{\frac{\lambda_g\mu}{s+\lambda}\mathcal{D}_{1,n-1}(s) - \lambda_g z^{-1}\mathcal{D}_{0,n-1}(s) + \lambda_g z^{-1}G_{n-1}(s,z)\right.\\ &\left.+\frac{\lambda_{\bar{g}}\mu}{s+\lambda}\mathcal{D}_{1,n}(s) - \lambda_{\bar{g}}z^{-1}\mathcal{D}_{0,n}(s) + (\lambda_{\bar{g}}z^{-1}+\mu z)G_n(s,z)\right\}.\end{aligned} \tag{20}$$

From (19) for $i=0$ and (20) we have that

$$G_n(s,z) = \frac{\lambda_g[G_{n-1}(s,z) - G_{n-1}(s,0)] + (\mu z - \lambda_{\bar{g}})G_n(s,0)}{(s+\lambda+\mu)z - \mu z^2 - \lambda_{\bar{g}}}. \tag{21}$$

In (21) we obtained a recursion for the computation of the power series $G_n(s,z), n \geqslant 1$. Once the power series $G_{n-1}(s,z)$ has been obtained. We still have to determine the boundary function $G_n(s,0)$ in order to uniquely determine the power series $G_n(s,z)$. Using Rouche's theorem (see, e.g., Ahlfors [1, p. 153]) it follows that the denominator of (21) has a unique root inside the unit disk $|z| < 1$. Then, using geometrical considerations it follows that this root is given by

$$z^*(s) = \frac{s+\lambda+\mu-\sqrt{(s+\lambda+\mu)^2-4\lambda_{\bar{g}}\mu}}{2\mu} \qquad \mathrm{Re}(s) \geqslant 0 . \tag{22}$$

Since the power series $G_n(s,z)$ is analytic inside the unit circle $|z| < 1$, then the numerator must vanish at $z = z^*(s)$, and we have that

$$G_n(s,0) = \frac{\lambda_g[G_{n-1}(s,z^*(s)) - G_{n-1}(s,0)]}{\lambda_{\bar{g}} - \mu z^*(s)} . \tag{23}$$

The procedure for the computation of the power series $G_n(s,z)$ proceeds as follows. First the power series $G_0(s,z)$ is obtained from (20) for $n = 0$. In step $k, k = 1, 2, \ldots, n$, the power series $G_k(s,z)$ is obtained by substituting $G_{k-1}(s,z)$ (which was obtained at step $k-1$) and $G_n(s,0)$ (which is obtained by substituting $G_{k-1}(s,z^*(s))$ in (23)) in (21) for $n = k$. Finally, the LST of the message delay $\mathcal{D}_n(s)$ is obtained by substituting $G_n(s,\rho)$ and $G_n(s,0)$ in (18). The difficulty in this procedure is the computation of $G_{k-1}(s,z^*(s))$ in step k, because the numerator (and the denominator) of (21) vanishes at $z^*(s)$ and hence we have to use L'Hôpital's law in order to compute this quantity.

In appendix A we describe an explicit computational procedure for the computation of the power series $G_n(s,z)$. We use subsequent substitutions of the recursion (21) to present the power series $G_n(s,z)$ as an explicit function of the boundary functions $G_j(s,0), 1 \leqslant j \leqslant n$. We then use the analytic properties of the power series to derive *n recursive* equations for the computation of the boundary functions.

5. Single session systems: Variable message size

In this section we consider systems with variable length messages, namely, packets that arrive to the system belong to messages of length which is independent of and geometrically distributed with parameter q. Variable message size is typical in data applications where the message can be a document, an e-mail message, or an arbitrary file. This model also assumes a variable size packet which may correspond to some natural partition of the message (i.e. sections of a document, paragraphs of the e-mail message, etc.). We confine ourselves in this section to the analysis of a single session system with Poisson arrival process with rate λ and transmission time exponentially distributed with rate μ. The average load ρ is defined as before. The extensions of the analysis to generally distributed transmission time and to multiple session systems are similar to the extensions in sections 3 and 4 and are not presented here.

Consider an arbitrary message arriving to the system in steady-state. Let d be a r.v. of the message delay and denote by $\mathcal{D}(s)$ its LST. Let d_i be a r.v. of the time delay from an arbitrary epoch to the departure epoch of the last packet of the message from the system, given that i packets are present in the system at that arbitrary epoch. Denote by $\mathcal{D}_i(s) \overset{\Delta}{=} E[\mathrm{e}^{-sd_i}]$ the LST of the r.v. d_i. Since the first packet of a message is arbitrary, we have that

$$\mathcal{D}(s) = \sum_{i=0}^{\infty} \Pi(i)\left(q\left(\frac{\mu}{\mu+s}\right)^{i+1} + \bar{q}\mathcal{D}_{i+1}(s)\right)$$
$$= \frac{q\mu(1-\rho)}{s+\mu-\lambda} + \frac{\bar{q}(1-\rho)}{\rho}\sum_{i=1}^{\infty}\rho^i\mathcal{D}_i(s)\,, \tag{24}$$

where in (24) we first conditioned on the number of packets found in the system at the arrival epoch of the first packet of the message, and then on the event that this packet is the last packet in the message (with probability q this is the last packet of the message and with probability $\bar{q} = 1 - q$ the next packet belongs to the message).

We proceed to obtain the quantity $\sum_{i=1}^{\infty} \rho^i\mathcal{D}_i(s)$. In order to do that, we define the power series in the complex variable z, $G(s, z) \overset{\Delta}{=} \sum_{i=0}^{\infty} \mathcal{D}_i(s)z^i$ for $|z| < 1$ and for every s, $\mathrm{Re}(s) \geqslant 0$. Since $|\mathcal{D}_i(s)| \leqslant 1, i \geqslant 0, \mathrm{Re}(s) \geqslant 0$, then using Abel's theorem (see, e.g., Ahlfors [1, p. 38]) it follows that the power series $G(s, z)$ for every s, $\mathrm{Re}(s) \geqslant 0$, converges absolutely and is an analytic function in the complex variable z inside the unit disk $|z| < 1$. Now, using this definition it follows directly from (24) that

$$\mathcal{D}(s) = \frac{q\mu(1-\rho)}{s+\mu-\lambda} + \frac{\bar{q}(1-\rho)}{\rho}(G(s,\rho) - G(s,0))\,. \tag{25}$$

We proceed to obtain an expression for the power series $G(s, z)$. In order to do that, we first obtain a set of recursive equations for the LSTs $\mathcal{D}_i(s), i \geqslant 0$. Consider an arbitrary epoch in which i packets are present in the system. Condition on the next event: an arrival of a packet (which belongs to the message) before the departure of the next packet from the system or on its complement, and on whether the next arrival is the last packet in the message. Then using the well-known properties of independent and exponentially distributed r.v.s (as in section 2), we have that

$$\mathcal{D}_0(s) = \frac{q\lambda\mu}{(s+\lambda)(s+\mu)} + \bar{q}\frac{\lambda}{s+\lambda}\mathcal{D}_1(s)\,,$$

$$\mathcal{D}_i(s) = \frac{q\lambda}{s+\lambda+\mu}\left(\frac{\mu}{s+\mu}\right)^{i+1} + \frac{\bar{q}\lambda}{s+\lambda+\mu}\mathcal{D}_{i+1}(s) + \frac{\mu}{s+\lambda+\mu}\mathcal{D}_{i-1}(s) \quad i \geqslant 1\,. \tag{26}$$

Now, substituting the LSTs $\mathcal{D}_i(s), i \geqslant 0$, from (26) in the power series $G(s, z)$ and using simple algebra, we have that

$$G(s,z)=\frac{q\mu\lambda z/(s+\mu-\mu z)+(\mu z-\bar{q}\lambda)G(s,0)}{(s+\lambda+\mu)z-\mu z^2-\bar{q}\lambda} \quad |z|<1. \tag{27}$$

In order to uniquely determine the power series $G(s,z)$ we will have to determine the boundary power series $G(s,0)$. Using Rouche's theorem (see, e.g., Ahlfors [1, p. 153]) it follows that the denominator of (27) has a unique root inside the unit disk $|z|<1$. Then, using geometrical considerations it follows that this root is given by

$$z^*(s)=\frac{s+\lambda+\mu-\sqrt{(s+\lambda+\mu)^2-4\bar{q}\lambda\mu}}{2\mu}. \tag{28}$$

Since the power series $G(s,z)$ is analytic inside the unit disk $|z|<1$, then the numerator must vanish at $z=z^*(s)$, and we have that

$$G(s,0)=\frac{-q\lambda\mu z^*(s)}{(s+\mu-\mu z^*(s))(\mu z^*(s)-\bar{q}\lambda)}. \tag{29}$$

By cancelling the term $z-z^*(s)$ in (27), we have that

$$G(s,z)=\frac{\mu^2 z^*(s)z-\bar{q}\lambda(s+\mu)G(s,0)}{\mu(s+\mu-\mu z)(z^*(s)z-\bar{q}\rho)}. \tag{30}$$

Finally, the LST of the message delay $\mathcal{D}(s)$ is obtained by substituting $G(s,\rho)$ from (27) in (25) to obtain,

$$\mathcal{D}(s)=\frac{q\mu(1-\rho)}{s+\mu-\lambda}+\frac{\bar{q}(1-\rho)}{\rho}\frac{q\lambda^2/(s+\mu-\lambda)-\rho sG(s,0)}{\rho s+q\lambda}. \tag{31}$$

Note also that the LSTs $\mathcal{D}_i(s), i\geqslant 0$, can be obtained explicitly by taking the inverse z-transform of (30).

The average message delay is obtained by taking the derivative of $\mathcal{D}(s)$ at $s=0$, and is equal to $\bar{q}/q\lambda+1/(\mu-\lambda)$. Using an independence assumption as in section 2, the LST of the message delay can be approximated by the LST of the sum of two independent r.v.s. The first stands for the time elapsing between the arrival epoch of the first packet of the message to the system until the arrival epoch of the last packet of that message to the system, and its LST is given by $q(\lambda+s)/(s+q\lambda)$. The second r.v. stands for the time delay of an arbitrary packet (stands for the last packet in that message) in the system which is exponentially distributed with rate $\mu-\lambda$. The average message delay can be obtained directly from the sum of the averages of these two r.v.s and it is the same as above.

NUMERICAL EXAMPLE

The variance of the message delay can be approximated using the independence assumption, and is given by $\bar{q}(1+q)/(q\lambda)^2+1/(\mu-\lambda)^2$. The relative variance error of the message delay, is plotted in fig. 3 versus the parameter q for $\mu=1$ and

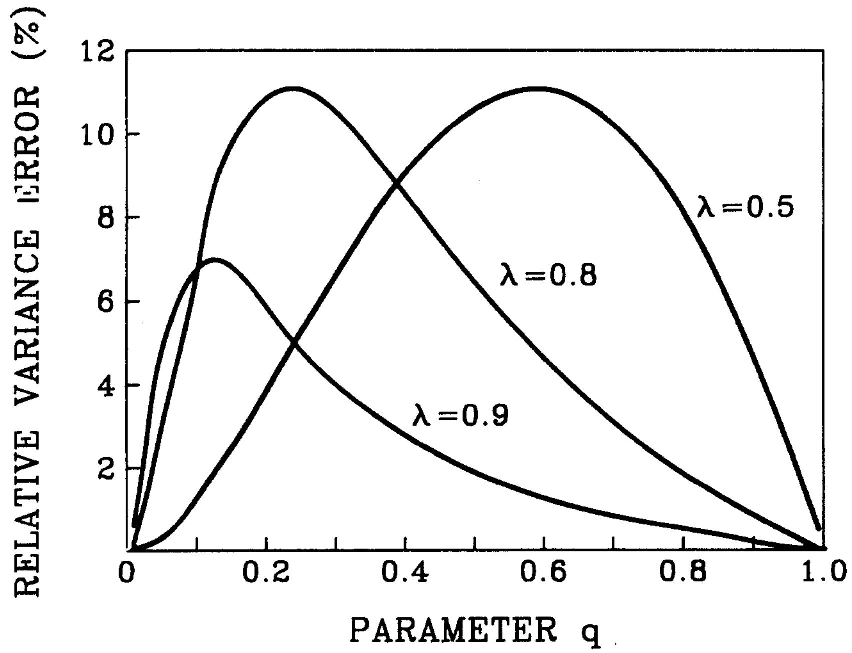

Fig. 3. The relative variance error of the message delay versus the parameter q.

for different values of λ ($\lambda = 0.5, 0.8, 0.9$). For all cases observe that the approximated variance of the message delay is larger than in the exact one.

6. Bursty traffic

Here, the generation of messages from the source is governed by an ON-OFF source model. This model is widely used in the literature to represent bursty and correlated cell arrivals, where a source may stay for relatively long durations in active (ON) and silent (OFF) periods (see e.g., Woodruff et al. [10]). We define the "active periods" and the "silent periods" of the source as the time periods during which the source generates packets or is idle, respectively. We assume that packets are generated by the source during active periods according to a Poisson process with rate λ. The duration of the active periods and the silent periods are assumed to be two independent sets of independent and identically distributed r.v.s, exponentially distributed with (positive) parameters α and β, respectively. We also assume that each active period corresponds to one message, and that a packet (the first packet of a message) is generated when the source goes from OFF to ON. We shall consider the case $\lambda \neq \beta$ as the case $\lambda = \beta$ degenerates to the standard $M/M/1$ case analyzed in section 2.

In appendix B, we derive the stationary probability of having i packets in the system at the arrival epoch of the first packet of a message, $\Pi(i), i \geqslant 0$, and its moment generating function, $F(z) \triangleq \sum_{i=0}^{\infty} \Pi(i) z^i$, $|z| \leqslant 1$,

$$\Pi(0) = \frac{(1-\rho)(\alpha+\beta)\sigma_1}{\lambda\sigma_1^2 - (\lambda+\mu)\sigma_1 + \mu},$$

$$\Pi(i) = \frac{\beta(\lambda+\alpha)\Pi(0)}{\lambda(\mu+\beta)+\alpha\beta}\left(\frac{1}{\sigma_2}\right)^i \quad i \geqslant 1,$$

$$F(z) = \mu\Pi(0)\frac{\lambda\sigma_1 z - \mu}{\sigma_1[\lambda(\mu+\beta)+\alpha\beta]z - \mu^2}, \tag{32}$$

where

$$\sigma_1 = \mu\frac{\alpha+\lambda+\mu+\beta-\sqrt{(\alpha+\lambda+\mu+\beta)^2 - 4[\lambda(\mu+\beta)+\alpha\beta]}}{2[\lambda(\mu+\beta)+\alpha\beta]},$$

$$\sigma_2 = \mu\frac{\alpha+\lambda+\mu+\beta+\sqrt{(\alpha+\lambda+\mu+\beta)^2 - 4[\lambda(\mu+\beta)+\alpha\beta]}}{2[\lambda(\mu+\beta)+\alpha\beta]}, \tag{33}$$

and $\rho = \beta(\lambda+\alpha)/\mu(\alpha+\beta) < 1$, is the steady-state condition of the system. Using the steady-state condition, it follows that, σ_1 and σ_2 are real, and $|\sigma_1| < 1$, $|\sigma_2| > 1$.

Now we proceed to obtain an expression for the LST of the message delay in steady-state. Note that message sizes are independent and geometrically distributed with parameter $q = \alpha/(\lambda+\alpha)$. Define $\gamma \triangleq \beta(\lambda+\alpha)\Pi(0)/(\lambda(\mu+\beta)+\alpha\beta)$. Then, using the same definitions and notations as in section 5, we have that the LST of the message delay, $\mathcal{D}(s)$, is given by

$$\begin{aligned}\mathcal{D}(s) &= \sum_{i=0}^{\infty} \Pi(i)\left(q\left(\frac{\mu}{\mu+s}\right)^{i+1} + \bar{q}\mathcal{D}_{i+1}(s)\right) \\ &= \frac{q\mu}{s+\mu}\left(F\left(\frac{\mu}{s+\mu}\right) - \Pi(0) + \gamma\right) + \left(\frac{(s+\lambda)(\Pi(0)-\gamma)}{\lambda} - \bar{q}\gamma\sigma_2\right)G(s,0) \\ &\quad + \bar{q}\gamma\sigma_2 G(s, 1/\sigma_2),\end{aligned} \tag{34}$$

where in (34), we used the definition of the power series $G(s,z)$ from section 5, and (26) for $i = 0$. (Note that, the analysis in section 5, and especially equations (26)–(30), holds also in this case). Now, substituting $G(s,z)$ from eqs. (28)–(30) in (34), we obtain the LST of the message delay, $\mathcal{D}(s)$.

Appendix A

EXPLICIT SOLUTION FOR MULTIPLE SESSION SYSTEM

In what follows, we describe an explicit computational procedure for the computation of the power series $G_n(s,z)$. From the recursion (21) we have by subsequent substitutions that

$$\begin{aligned} G_n(s,z) = \Big\{ & \lambda_g^n \mu z + (s+\mu-\mu z)(\mu z - \lambda_{\bar{g}} - F(z)) \\ & \sum_{j=1}^{n-1} \lambda_g^{n-j} (F(z))^{j-1} G_j(s,0) \\ & + (s+\mu-\mu z)(\mu z - \lambda_{\bar{g}})(F(z))^{n-1} G_n(s,0) \Big\} \\ & \times (s+\mu-\mu z)^{-1} (F(z))^{-n}, \end{aligned} \tag{35}$$

where $F(z) \triangleq (s+\lambda+\mu)z - \mu z^2 - \lambda_{\bar{g}}$.

In order to uniquely determine the power series $G_n(s,z)$, we still have to determine the boundary functions $G_j(s,0)$, $1 \leqslant j \leqslant n$.

DETERMINATION OF THE BOUNDARY FUNCTIONS $G_j(s,0)$, $1 \leqslant j \leqslant n$

The numerator of (35) is analytic in the unit disk $|z| < 1$ which contains $z^*(s)$ and hence it can be expanded to its *Taylor* series around this point, $\sum_{j=0}^{2n} a_j(s)(z - z^*(s))^j$. The coefficients $a_j(s)$, $0 \leqslant j \leqslant 2n$, are given by

$$a_0(s) = \lambda_g^n \mu z^*(s) + \gamma(s) S_0,$$

$$a_1(s) = \lambda_g^n \mu + (\beta(s) - \alpha(s)\delta(s)) S_0 + \gamma(s) S_1,$$

$$\begin{aligned} a_j(s) = \Big\{ & 1\{j \geqslant n-1\} \gamma(s) \binom{n-1}{j-n+1} + 1\{j \geqslant n\} \beta(s)\alpha(s) \binom{n-1}{j-n} \\ & - 1\{j \geqslant n+1\} \mu^2 (\alpha(s))^2 \binom{n-1}{j-n-1} \Big\} (-1)^{n-1} (\alpha(s))^{2n-j-2} G_n(s,0) \\ & - 1\{3 \leqslant j \leqslant 2n-1\} \mu S_{j-3} + 1\{j \leqslant 2n-2\}(\delta(s) + \mu\alpha(s) - \mu^2) S_{j-2} \\ & + 1\{j \leqslant 2n-3\}(\beta(s) - \alpha(s)\delta(s)) S_{j-1} + 1\{j \leqslant 2n-4\} \gamma(s) S_j \quad 2 \leqslant j \leqslant 2n, \end{aligned} \tag{36}$$

where in (36) we used the following notations. $1\{\cdot\}$ is an indicator function,

$$\alpha(s) \triangleq 2z^*(s) - \frac{s+\lambda+\mu}{\mu},$$

$$\beta(s) \overset{\Delta}{=} \mu(s + \mu + \lambda_{\bar{g}} - 2\mu z^*(s)),$$

$$\delta(s) \overset{\Delta}{=} s + \mu - \mu z^*(s),$$

$$\gamma(s) \overset{\Delta}{=} (\mu z^*(s) - \lambda_{\bar{g}})\delta(s),$$

$$S_j \overset{\Delta}{=} \sum_{k=\lceil \frac{j}{2} \rceil + 1}^{\min(j+1,n-1)} \lambda_g^{n-k}(-1)^{k-1}(\alpha(s))^{2k-j-2}\binom{k-1}{j-k+1} G_k(s,0) \quad 0 \leqslant j \leqslant 2n-4, \tag{37}$$

where an empty sum vanishes, $0! \overset{\Delta}{=} 1$ and $S_{-1} \overset{\Delta}{=} 0$.

In steady-state $(\rho < 1)$, the denominator of (35) has exactly one root $(z^*(s))$ of order n inside the unit disk $|z| < 1$. Since the power series $G_n(s, z)$ is analytic inside the unit disk $|z| < 1$, then the numerator must also have a root of order n at $z = z^*(s)$. This implies that $a_j(s) = 0, 0 \leqslant j \leqslant n-1$, which gives n *recursive* equations for the computation of the boundary functions $G_j(s,0), 1 \leqslant j \leqslant n$,

$$G_1(s,0) = \frac{-\lambda_g \mu z^*(s)}{\gamma(s)},$$

$$G_2(s,0) = \frac{\lambda_g^2 \mu + (\beta(s) - \alpha(s)\delta(s))\lambda_g G_1(s,0)}{\alpha(s)\gamma(s)},$$

$$\begin{aligned} G_j(s,0) = &\Big\{ (\beta(s) - \alpha(s)\delta(s))S_{j-2} + (\delta(s) + \mu\alpha(s) - \mu^2)S_{j-3} - \mu S_{j-4} \\ &+ \gamma(s) \sum_{k=\lceil \frac{j-1}{2} \rceil + 1}^{j-1} \lambda_g^{n-k}(-1)^{k-1}(\alpha(s))^{2k-j-1}\binom{k-1}{j-k} G_k(s,0) \Big\} \\ &\times \left(\gamma(s)\lambda_g^{n-j}(-1)^j(\alpha(s))^{j-1}\right)^{-1} \quad 3 \leqslant j \leqslant n-1 \end{aligned}$$

$$G_n(s,0) = \frac{[\gamma(s)S_{n-1} + (\beta(s) - \alpha(s)\delta(s))S_{n-2} + (\delta(s) + \mu\alpha(s) - \mu^2)S_{n-3} - \mu S_{n-4}]}{\gamma(s)(-1)^n(\alpha(s))^{n-1}}$$

$$n \geqslant 3. \tag{38}$$

Once the boundary functions $G_j(s,0), 1 \leqslant j \leqslant n$, have been obtained from the recursions in (38), the power series $G_n(s,z)$ can be obtained from (35). Then, the LST of the message delay, $\mathcal{D}_n(s)$, is obtained from (18).

Appendix B

Consider the system described in section 6. Denote by $\Pi_i^j, j = 0, 1, i \geqslant 0$, the stationarity probability of i packets in the system when the source is OFF and ON, respectively. Denote by $F_j(z) \triangleq \sum_{i=0}^{\infty} \Pi_i^j z^i$ its moment generating function. The state diagram of the system in steady-state is plotted in fig. 4. From this figure we obtain the following equilibrium equations:

$$\Pi_1^0 = \frac{\beta}{\mu}\Pi_0^0 - \frac{\alpha}{\mu}\Pi_0^1,$$

$$\Pi_i^0 = \left(\frac{\beta+\mu}{\mu}\right)\Pi_{i-1}^0 - \frac{\alpha}{\mu}\Pi_{i-1}^1 \quad i \geqslant 2,$$

$$\Pi_i^1 = \frac{\lambda}{\mu}\Pi_{i-1}^1 + \frac{\beta}{\mu}\Pi_{i-1}^0 - \Pi_i^0 \quad i \geqslant 1. \tag{39}$$

From (39), we have that

$$F_0(z) = (1-z)\Pi_0^0 + \left(\frac{\beta+\mu}{\mu}\right)zF_0(z) - \frac{\alpha}{\mu}zF_1(z),$$

$$F_1(z) = \Pi_0^0 + \Pi_0^1 + \left(\frac{\beta}{\mu}z - 1\right)F_0(z) + \frac{\lambda}{\mu}zF_1(z), \tag{40}$$

From (40), we have that

$$F_0(z) = \frac{\mu(\lambda z^2 - (\lambda+\mu+\alpha)z + \mu)\Pi_0^0 - \mu\alpha z\Pi_0^1}{[\lambda(\mu+\beta)+\alpha\beta]z^2 - \mu(\alpha+\lambda+\mu+\beta)z + \mu^2}. \tag{41}$$

From the normalization condition $F_0(1) = \alpha/(\alpha+\beta)$, we have that

$$\Pi_0^0 + \Pi_0^1 = 1 - \rho, \tag{42}$$

where $\rho = \beta(\lambda+\alpha)/\mu(\alpha+\beta)$, and $\rho < 1$ is the steady-state condition of the system.

In order to uniquely determine the moment generating function $F_0(z)$, we still have to determine the probabilities $\Pi_0^j, j = 0, 1$. Using the steady-state condition $\rho < 1$, it can easily be shown that the denominator of (41) has two real zeroes,

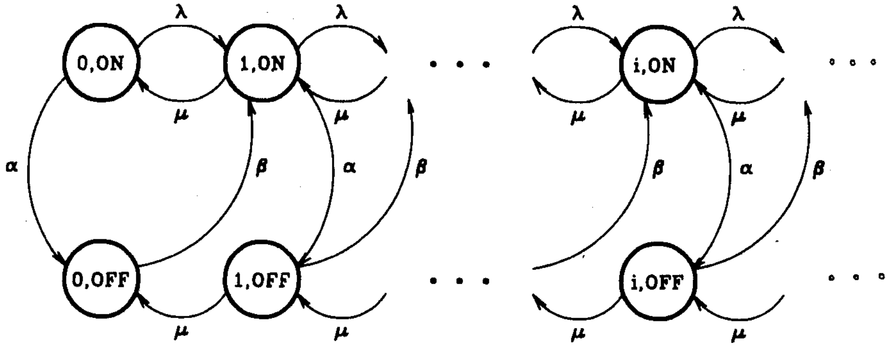

Fig. 4. The state diagram of the bursty traffic, single session system.

$$\sigma_{1,2} = \mu \frac{\alpha + \lambda + \mu + \beta \pm \sqrt{(\alpha + \lambda + \mu + \beta)^2 - 4[\lambda(\mu + \beta) + \alpha\beta]}}{2[\lambda(\mu + \beta) + \alpha\beta]}, \tag{43}$$

where $|\sigma_1| < 1$ (the "minus" zero) and $|\sigma_2| > 1$.

Since $F_0(z)$ is analytic inside the unit disk $|z| < 1$, the numerator of (41) must vanish at σ_1, then,

$$(\lambda\sigma_1^2 - (\lambda + \mu + \alpha)\sigma_1 + \mu)\Pi_0^0 - \alpha\sigma_1\Pi_0^1 = 0. \tag{44}$$

From (42) and (44), we have that

$$\Pi_0^0 = \frac{(1 - \rho)\alpha\sigma_1}{\lambda\sigma_1^2 - (\lambda + \mu)\sigma_1 + \mu}, \tag{45}$$

where the denominator of (45) is greater than zero for $\beta \neq \lambda$.

By canceling the term $z - \sigma_1$ in (41), we have that

$$F_0(z) = \mu\Pi_0^0 \frac{\lambda\sigma_1 z - \mu}{\sigma_1[\lambda(\mu + \beta) + \alpha\beta]z - \mu^2}. \tag{46}$$

Using the inverse z-transform of $F_0(z)$, we obtain the steady-state probabilities,

$$\Pi_i^0 = \frac{\beta(\lambda + \alpha)\Pi_0^0}{\lambda(\mu + \beta) + \alpha\beta}\left(\frac{1}{\sigma_2}\right)^i \quad i \geqslant 1. \tag{47}$$

Denote by $\Pi(i|\mathrm{OFF})$ the stationary probability of i packets in the system given that the source is silent. Then,

$$\Pi(i|\mathrm{OFF}) = \frac{\alpha + \beta}{\alpha}\Pi_i^0 \quad i \geqslant 0. \tag{48}$$

Now, given the source is silent, the first packet of a message arrives according to a Poisson process with rate β. Then, using the PASTA property, the probability $\Pi(i|\mathrm{OFF})$ corresponds to the stationary probability of having i packets in the system at the arrival epoch of the first packet of the message.

Acknowledgement

The authors would like to thank Opher Yaron for interesting discussions.

References

[1] L.V. Ahlfors, *Complex Analysis*, 3rd ed. (McGraw–Hill International Editions, 1979).
[2] D. Bertsekas and R. Gallager, *Data Networks* (Prentice–Hall International Editions, 1987).

[3] CCITT SG XVIII, Draft Recommendation I.121: Broadband Aspects of ISDN, Geneva, (January 1990).

[4] D. Comer, *Internetworking with TCP/IP, Principles, Protocols, and Architectures* (Prentice–Hall, Englewood Cliffs, NJ, 1988).

[5] I. Cidon, A. Khamisy and M. Sidi, On packet loss processes in high-speed networks, IEEE Trans. Information Theory 39 (1) (1993) 98–108.

[6] S. Halfin, Batch delays versus customer delays, Bell Syst. Tech. J. 62 (September 1983).

[7] L. Kleinrock, *Queueing Systems*, Vol. 1 (Wiley, New York, 1975).

[8] R. Rom and M. Sidi, *Multiple Access Protocols; Performance and Analysis* (Springer, Berlin, 1990).

[9] K. Sohraby, Delay analysis of a single server queue with poisson cluster arrival process arising in ATM networks, in: *GLOBECOM'89* (1989) pp. 611–616.

[10] G.M. Woodruff, R.G.H. Rogers and P.S. Richards, A congestion control framework for high-speed integrated packetized transport, in: *GLOBECOM'88* (1988) pp. 203–207.